electronics

MDPI

Article

Stabilized Temporal 3D Face Alignment Using Landmark Displacement Learning

Seongmin Lee [1], Hyunse Yoon [1], Sohyun Park [2], Sanghoon Lee [1] and Jiwoo Kang [2,3,*]

[1] Department of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, Republic of Korea; lseong721@yonsei.ac.kr (S.L.); hsyoon97@yonsei.ac.kr (H.Y.); slee@yonsei.ac.kr (S.L.)
[2] Division of Artificial Intelligence Engineering, Sookmyung Women's University, Seoul 04310, Republic of Korea; so209@sookmyung.ac.kr
[3] Artificial Intelligence Innovation Research Center, Sookmyung Women's University, Seoul 04310, Republic of Korea
* Correspondence: jwkang@sookmyung.ac.kr

Abstract: One of the most crucial aspects of 3D facial models is facial reconstruction. However, it is unclear if face shape distortion is caused by identity or expression when the 3D morphable model (3DMM) is fitted into largely expressive faces. In order to overcome the problem, we introduce neural networks to reconstruct stable and precise faces in time. The reconstruction network extracts the 3DMM parameters from video sequences to represent 3D faces in time. Meanwhile, our displacement networks learn the changes in facial landmarks. In particular, the networks learn changes caused by facial identity, facial expression, and temporal cues, respectively. The proposed facial alignment network exhibits reliable and precise performance in reconstructing static and dynamic faces by leveraging these displacement networks. The 300 Videos in the Wild (300VW) dataset is utilized for qualitative and quantitative evaluations to confirm the effectiveness of our method. The results demonstrate the considerable advantages of our method in reconstructing 3D faces from video sequences.

Keywords: face alignment; face tracking; face displacement; temporal stability; video-based alignment

Citation: Lee, S.; Yoon, H.; Park, S.; Lee, S.; Kang, J. Stabilized Temporal 3D Face Alignment Using Landmark Displacement Learning. *Electronics* **2023**, *12*, 3735. https://doi.org/10.3390/electronics12173735

Academic Editors: Byung-Gyu Kim and Jan Platoš

Received: 9 August 2023
Revised: 31 August 2023
Accepted: 3 September 2023
Published: 4 September 2023

1. Introduction

Three-dimensional (3D) facial models find widespread applications in various facial tasks, including facial animation, facial synthesis, facial reconstruction, facial recognition, and facial tracking. A crucial pre-processing step for utilizing 3D facial models is facial alignment, which involves moving and deforming a facial model to match an image. The regularized structure of facial components, such as the eyes, lips, and nose in human faces, serves as a valuable prior for efficient facial alignment. However, conventional facial alignment methods exhibit instability when dealing with large pose and expression changes. In such scenarios, it becomes challenging to distinguish whether the observed facial shape changes derive from identity, expression, or pose variations. Furthermore, when this ambiguity extends to the temporal domain, it results in unnatural facial shape changes and jittering artifacts, leading to significant visual quality degradation. To overcome these issues, this paper introduces a facial reconstruction framework that learns facial movements, i.e., displacements, according to facial identity, expression, and temporal cues.

The 3D morphable model (3DMM) stands as the most widely utilized statistical representation for obtaining 3D faces from facial images in diverse face-related applications. Since its initial introduction [1], various adaptations of the 3DMM have been developed by employing principal component analysis (PCA) to decompose facial scans of different identities and expressions, enabling the representation of arbitrary human faces. Consequently, it efficiently captures the 3D facial shape from a given facial image. However, challenges arise when fitting the 3DMM to facial images exhibiting large expression or pose variations.

In such cases, there exists ambiguity in the facial shape, making it difficult to determine whether the facial shape deformation is due to identity or expression changes. While this ambiguity may not result in substantial visual degradation in a static context, it becomes evident in a temporal domain, leading to apparent visual artifacts, such as unnatural facial shape changes and jittering artifacts. To address the problem, we separately modeled identity shape, expression shape, and temporal movements. In facial parametric models, such as 3DMM [1] or FLAME [2], it is demonstrated that the statistical shape variations (i.e., movements or displacements) caused by facial identity and expressions are independent of each other. Thus, the proposed method models these movements separately to effectively reduce the ambiguity of the facial movement.

Recently, with the expansion of the generative adversarial network (GAN) in deep learning, it has been found that using discriminators leads to a network with higher performance [3]. The GAN is composed of two networks: a generator and a discriminator. The discriminator is trained to determine whether the input data distribution is close to the ground-truth data distribution or the generated data distribution. At the same time, the generator is trained to fool the discriminator, by generating more accurate data. Recently, thanks to the powerful performance of the discriminator, the discriminator has been widely adopted in various temporal data generation tasks [4,5]. Motivated by this, we propose a stable and accurate facial alignment framework by introducing displacement discriminators that determine that the regressed camera and facial shape parameters are stable. We train a discriminator to evaluate whether the distribution of the 3D face alignment results is similar to ground-truth 3D face movements. Thus, this discriminator learns the distribution difference between alignment results and the ground-truth movements. Then, the 3D facial alignment network is trained to produce a stable 3D face alignment using the distribution difference trained from the discriminator as guidance. Here, to learn the distribution difference more precisely, we present three displacement discriminators that separately discriminate the facial movements according to personal identity, expression, and temporal cues. The identity and expression displacement discriminators are trained to discriminate whether the facial deformations generated from the estimated facial identity and expression parameters are stable. This enables the facial alignment network to estimate the accurate facial identity and expression parameters. The temporal displacement discriminator is trained to discriminate whether the facial temporal displacement is stable, which allows the alignment network to achieve temporally stable alignment results. Using these displacement discriminators, the proposed facial alignment network shows accurate and stable facial alignment performance in both the static and temporal domains.

