In-Domain Inversion for Improved 3D Face Alignment on Asymmetrical Expressions

Jilliam M. Díaz Barros[1,2], Jason Rambach[1], Pramod Murthy[1] and Didier Stricker[1,2]
[1] German Research Center for Artificial Intelligence (DFKI)
[2] RPTU Kaiserslautern

***Abstract*— Facial landmark detection, often termed as face alignment, is a well-studied research problem in computer vision. Nonetheless, face alignment on asymmetrical expressions has been overlooked in the literature, particularly for unusual gestures observed in individuals with unilateral facial paralysis. In this paper, we explore in-domain inversion in a semi-supervised approach for face alignment and target the detection of 3D landmarks on symmetrical and extremely asymmetrical facial expressions due to paralysis. Our approach first leverages unlabeled face data to synthesize face images, while learning a compressed representation in the latent space. Then, it integrates in-domain inversion in the self-supervised stage, to make the latent space semantically meaningful. This is exploited in the supervised stage by a 2D face landmark detector, trained on labeled data. Finally, we extend the pipeline to 3D face alignment and regress the depth coordinate from the intermediate latent space and the predicted 2D landmarks. We evaluate and compare our method to related work on publicly available datasets, and demonstrate that our approach outperforms the state of the art in the detection of 3D facial landmarks in our newly introduced dataset of facial paralysis, ParFace. Our implementation and dataset are available at https://github.com/jilliam/ParFace.**

I. INTRODUCTION

Face alignment aims to register a predefined set of landmarks on a face image and is a key step to other face analysis tasks, such as head pose estimation [21], face synthesis [102], reconstruction [77], animation [22] and palsy assessment [42]. Many of these landmarks are semantically meaningful, referring, *e.g.*, to the corners of the eyes and lips, the tip of the nose and the contours of the eyebrows.

In the past years, many researchers strove to unify and standardize the set of keypoints used for face alignment [20], [74], [75], [76], [97]. The most common set defines 68 fiducial points on the eyes, nose, lips, eyebrows and around the boundary of the face, following the convention proposed in Multi-PIE [31]. This number differs for profile faces, where 39 fiducial points are annotated instead. These landmarks, referred to as 2D facial landmarks, are defined around the face contour and do not always correspond to the projection of 3D landmarks onto a 2D image, specifically when the face is not frontal [45]. Although this convention is useful for tasks such as face segmentation, it is error prone for optimization problems, *e.g.*, when minimizing the reprojection error [12], [49]. 3D Morphable Models (3DMMs) [8] and deep architectures have enabled the collection of datasets

This work was partially funded by the EU Horizon Europe Framework Program under Grant Agreement 101070192 (CORTEX2).

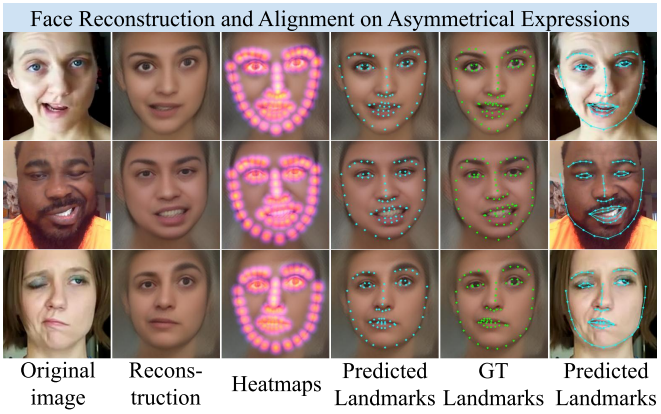

Fig. 1. Our approach learns to synthesize faces from unlabeled datasets and exploits the latent code to predict the landmarks.

with additional annotations, such as 3D landmarks [9], [96], [107]. Other annotations, such as the projected 3D landmarks in the image space, namely 3DA-2D, have also become available [9], [11], [96].

With the introduction of large-scale datasets [89], [97], [107] for training deep neural networks (DNN), 2D face alignment gained a performance boost *w.r.t.* traditional computer vision approaches, especially for challenging images with varying illumination, large head poses and occlusion. These datasets, however, have relatively few samples of large asymmetrical expressions and even less of peripheral facial paralysis, or palsy, affecting current face alignment approaches (see Fig. 2). This limitation has a negative impact on palsy assessments that rely on face alignment [2], [34]. Such assessments usually require the patient to follow predefined facial expressions, *e.g.* raising the eyebrows, closing the eyes and smiling. Then, an asymmetry index is computed based on measurements between specific areas in the affected side *w.r.t.* the unaffected side or the face at rest. An automatic method for extracting features or parts of the face used in the evaluations would reduce the associated costs and observer dependence inherent to manual assessment [37], [55]. In addition, 2D-landmark-based palsy assessment requires fully frontal face images [34] or pose correction techniques [37], [71], while the assessment with 3D landmarks is less prone to measurement errors since distances are not affected by the face orientation.

In this work, we aim to detect 3D facial landmarks and

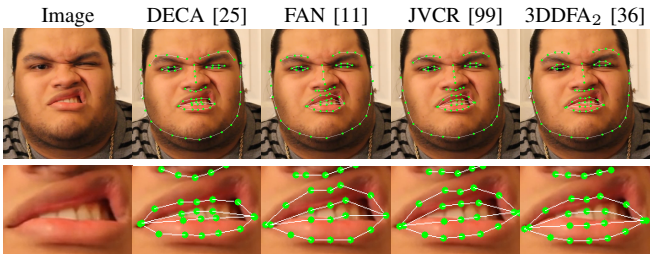

Fig. 2. Face alignment on patient with palsy. Top row: landmarks extracted from SOTA architectures. Bottom row: Close-up of the landmarks in the mouth. Note that the landmarks are defined around the contours of the lips.

address the limitation of current approaches, to target cases with large asymmetrical facial expressions from patients with facial palsy, alongside healthy subjects. Our approach exploits unlabeled face data with and without facial paralysis, to train an autoencoder and create an intermediate representation in a latent vector. In this stage, an in-domain-inversion module is incorporated to ensure a smooth latent space and enhance the representation of the expressions. In the supervised stage, we integrate interleaved transfer layers to the decoder to regress 3DA-2D landmarks, inspired by the state of the art (SOTA) 2D face alignment method, 3FabRec [10]. Our approach additionally enables the detection of 3D landmarks by means of a newly proposed 3D landmark detector. By relying on unlabeled data, our approach seeks to alleviate the cumbersome landmark annotation task, particularly for clinical data.

The proposed approach is supported by multiple experiments and evaluation on public face alignment datasets, in addition to a newly introduced facial palsy dataset.

The main contributions of this work are:

- A novel approach for 3D face alignment which encompasses cases with large facial asymmetry (see Fig. 1).
- The novel integration of in-domain GAN inversion in the self-supervised stage, to enhance the detection of the facial landmarks.
- ParFace, a 3D face alignment dataset on patients with palsy. ParFace and our source code is publicly available for research purposes.
- Evaluation on public face alignments datasets, in addition to the proposed facial palsy dataset, with improvements *w.r.t.* the state of the art in 3D face alignment.

II. RELATED WORK

Face alignment has been widely studied in the computer vision community. We classify face alignment approaches based on the type of landmarks: 2D and 3D.

A. 2D Face Alignment

These methods consider the face as a 2D object and detect only visible landmarks. Classical approaches use Active Appearance Models (AAMs) [14], [50], [78], Active Shape Models (ASMs) [15], [16], [67], Constrained Local Models (CLMs) [79], [94] and Cascaded Regression Methods (CRMs) [1], [54], [85], [106]. Since 2D landmarks do not maintain a one-to-one correspondence across large head poses, they are not robust against extreme head rotations.

