Abstract

Accurate facial landmark detection on wild images plays an essential role in human-computer interaction, entertainment, and medical applications. Existing approaches have limitations in enforcing 3D consistency while detecting 3D/2D facial landmarks due to the lack of multi-view in-the-wild training data. Fortunately, with the recent advances in generative visual models and neural rendering, we have witnessed rapid progress towards high quality 3D image synthesis. In this work, we leverage such approaches to construct a synthetic dataset and propose a novel multi-view consistent learning strategy to improve 3D facial landmark detection accuracy on in-the-wild images. The proposed 3D-aware module can be plugged into any learning-based landmark detection algorithm to enhance its accuracy. We demonstrate the superiority of the proposed plug-in module with extensive comparison against state-of-the-art methods on several real and synthetic datasets.

1. Introduction

Accurate and precise facial landmark plays a significant role in computer vision and graphics applications, such as face morphing [54], facial reenactment [58], 3D face reconstruction [17, 18, 30], head pose estimation [38], face recognition [1, 10, 13, 19, 32, 41, 71], and face generation [11, 21, 60, 69]. In these applications, facial landmark detection provides great sparse representation to ease the burden of network convergence in different training stages and is often used as performance evaluation metric. For instance, as a facial prior, it provides good initialization for subsequent training [66, 67, 69, 76], good intermediate representation to bridge the gap between different modalities for content generation [11, 27, 51, 79], loss terms which reg-
ularize the facial expression [11, 52], or evaluation metrics to measure the facial motion quality [53, 73, 78].

The aforementioned applications require the estimated facial landmarks to be accurate even with significantly varied facial appearance under different identities, facial expressions, and extreme head poses. Tremendous efforts have been devoted to address this problem [15, 22–24, 29, 34, 40, 56, 63, 74, 75, 77, 82, 84]. These approaches often rely on manually annotated large-scale lab-controlled or in-the-wild image datasets [4, 34] to handle various factors such as arbitrary facial expressions, head poses, illumination, facial occlusions, etc.

However, even with the high cost of human labeling, consistent and accurate manual annotation of landmarks remains challenging [22, 23, 34]. It is very difficult, if not impossible, to force a person to annotate the facial landmark keypoints at the same pixel locations for faces of different poses, let alone different annotators under different labeling environments. Such annotation inconsistency and inaccuracy in training images are often the killing factor to learn an accurate landmark localization model. This is particularly a major problem in non-frontal faces where annotation becomes extremely challenging. As shown in Fig. 1(a) a small annotation variation in view #1, results in a significant inaccuracy in view #2. This multi-view inconsistency and inaccuracy can ultimately lead to poor landmark detection accuracy, especially for facial images with extreme head pose.

To mitigate this annotation inconsistency and inaccuracy issue, we propose to learn facial landmark detection by enforcing multi-view consistency during training. Given the images of the same facial identity captured with different head poses, instead of detecting facial landmark at each separate facial image, we propose a multi-view consistency supervision to locate facial landmark in a holistic 3D-aware manner. To enforce multi-view consistency, we introduce self-projection consistency loss and multi-view landmark loss in training. We also propose an annotation generation procedure to exploit the merits of lab-controlled data (e.g., multi-view images, consistent annotations) and in-the-wild data (e.g., wide range of facial expressions, identities). Thanks to this synthetic data, our method does not rely on human annotation to obtain the accurate facial landmark locations. Therefore, it alleviates the problem of learning from inaccurate and inconsistent annotations.

We formulate our solution as a plug-in 3D aware module, which can be incorporated into any facial landmark detector and can boost a pre-trained model with higher accuracy and multi-view consistency. We demonstrate the effectiveness of our approach through extensive experiments on both synthetic and real datasets. The main contributions of our work are as follows:

- We show, for the first time, how to combine the merits of lab captured face image data (e.g., multi-view) and the in-the-wild face image datasets (e.g., appearance diversity). Using our proposed approach we produce a large-scale synthetic, but realistic, multi-view face dataset, titled DAD-3DHeads-Syn.

- We propose a novel 3D-aware optimization module, which can be plugged into any learning-based facial landmark detection methods. By refining an existing landmark detection algorithm using our optimization module, we are able to improve its accuracy and multi-view consistency.