For the qualitative and quantitative evaluations, we use the 300 Videos in the Wild (300VW) dataset [6], which provides large-scale facial tracking data. In the experimental results, the proposed method shows significant improvements over state-of-the-art methods for temporal facial alignment. The results demonstrate that the proposed method enables accurate facial tracking with multiple discriminators by stabilizing facial locations and shapes over time.

2. Related Works

2.1. 3D Morphable Model

Since the pioneering work of Blanz [1] introducing the first 3D morphable model (3DMM), several subsequent 3DMMs have been proposed [7–9]. These models are constructed by encoding the features of 3D facial scans pertaining to identity, expression, and texture through PCA decomposition, leveraging data collected from multiple subjects. Due to the distinct topology of each facial scan, mesh registration is essential to establish vertex correspondences among them. In Blanz's work [1], optical methods were employed to determine the vertex correspondences between facial scans. Paysan et al. [8] proposed a non-rigid registration approach utilizing warping based on thin-plate splines (TPS) [10], and a non-rigid iterative closest point (ICP) [11] was utilized to achieve accurate alignment.

Vlasic et al. [9] presented a multilinear facial mode, representing facial identity and expression using singular value decomposition (SVD). Subsequently, Cao et al. [7] proposed a bilinear facial model, building upon the multilinear model by deforming the facial scan into a template model with expression. Thanks to the considerable efforts devoted to constructing accurate 3DMMs, an arbitrary 3D face can now be effectively and precisely represented using these models.

2.2. 3D Face Alignment

3D facial alignment is a task that fits the 3D facial shape into the input facial images. Due to the powerful representation performance of the 3DMM, it is widely used for face alignment. The first method for 3D facial alignment [12] performed alignment of the 3DMM to the input image by minimizing the pixel-wise difference between the target facial image and a rendered image of the 3DMM. In recent years, regression-based 3D facial alignment techniques have been introduced [13–17], which minimize the discrepancy between the target 2D landmarks and the projected 2D landmarks of the 3DMM. While these approaches demonstrated performance improvements, two major challenges remain.

Firstly, self-occlusion becomes a concern when dealing with large pose or expression variations. Self-occlusion leads to the loss of facial semantic information, resulting in unreliable facial alignment. Secondly, in temporal sequences, temporal instability becomes pronounced during rapid and substantial facial motion. While facial alignment results may appear reliable in static shots, jittering artifacts often emerge in the temporal domain. To address these issues, this paper introduces novel stabilization discriminators that effectively guide changes in the stabilized facial shape, particularly when dealing with large poses, expressions, and motion.

3. Method

The proposed method is composed of the facial alignment network and the displacement discriminators. For the facial alignment network, we employ the 3DMM for efficient facial shape alignment. In addition, to ensure consistent facial alignments for an individual's identity and expression over time, multiple sub-discriminators are integrated into the displacement discriminators. Figure 1 provides an overview of the entire framework of the proposed method consisting of the facial alignment network and displacement discriminators: the identity displacement discriminator (IDD), expression displacement discriminator (EDD), and temporal displacement discriminator (TDD).

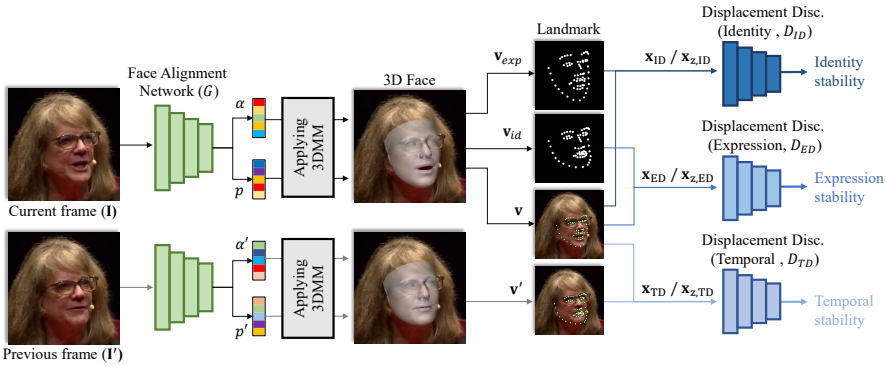


Figure 1. Overall framework of the proposed facial alignment method. The proposed method is composed of the facial alignment network (G), identity displacement discriminators (D_{ID}), expression displacement discriminator (D_{ED}), and temporal displacement discriminator (D_{TD}). For the face alignment, the facial alignment network estimates the 3DMM parameters α and camera parameters p corresponding to the current image. The identity and expression displacement discriminators (D_{ID}, D_{ED}) are trained the assess whether facial shape changes align with an individual's identity and expression, respectively. In addition, the temporal displacement discriminator is trained to determine whether the temporal facial shape change is stable or not.

3.1. Facial Alignment Network

A 3DMM represents an arbitrary 3D face (**S**) using bases decomposed through PCA. Using the 3DMM, the 3D face (**S**) can be represented by parameters for both identity and expression, $\alpha = [\alpha_{id}, \alpha_{exp}]$. Given a 2D image **I**, the facial reconstruction network **G** finds the shape parameters α. Then, the projected landmark of the 3D face is estimated using a landmark index vector $\mathbf{l} \in \mathbb{R}^{68}$. The reconstruction network is trained using the $\mathcal{L}_{land}$ loss. $\mathcal{L}_{land}$ is defined as follows:

$$\mathcal{L}_{land} = |\mathbf{v}(:, \mathbf{l}) - \mathbf{U}|_2, \tag{1}$$

where **U** is the labeled ground-truth 2D landmark location of the input image.