More recently, several DNN architectures have been proposed for this task. They can be categorized as coordinate-regression or heatmap-based methods. The former includes architectures that regress the mapping between the image and the 2D coordinates and the latter regresses heatmaps for every landmark. Coordinate-based methods are more computationally efficient, while heatmap-based approaches usually have higher accuracy [47]. Coordinate-based methods span DAN [57], DeCaFa [17] and DTLD [60], and heatmap-based methods include SAN [23], HRNetV2 [82], LUVLi [58], 3FabRec [10], PIPNet [47], H3R [93], LDEQ [66] and [104].

B. 3D Face Alignment

These methods integrate 3D face models, either implicitly or explicitly, to recover a sparse or dense set of 3D facial landmarks. Some approaches jointly perform face alignment and reconstruction, with the aid of 3DMMs and large datasets of 3D faces [25], [32], [36], [49], [61], [62], [72], [84], [73], [88], [103], [107]. In general, they are more robust to large head poses and occlusion [45], but show poor generalization capabilities when data is low in quantity or variability [7]. Classical approaches such as [32], register a 3DMM to a face image. The alignment is formulated as a Bayesian inference problem and is solved using the Expectation-Maximization algorithm. CRMs have been extended to 3D face alignment as well [61], [62], [91].

Recent DNN use cascades of CNN regressors either in model-free [11], [24], [44], [95] or model-based approaches [49], [84], [103], [107], or 3D model warping functions [7], [80]. 3DDFA_V2 [36] leverages MobileNet [41] for 3DMM fitting, while SynergyNet [88] uses [36]. [36] adds layers for landmark regression and regularization, while [88] extracts the landmarks from the model and refines them 2DASL [84] uses self-supervised-learning to integrate datasets with only 2D or 3D annotations. FAN [11] uses stacked hourglass (HG) for 2D face alignment and an additional ResNet [39] to estimate the depth. [44] and JVCR [99] regress a volumetric representation of the face from a CNN based on stacked HG. JVCR additionally uses a 3D CNN to regress 3D coordinates. [24] exploits StyleGAN2 [52] to detect 3DA-2D landmarks. The generator is modified based on [10]. 3DSTN [7], a spatial transformer network, uses a generic 3D model along with Thin Plate Spline warping to handle unseen faces. [95] extends CLMs, where a CNN-based local detector exploits the advantages of mixture of experts. [30] incorporates an attention mechanism from a spatial transformer to the regression pipeline network in [53], to refine the landmark detection in eyes, irises and lips. [13] introduces a queried landmark predictor, allowing detection of 3D landmark configurations using a 3D face model reference. [98] proposes a multi-view consistent pipeline for landmark detection that leverages a multi-view dataset built using Neural Radiance Field (NeRF) [68].

C. Face Alignment for Palsy

Facial palsy assessment is performed either by segmenting susceptible regions of the face, such as eyebrows, eyes, nostrils and mouth [42], [46], [63], [70], by locating the muscle activation and exploiting action units (AUs) [4], [27], or by directly detecting facial landmarks. These landmarks have been used as well to locate AUs [27] or face regions heuristically [63], [83] or with more elaborated methods as in [42], [43], [46].

Face alignment for palsy assessment can be divided in two categories, 2D and 3D landmarks-based. 3D-based methods usually compute the landmarks from multi-camera systems [40], [101] and 3D sensors such as Kinect [26], deterring their implementation in a clinical setting. 2D landmarks, on the other hand, are extracted from grayscale or RGB images, captured from easily accessible cameras in smartphones [55], web [37] or digital cameras [29].

Relevant to this work are pipelines based on 2D images, which have been achieved with AAMs [18], [92], ASMs [86], CLMs [2], CRMs such as supervised descent method (SDM) [37], a parallel cascade of linear regressors [55] from [1], an ensemble of regression trees [2], [3], [34], [46], [64] from [54], supervised face alignment networks such as FAN [2], [33], [42], DAN [38], SAN [83] and other DNN [43], [90], [92]. In most cases, the alignment is performed with models that have been trained on images with healthy subjects, with very few or no samples of large asymmetrical facial expressions, limiting their scope. The bias in face alignment on palsy can be tackled with incremental learning on discriminative models such as [1], and retraining or fine-tuning existing regression-based pipelines [34] or face alignment networks, as in [2], [38], [42], with dedicated datasets from the target population, via transfer learning. Note that previous works on palsy face alignment use supervised approaches, while our method is semi-supervised and does not require large labeled datasets with asymmetrical expressions.

III. METHOD

In this work, we explore the semantically meaningful latent space in a reconstruction-based architecture, to improve the detection of facial landmarks in faces with a varying range of expressions. The proposed semi-supervised architecture ParFace-Net is shown in Fig. 3. In the self-supervised stage, an autoencoder (AE) is trained with unlabeled face datasets, where the encoder E learns the mapping from the input data to a low-dimensional intermediate vector z. This latent code is further enforced to be semantically meaningful, through the feature disentanglement process introduced by in-domain inversion. This is achieved by means of the discriminator D and adversarial training on the E and D, while the decoder G is frozen. In the supervised stage, a 2D landmark detector learns to regress 3DA-2D landmark heatmaps from the semantically rich latent code, which in turn are used to predict the depth coordinate. We further fine-tune the encoder with the gradients from the landmark heads to improve the results.

A. Self-Supervised Stage

This stage consists of an adversarial AE, trained on large-scale face datasets. The encoder E learns to capture the most important facial attributes in an intermediate latent vector z, while the decoder G is posed as a generator of a GAN that reconstructs the original image from the latent code. The AE is trained on a combination of three losses, as follows:

$$\mathcal{L}_{AE} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{adv}\mathcal{L}_{adv}, \tag{1}$$

where $\mathcal{L}_{rec}$ is the reconstruction loss, given by the $L1$ or $L2$ pixel-wise distance between the input x and the reconstruction $\hat{x}$ at $G(E(x))$; $\mathcal{L}_{perc}$ is the perceptual loss [48], in eq. (2); $\mathcal{L}_{adv}$ is an adversarial image loss [28], which enforces the AE to produce realistic faces based on the output from D; and $\lambda_{(\cdot)}$ is the respective weight of each loss.

$$\mathcal{L}_{perc}(x, \hat{x}) = \sum_{i}^{\phi} \frac{1}{C_i H_i W_i} ||V_i(\hat{x}) - V_i(x)||_2^2. \tag{2}$$

C_i, H_i, and W_i are the depth, height and width of the feature map $V_i(\cdot)$ at layer i of a VGG network [81]; x and $\hat{x}$ are the input and reconstructed images; and ϕ is the set of layers from the VGG.

We introduce a discriminator D during the face reconstruction phase, trained with the Wasserstein Loss with Gradient Penalty (WGAN-GP) [35], formulated as

$$\mathcal{L}_D = \mathbb{E}[D(\hat{x})] - \mathbb{E}[D(x)] + \frac{1}{2}\gamma\mathbb{E}[\nabla D], \tag{3}$$

where the last term is the gradient regularization and the hyper-parameter $\gamma = 10$. The AE and D are trained using a similar procedure to GANs, alternating gradient updates.

B. In-Domain Inversion

We leverage the generative capabilities of the AE by incorporating in-domain GAN inversion. Inspired by Zhu *et al.* [105], we follow a domain-regularized approach that pushes the encoder to create latent code in the semantic domain. In [105], this module enables semantic editing of facial attributes such as expression and pose, while an additional optimization stage improves the reconstructed face in the pixel level. Unlike [105], our approach does not seek to edit facial attributes nor aims to create a faithful reconstruction of the face. Instead, we propose to encode facial attributes in the latent vector that boost the alignment in the landmark detectors for a wide range of expressions.

