CONFIDENCE-AWARE CLUSTERED LANDMARK FILTERING FOR HYBRID 3D FACE TRACKING

Jilliam María Díaz Barros[♭§] *Chen-Yu Wang*[♭] *Didier Stricker*[♭§] *Jason Rambach*[§]

[♭] RPTU Kaiserslautern [§] German Research Center for Artificial Intelligence (DFKI)

ABSTRACT

The detection of facial landmarks in 2D images has received a great attention in the last decade, as it is a key step for several computer-vision-related applications. Most of the approaches are focused on still images, and are extended to videos by using a tracking-by-detection scheme. In this work, we propose a frame-to-frame tracking module based on grouped-landmark Kalman filters that can be integrated into existing deep-learning-based 3D face alignment pipelines. This method improves the landmark accuracy in cases with large occlusion, extreme head poses and blurriness that affect existing approaches. Our experiments on the Menpo 3DA-2D benchmark show improvements on model-free and 3D-model-based face alignment approaches.

***Index Terms*—** Face alignment, tracking, landmarks, Kalman filter

1. INTRODUCTION

Face analysis is a widely researched field in computer vision, with multiple applications in medicine, driving assistance, social networking, among others. Facial landmark detection, often termed as face alignment (FA), is relevant for applications in face analysis such as face reconstruction [1, 2, 3, 4], head pose estimation and performance capture [5, 6]. In such cases, a sparse [7, 8] or dense set [2, 9, 10, 11] of fiducial facial points (landmarks) are registered in 2D images or videos.

The landmarks can be defined in 2D or 3D coordinates. 3D landmarks maintain correspondence across multiple poses, while 2D landmarks suffer from discrepancies, for example, for frontal and profile faces [2]. To alleviate this problem, a new set of landmarks referred to as 3DA-2D were defined [12]. These landmarks correspond to the projection of the 3D landmarks in the image space and also have one-to-one correspondences across different poses [4].

In the literature, landmarks are represented with coordinates, heatmaps [7] or random variables [4]. The last two provide information about the position and uncertainty of each landmark. This is valuable in FA, to indicate, for instance, that landmarks are occluded due to large head rotations.

Most FA approaches work on a tracking-by-detection basis, where a face detector retrieves the respective bounding box in every frame. While the current state of the art on FA for near-frontal and partially-occluded faces has reached high robustness, the detection and alignment still fails on cases with large occlusion, extreme head poses and blurriness [13].

Motivated by the previous limitations, we introduce a novel approach based on Kalman Filter (KF) that integrates landmark tracking into existing deep-neural-network (DNN) face alignment pipelines, improving the alignment on video sequences. Our approach exploits the correlation between landmark motion by using grouped landmark filters, as well as the uncertainty of FA predictors by adapting the KF measurement noise. Furthermore, our method is modular and can be combined with any off-the-shelf FA approach which retrieves landmark coordinates.

The main contributions of our work are:

- A novel facial landmark hybrid (DNN + KF) tracking pipeline using grouped landmark KFs, to capture the inter-landmark correlations in the filter state.
- An adaptive KF measurement noise derived from the uncertainty of the DNN predictions.
- A mechanism to override the face detection in every frame, based on the behaviour of the KF state covariance.
- An evaluation of our modules combined with existing state-of-the-art FA methods, showing significant improvement in face tracking.

2. RELATED WORK

In this section, we review the relevant work on 3D FA and tracking. For a detailed survey on 2D alignment and tracking, we refer the reader to [14] and [15], respectively.

3D Face Alignment. Recent 3D FA methods can be grouped in two main categories: (a) model-free [4, 5, 7, 8] and (b) model-based approaches [1, 10, 16, 17, 18]. In the latter, the alignment is aided by a 3D model, such as 3D Morphable Model (3DMM) [19]. These methods are less sensitive to large rotations and occlusion in terms of pose, but are less flexible in FA, particularly for unseen facial expressions and shapes. On the contrary, model-free approaches are

This work was partially funded by the German Ministry of Education and Research (BMBF) under Grant Agreement 01IW20002 (SocialWear). Many thanks to Jiankang Deng, for his effort providing the Menpo dataset.

more flexible, but suffer in cases with (self-)occlusion.