- We demonstrate the performance improvements of our module built on top multiple baseline methods on simulated dataset, lab-captured datasets, and in-the-wild datasets.

2. Related Work

In this section, we review face landmark datasets and detection algorithms that are most related to our approach. We also provide a brief review of data simulation tools related to our work.

2.1. Face Landmark Detection Dataset

Lab-controlled dataset.Datasets under “controlled” conditions [8, 20, 36, 39, 46, 48, 64, 65, 72] typically collect video/images from indoor scenarios with certain restrictions, e.g. pre-defined expressions, head poses, etc. For example, FaceScape dataset [65] contains 938 individuals and each with 20 expressions using an array of 68 cameras under controlled illumination and positions. Thus, it contains aligned and consistent multi-view images and facial landmark annotations. However, the identities, poses, and expressions are limited. In addition, the environment conditions are fully controlled. These result in limited generalization capability of models trained on this dataset. Moreover, the annotation workflow of such a dataset is expensive and hard to scale.

In-the-wild dataset. The boom of internet image sharing has enabled the creation of many “in-the-wild” facial landmark datasets [3, 7, 32, 49, 85], collected from the web, to facilitate facial landmark detection research. However, manually annotating facial landmarks on in-the-wild images is a time-consuming process and not scalable. Zhu et al. [83] release 300W-LP by extending the original 300W dataset with synthetic images with extreme pose through image profiling of frontal pose images. However, the novel view images are generated by simply applying rotation matrix on the original images, which leads to limited view range and poor image quality. Meanwhile, 300W-LP lacks diversity in face appearance and expression because of the intrinsic limitations of 300W. Recently, Martyniuk et al. [34] introduce a
new dataset, DAD-3DHeads, by proposing a novel annotation scheme. Specifically, their approach allows the annotator to adjust the landmarks by looking at how well the mesh, generated from the landmarks, fits the input image. The proposed scheme addresses the problems exhibited by existing labeling tools, such as “guessing” the positions of the correct landmarks for invisible parts of the head, thus enabling accurate annotations. DAD-3DHeads dataset contains 44,898 in-the-wild images, covering extreme facial expressions, poses, and challenging illuminations. However, the DAD-3DHeads still has some drawbacks. First, even with the mesh fitting guidance, the annotations can be inaccurate. As shown in Fig. 1 (a), even a small inaccuracy in one view could result in a significant inconsistency when projected to another view. This inconsistency could negatively affect the training of the detection network. Second, since the depth is estimated by FLAME [33], annotation accuracy is limited by the FLAME model. Third, this dataset lacks multi-view images, and thus cannot be used to enforce multi-view consistency.

2.2. Data Simulation

Simulation [26, 28, 35, 42, 44, 45, 50, 59, 61, 62, 70] is a useful tool in situations where training data for learning-based methods is expensive to annotate or even hard to acquire. For example, Zeng et al. [70] and Richardson et al. [42] use 3D Morphable Model (3DMM) to render training data with different lighting conditions, identities, expressions, and texture basis elements for reconstructing detailed facial geometry. However, the simulated images produced by these approaches lack realism and have severe domain gaps compared with real-world captures, limiting their usage. Bak et al. [2] adapt synthetic data using a CycleGAN [81] with a regularization term for preserving identities. Ayush et al. [57] use the images and latent code generated by StyleGAN [81] to train a controllable portrait image generation model. However, it is hard to control the attribute consistencies of images simulated by generative models, which limits the usage of the generated datasets.

2.3. Face Landmark Detection Algorithms

Traditional facial landmark detection methods leverage either holistic facial appearance information [12], or the global facial shape patterns [31, 85]. They yield reasonable results for images captured in lab-controlled environments with frontal faces and good lighting, however the performance on most of in-the-wild images is inferior.