The inversion is achieved in [105] by introducing a domain-guided encoder to the GANs-based formulation. We instead exploit the pre-trained encoder E from the previous step, as shown in red in Fig. 3. The discriminator D is then used to compete with E, which acts as the domain-guided encoder and refines the latent space z to be aligned with the semantic latent space of the reconstruction process. During this stage, the decoder G is fixed, and E and D take turns to train with the loss functions in (1) and (3), respectively. To that end, the same unlabeled data as in the self-supervised stage is used, where the input of E corresponds to the image

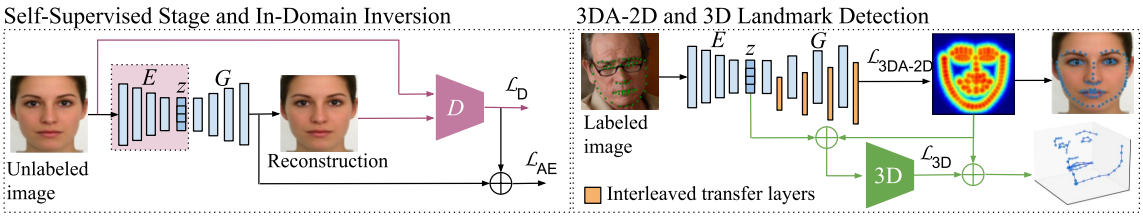

Fig. 3. Architecture of ParFace-Net (PF-Net). Our pipeline consists of a self-supervised stage to train an autoencoder, where the latent code z is disentangled with in-domain inversion. In the supervised stage, z is leveraged by the landmark detector to retrieve 3DA-2D and 3D landmarks from dedicated networks.

x, and the input of D is given by x and the reconstruction $\hat{x}$.

The asymmetrical features in the latent code from palsy patients are refined in this stage, without affecting the reconstruction of symmetrical faces. Hence, the same trained model could be used to align the landmarks during different levels of palsy, to continuously track the recovery.

In contrast to [105], we do not apply the final optimization step to enhance the output of the reconstruction, since we do not aim to create an accurate reconstruction in the pixel level.

C. Supervised Stage

The supervised stage is composed of a 3DA-2D and a 3D landmark detector, where all the face information learned in the self-supervised stage and refined through the in-domain inversion module is available for generalized usage across various landmark datasets.

1) 3DA-2D Landmark Detector: In this stage, the landmark detector maps the disentangled latent code z to 2D heatmaps that represent the probability map of each landmark location. During training, the parameters of the autoencoder are fixed and the layers of the decoder G are interleaved with 3×3 convolutional layers, inspired by 3FabRec [10]. The last convolutional layer that produces the face image is then superseded by a convolutional layer that provides the heatmaps, as shown in Fig. 3.

We propose to adopt the adaptive wing loss (AWing) [87] as the heatmap prediction loss, instead of the mean squared error (MSE) from [10]. Since background pixels on a heatmap dominate over foreground pixels, this loss function penalizes small errors on foreground pixels while tolerating small errors on background pixels. It is formulated as

$$\mathcal{L}_{2D}(h,\hat{h}) = \begin{cases} \omega \ln\left(1+\left|\frac{h-\hat{h}}{\epsilon}\right|^{\alpha-h}\right) & \text{if } |(h-\hat{h})| < \theta \\ A|h-\hat{h}|-C & \text{otherwise,} \end{cases} \tag{4}$$

where h and $\hat{h}$ denote the ground truth and predicted heatmap pixel values, and ω, θ, α, and ϵ are positive values. A and C are added to smooth the loss function at $|h-\hat{h}| = \theta$.

2) 3D Landmark Detector.: We introduce a 3D landmark detector to regress the depth coordinate of the 3DA-2D landmarks. It takes as input the concatenation of the intermediate latent vector and the predicted 3DA-2D landmark heatmaps.

TABLE I

PUBLICLY AVAILABLE DATASETS USED FOR TRAINING THE AUTOENCODER (AE), THE 2D AND 3D LANDMARK DETECTORS, AND FOR TESTING THE CURRENT MODEL.

Dataset	Images / Frames	Train			Test		
		AE	2D	3D	AE	2D	3D
CelebA [65]	202599	✓	-	-	-	-	-
AffectNet [69]	291650	✓	-	-	-	-	-
Menpo 2D [97]	8954	✓	-	-	-	-	-
LS3D-W [11]	7200	✓	-	-	-	-	-
FFHQ [51]	70000	-	-	-	✓	-	-
NeuroFace [2]	3306	✓	-	-	-	-	-
MEEI [29]	12050	✓	-	-	-	-	-
ParFace (Ours)	4200	✓	-	-	-	-	✓
300W-LP [107]	61225	✓	-	✓	-	-	-
AFW [108]	337	-	✓	-	-	-	-
HELEN [59]	2330	-	✓	-	-	✓	-
LFPW [6]	1035	-	✓	-	-	✓	-
300W [76], [75], [74]	600	-	-	-	-	✓	-
iBUG [76], [75], [74]	135	-	-	-	-	✓	-
WFLW [89]	10000	-	✓	-	-	✓	-
AFLW2K-3D [107]	2000	-	-	-	-	-	✓

This detector uses the MSE loss, defined as

$$\mathcal{L}_{3D}(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2, \tag{5}$$

where N is the number of landmarks, and y and $\hat{y}$ are the ground truth and predicted depth values, respectively.

3) Encoder Fine-tuning.: This strategy proposes to further optimize the encoder E along with the Interleaved Transfer Layers (ITL) in tandem [10]. The fine-tuning encourages the encoder to embed more features in the latent code that enhance the landmark predictions. By integrating this step, the identity of the reconstructed face no longer resembles the original image and the reconstruction tends towards an average face, as shown in Fig. 1. Nonetheless, other attributes such as the expression and pose are enhanced.

IV. EXPERIMENTS AND RESULTS

ParFace-Net was implemented in Python using PyTorch. The AE was trained on a Nvidia A100, while the face alignment networks were trained on a Nvidia RTX2080-Ti.

A. Datasets

ParFace-Net is trained on well-known public datasets on face analysis. Table I lists the datasets and in which stage they were used. In the self-supervised stage and during the

in-domain inversion, the AE is trained with multiple datasets, without any type of landmark annotation. We introduced palsy datasets in these stages, such as the Toronto NeuroFace, MEEI and the unlabeled set of ParFace. The 3DA-2D and 3D landmark detectors are trained with 300W-LP. We separately train the 3DA-2D detector with 2D landmarks, to investigate the performance on 2D face alignment. These results are reported in the Supplementary Material.

Palsy Dataset. We introduce ParFace, the first dataset on palsy face alignment with 3D landmarks annotations in video sequences. We collected 28 videos from YouTube of 150 frames each, where the subjects are usually talking to the camera or making a wide range of facial expressions. The videos have varying resolution and cover a wide range of ages, ethnicity, poses, illumination settings and backgrounds. We provide 68 landmarks annotations for 1350 frames in 9 videos, for a total of $\sim$92K annotations.

We developed an annotation tool, which provides an initial 3D landmark estimation by 3D-FAN [11]. Since 3D-FAN was trained on datasets without palsy, each landmark was manually refined to match the asymmetrical facial expressions and provide high quality annotations. This refinement affected most of the 3DA-2D landmarks, and less the depth coordinate. Some sample images are shown in Figure 5 and in the Supplementary Material.

The annotated set of ParFace can be used as a benchmark to evaluate palsy alignment, as in Section IV-E, or to fine-tune semi- or fully supervised approaches as in Section IV-G. The unlabeled set of ParFace can be used for training semi- or self-supervised architectures, similarly to ParFace-Net.

B. Implementation Details

The AE takes as input a cropped version of the face. For labeled datasets, we use the ground truth landmarks to compute the bounding box, following related works. Otherwise, we use the MTCNN face detector [100]. Faces with a height less than 100px are discarded. The data is augmented with random horizontal flipping (50%), translation ($\pm 4\%$), scale jittering (94% to 103%) and rotation (between $\pm 45°$).