3.2. Displacement Discriminators

To achieve stability in both the temporal and static domains during training of the facial alignment network, we propose the use of three displacement discriminators: identity, expression, and temporal cues. The identity and expression displacement discriminators play a vital role in stabilizing the facial alignment network in the static domain. This is accomplished by distinguishing between the changes in facial shape estimated based on the identity and expression parameters. Consequently, the network can better understand and differentiate the influences of identity and expression on facial shape variations. On the other hand, the temporal displacement discriminator ensures stability in facial alignment over time by discerning changes in facial shape across consecutive frames. This helps the network maintain consistent facial alignments throughout a temporal sequence.

3.2.1. Identity Displacement Discriminator

The identity displacement discriminator (IDD) is to determine whether the estimated changes in facial shape align with the desired facial shape corresponding to the regressed facial identity parameter. To train the IDD, we calculate the difference between the facial landmarks and the estimated landmarks without considering identity information. To compute this calculation, we estimate the landmark displacement depending on the identity parameter as follows:

$$\mathbf{S}_{exp} = \bar{\mathbf{S}} + \mathbf{A}_{exp}\alpha_{exp}, \tag{2}$$

$$\mathbf{v}_{exp} = f \cdot \mathbf{P} * \mathbf{R} * \mathbf{S}_{exp} + \mathbf{t}, \tag{3}$$

Facial landmarks are detected from the projected facial vertices, which are in image coordinates. To facilitate comparison, both the ground-truth and estimated landmarks are normalized to the range of [0, 1] before computing the difference. The input for the IDD is then obtained by calculating the discrepancy between the normalized landmarks. This process ensures that the IDD can effectively discern facial shape changes due to variations in identity. The difference to be used as input for the IDD is computed by using the normalized landmarks as follows:

$$\mathbf{x}_{ID} = \mathbf{U} - \mathbf{v}_{exp}(:, \mathbf{l}), \tag{4}$$

$$\mathbf{x}_{z,ID} = \mathbf{v}(:, \mathbf{l}) - \mathbf{v}_{exp}(:, \mathbf{l}), \tag{5}$$

where $\mathbf{x}_{ID}$ is the landmark difference between the ground-truth and estimated $\mathbf{x}_{z,ID}$ landmarks. To make the IDD learn the stabilized displacement based on the identity parameter, we use $\mathbf{x}_{ID}$ as the real distribution, and we use $\mathbf{x}_{z,ID}$ as the fake distribution. Therefore, the loss for the IDD is defined as follows:

$$\mathcal{L}_{D_{ID}} = \mathbb{E}_{\mathbf{x}_{ID}}[\log(D_{ID}(\mathbf{x_{ID}})] + \mathbb{E}_{\mathbf{x}_{z,ID}}[\log(D_{ID}(\mathbf{x_{z,ID}})] \tag{6}$$

3.2.2. Expression Displacement Discriminator

Similarly to the IDD, the expression displacement discriminator (EDD) is trained to distinguish facial shape changes based on the validity of the expression parameter. Similarly to Equation (3), we calculate the facial shape displacement without expression $\mathbf{S}_{ID}$ by replacing $\mathbf{A}_{exp}$ and αexp with $\mathbf{A}_{id}$ and α_{id}. Thus, the expression-based facial shape displacement is defined as follows:

$$\mathbf{S}_{id} = \bar{\mathbf{S}} + \mathbf{A}_{id}\alpha_{id}, \tag{7}$$

$$\mathbf{v}_{id} = f \cdot \mathbf{P} * \mathbf{R} * \mathbf{S}_{id} + \mathbf{t}, \tag{8}$$

Then, the expression-based landmark displacement $\mathbf{x}_{exp}$ and $\mathbf{x}_{z,ED}$ are defined as follows:

$$\mathbf{x}_{ED} = \mathbf{U} - \mathbf{v}_{id}(:, \mathbf{l}), \tag{9}$$

$$\mathbf{x}_{z,ED} = \mathbf{v}(:, \mathbf{l}) - \mathbf{v}_{id}(:, \mathbf{l}), \tag{10}$$

During the training of the EDD, the differences between the calculated landmarks without expression $\mathbf{x}_{ED}$ are used as the real data distribution, and the estimated landmarks $\mathbf{x}_{z,ED}$ are used as the fake distribution. The loss for the EDD is defined as follows:

$$\mathcal{L}_{D_{ED}} = \mathbb{E}_{\mathbf{x}_{ED}}[\log(D_{ED}(\mathbf{x_{ED}})] + \mathbb{E}_{\mathbf{x}_{z,ED}}[\log(D_{ED}(\mathbf{x_{z,ED}})] \tag{11}$$

3.2.3. Temporal Displacement Discriminator

The IDD and EDD are responsible for stabilizing the facial alignment network in a static domain. To further enhance the temporal stabilization performance, we introduce a temporal displacement discriminator (TDD) to guide the changes in the temporal facial shape through the frames. The input for the TDD is derived from the variation in facial landmarks between the current and previous frames. Facial temporal changes are assessed by calculating the difference between the landmarks of the current frame and those of the previous frame as follows:

$$\mathbf{x}_{TD} = \mathbf{U} - \mathbf{U}', \tag{12}$$

$$\mathbf{x}_{z,TD} = \mathbf{v}(:, \mathbf{l}) - \mathbf{v}'(:, \mathbf{l}), \tag{13}$$

where $\mathbf{v}'$ and $\mathbf{U}'$ are the projected vertices and the ground-truth landmark of the previous frame, respectively. The temporal discriminator loss is defined as follows:

$$\mathcal{L}_{D_{TD}} = \mathbb{E}_{\mathbf{x}_{TD}}[\log(D_{TD}(\mathbf{x}_{TD})] + \mathbb{E}_{\mathbf{x}_{z,TD}}[\log(D_{TD}(\mathbf{x}_{z,TD})] \tag{14}$$