JVCR [8] exploits stacked hourglass networks (SHN) to regress the 3D landmarks, represented as voxels. An additional sub-network retrieves the coordinates from the voxels. [20] proposes to leverage Constrained Local Models (CLMs) and integrates a CNN for local landmark detection. [5] introduced an attention-based pipeline with spatial transformers for dense FA. FAN [7] consists of SHN to regress 2D heatmaps, and a ResNet [21] to compute the z coordinate.

3D FA and face reconstruction are closely related, and many methods perform both tasks simultaneously [10], cyclically [18] or consecutively. [2, 4] leverage the 3D landmarks to fit a 3DMM, while [1, 16, 17] derive the landmarks from the reconstructed face. [4] proposes a dense FA approach, where the landmarks are extracted with a CNN and used to fit a 3DMM via optimization. [1, 16, 18] propose analysis-by-synthesis approaches with dedicated networks to learn the 3DMMs parameters. [16] uses cascades of CNN regressors, while [17] uses a backbone based on MobileNet [22], with additional layers for landmark regression and regularization. [1] introduces a UV displacement map, which models dynamic features such as wrinkles and SynergyNet [18] uses [17] for 3DMM fitting. The landmarks are then extracted from the model and refined with multi-attribute feature aggregation.

3D Landmark Tracking. Classical approaches for 2D FA, such as Cascaded Regression Methods (CRMs) [23, 24, 2] and CLMs [25] have been extended to 3D landmark tracking. While [23] does not work on still images, since it needs an approximate ground truth shape to predict the landmarks, [25] was aided by a 3D depth sensor. Some methods use a tracking-by-detection approach [2, 9, 26], where a face detector extracts the face in every frame before the FA. [2] registers a dense 3D shape frame by frame, using a cascade regression framework trained on a large dataset of 3D face scans.

DNN architectures use CNN regressors to either estimate the parameters in 3DMMs [9] or use model-free schemes [26, 27]. [9] introduced an updated U-Net [28] to estimate frame by frame dense 3D landmarks, that are then smoothed over consecutive frames. [26] combines local heatmap with global shape regression to design a local-to-global pipeline. A face detector along with the mean shape of 3D landmarks are employed to initialize every frame. [27] proposes a multi-view SHN, where the bounding box from the previous frame is used to initialize the next frame. Based on Re^3 [29], [30] introduced a tracking approach which integrated the temporal dependency with Long Short Term Memory (LSTM).

Some prior works explored the idea of combining 2D FA with statistical filtering. [31] and [32] propose to use a KF per landmark and a constant acceleration model. [31] uses Active Shape Model for FA and additionally introduced an approach for head pose tracking with KF. [32] included a 3D model for aiding the landmark tracking. Nonetheless, these methods are based on 2D landmark tracking, which are inconsistent across different poses, *e.g.* for profile and frontal faces.

In contrast to the prior work, we introduce for the first time a 3D FA approach using KF. Additionally, we propose the novel concepts of grouped-landmark KFs and FA uncertainty integration into the KF measurement update.

3. PROPOSED METHOD

In this work, we propose a hybrid 3D landmark tracking pipeline, which combines the landmarks retrieved from a DNN-based FA method with a KF. Based on a motion model, the filter predicts the landmark positions, which are later corrected by the measurements from the DNN. Additionally, our pipeline leverages uncertainty information from heatmap or random-variable-based FA, by adapting the KF measurement noise based on the confidence of the FA detection.

3.1. Noise-Adaptive Clustered Landmark Kalman Filters

We perform frame-to-frame tracking of facial landmarks using linear KFs [33]. A KF represents the belief for the state $\mathbf{x}_k$ at time-step k, by its mean $\hat{\mathbf{x}}_k$ and covariance $\mathbf{P}_k$. In the prediction or time-update step of a KF between frames $k-1$ and k, the state $\hat{\mathbf{x}}_k$ is updated based on an underlying motion model, described by a transition matrix $\mathbf{A}$ as

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{A}\hat{\mathbf{x}}_{k-1|k-1} \tag{1}$$

and the covariance as

$$\mathbf{P}_{k|k-1} = \mathbf{A}\mathbf{P}_{k-1|k-1}\mathbf{A}^\top + \mathbf{Q}_k \tag{2}$$

where $\mathbf{Q}_k$ denotes the process noise covariance. Subsequently, in the KF update step, the measurement $\mathbf{y}_k$ for frame k is incorporated. In our case, this measurement is given by the landmark location predictions from the DNN. First, the Kalman gain $\mathbf{K}_k$ and the innovation $\mathbf{z}_k$ are computed as