1) Model Architecture.: The autoencoder consists of a ResNet-18 [39], which encodes a 99-dimensional latent vector, and an inverted ResNet-18 [5] for decoding. The perceptual loss uses layers ϕ [3, 8, 15, 22] of a VGG-19 [81] pre-trained on ImageNet [19]. In the supervised stage, the 3DA-2D landmark detector is an inverted ResNet-18 that outputs landmark heatmaps of size 128 × 128 with N channels, where N is the number of landmarks. The 3D landmark detector is a ResNet-18, which regresses the depth. This coordinate is normalized to lie between [-1, 1] to achieve faster convergence and numerical stability.

2) Training Details.: We use the Adam optimizer [56] with a learning rate of 2e-5, $\beta_1 = 0.0$ and $\beta_2 = 0.999$. The autoencoder is trained with input and output images of size 256 × 256. We train for 50 epochs with (1), where $\mathcal{L}_{rec}$ is the L2 loss, followed by 50 epochs with the L1 loss as $\mathcal{L}_{rec}$. After that, we fix the decoder G and optimize the encoder E for feature disentanglement against the discriminator D with the L2 loss as the $\mathcal{L}_{rec}$, for 50 epochs.

The 3DA-2D landmark detector is trained for 100 epochs to predict the heatmaps. We fine-tune the encoder with gradients from the landmark head for 100 epochs. A similar procedure is followed in the experiments to train with 2D landmarks. For 3D face alignment, the 3D landmark detector is trained with the ground truth 3DA-2D landmark heatmaps for 50 epochs.

C. Evaluation Metrics

Following the standard protocol, we adopt the normalized mean error (NME) to evaluate 3DA-2D face alignment on ALFW2000-3D and ParFace. We additionally report the failure rate (FR) and area under the curve (AUC) at 10% of the Cumulative Error Distribution (CED) on ParFace. 3D face alignment is evaluated using the ground truth error (GTE) on AFLW2000-3D and ParFace. The GTE is equivalent to the NME, but evaluates the full 3D coordinates. The GTE is normalized by the inter-ocular (IO) distance, while the NME is normalized by the square-root of the bounding box size enclosing the landmarks, following related works. We report the standard deviations σ of the NME and GTE in ParFace.

D. Evaluation on AFLW2000-3D

3DA-2D and 3D face alignment are evaluated on the widely used benchmark AFLW2000-3D, where the landmark detectors are trained using 300W-LP. We employ the AE with in-domain inversion to train the landmark detectors. Furthermore, we refine the 3DA-2D landmarks with encoder fine-tuning. The results are shown in Table II.

We observed that our models outperform the SOTA in 3DA-2D face alignment (NME) for frontal and near frontal faces (0 to 30°) and our ParFace-Net with the AWing loss has the 2nd best GTE on 3D face alignment for the reported methods. For larger poses, we noticed a decreased performance of ParFace-Net. This could be attributed to the small portion of non-frontal faces in the self-supervised stage, where face semantics are learned mostly for near-frontal poses. We also observed that model-based approaches tend to be more robust to large head poses, since they are trained with additional 3DMM parameters such as head orientation and face shape. However, as shown in the next section, model-based methods do not cope well with a wide range of facial expressions, including asymmetrical expressions.

E. Evaluation on ParFace

The annotated set of ParFace is employed to evaluate our models from Section IV-D, with no annotations for palsy. The results are shown in Table III. We additionally report the performance of different SOTA model-based and model-free methods for 3D face alignment, which have been trained on 300W-LP or 3DMMs datasets. To discard alignment errors due to face detection inaccuracies, we replaced the face detectors in every method and provided the bounding box from the ground truth landmarks, to crop the input images.

The CED curves for the normalized 3DA-2D and 3D RMSE are shown in Figure 4. Our models achieve the lowest

TABLE II

NME (3DA-2D) AND GTE (3D) ON AFLW2000-3D, FOR DIFFERENT YAW ANGLES. † REPORTED IN [99]. METHODS WITH * ARE SEMI-SUPERVISED.

	Method	NME↓ 0 to 30	30 to 60	60 to 90	All	GTE ↓
Model-based	3DDFA [107]	2.84	3.57	4.96	3.79	-
	SPDT [72]*	3.56	4.06	**4.11**	3.88	-
	3DDFA_V2 [36]	2.63	3.42	4.48	3.51	-
	2DAL [84]*	2.75	3.46	4.45	3.55	-
	SADRNet [73]	2.66	**3.30**	4.42	3.46	-
	SynergyNet [88]	2.66	**3.30**	4.27	**3.41**	-
Model-free	SDM [91]†	3.67	4.94	9.76	6.12	-
	3D-FAN [11]†	2.77	3.48	4.60	3.62	7.45
	JVCR [99]	2.94	3.46	4.53	3.64	**7.28**
	StyleGAN-FA[24]*	2.65	3.62	4.89	3.72	-
	PF-Net$_{MSE}$*	2.62	3.65	4.80	3.69	7.42
	PF-Net$_{Awing}$*	**2.61**	3.67	4.74	3.67	7.38

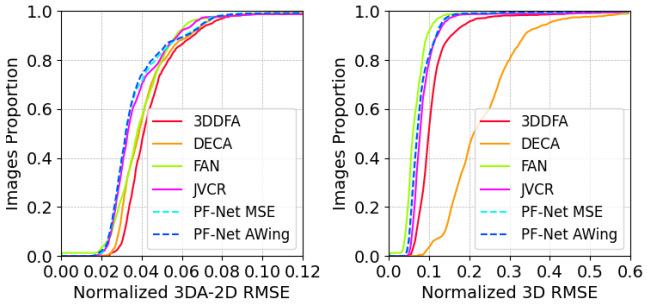

Fig. 4. CED curves for 3DA-2D and 3D face alignment of models tested on ParFace. Our models are 'PF-Net: MSE' and 'PF-Net: Awing', represented with dotted lines.

NME and FR, the highest AUC and the 2nd and 3rd lowest GTE on ParFace. As mentioned in Section IV-A, for labelling ParFace, an initial landmark prediction was computed using 3D-FAN. However, the 3DA-2D landmarks were heavily refined, while the z coordinates were refined to a lesser extent. As expected, 3D-FAN has the lowest GTE in this dataset. Qualitative results are shown in Figures 5 and 6. Table III and Figure 5 show that model-free methods have in general a better performance and are more flexible on asymmetrical expressions than model-based pipelines.

TABLE III

PERFORMANCE ON PARFACE. THE NME, AUC AND FR EVALUATE 3DA-2D LANDMARKS, WHILE GTE EVALUATES 3D ALIGNMENT.

	Method	NME $\pm\sigma$ ↓	AUC$_{10}$↑	FR$_{10}$↓	GTE $\pm\sigma$↓
Model-based	3DDFA_V2 [36]	4.65 ± 2.51	55.14	1.11	11.18 ± 5.98
	SynergyNet [88]	5.49 ± 4.96	49.48	1.70	13.89 ± 8.48
	DECA [25]	4.24 ± **1.19**	57.81	0.88	23.27 ± 6.27
Model-free	3D-FAN [11]	4.26 ± 3.64	59.83	0.96	**7.06** ± 6.20
	JVCR [99]	4.00 ± 2.79	61.87	1.26	8.79 ± 4.91
	PF-Net$_{MSE}$	3.83 ± 3.19	62.74	0.96	8.14 ± 4.60
	PF-Net$_{AWing}$	**3.79** ± 2.98	**62.81**	**0.82**	8.05 ± **4.37**

F. Runtime and Model Parameters

We measured for ParFace-Net a runtime of ∼230FPS on average for 1K repetitions, for 2D and 3DA-2D face alignment, on a Nvidia RTX2080-Ti. To estimate the full 3D coordinates, ParFace-Net runs at ∼156FPS.

ParFace-Net is composed of two ResNet-18 and an inverted ResNet-18, with a total of ∼24.14M parameters. 3D-FAN is composed of 4 HG networks with ∼24M parameters and a ResNet-152 with ∼58.5M parameters to compute the depth coordinate (in total ∼82.5M). Likewise, JVCR uses 4 stacked HG and an additional network to map the voxels to coordinates, with 32.47M parameters in total. SynergyNet has 4.6M parameters, 3DDFA_V2 3.27M and DECA uses two ResNet-50 with more than 25M parameters each and multiple decoders to retrieve the parameters of the 3DMM.