3.2.4. Adversarial Loss Function

These multiple discriminators (i.e., IDD, EDD, and TDD) are trained to discern the validity of identity, expression, and temporal changes in facial shape. Concurrently, the facial alignment network is trained to deceive these discriminators. The overall adversarial losses for these discriminators, denoted as $\mathcal{L}_{D_{ID}}$, $\mathcal{L}_{D_{ED}}$, and $\mathcal{L}_{D_{TD}}$, are defined as follows:

$$\mathcal{L}_D = \lambda_{ID}\mathcal{L}_{D_{ID}} + \lambda_{ED}\mathcal{L}_{D_{ED}} + \lambda_{TD}\mathcal{L}_{D_{TD}} \tag{15}$$

where λ_{ID}, λ_{ED}, and λ_{TD} are factors for balancing between each loss term. Thus, the loss function in Equation (15) is used to train the IDD, EDD, and TDD.

The total loss for the facial alignment network (G) is defined by combining the alignment and adversarial losses as follows:

$$\mathcal{L}_{G_{ID}} = \mathbb{E}_{\mathbf{x}_{ID}}[\log(D_{ID}(\mathbf{x}_{z,ID})], \tag{16}$$

$$\mathcal{L}_{G_{ED}} = \mathbb{E}_{\mathbf{x}_{z,ED}}[\log(D_{ED}(\mathbf{x}_{z,ED})], \tag{17}$$

$$\mathcal{L}_{G_{TD}} = \mathbb{E}_{\mathbf{x}_{z,TD}}[\log(D_{TD}(\mathbf{x}_{z,TD})], \tag{18}$$

$$\mathcal{L}_G = \mathcal{L}_{G_{align}} + \lambda_{ID}\mathcal{L}_{G_{ID}} + \lambda_{ED}\mathcal{L}_{G_{ED}} + \lambda_{TD}\mathcal{L}_{G_{TD}} \tag{19}$$

Similarly to the conventional GAN training, we freeze the discriminators, i.e., IDD, EDD, and TDD, when training the facial alignment network. Thus, the discriminators and facial alignment network are trained alternately. In our experiments, we used balancing factors λ_{ID}=$\lambda_{ED} = \lambda_{TD} = 0.1$. In our experiments, the same network architecture was employed for all discriminators. To assess stability, we utilized the landmark difference and passed it through three fully connected layers, which ultimately produced a single scalar value ranging from 0 to 1.

For better understanding, we represent a block diagram of the training procedure of the proposed displacement learning method in Figure 2. The IDD (D_{ID}) and EDD (D_{ED}) are trained to judge the estimated facial landmark displacement ($\mathbf{x}_{z,ID}$ and $\mathbf{x}_{z,ED}$) as unstable (0) and the ground-truth landmark displacement ($\mathbf{x}_{ID}$, $\mathbf{x}_{ED}$) as stable (1). The TDD (D_{TD}) is trained to discriminate the temporal displacement of the estimated faces $\mathbf{x}_{z,TD}$ as unstable (0) and that of the ground-truth faces $\mathbf{x}_{TD}$ as stable (1). To deceive these identity, expression, and temporal displacement discriminators, the facial alignment network is trained for these discriminators to output stable (1) from the estimated face ($\mathbf{x}_{z,ID}$, $\mathbf{x}_{z,ED}$, and $\mathbf{x}_{z,TD}$). In short, similarly to the conventional GAN training procedure, we alternately train the facial alignment network and the displacement discriminators. In a single training iteration, we first train the displacement discriminators by freezing the facial alignment network and then train the facial alignment network by freezing the displacement discriminators.

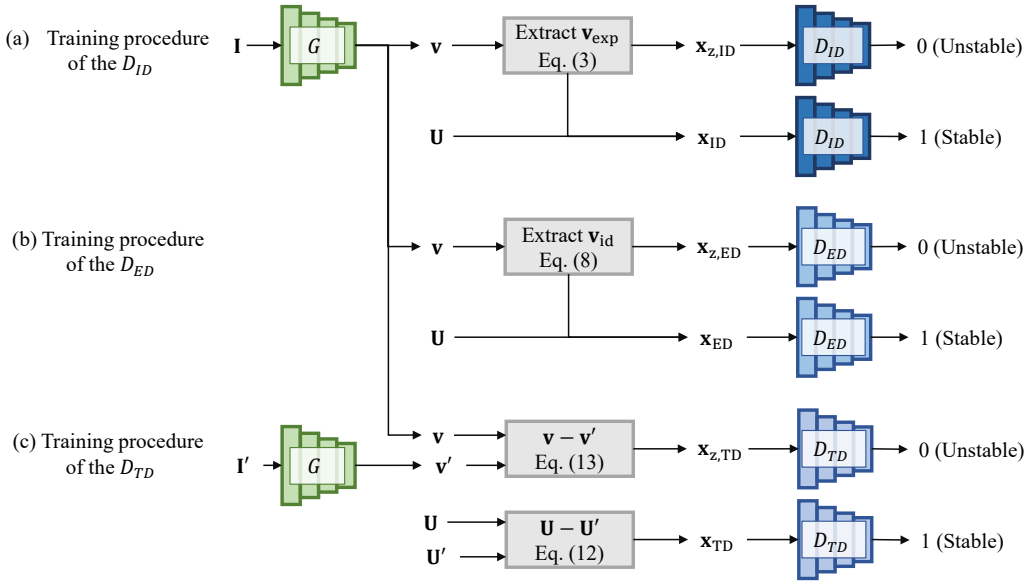


Figure 2. Block diagram of the training procedure of the proposed displacement learning method. **I** and **I**′ are the image of the current and previous frames, respectively. The face alignment network *G* has shared weight when aligning the current 3D face **v** and previous 3D face **v**′.