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{C}^\top\left(\mathbf{C}\mathbf{P}_{k|k-1}\mathbf{C}^\top + \mathbf{W}_k\right)^{-1} \tag{3}$$

$$\mathbf{z}_k = \mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_{k|k-1} \tag{4}$$

where $\mathbf{C}$ is the correction function and $\mathbf{W}_k$ the measurement noise. Finally, the filter state and covariance are updated as

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k\mathbf{z}_k \tag{5}$$

$$\mathbf{P}_{k|k} = \mathbf{P}_{k|k-1} - \mathbf{K}_k\mathbf{C}\mathbf{P}_{k|k-1}. \tag{6}$$

Existing works on landmark tracking with KFs consider each landmark separately, by introducing one KF per landmark [31, 32]. However, taking into account the strong correlations in the motion between face landmarks due to the connectivity of face parts, we propose to use clustered KFs that contain information of more than one landmark in their state. We propose two landmark grouping options shown in Fig. 1.

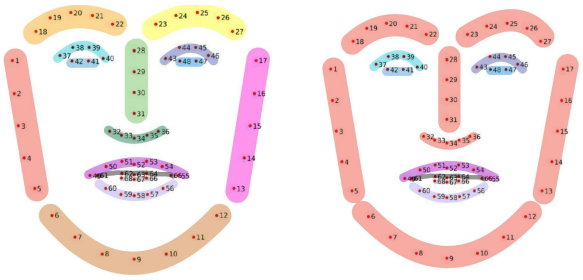

Fig. 1. Landmark grouping in KFs. The right group (G2) separates highly expressive regions (eyes, mouth) from the rest of the face, while the left group (G1) is even more granular.

For our KFs, we implemented two motion models: with constant velocity and with constant acceleration. For the constant velocity model, the KF state is given by:

$$\mathbf{x}_{\text{cv}} = \left[\ell_1^\top \ell_2^\top \cdots \ell_M^\top \dot{\ell}^\top\right]^\top, \in \mathbb{R}^{(3M+3)\times(1)} \tag{7}$$

where $\ell_i^\top$ is a 3D landmark position, M is the number of landmarks in the KF and $\dot{\ell}^\top$ is a common landmark velocity vector for all landmarks. The transition matrix is then defined as:

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}_{3M} & \mathbf{T} \\ \mathbf{0}_3 & \mathbf{I}_3 \end{bmatrix}, \in \mathbb{R}^{(3M+3)\times(3M+3)} \tag{8}$$

where $\mathbf{T} = [\mathbf{I}_3\Delta t \cdots \mathbf{I}_3\Delta t]^\top, \in \mathbb{R}^{3M\times 3}$, and Δt is the time between two consecutive frames, calculated from the respective dataset frame-rate. The measurement vector $\mathbf{y}_k^\top = \left[\bar{\ell}_1^{k\top} \cdots \bar{\ell}_M^{k\top}\right], \in \mathbb{R}^{3M\times 1}$ contains the landmark estimates $\bar{\ell}_i$ produced by the FA network. Thereby, the correction function is given by:

$$\mathbf{C} = [\mathbf{I}_{3M} \quad \mathbf{0}_{3M\times 3}], \in \mathbb{R}^{(3M)\times(3M+3)} \tag{9}$$

For the constant acceleration model, the KF definition follows the same concept, with the addition of a common acceleration vector for all landmarks, making the state vector

$$\mathbf{x}_{\text{ca}} = \left[\ell_1^\top \ell_2^\top \cdots \ell_M^\top \dot{\ell}^\top \ddot{\ell}^\top\right]^\top, \in \mathbb{R}^{(3M+6)\times(1)}. \tag{10}$$

For brevity, we exclude the definition of the matrices $\mathbf{A}, \mathbf{C}$ for the constant acceleration model.