G. Ablation Study

We evaluate the contribution of each module in the face alignment process.

1) Training the Self-Supervised Stage: The impact of the self-supervised stage in the landmark detection task is analyzed. For that purpose, we trained the landmark detectors omitting the self-supervised stage and only the encoder is pre-trained on ImageNet. Since the latent code does not encode face information, the in-domain inversion is not applied either. The encoder is later fine-tuned after training the 3DA-2D landmark detector, as detailed in Section III-C. The results are shown in Table IV, without check marks in the categories 'Self-Supervision' and 'In-Domain Inversion'. For every metric, there is a large decline in the performance when the self-supervised stage is omitted.

2) Training with In-Domain Inversion: We investigated the effect of the in-domain inversion module as well. To that end, we trained the landmark detectors before and after the in-domain inversion is applied. The results for 3DA-2D and 3D face alignment are shown in Table IV. We observed that in-domain inversion boosted the performance in every metric *w.r.t.* the model without inversion. The improvement is more noticeable for ParFace, both in the NME and GTE.

3) Effect of AWing Loss: We additionally examined the performance of the 3DA-2D landmark detector using the MSE loss and the proposed AWing loss. The results are reported in Table IV. During the experiments, we observed that the MSE loss converged faster, but in overall the AWing loss leads to improved accuracy in most of the metrics. We hypothesize that this is due to AWing loss being more sensitive to foreground pixels than background pixels, considering that background pixels predominate in the heatmaps.

4) Training the AE with Portions of the Data: As an additional ablation study, we explore how the performance of the landmark detectors are affected when the self-supervised stage is trained with different portions of the data. The quantitative results for AFLW2000-3D and ParFace are reported in Table V. To that end, we trained the AE with multiple combinations of the datasets from Table I, where the total amounts to ∼590K images. Note that only the models with 3% and 100% included palsy data, and that all the models

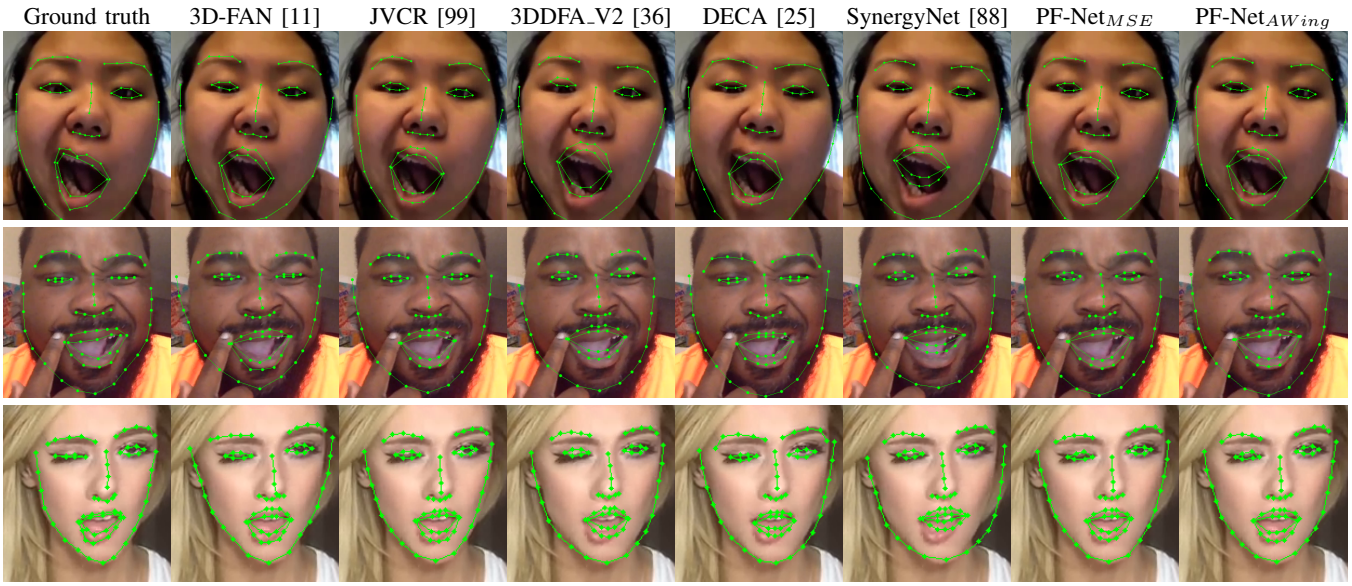

Fig. 5. Face alignment on ParFace. 3D-FAN, JVCR, 3DDFA_V2 and SynergyNet were trained with 300W-LP, while DECA uses pseudo ground truth from 3D-FAN. 3DDFA_V2, DECA and SynergyNet are trained on datasets for face reconstruction to fit 3DMMs as well.

TABLE IV

ABLATION STUDY ON AFLW2000-3D AND PARFACE. WE EVALUATE THE IMPACT OF THE SELF-SUPERVISED STAGE, IN-DOMAIN INVERSION AND THE LOSS FUNCTION FOR LANDMARK DETECTION. THE NME AND GTE ARE AVERAGED IN EACH DATASET ON THE TOTAL NUMBER OF IMAGES.

Self-Supervision	In-Domain Inversion	Alignment Loss	AFLW2000-3D				ParFace			
			NME $\pm\sigma$ $\downarrow$	AUC$_{10}^{\uparrow}$	FR$_{10}^{\downarrow}$	GTE $\pm\sigma$ $\downarrow$	NME $\pm\sigma$ $\downarrow$	AUC$_{10}^{\uparrow}$	FR$_{10}^{\downarrow}$	GTE $\pm\sigma$ $\downarrow$
-	-	MSE	3.91 $\pm$ 3.94	65.31	3.56	8.81 $\pm$ 6.87	4.24 $\pm$ 3.27	58.39	1.11	9.16 $\pm$ 4.89
✓	-	MSE	3.23 $\pm$ **2.90**	69.40	1.50	7.77 $\pm$ 6.22	3.96 $\pm$ 3.34	61.75	0.96	8.40 $\pm$ 4.87
✓	✓	MSE	3.15 $\pm$ 2.96	70.17	1.50	7.42 $\pm$ 5.33	3.83 $\pm$ 3.19	62.74	0.96	8.14 $\pm$ 4.60
-	-	AWing	3.96 $\pm$ 3.97	64.81	3.65	9.03 $\pm$ 6.93	4.46 $\pm$ 3.56	56.17	1.04	9.47 $\pm$ 5.27
✓	-	AWing	3.19 $\pm$ 2.95	69.78	1.45	7.45 $\pm$ 5.07	3.94 $\pm$ 3.21	61.76	0.96	8.41 $\pm$ 4.75
✓	✓	AWing	**3.14** $\pm$ 2.92	**70.22**	**1.35**	**7.38** $\pm$ **5.05**	**3.79** $\pm$ **2.98**	**62.81**	**0.82**	**8.05** $\pm$ **4.37**

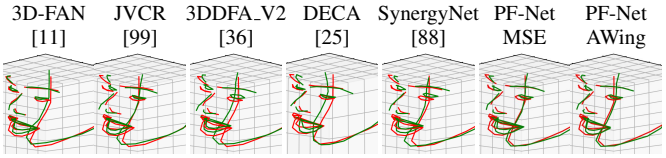

Fig. 6. 3D face alignment on ParFace with different methods of the SOTA. Ground truth in red and predictions in green.

were trained with in domain inversion, the AWing loss and encoder fine-tuning.