4. Experimental Results

4.1. Implementation Details

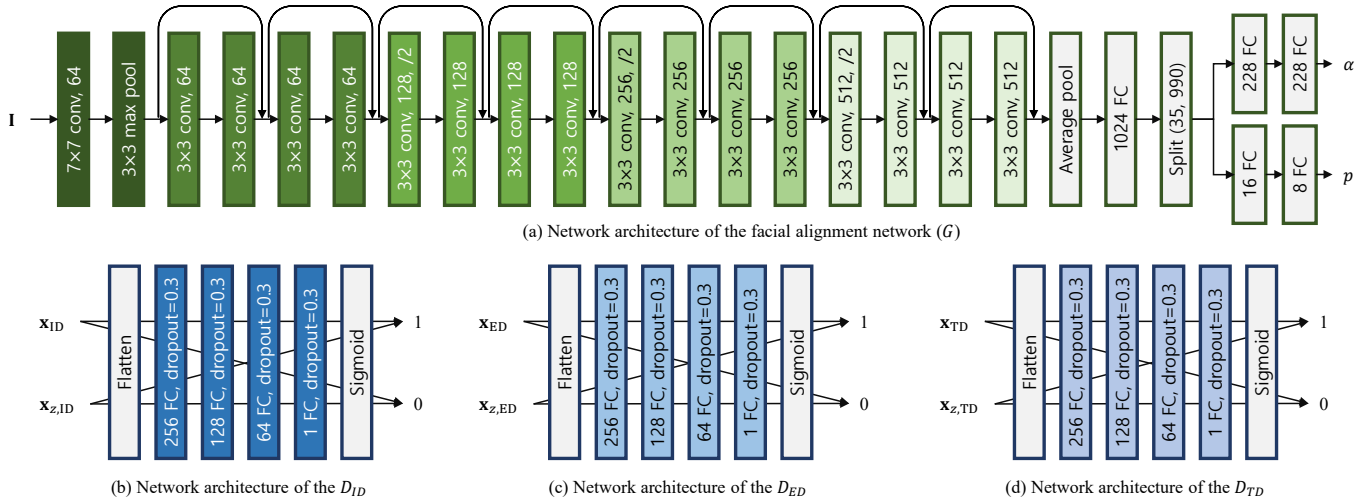
In our experiments, we utilized the 300VW dataset [6], a large-scale facial tracking dataset containing 114 videos with a total of 218,595 frames, each annotated with 68-point

landmark labels. Among these videos, 50 were used for training, and the remaining 64 were designated for testing. The test videos were further categorized into A, B, and C sets, with C being the most challenging test subset.

During the training phase, each frame was cropped using a ground-truth landmark and resized to 256 × 256 pixels, serving as input for the facial alignment network. We employ the ResNet 18 backbone [18] for the facial alignment network. Figure 3 represents the details of the facial alignment network, identity displacement discriminator, expression displacement discriminator, and temporal displacement discriminator for reproducibility. In the facial alignment network, we add the splitting layer and four fully connected layers at the end of the ResNet backbone to estimate the 3DMM and camera parameters. In the splitting layer, the output feature vector is split into 990-dimensional and 35-dimensional feature vectors by proportionally dividing them based on the anticipated number of 228 3DMM parameters and 8 camera parameters. The 990-dimensional feature vector is fed into two fully connected layers, which are composed of 228 and 228 nodes, to estimate the 3DMM parameters. The output for the 3DMM parameter has a 228-dimensional vector. Here, 199 dimensions are used for identity parameters, and 29 dimensions are used for expression parameters. Similarly, the 35-dimensional feature vector is fed into two fully connected layers to estimate the camera parameters. For the camera parameter estimation, each fully connected layer has 16 and 8 nodes, respectively. In all layers of the facial alignment network, the RELU activation functions are employed except for the last layer. In the last layer, no activation functions are employed. For the displacement discriminators, we use four fully connected layers. Each fully connected layer has 256, 128, 64, and 1 node, respectively. All of the displacement discriminators are constructed with the same architecture. Table 1 summarizes the architecture of the displacement discriminator.

Table 1. Network architecture of the displacement discriminator.

Layer	Configuration	Size
Input	Flatten landmark	B × 136
FC1	Node = 256, Activation = RELU, dropout = 0.3	B × 256
FC2	Node = 128, Activation = RELU, dropout = 0.3	B × 128
FC3	Node = 64, Activation = RELU, dropout = 0.3	B × 64
FC4	Node = 1, Activation = Sigmoid, dropout = 0.3	B × 1



(a) Network architecture of the facial alignment network (G)

(b) Network architecture of the D_{ID}

(c) Network architecture of the D_{ED}

(d) Network architecture of the D_{TD}

Figure 3. Detailed architecture of the facial alignment network G, identity displacement discriminator D_{ID}, expression displacement discriminator D_{ED}, and temporal displacement discriminator D_{TD}.