Confidence adaptive KF: Some FA networks such as FAN provide additional information on their confidence for each landmark detection. We incorporate this information into the KF and make its measurement noise covariance dynamically-adaptive based on this confidence. For a landmark measurement $\bar{\ell}_i$, the noise covariance is computed as:

$$\sigma_{\ell_i}^2 = \begin{bmatrix} \sigma_{\ell_i,x}^2 \\ \sigma_{\ell_i,y}^2 \\ \sigma_{\ell_i,z}^2 \end{bmatrix} = \begin{bmatrix} \left(\frac{\bar{\ell}_{i,x}(1-c_{l_i})}{|\bar{\ell}_{i,x}|}\right)^2 \\ \left(\frac{\bar{\ell}_{i,y}(1-c_{l_i})}{|\bar{\ell}_{i,y}|}\right)^2 \\ \left(\frac{\bar{\ell}_{i,z}(1-c_{l_i})}{|\bar{\ell}_{i,z}|}\right)^2 \end{bmatrix} \tag{11}$$

where c_{l_i} is the confidence value between $[0, 1]$ for landmark i from the FA network. Using Eq. 11, the measurement noise matrix of a KF is updated for every frame k as

$$\mathbf{W}_k = diag\left(\left[\sigma_{\ell_1}^{2\top} \quad \sigma_{\ell_2}^{2\top} \quad \cdots \quad \sigma_{\ell_M}^{2\top}\right]\right). \tag{12}$$

3.2. Activation of Face Detector

FA is performed on images cropped around the face. To that end, a face detector is employed to retrieve the bounding box on every frame in tracking-by-detection approaches. Other methods propose to use the landmarks of the previous frame to estimate the bounding box of the current frame, under the assumption that the motion between frames is not large. In this work, we introduce a novel mechanism for reducing the times a face detector is required in video sequences, under the same assumption. This mechanism follows the behaviour of the KF state covariance in consecutive frames as a cue to activate the face detector. When this covariance consistently increases over μ number of frames, as a result of a decrease on the confidence score of the landmarks, it is an indicator of divergence, and hence the face detector is employed in the current frame. Otherwise, the bounding box is continuously computed from the landmarks detected in the previous frame.

4. EXPERIMENTS AND RESULTS

Dataset. Our method was evaluated on the 3DA-2D Menpo tracking benchmark [12], currently the only video sequence-based dataset for 3D facial landmark tracking. The training and test set consist of 90 and 35 videos, respectively, of 1K frames each, with annotations for 84 landmarks. In our case, the training set was used for parameter fine-tuning and evaluation of different KF architectures. The test set was used to evaluate the face detector activation mechanism and the performance of the KF with different FA methods.

Metrics. Following [12], the performance is evaluated using the normalized point-to-point root mean square error (RMSE), Cumulative Error Distribution (CED) curve, failure rate (FR) based on an error threshold of 5% and area under the curve (AUC). The RMSE is normalized by the face diagonal of the bounding box that tightly encloses the ground truth landmarks.

Implementation details. The proposed pipeline was implemented in Python. The code is publicly available at https://github.com/jilliam/FLTrack.

4.1. Experiments

We performed multiple experiments and ablation study to investigate the influence of different KF configurations, the performance of the proposed face detector activation mechanism and the integration of KFs with different type of FA methods. For the experiments with a face detector, we used FaceBoxes [34] to crop the face.

Method	RMSE$^{\downarrow}$
FAN [7]	0.02193
FAN + KF: Constant acceleration	0.0218
FAN + KF: Constant velocity	0.02172
FAN + KF: G1 + const. velocity	0.02154
FAN + KF: G1 + const. velocity + adapt. covariance	0.02149
FAN + KF: G2 + const. velocity + adapt. covariance	0.02148

Table 1. Performance of different KF configurations on the training set of Menpo 3DA-2D benchmark.

KF design. We evaluated several KFs configurations, as described in Section 3.1. In this experiment, we used FAN [7] for landmark detection and provided the ground truth landmarks to crop the face. The results in Table 1 show that the KF has a positive effect on the face tracking, particularly when using landmark grouping. The results further improve when integrating the FA confidence as adaptive covariance.

In the following, we selected the best performing configuration with grouping G2. For FA methods without confidence scores, we excluded the adaptive covariance.

Activation of face detector. We additionally tested the proposed face detector activation mechanism with FAN (Section 3.2). We compared it to using face detection in every frame, with and without KF. The results are shown in Table 2.

Method	RMSE$^{\downarrow}$	AUC$_5^{\uparrow}$	FR$_5^{\downarrow}$(%)
Face detector	0.0245	0.696	1.72
KF + Face detector	0.0240	0.758	1.63
Activation with KF ratio 1/3	0.0189	0.761	0.73
Activation with KF ratio 1/5	**0.0181**	**0.771**	**0.46**

Table 2. Performance of face detection activation mechanism on Menpo 3DA-2D benchmark.