As part of the experiments, we trained the AE only with palsy datasets at our disposal: Toronto NeuroFace, MEEI and the unlabeled set of ParFace. The results correspond to 3% of the total data in Table V. From this experiment, we observed a comparable performance on ParFace with the model trained with 1%, with a minimal improvement on the model trained with palsy data. However, on AFLW2000-3D, the model trained with palsy data showed in general a slightly lower performance than the model trained with only 1% of the data. The main reason is that the Toronto NeuroFace and MEEI are clinical datasets collected in controlled conditions, with little diversity in terms of pose, lighting and background setting. Therefore, a model trained with relatively few images in the wild (in this case from ParFace), would not perform well on challenging images with large poses, occlusion and varying lighting, such as in AFLW2000-3D, due to insufficient data to generate a compact face representation embedded in the latent code.

Overall, the alignment performance shows a gradual improvement as more data is added to the self-supervised stage. These results lead to the assumption that the landmark detectors can be further enhanced as more unlabeled data with large diversity is used for training the AE.

5) Training with Labeled Palsy Data: The results in Section IV-E were computed from models that were not trained using labeled data from ParFace. To evaluate the influence of labeled palsy data in our approach, we split the dataset into a training and test set and fine-tune the previously trained models with a portion of the data. The results are shown in Table VI. We use 6 sequences for training and 3 for testing. We split the training set into 6 parts, each containing N number of sequences, where N is in the range [1,6]. The number of sequences used are added as numerals in the Table. MSE-0 and Awing-0 use the models trained

TABLE V

ABLATION STUDY ON 3DA-2D AND 3D FACE ALIGNMENT AFTER TRAINING THE SELF-SUPERVISED STAGE WITH PORTIONS OF THE DATA. THE AE IN * WAS TRAINED ONLY WITH PALSY DATA.

%	AFLW2000-3D				ParFace			
	NME $\pm \sigma^{\downarrow}$	$AUC_{10}^{\uparrow}$	$FR_{10}^{\downarrow}$	GTE $\pm \sigma^{\downarrow}$	NME $\pm \sigma^{\downarrow}$	$AUC_{10}^{\uparrow}$	$FR_{10}^{\downarrow}$	GTE $\pm \sigma^{\downarrow}$
0	3.96 ± 3.97	64.81	3.65	9.03 ± 6.93	4.46 ± 3.56	56.17	1.04	9.47 ± 5.27
1	3.57 ± 3.36	67.31	2.45	8.20 ± 5.94	4.20 ± 3.42	58.87	1.19	9.06 ± 5.15
3*	3.58 ± 3.24	66.97	2.65	8.16 ± 5.59	4.19 ± 3.31	58.86	0.89	9.06 ± 4.94
13	3.35 ± 3.09	68.94	1.70	7.70 ± 5.42	3.84 ± 2.85	61.89	0.89	8.65 ± 4.34
34	3.26 ± 2.94	69.35	1.55	7.55 ± 5.07	3.96 ± 3.33	61.52	0.82	8.30 ± 4.94
97	3.17 ± 2.96	70.07	1.30	7.46 ± 5.14	3.82 ± 2.98	62.38	0.82	8.15 ± 4.41
100	**3.14** ± 2.92	**70.22**	**1.35**	**7.38 ± 5.05**	**3.79 ± 2.98**	**62.81**	**0.82**	**8.05 ± 4.37**

TABLE VI

PERFORMANCE METRICS ON PARFACE AFTER FINE-TUNING WITH TRAINING SETS FROM PARFACE. THE NUMBER OF SEQUENCES USED FOR FINE-TUNING ARE ADDED AS NUMERALS ADJACENT TO THE LOSS.

Method	NME$^{\downarrow}$	$AUC_{10}^{\uparrow}$	$FR_{10}^{\downarrow}$	GTE$^{\downarrow}$
MSE-0	3.99	61.10	1.11	8.79
MSE-1	4.41	56.81	1.11	9.70
MSE-2	4.13	58.63	0.44	8.98
MSE-3	3.60	63.70	0.44	8.34
MSE-4	3.45	65.22	0.44	8.03
MSE-5	3.43	65.39	0.44	7.93
MSE-6	3.22	67.43	0.44	7.99
AWing-0	3.92	61.45	0.89	8.72
AWing-1	4.17	58.70	0.89	8.94
AWing-2	4.08	59.06	0.44	8.50
AWing-3	3.77	62.11	0.44	8.44
AWing-4	3.55	64.12	**0.22**	7.87
AWing-5	3.53	64.47	0.44	8.10
AWing-6	**3.18**	**67.77**	**0.22**	**7.44**

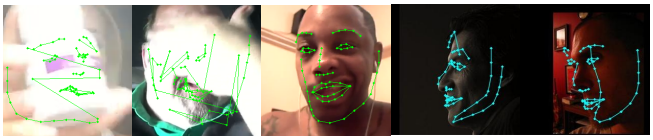

Fig. 7. Failure cases.

without palsy data, to evaluate the test set of ParFace (450 images in total).

The results show an overall improvement when more data is added to fine-tune the models. By adding $6 * 150 = 900$ training images with palsy data to a model trained with more than 61K labeled images (representing less than 2%), we obtained performance gains of around 20% in the NME and more than 10% in the GTE.

These experiments validate the use of ParFace to fine-tune semi- or fully supervised DNN for 3DA-2D and 3D face alignment on asymmetrical facial expressions.

H. Discussion and Limitations

As most of the heatmap-based methods for face alignment, our approach fails under extreme occlusions, as shown in Figure 7. Other failure cases occur when the face synthesis fails due to unusual facial expressions, large head poses, lighting and low contrast. 3DMMs-based methods are more robust in such cases, by keeping the spatial structure of the landmarks, even if the face is not properly aligned. However, they are not able to align the landmarks correctly for unseen faces, such as in asymmetrical expressions (see DECA in Fig. 5).

We noticed that our method heavily depends on the training data in the self-supervised stage. By using datasets with less diversity in terms of pose, expression, occlusion and illumination, the performance of the landmark detectors drops. The dependency on unlabeled data is not to be seen as a drawback, since the generation of such datasets is much less expensive than labeled data. We also observed that using a small set of unlabeled palsy faces (∼3% of the total amount, see Table I) to train the AE enabled the in-domain inversion module to encode asymmetrical features in the latent vector, improving the landmark detection.

Dedicated architectures for HQ face reconstruction such as StyleGAN2 [52] could replace the inverted ResNet-18, as in [24], to improve the reconstruction. This comes at a cost of increased complexity and trainable parameters in the pipeline. While StyleGAN2 has ∼28M parameters and a computational complexity of 143.15 Giga Multiply-Accumulate Operations (GMACs), ResNet-18 has ∼11M parameters and a complexity of 1.82GMAC.

V. CONCLUSIONS

This work introduced a pipeline for 3D face alignment, targeted to faces with symmetrical and asymmetrical expressions. We propose a semi-supervised architecture which exploits large unlabeled datasets and integrates face alignment with smaller labeled datasets. We explore the latent space in the self-supervised stage, and optimize the encoder to produce a disentangled latent space with in-domain inversion. Our landmark detector uses the AWing loss to regress 3DA-2D landmark heatmaps and a newly introduced separate branch computes the depth of the 3D landmarks. A future direction would be to exploit additional 3DMMs parameters, to enable the autoencoder to learn the pose, expression, and shape from 2D images under large head poses and extreme occlusion.

REFERENCES

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *CVPR*, pages 1859–1866. IEEE/CVF, 2014.

[2] A. Bandini, S. Rezaei, D. L. Guarin, M. Kulkarni, D. Lim, M. I. Boulos, L. Zinman, Y. Yunusova, and B. Taati. A new dataset for facial motion analysis in individuals with neurological disorders. *JBHI*, 25(4):1111–1119, 2020.

[3] J. Barbosa, W. Seo, and J. Kang. parafacetest: an ensemble of regression tree-based facial features extraction for efficient facial paralysis classification. *BMC Medical Imaging*, 19(1):1–14, 2019.

[4] G. Barrios Dell'Olio and M. Sra. Farapy: An augmented reality feedback system for facial paralysis using action unit intensity estimation. In *ACM Symposium on User Interface Software and Technology*, pages 1027–1038, 2021.