To improve the network's temporal robustness, the frame interval between the current and previous frames was randomly increased within the range from one to six. After the

second frame, each frame was cropped using the landmarks estimated from the previous frame. In the testing phase, the first frame was cropped based on landmarks detected using a conventional landmark detection algorithm called MTCNN [19]. For subsequent frames, each one was cropped using landmarks estimated from the previous frame. The proposed method used in all experiments was trained for 500 epochs using TensorFlow (version 2.10.0), CUDNN (version 8.1), and CUDA (version 11.2). We employed the Adam optimizer for optimization and trained the model on a single NVIDIA 2080Ti (11 GB) GPU with a batch size of 20. The learning rate was set to 0.001 during the initial training phase, and it gradually decreased to 0.00001 over time.

4.2. Performance Evaluation

For this evaluation, we conducted a comparison of our method against other state-of-the-art facial alignment techniques, namely 3DDFA [20], RingNet [21], DSFNet [22], and SADRNet [23]. To quantitatively assess the performance, we measured the normalized mean error (NME) of the 2D facial landmarks. The NME is calculated as the average normalized landmark error divided by the facial bounding size, as per previous facial alignment methods [24,25]. The facial bounding box's size is defined as the square root of the product of the width and height of the rectangular hull formed by all the landmarks.

For the qualitative comparison, we visualize some examples of the 3D face alignment outcomes on 300VW-A, 300VW-B, and 300VW-C in Figure 4. On the 300VW-A set, the easiest dataset, all comparison methods, including ours, show similar alignment performance. In contrast, our method shows significantly better performance than the comparison methods on the 300VW-C dataset. Note that 300VW-C is the most challenging dataset because of the fast motion and extreme light conditions. Specifically, our method shows more accurate alignment results in the face contour and the mouth. In summary, the results show that our method outperforms the state-of-the-art face alignment method in all cases of the 300VW dataset.

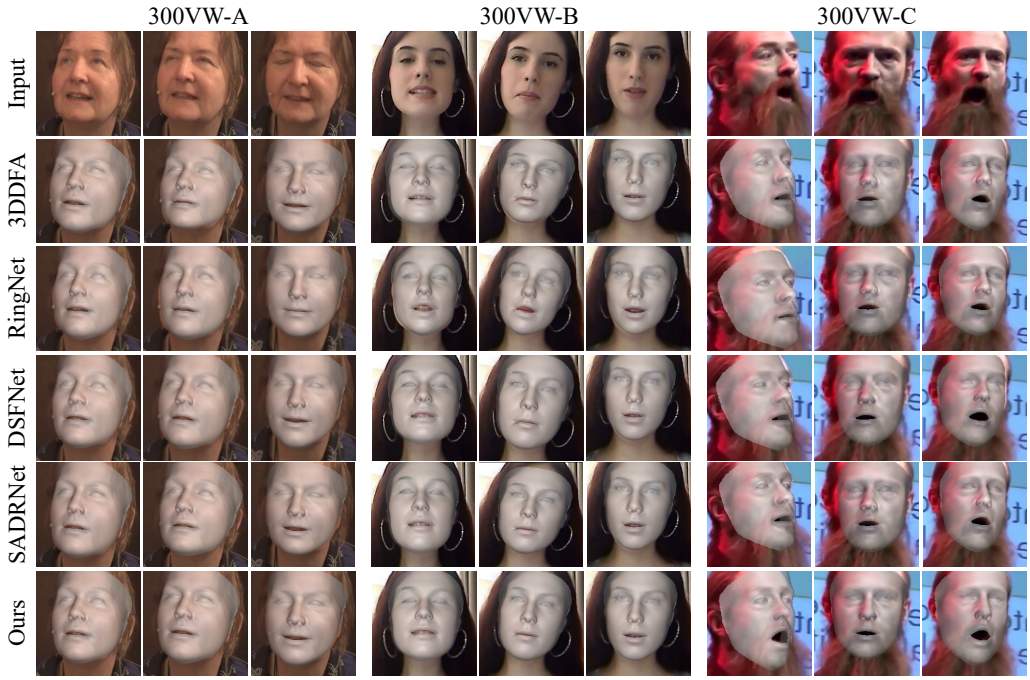


Figure 4. Qualitative evaluation of temporal 3D face alignment performance on 300VW dataset.

We also evaluate the performance of face alignment in cases where a part of the face is occluded. Figure 5 demonstrates the results of face alignment under occlusion. 3DDFA [20] and RingNet [21] often fail to align the 3D face when significant occlusion occurs. DFSNet [22] and SADRNet [22] exhibit substantial alignment errors when the

face is occluded. In particular, when most of the facial region is occluded, both methods have large alignment errors in rotation, translation, and scale estimations. In contrast, our proposed method demonstrates stable face alignment even under significant occlusion, providing accurate results for rotation, translation, and scale estimation. This result shows that our method outperforms other methods by achieving a stable alignment performance in extreme cases.

Figure 5. Qualitative evaluation of temporal 3D face alignment performance on occlusion case.

In addition, we quantitatively evaluate face alignment accuracy by measuring the normalized mean error (NME) of the 2D facial landmarks. The NME is the landmark error normalized by the size of the facial bounding box [15]. The size of the facial bounding box is defined by the $\sqrt{height \times width}$ of the rectangular hull calculated from all landmarks. The quantitative evaluation is summarized in Table 2. Here, the accuracy is the percentage of the bounding box size. It shows that our method outperforms other state-of-the-art face alignment methods. On the 300VW-A set, our method achieves a 14.34% accuracy improvement over the 3DDFA [20], which has the lowest accuracy. In addition, our method achieves accuracy improvements of 16.01% and 19.27% on the 300VW-B and 300VW-C datasets over the 3DDFA [20], respectively. This shows that our method demonstrated a distinct advantage in the challenging tracking case (300VW-C) compared to the other comparison methods, and it is consistent with the result in Figure 4.