We investigated with $\mu = 3$, and varied the ratio of KFs for which the covariance constantly increases between 1/3 and 1/5 of all KFs. The activation mechanism improves the results significantly. With the KF ratio of 1/3, the face detector was used in average 18.7% of the total number of frames, while with 1/5, 52.4%. This also indicates that the activation mechanism has a positive effect on the computational load of the method.

Performance with different FA methods. Finally, we evaluated how the proposed pipeline performed with different DNN-based FA approaches. Our tracking approach was integrated and tested with 4 different FA methods: (a) a heatmap-based method, FAN [7]; (b) a volumetric approach, JVCR [8]; (c) a 3DMM-based pipeline, 3DDFA [17]; and (d) a hybrid approach, based on 3DMM and landmark regression, SynergyNet [18]. Note that these approaches were trained on a smaller dataset, 300W-LP [16], which consists of ∼61K static images with annotations for 68 landmarks. As it does not contain videos, these approaches were not trained for face tracking. Note that mapping from the 84 landmarks from Menpo to 68 and back is a straightforward task, since the extra landmarks correspond to the midpoint in each pair of consecutive landmarks in the jaw. The results of this experiment are shown in Figure 2 and Table 3.

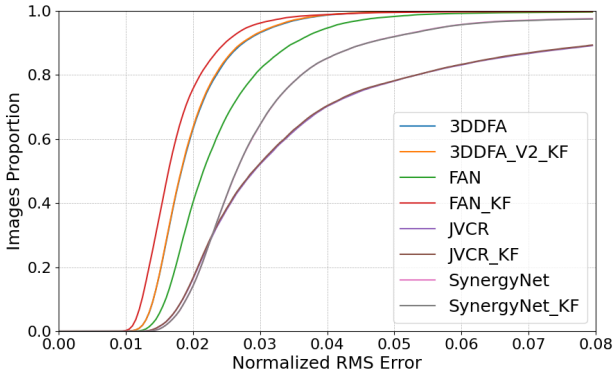

Fig. 2. CED curve for FA alignment methods with and without KF.

	Without KF			With KF		
Method	RMSE$^{\downarrow}$	AUC$_5^{\uparrow}$	FR$_5^{\downarrow}$	RMSE$^{\downarrow}$	AUC$_5^{\uparrow}$	FR$_5^{\downarrow}$
FAN [7]	0.0245	0.696	1.72	0.0181	0.771	0.46
JVCR [8]	0.0416	0.538	21.7	0.0410	0.540	21.5
3DDFA-V2 [17]	0.0199	0.746	0.21	0.0198	0.747	0.19
SynergyNet [18]	0.0342	0.620	7.89	0.0340	0.620	7.93

Table 3. RMSE, area under the curve and failure rate (%) on Menpo 3DA2D tracking test set.

In Table 3, we observe that there is a consistent improvement of the RMSE score when our KF is added. The same holds for the AUC metric and the failure rate, except in SynergyNet. This could be attributed to the sensitivity of SynergyNet to the choice of face detector.

In Fig. 2 we observed that 3DDFA-V2, a 3DMM-based method, is more robust to video sequences than model-free methods such as FAN and JVCR. Nonetheless, the improvement gained by our KF method is more noticeable when we combine it with FAN, where the confidence score is provided.

5. CONCLUSION

In this work, we introduced a novel mechanism to integrate tracking in FA approaches. Our method is based on KF and exploits the uncertainty from the detected facial landmarks. Our results show how this mechanism benefits the alignment in videos, particularly in cases with large poses and occlusion.

Our experiments were evaluated on sparse FA approaches, but our pipeline can be extended to dense FA methods where the landmarks can be grouped semantically and with uncertainty as in [4].

6. REFERENCES

[1] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images," *TOG*, vol. 40, no. 4, pp. 1–13, 2021.

[2] L.A. Jeni, J.F. Cohn, and T. Kanade, "Dense 3D face alignment from 2D video for real-time use," *Image and Vision Computing*, vol. 58, pp. 13–24, 2017.

[3] E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, T.J. Cashman, and J. Shotton, "Fake it till you make it: face analysis in the wild using synthetic data alone," in *ICCV*, 2021, pp. 3681–3691.

[4] E. Wood, T. Baltrušaitis, C. Hewitt, M. Johnson, J. Shen, N. Milosavljević, D. Wilde, S. Garbin, T. Sharp, I. Stojiljković, et al., "3D face reconstruction with dense landmarks," in *ECCV*. Springer, 2022, pp. 160–177.