[5] J. Behrmann, W. Grathwohl, R. T. Chen, D. Duvenaud, and J. Jacobsen. Invertible residual networks. In *ICML*, pages 573–582. PMLR, 2019.

[6] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *PAMI*, 35(12):2930–2940, 2013.

[7] C. Bhagavatula, C. Zhu, K. Luu, and M. Savvides. Faster than real-time facial alignment: A 3D spatial transformer network approach in unconstrained poses. In *ICCV*, pages 3980–3989. IEEE, 2017.

[8] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[9] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou. 3D face morphable models "in-the-wild". In *CVPR*, pages 5464–5473. IEEE/CVF, 2017.

[10] B. Browatzki and C. Wallraven. 3FabRec: Fast few-shot face alignment by reconstruction. In *CVPR*, pages 6110–6120. IEEE, 2020.

[11] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem?(and a dataset of 230,000 3D facial landmarks). In *ICCV*, pages 1021–1030. IEEE, 2017.

[12] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3D shape regression for real-time facial animation. *TOG*, 32(4), 2013.

[13] P. Chandran, G. Zoss, P. Gotardo, and D. Bradley. Continuous landmark detection with 3d queries. In *CVPR*, pages 16858–16867. IEEE/CVF, 2023.

[14] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *PAMI*, 23(6):681–685, 2001.

[15] T. Cootes and C. Taylor. Active shape models – 'smart snakes'. In *BMVC*, pages 266–275. Springer, 1992.

[16] T. Cootes and C. Taylor. Active shape model search using local grey-level models: A quantitative evaluation. In *BMVC*, volume 93, pages 639–648. Springer, 1993.

[17] A. Dapogny, K. Bailly, and M. Cord. Decafa: Deep convolutional cascade for face alignment in the wild. In *ICCV*, pages 6893–6901. IEEE, 2019.

[18] J. Delannoy and T. Ward. A preliminary investigation into the use of machine vision techniques for automating facial paralysis rehabilitation therapy. In *Irish Signals and Systems Conference (ISSC 2010)*. IET, 2010.

[19] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE/CVF, 2009.

[20] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, and S. Zafeiriou. The menpo benchmark for multi-pose 2D and 3D facial landmark localisation and tracking. *IJCV*, pages 1–26, 2018.

[21] D. Derkach, A. Ruiz, and F. M. Sukno. Head pose estimation based on 3-D facial landmarks localization and regression. In *FG*, pages 820–827. IEEE, May 2017.

[22] J. M. Díaz Barros, V. Golyanik, K. Varanasi, and D. Stricker. Face it!: A pipeline for real-time performance-driven facial animation. In *ICIP*, pages 2209–2213. IEEE, 2019.

[23] X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Style aggregated network for facial landmark detection. In *CVPR*, pages 379–388. IEEE/CVF, 2018.

[24] M. Dornier, P. Gosselin, P. Raymond, Y. Ricquebourg, and B. Coüasnon. Stylegan-based heatmap generator for face alignment with limited training data. *hal-03778322v2*, 2023.

[25] Y. Feng, H. Feng, M. J. Black, and T. Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *TOG*, 40(4):1–13, 2021.

[26] A. Gaber, M. F. Taher, and M. A. Wahed. Quantifying facial paralysis using the kinect v2. In *EMBC*, pages 2497–2501. IEEE, 2015.

[27] X. Ge, J. M. Jose, P. Wang, A. Iyer, X. Liu, and H. Han. Adaptive local-global relational network for facial action units recognition and facial paralysis estimation. *arXiv preprint arXiv:2203.01800*, 2022.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *NIPS*, 27, 2014.

[29] J. J. Greene, D. L. Guarin, J. Tavares, E. Fortier, M. Robinson, J. Dusseldorp, O. Quatela, N. Jowett, and T. Hadlock. The spectrum of facial palsy: The meei facial palsy photo and video standard set. *The Laryngoscope*, 130:32–37, 2020.

[30] I. Grishchenko, A. Ablavatski, Y. Kartynnik, K. Raveendran, and M. Grundmann. Attention mesh: High-fidelity face mesh prediction in real-time. *arXiv preprint arXiv:2006.10962*, 2020.

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

[32] L. Gu and T. Kanade. 3D alignment of face in a single image. In *CVPR*, volume 1, pages 1305–1312. IEEE/CVF, 2006.

[33] D. Guarin, J. Dusseldorp, T. Hadlock, and N. Jowett. A machine learning approach for automated facial measurements in facial palsy. *JAMA facial plastic surgery*, 20(4):335–337, 2018.

[34] D. L. Guarin, Y. Yunusova, B. Taati, J. R. Dusseldorp, S. Mohan, J. Tavares, M. M. van Veen, E. Fortier, T. A. Hadlock, and N. Jowett. Toward an automatic system for computer-aided assessment in facial palsy. *FPSAM*, 22:42–49, 2020.

[35] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. *NIPS*, 30, 2017.

[36] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li. Towards fast, accurate and stable 3D dense face alignment. In *ECCV*, pages 152–168. Springer, 2020.

[37] Z. Guo, G. Dan, J. Xiang, J. Wang, W. Yang, H. Ding, O. Deussen, and Y. Zhou. An unobtrusive computerized assessment framework for unilateral peripheral facial paralysis. *JBHI*, 22(3):835–841, 2017.

[38] Z. Guo, W. Li, J. Dai, J. Xiang, and G. Dan. Facial imaging and landmark detection technique for objective assessment of unilateral peripheral facial paralysis. *Enterprise Information Systems*, pages 1–17, 2021.

[39] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE, 2016.

[40] B. Hontanilla and D. Marré. Comparison of hemihypoglossal nerve versus masseteric nerve transpositions in the rehabilitation of short-term facial paralysis using the facial clima evaluating system. *Plastic and Reconstructive Surgery*, 130(5):662e–672e, 2012.

[41] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for MobileNetv3. In *CVPR*, pages 1314–1324, 2019.

[42] G. J. Hsu, W. Huang, and J. Kang. Hierarchical network for facial palsy detection. In *CVPR-W*, pages 580–586, 2018.

[43] G. J. Hsu, J. Kang, and W. Huang. Deep hierarchical network with line segment learning for quantitative analysis of facial palsy. *IEEE Access*, 7:4833–4842, 2018.

[44] A. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3D face reconstruction from a single image via direct volumetric cnn regression. In *ICCV*, pages 1031–1039. IEEE, 2017.

[45] L. Jeni, J. Cohn, and T. Kanade. Dense 3D face alignment from 2D video for real-time use. *IMAVIS*, 58:13–24, 2017.

[46] C. Jiang, J. Wu, W. Zhong, M. Wei, J. Tong, H. Yu, and L. Wang. Automatic facial paralysis assessment via computational image analysis. *Journal of Healthcare Engineering*, 2020, 2020.

[47] H. Jin, S. Liao, and L. Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *IJCV*, 129(12):3174–3194, 2021.

[48] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[49] A. Jourabloo and X. Liu. Large-pose face alignment via CNN-based dense 3D model fitting. In *CVPR*, pages 4188–4196. IEEE/CVF, 2016.

[50] F. Kahraman, M. Gokmen, S. Darkner, and R. Larsen. An active illumination and appearance (AIA) model for face alignment. In *CVPR*, pages 1–7. IEEE/CVF, 2007.

[51] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410. IEEE/CVF, 2019.

[52] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119. IEEE/CVF, 2020.