Table 2. Shows 2D facial alignment accuracy (%) on 300VW dataset.

Method	300VW-A	300VW-B	300VW-C
3DDFA [20]	2.913	3.035	3.387
RingNet [21]	2.845	2.983	3.343
DSFNet [22]	2.799	2.878	3.214
SADRNet [23]	2.745	2.770	2.858
Ours	2.495	2.549	2.734

4.3. Ablation Study

The proposed method is inspired by the fact that statistical facial parameter models, such as FLAME and the 3DMM, independently model facial movements with respect to identity, expression, and temporal aspects. Thus, the proposed method separately models identity shape, expression shape, and temporal movements with three different discriminators. To assess this, we conducted eight ablation tests according to the use of the discriminator.

In the baseline experiment, we trained the facial alignment network without incorporating any discriminator. Subsequently, for each discriminator, we evaluated its individual performance on the baseline model. The ablation tests were performed by measuring the accuracy of facial alignment, represented by the NME. The results of these tests are summarized in Table 3.

Table 3. Ablation tests of 2D facial alignment accuracy (%) depending on the displacement discriminator.

Method	300VW-A	300VW-B	300VW-C
Without all	3.731	4.060	4.677
With D_{ID}	3.315	3.575	3.912
With D_{ED}	3.133	3.284	3.665
With D_{TD}	3.078	3.192	3.504
With D_{ID}, D_{ED}	3.099	3.227	3.601
With D_{ID}, D_{TD}	2.912	3.085	3.311
With D_{ED}, D_{TD}	2.721	2.801	3.057
With D_{ID}, D_{ED}, D_{TD}	2.495	2.549	2.734

The results indicate that each individual discrimination of identity, expression, and temporal changes contributes significantly to performance improvements. Notably, temporal discrimination plays the most crucial role in achieving stable facial alignments over time, while identity discrimination has the least impact. By comparing the outcomes in Table 3, it is evident that employing multiple discriminators for temporal, identity, and expression simultaneously provides substantial benefits in obtaining stable 3D facial alignments. Therefore, the ablation tests show this independence and orthogonality of facial identity, expression, and temporal movements, and it is demonstrated that using all discriminators can significantly boost performance.

5. Discussion and Conclusions

In this paper, we present a robust and precise facial alignment framework by introducing multiple stability discriminators. These discriminators effectively determine the camera, face identity, and expression parameters from an input image simultaneously. The proposed framework comprises a facial alignment network and three displacement discriminators: identity (IDD), expression (EDD), and temporal (TDD) discriminators. The previous temporal smoothing scheme uses the local average to reduce the outlier alignment result. It effectively reduces the alignment error of the outlier frame but causes unwanted alignment errors in nearby frames due to the local averaging scheme. In contrast, the proposed discriminator-based method can effectively reduce the alignment error in the outlier frame without causing unwanted alignment error propagation. This is possible because the discriminator accurately distinguishes unnatural and unstabilized facial movements based on facial identity, expression, and temporal cues using a comparison with the ground truth of the facial movement. To evaluate the performance of the proposed discriminators, we conducted qualitative and quantitative assessments using the 300VW dataset, a large-scale facial tracking dataset. The experimental results demonstrate significant improvements over state-of-the-art methods, showcasing the effectiveness of our approach in achieving accurate and stable facial alignment over time.

However, the main bottleneck in our method is that displacement discrimination is performed based on the 2D facial landmarks. This is because there is no publicly available

video-based dense 3D face dataset. Since a 2D facial landmark provides sparse information on the facial shape, more detailed facial deformation, such as facial wrinkles, cannot be represented using the facial landmark. Therefore, the loss of information in facial details is a limitation of our work. We believe that when the proposed method is trained using a video-based dense 3D face dataset, it will exhibit stable temporal alignment performance while generating facial details. In future research, we plan to extend landmark displacement discrimination to dense displacement discrimination by employing the self-supervised method. This may accurately represent changes in facial details as well as facial shape over time. Lastly, we hope that our work will be valuable in various facial applications, including facial recognition [26–28], facial animation [29,30], and VR communication [31,32].

Author Contributions: Conceptualization, J.K. and S.L. (Seongmin Lee); methodology, S.L. (Seongmin Lee) and H.Y.; software, H.Y. and S.P.; validation, S.L. (Seongmin Lee) and H.Y.; formal analysis, S.L. (Seongmin Lee) and H.Y.; writing—review and editing, J.K. and S.L. (Sanghoon Lee); supervision, J.K.; project administration, J.K.; funding acquisition, J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) (No. RS-2023-00229451, Interoperable Digital Human (Avatar) Interlocking Technology Between Heterogeneous Platforms).

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the editor and the reviewers for their contributions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of this study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

IDD	Identity displacement discriminator
EDD	Expression displacement discriminator
TDD	Temporal displacement discriminator
3D	Three-dimensional
2S	Two-dimensional
GAN	Generative adversarial network
PCA	Principal component analysis
SVD	Singular value decomposition
TPS	Thin-plate splines
3DMM	3D morphable model

References

1. Blanz, V.; Vetter, T. A morphable model for the synthesis of 3D faces. In Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA, 8–13 August 1999; pp. 187–194.
2. Li, T.; Bolkart, T.; Black, M.J.; Li, H.; Romero, J. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* **2017**, *36*, 194:1–194:17. [CrossRef]
3. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
4. Kim, J.; Oh, H.; Kim, S.; Tong, H.; Lee, S. A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3490–3500.
5. Sun, G.; Wong, Y.; Cheng, Z.; Kankanhalli, M.S.; Geng, W.; Li, X. DeepDance: Music-to-dance motion choreography with adversarial learning. *IEEE Trans. Multimed.* **2020**, *23*, 497–509. [CrossRef]
6. Shen, J.; Zafeiriou, S.; Chrysos, G.G.; Kossaifi, J.; Tzimiropoulos, G.; Pantic, M. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 50–58.