[5] 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.

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

[7] 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*. IEEE, 2017, pp. 1021–1030.

[8] H. Zhang, Q. Li, and Z. Sun, "Joint voxel and coordinate regression for accurate 3D facial landmark localization," in *ICPR*. IEEE, 2018, pp. 2202–2208.

[9] D. Crispell and M. Bazik, "Pix2face: Direct 3D face model estimation," in *ICCVW*, 2017, pp. 2512–2518.

[10] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3D face reconstruction and dense alignment with position map regression network," in *ECCV*, 2018, pp. 534–551.

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

[12] J. Deng, A. Roussos, G. Chrysos, E.s Ververas, I. Kotsia, J. Shen, and S. Zafeiriou, "The Menpo benchmark for multi-pose 2D and 3D facial landmark localisation and tracking," *IJCV*, vol. 127, no. 6, pp. 599–624, 2019.

[13] H. Jin, S. Liao, and L. Shao, "Pixel-in-pixel net: Towards efficient facial landmark detection in the wild," *IJCV*, vol. 129, no. 12, pp. 3174–3194, 2021.

[14] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *IJCV*, vol. 127, pp. 115–142, 2019.

[15] G.G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou, "A comprehensive performance evaluation of deformable face tracking 'in-the-wild'," *IJCV*, vol. 126, no. 2-4, pp. 198–232, 2018.

[16] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3D total solution," *PAMI*, vol. 41, no. 1, pp. 78–92, 2017.

[17] 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*. Springer, 2020, pp. 152–168.

[18] C. Wu, Q. Xu, and U. Neumann, "Synergy between 3DMM and 3D landmarks for accurate 3D facial geometry," in *3DV*. IEEE, 2021, pp. 453–463.

[19] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *SIGGRAPH*, 1999, pp. 187–194.

[20] A. Zadeh, Y. Chong Lim, T. Baltrusaitis, and L.P. Morency, "Convolutional experts constrained local model for 3D facial landmark detection," in *ICCVW*. IEEE, 2017, pp. 2519–2528.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*. 2016, pp. 770–778, IEEE.

[22] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al., "Searching for MobileNetv3," in *CVPR*, 2019, pp. 1314–1324.

[23] X. Xiong and F. De la Torre, "Global supervised descent method," in *CVPR*. IEEE, 2015, pp. 2664–2673.

[24] C. Cao, Y. Weng, S. Lin, and K. Zhou, "3D shape regression for real-time facial animation," *TOG*, vol. 32, no. 4, 2013.

[25] T. Baltrušaitis, P. Robinson, and L.P. Morency, "3D constrained local model for rigid and non-rigid facial tracking," in *CVPR*. IEEE, 2012.

[26] P. Xiong, G. Li, and Y. Sun, "Combining local and global features for 3D face tracking," in *ICCVW*. IEEE, 2017, pp. 2529–2536.

[27] J. Deng, Y. Zhou, S. Cheng, and S. Zaferiou, "Cascade multi-view hourglass model for robust 3D face alignment," in *FG*. IEEE, 2018, pp. 399–403.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image ssegmentation," in *MICCAI*. 2015, pp. 234–241, Springer.

[29] D. Gordon, A. Farhadi, and D. Fox, "Re^3: Real-time recurrent regression networks for visual tracking of generic objects," *RA-L*, vol. 3, no. 2, pp. 788–795, 2018.

[30] D. Aspandi, O. Martinez, F. Sukno, and X. Binefa, "Fully end-to-end composite recurrent convolution network for deformable facial tracking in the wild," in *FG*. IEEE, 2019.

[31] U. Prabhu, K. Seshadri, and M. Savvides, "Automatic facial landmark tracking in video sequences using kalman filter assisted active shape models," in *ECCVW*. Springer, 2010, pp. 86–99.

[32] Y. Jin, X. Guo, Y. Li, J. Xing, and H. Tian, "Towards stabilizing facial landmark detection and tracking via hierarchical filtering: A new method," *Journal of the Franklin Institute*, vol. 357, no. 5, pp. 3019–3037, 2020.

[33] R.E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, 1960.

[34] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S.Z. Li, "Faceboxes: A cpu real-time face detector with high accuracy," in *IJCB*. IEEE, 2017, pp. 1–9.