[53] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann. Real-time facial surface geometry from monocular video on mobile

gpus. *arXiv preprint arXiv:1907.06724*, 2019.
[54] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*. IEEE/CVF, 2014.
[55] H. S. Kim, S. Y. Kim, Y. H. Kim, and K. S. Park. A smartphone-based automatic diagnosis system for facial nerve palsy. *Sensors*, 15(10):26756–26768, 2015.
[56] D. Kingma. Adam: a method for stochastic optimization. In *ICLR*, 2014.
[57] M. Kowalski, J. Naruniec, and T. Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *CVPR-W*, pages 88–97. IEEE, 2017.
[58] A. Kumar, T. Marks, W. Mou, Y. Wang, M. Jones, A. Cherian, T. Koike-Akino, X. Liu, and C. Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *CVPR*, pages 8236–8246. IEEE/CVF, 2020.
[59] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. Huang. Interactive facial feature localization. In *ECCV*, pages 679–692. Springer, 2012.
[60] H. Li, Z. Guo, S. Rhee, S. Han, and J. Han. Towards accurate facial landmark detection via cascaded transformers. In *CVPR*, pages 4176–4185. IEEE/CVF, 2022.
[61] F. Liu, D. Zeng, Q. Zhao, and X. Liu. Joint face alignment and 3D face reconstruction. In *ECCV*, pages 545–560. Springer, 2016.
[62] F. Liu, Q. Zhao, D. Zeng, et al. Joint face alignment and 3D face reconstruction with application to face recognition. *PAMI*, 2018.
[63] X. Liu, Y. Xia, H. Yu, J. Dong, M. Jian, and T. D. Pham. Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation. *TNSRE*, 28(10):2325–2332, 2020.
[64] Y. Liu, Z. Xu, L. Ding, J. Jia, and X. Wu. Automatic assessment of facial paralysis based on facial landmarks. In *PRML*, pages 162–167. IEEE, 2021.
[65] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738. IEEE, 2015.
[66] P. Micaelli, A. Vahdat, H. Yin, J. Kautz, and P. Molchanov. Recurrence without recurrence: Stable video landmark detection with deep equilibrium models. In *CVPR*, pages 22814–22825. IEEE/CVF, 2023.
[67] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *ECCV*, pages 504–513. Springer, 2008.
[68] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
[69] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.*, pages 18–31, 2017.
[70] T. Ngo, M. Seo, N. Matsushiro, W. Xiong, and Y.-W. Chen. Quantitative analysis of facial paralysis based on limited-orientation modified circular gabor filters. In *ICPR*, pages 349–354. IEEE, 2016.
[71] G. Parra-Dominguez, R. Sanchez-Yanez, and C. Garcia-Capulin. Facial paralysis detection on images using key point analysis. *Applied Sciences*, 11(5):2435, 2021.
[72] J. Piao, C. Qian, and H. Li. Semi-supervised monocular 3d face reconstruction with end-to-end shape-preserved domain transfer. In *ICCV*, pages 9398–9407, 2019.
[73] Z. Ruan, C. Zou, L. Wu, G. Wu, and L. Wang. SADRNet: Self-aligned dual face regression networks for robust 3D dense face alignment and reconstruction. *Trans. on Image Processing*, 2021.
[74] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *IMAVIS*, 47:3–18, 2016.
[75] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV-W*, pages 397–403. IEEE, 2013.
[76] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPR-W*, pages 896–903. IEEE/CVF, 2013.
[77] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *CVPR*, pages 7763–7772. IEEE, June 2019.
[78] J. Saragih and R. Goecke. A nonlinear discriminative approach to aam fitting. In *ICCV*, pages 1–8. IEEE, 2007.
[79] J. Saragih, S. Lucey, and J. Cohn. Deformable model fitting by regularized landmark mean-shift. *IJCV*, 91(2):200–215, 2011.
[80] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *ICCV*, pages 1576–1585. IEEE, 2017.
[81] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.
[82] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.
[83] A. Taufique, A. Savakis, and J. Leckenby. Automatic quantification of facial asymmetry using facial landmarks. In *Western New York Image and Signal Processing Workshop (WNYISPW)*, pages 1–5. IEEE, 2019.
[84] X. Tu, J. Zhao, M. Xie, Z. Jiang, A. Balamurugan, Y. Luo, Y. Zhao, L. He, Z. Ma, and J. Feng. 3D face reconstruction from a single image assisted by 2D face images in the wild. *IEEE Transactions on Multimedia*, 23:1160–1172, 2021.
[85] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *CVPR*, pages 3659–3667. IEEE/CVF, 2015.
[86] T. Wang, J. Dong, X. Sun, S. Zhang, and S. Wang. Automatic recognition of facial movement for paralyzed face. *Bio-medical materials and engineering*, 24(6):2751–2760, 2014.
[87] X. Wang, L. Bo, and L. Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *ICCV*, pages 6971–6981. IEEE, 2019.
[88] C. Wu, Q. Xu, and U. Neumann. Synergy between 3DMM and 3D landmarks for accurate 3D facial geometry. In *3DV*, pages 453–463. IEEE, 2021.
[89] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, pages 2129–2138. IEEE, June 2018.
[90] Y. Xia, C. Nduka, R. Y. Kannan, E. Pescarini, J. E. Berner, and H. Yu. AFLFP: A database with annotated facial landmarks for facial palsy. *Transactions on Computational Social Systems*, 2022.
[91] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539. IEEE/CVF, 2013.
[92] H. Yoshihara, M. Seo, T. Ngo, N. Matsushiro, and Y. Chen. Automatic feature point detection using deep convolutional networks for quantitative evaluation of facial paralysis. In *CISP-BMEI*, pages 811–814. IEEE, 2016.
[93] B. Yu and D. Tao. Heatmap regression via randomized rounding. *PAMI*, 2021.
[94] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *ICCV*, pages 1944–1951. IEEE, 2013.
[95] A. Zadeh, Y. Chong Lim, T. Baltrusaitis, and L. Morency. Convolutional experts constrained local model for 3D facial landmark detection. In *ICCV Workshops*, pages 2519–2528. IEEE, 2017.
[96] S. Zafeiriou, G. Chrysos, A. Roussos, E. Ververas, J. Deng, and G. Trigeorgis. The 3D menpo facial landmark tracking challenge. In *ICCV-W*, pages 2503–2511, 2017.
[97] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *CVPR-W*, 2017.
[98] L. Zeng, L. Chen, W. Bao, Z. Li, Y. Xu, J. Yuan, and N. K. Kalantari. 3D-aware facial landmark detection via multi-view consistent training on synthetic data. In *CVPR*, pages 12747–12758. IEEE/CVF, 2023.
[99] H. Zhang, Q. Li, and Z. Sun. Joint voxel and coordinate regression for accurate 3D facial landmark localization. In *ICPR*, pages 2202–2208. IEEE, 2018.
[100] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, pages 1499–1503, 2016.
[101] Y. Zhao, G. Feng, H. Wu, S. Aodeng, X. Tian, G. F. Volk, O. Guntinas-Lichius, and Z. Gao. Prognostic value of a three-dimensional dynamic quantitative analysis system to measure facial motion in acute facial paralysis patients. *Head & face medicine*, 16(1):1–10, 2020.
[102] W. Zhong, C. Fang, Y. Cai, P. Wei, G. Zhao, L. Lin, and G. Li. Identity-preserving talking face generation with landmark and appearance priors. In *CVPR*, pages 9729–9738, 2023.
[103] Y. Zhou, J. Deng, I. Kotsia, and S. Zafeiriou. Dense 3D face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In *CVPR*, pages 1097–1106. IEEE/CVF, 2019.
[104] Z. Zhou, H. Li, H. Liu, N. Wang, G. Yu, and R. Ji. Star loss: Reducing semantic ambiguity in facial landmark detection. In *CVPR*, pages 15475–15484, 2023.
[105] J. Zhu, Y. Shen, D. Zhao, and B. Zhou. In-domain gan inversion for real image editing. In *ECCV*, pages 592–608. Springer, 2020.
[106] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, pages 4998–5006. IEEE/CVF, 2015.

[107] X. Zhu, X. Liu, Z. Lei, and S. Z. Li. Face alignment in full pose range: A 3D total solution. *PAMI*, 41(1):78–92, 2017.
[108] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886. IEEE, 2012.