Recently, deep learning-based algorithms have made promising progress on 2D facial landmark localization [15, 22–24, 29, 34, 40, 56, 63, 74, 75, 77, 82, 84] in terms of robustness, generalizability, and accuracy. FAN [6] constructs, for the first time, a very strong baseline by combining a state-of-the-art residual block and a state-of-the-art architecture for landmark localization and trains it on a very large yet synthetically expanded 2D facial landmark dataset. To address self-occlusion and large appearance variation, Zhu et al. [82] propose a cascaded convolutional neural network and optimized weighted parameter distance cost loss function to formulate the priority of 3DMM parameters during training instead of predicting facial landmark keypoints. To further address the problems of shape reconstruction and pose estimation simultaneously, Martyniuk et al. propose an end-to-end trained DAD-3DNet [34] to regress 3DMM parameters and recover the 3D head geometry with differential FLAME decoder. However, due to the intrinsic limitation of the manually annotated in-the-wild dataset, the detection results are affected by the annotation noise and the 3D inconsistency of the single view images. In this paper, we mainly focus on improving the performance of deep-learning based methods.

3. Balanced and Realistic Multi-view Face Dataset

We believe there are five desired properties that a good facial landmark dataset should fulfill: (1) contain full range of multi-view images; (2) bridge the domain gap between the dataset and the real-world captured images; (3) contain diverse facial appearance including different poses, expressions, illuminations, and identities; (4) have consistent and accurate annotations across the whole dataset; (5) be
easy to obtain and scalable. The existing datasets can are either lab-controlled captures \[64, 65\] or in-the-wild collected \[34, 47, 68\]. Unfortunately, these datasets lack one or more desired attributes. In contrast, our dataset meets all of these criteria (Fig. 2).

Unlike previous graphics or generative model-based data synthesis approaches described in Sec. 2.2, we propose a novel facial dataset simulation scheme by leveraging Neural Radiance Field (NeRF) \[37\] to facilitate training a facial landmark detection network. Fig. 3 shows our dataset creation pipeline. We generate multiview images with consistent landmarks using a single in-the-wild image along with annotated landmark as input.

Specifically, we choose DAD-3DHeads \[34\] as our initial dataset since it contains images under a variety of extreme poses, facial expressions, challenging illuminations, and severe occlusions cases. Given an image and its landmarks from this dataset, our goal is to reconstruct multiview images with their corresponding landmarks. Inspired by GAN inversion \[80\], we first fit a latent code to each image in DAD-3DHeads datasets using EG3D \[9\] as decoder by following Pivotal Tuning Inversion (PTI) \[43\]. Note that, EG3D GAN inversion requires the camera pose of the input image, which we estimate using Deep3DFace \[18\]. Then we can use EG3D to decode the optimized latent code to NeRF. Next, we use volume rendering on the NeRF with 512 uniformly sampled camera views from a large view range, producing 512 multi-view images.

To obtain the landmarks for each image, we start with the well-annotated groundtruth 2D landmarks of the original images from the DAD-3DHeads dataset. Then we use the estimated camera pose of the input image to unproject the annotated landmarks to 3D space. At last, we project the 3D landmarks to the 512 sampled camera views to obtain landmark annotation on the simulated views. The simulated dataset not only inherits the merits of DAD-3DHeads (e.g., diverse identities, expressions, poses, and illuminations), but also comes with a lot of new features (e.g., balanced head pose, consistent annotation, and multi-view images). In total, there are 2,150,400 training pairs and 204,800 testing pairs in our extended dataset, called DAD-3DHeads-Syn.

4. 3D-Aware Multi-view Consistency Training

4.1. Overview

The state-of-the-art landmark detectors \[5, 34\] can output reasonable results on in-the-wild images. However, we may observe that the predicted landmark are floating on the face surface instead of fitting the face perfectly in a lot of cases. We can easily verify if the detected landmark fits the face by projecting the detected landmark to another view (see Fig. 1(a)). Armed by this observation of multi-view in-
Sec. 3). Moreover, pre-defined for view synthesis during volume rendering (see are the corresponding camera extrinsic matrices which are denoted as green, blue, red, and yellow points respectively. The processes of calculating 3D landmark \( L \) and the projection procedure are shown as light blue and pink arrows, respectively. \( \mathcal{L}_{\text{Self-Cons}} \) and \( \mathcal{L}_{\text{MultiView}} \) are represented as red and light green lines, respectively.