7. Cao, C.; Weng, Y.; Zhou, S.; Tong, Y.; Zhou, K. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Trans. Vis. Comput. Graph.* **2013**, *20*, 413–425.
8. Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; Vetter, T. A 3D face model for pose and illumination invariant face recognition. In Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, Genova, Italy, 2–4 September 2009; pp. 296–301.
9. Vlasic, D.; Brand, M.; Pfister, H.; Popovic, J. Face transfer with multilinear models. In Proceedings of the SIGGRAPH06: Special Interest Group on Computer Graphics and Interactive Techniques Conference, Boston, MA, USA, 30 July–3 August 2006; p. 24.
10. Bookstein, F.L.; Green, W. A thin-plate spline and the decomposition of deformations. *Math. Methods Med. Imaging* **1993**, *2*, 3.
11. Amberg, B.; Romdhani, S.; Vetter, T. Optimal step nonrigid ICP algorithms for surface registration. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
12. Blanz, V.; Vetter, T. Face recognition based on fitting a 3D morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 1063–1074. [CrossRef]
13. Jourabloo, A.; Liu, X. Pose-invariant face alignment via CNN-based dense 3D model fitting. *Int. J. Comput. Vis.* **2017**, *124*, 187–203. [CrossRef]
14. Kang, J.; Lee, S. A greedy pursuit approach for fitting 3d facial expression models. *IEEE Access* **2020**, *8*, 192682–192692. [CrossRef]
15. Kang, J.; Lee, S.; Lee, S. Competitive learning of facial fitting and synthesis using uv energy. *IEEE Trans. Syst. Man Cybern. Syst.* **2021**, *52*, 2858–2873. [CrossRef]
16. Kang, J.; Song, H.; Lee, K.; Lee, S. A Selective Expression Manipulation with Parametric 3D Facial Model. *IEEE Access* **2023**, *11*, 17066–17084. [CrossRef]
17. Kang, J.; Lee, S.; Lee, S. UV Completion with Self-referenced Discrimination. In Proceedings of the EUROGRAPHICS 2020, Norrköping, Sweden, 25–29 May 2020; pp. 61–64.
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
19. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
20. Guo, J.; Zhu, X.; Yang, Y.; Yang, F.; Lei, Z.; Li, S.Z. Towards fast, accurate and stable 3d dense face alignment. In *Computer Vision—ECCV 2020, Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020; pp. 152–168.
21. Sanyal, S.; Bolkart, T.; Feng, H.; Black, M.J. Learning to regress 3D face shape and expression from an image without 3D supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7763–7772.
22. Li, H.; Wang, B.; Cheng, Y.; Kankanhalli, M.; Tan, R.T. DSFNet: Dual Space Fusion Network for Occlusion-Robust 3D Dense Face Alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 4531–4540.
23. Ruan, Z.; Zou, C.; Wu, L.; Wu, G.; Wang, L. Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction. *IEEE Trans. Image Process.* **2021**, *30*, 5793–5806. [CrossRef] [PubMed]
24. Zhu, X.; Liu, X.; Lei, Z.; Li, S.Z. Face alignment in full pose range: A 3d total solution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *41*, 78–92. [CrossRef] [PubMed]
25. Jourabloo, A.; Liu, X. Large-pose face alignment via CNN-based dense 3D model fitting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4188–4196.
26. Dong, J.; Zhang, Y.; Fan, L. A Multi-View Face Expression Recognition Method Based on DenseNet and GAN. *Electronics* **2023**, *12*, 2527. [CrossRef]
27. Białek, C.; Matiolański, A.; Grega, M. An Efficient Approach to Face Emotion Recognition with Convolutional Neural Networks. *Electronics* **2023**, *12*, 2707. [CrossRef]
28. Lee, S.; Lee, J.; Kim, M.; Lee, S. Region Adaptive Self-Attention for an Accurate Facial Emotion Recognition. In Proceedings of the 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, 7–10 November 2022; pp. 791–796.
29. Bi, S.; Lombardi, S.; Saito, S.; Simon, T.; Wei, S.E.; Mcphail, K.; Ramamoorthi, R.; Sheikh, Y.; Saragih, J. Deep relightable appearance models for animatable faces. *ACM Trans. Graph. (TOG)* **2021**, *40*, 1–15. [CrossRef]
30. Sevastopolsky, A.; Ignatiev, S.; Ferrer, G.; Burnaev, E.; Lempitsky, V. Relightable 3d head portraits from a smartphone video. *arXiv* **2020**, arXiv:2012.09963.
31. Zhang, Y.; Yang, J.; Liu, Z.; Wang, R.; Chen, G.; Tong, X.; Guo, B. Virtualcube: An immersive 3d video communication system. *IEEE Trans. Vis. Comput. Graph.* **2022**, *28*, 2146–2156. [CrossRef] [PubMed]
32. Ahn, S.J.; Levy, L.; Eden, A.; Won, A.S.; MacIntyre, B.; Johnsen, K. IEEEVR2020: Exploring the first steps toward standalone virtual conferences. *Front. Virtual Real.* **2021**, *2*, 648575. [CrossRef]