points in Fig. 4. We then randomly select \( P \) predicted landmarks \( \hat{L}_{1,...,P} \in \mathbb{R}^{P \times 68 \times 2} \) from \( \hat{L}_{1,...,N} \) to calculate the “canonical” 3D landmark \( \hat{L} \in \mathbb{R}^{6 \times 3} \), as shown by the blue point in Fig. 4. We calculate each keypoint of the “canonical” 3D landmark \( \hat{L}(k) \in \mathbb{R}^{3}, 1 \leq k \leq 68 \) through Direct Linear Transformation (DLT) \([16, 25]\), as follows:

\[
\mu_p = M_p[0, :] - M_p[2, :] \cdot \hat{L}_p[0] \in \mathbb{R}^4, \tag{2}
\]

\[
v_p = M_p[1, :] - M_p[2, :] \cdot \hat{L}_p[1] \in \mathbb{R}^4, \tag{3}
\]

\[
A = [\mu_1 | \mu_2 | ... | \mu_p | v_1 | v_2 | ... | v_p]^T \in \mathbb{R}^{2P \times 4}, \tag{4}
\]

\[
\hat{L}(k) = \left( A[:, : 3]^T A[:, : 3] \right)^{-1} A[:, : 3]^T (-A[:, : 3]), \tag{5}
\]

where, \( p, 1 \leq p \leq P \), is the index of views, and \( M_{1,...,P} \) are the corresponding camera extrinsic matrices which are pre-defined for view synthesis during volume rendering (see Sec.3). Moreover, \( M_p[; :, i] \) indicates the \( i \)-th row of \( M_p \), \( A[; :, i] \) indicates columns 0 to \( i - 1 \) of \( A \), and \( A[; :, i] \) indicates the \( i \)-th column of \( A \). By Eq. 2 and Eq. 3, we first calculate the projection constraints for \( \hat{L}(k) \), i.e., \( \mu_p[:, 3] \cdot \hat{L}(k) + v_p[3] = 0 \), where ‘.’ indicates the dot product. Then we stack all of the constraints into \( A \in \mathbb{R}^{2P \times 4} \) by Eq. 4. At last, we compute \( \hat{L}(k) \) with a least square approach (Eq. 5).

After obtaining the “canonical” 3D landmark \( \hat{L} \), we project it onto the image planes of rest of \( Q = N - P \) views to obtain the projected landmark \( \hat{L}_{1,...,Q} \), shown as red points in Fig. 4, by the following equations:

\[
s = M_q[:, 3] \hat{L}(k) + M_q[:, 3] \in \mathbb{R}^{3 \times 1}, \tag{6}
\]

\[
\hat{L}_q(k) = \begin{bmatrix} s[0] / s[2] \\ s[1] / s[2] \end{bmatrix} \in \mathbb{R}^{2 \times 1}, \tag{7}
\]

where, in our case, \( 1 \leq q \leq Q \). Eq. 6 transforms 3D landmark from “canonical” space to the camera space of view \( q \), and Eq. 7 transforms it from camera space to image space.

**Self-Projection Consistency Loss.** Since all \( M \) views are sampled from one NeRF with different camera views, the predicted landmarks \( \hat{L}_{1,...,Q} \) and the projected landmarks \( \hat{L}_{1,...,Q} \) should be consistent. Therefore, we propose to minimize the error between the predicted and projected landmarks as follows:

\[
\mathcal{L}_{\text{Self-Cons}} = \sum_{q=1}^{Q} || \hat{L}_q - \hat{L}_q ||_1. \tag{8}
\]

**Mesh Consistency Loss** Besides the self-projection consistency, all the \( N \) views also share one mesh topology in the canonical space. Therefore, we apply a mesh consistency loss in canonical space calculated by:

\[
\mathcal{L}_{\text{Mesh-Cons}} = \sum_{n=1}^{N} || \hat{M}_n - \hat{M} ||_2, \tag{9}
\]

where \( \hat{M}_n \) is the predicted mesh of view \( n \) in the canonical space, and \( \hat{M} \) is the ground truth mesh of the original reference image.

**Multiview Landmark Loss.** We also minimize the distance between the predicted 2D facial landmarks and the corresponding multi-view ground truth landmarks we obtained in Sec. 3, which are denoted as yellow points in Fig. 4. The loss can be formulated as follows:

\[
\mathcal{L}_{\text{MultiView}} = \sum_{q=1}^{N} || \hat{L}_q - L_q ||_1. \tag{10}
\]

We also incorporate the original loss of the baseline method computed with the image and landmark pairs \( \{ I, L \}_{1,...,M} \) from dataset \( D \) to stabilize our 3D-aware training. The overall loss is:

\[
\mathcal{L} = \lambda_1 \mathcal{L}_{\text{Self-Cons}} + \lambda_2 \mathcal{L}_{\text{Mesh-Cons}} + \lambda_3 \mathcal{L}_{\text{MultiView}} + \mathcal{L}_{\text{Original}}, \tag{11}
\]

where \( \lambda_{1,2,3} \) are hyper parameters that control the contribution of each components. We set \( \lambda_{1,2,3} \) to 0.1 empirically.

Note that our training is a plug-in module and can be incorporated into any existing facial landmark detector easily. For different pretrained models, we just need to change \( \mathcal{L}_{\text{Original}} \), while the other novel loss components calculated on our balanced synthetic dataset \( \mathcal{D} \) can be applied directly. We show this plug-in capability on top of different baseline methods (e.g., DAD-3DNet [34] and 3DDFA [22]), and demonstrate that our 3D-aware training indeed improves their performance (see Sec. 5).

*We can apply it depending on whether the baseline network outputs mesh. In our case, the 3DDFA [22] and DAD-3DNet [34] both do.*
Dataset. Besides DAD-3DHeads, we use two additional A6000 GPUs.

Training Details. We implement our algorithm in Pytorch and adopt ADAM to optimize the baseline networks. We run our 3D-aware training for 100 epochs with a batch size of 4, and a learning rate of \(1 \times 10^{-4}\) on each baseline network. As to computational cost, fine-tuning DAD-3DNet takes about 16.25 hours on 4 NVIDIA RTX A6000 GPUs.

5. Experiments

5.1. Experimental Settings

Training Details. We implement our algorithm in Pytorch and adopt ADAM to optimize the baseline networks. We run our 3D-aware training for 100 epochs with a batch size of 4, and a learning rate of \(1 \times 10^{-4}\) on each baseline network. As to computational cost, fine-tuning DAD-3DNet takes about 16.25 hours on 4 NVIDIA RTX A6000 GPUs.

Dataset. Besides DAD-3DHeads, we use two additional datasets to conduct the evaluations.

- DAD-3DHeads [34] is the state-of-the-art in-the-wild 3D head dataset, which contains dense, accurate annotations, and diverse facial appearances. It consists of 44,898 images collected from various sources (37,840 in the training set, 4,312 in the validation set, and 2,746 in the test set).

- FaceScape [65] is a large-scale high-quality lab-controlled 3D face dataset, which contains 18,760 examples, captured from 938 subjects and each with 20 specific expressions.

- MultiFace [64] is a new multi-view, high-resolution human face dataset collected from 13 identities for neural face rendering.

Training and Testing Split. In all the experiments, we only refine the baseline models with the training set of our DAD-3DHeads-Syn and their original training dataset. We use the test sets of DAD-3DHeads-Syn and DAD-3DHeads [34], and use the full datasets of FaceScape [65] and MultiFace [63] for performance evaluation. All the comparison methods have not been trained on the split test sets.

Evaluation Metrics. We evaluate the facial landmark distance by calculating the Normalized Mean Error (NME). We normalize the landmark error by dividing its image resolution instead of the eye distance [55], since all the test images are aligned with offline tools. We calculate the head pose error by the absolute distance of the Euler angle values.

5.2. Quantitative Evaluation

Landmark Detection Results. The quantitative landmark detection results on DAD-3DHeads [34], FaceScape [65], and MultiFace [64] are shown in Tab. 1. We can find that the DAD-3DNet refined by our 3D-aware multi-view consistency training achieves the best performance on all three datasets. Moreover, according to the results of 3DDFA [22], 3DDFA+, DAD-3DNet [34], and DAD-3DNet+, we find that after refinement, the new models (3DDFA+ and DAD-3DNet+) achieve much better results than the baseline models. For example, the detection error of DAD-3DNet [34] drops 0.631 and 0.306, a 9% and 5% improvement, on FaceScape and MultiFace datasets, respectively. Similarly, we improve the 3DDFA [22] by 0.298 (7%), 0.563 (7%), and 0.816 (10%) on DAD-3DHeads, FaceScape and MultiFace datasets, respectively. We attribute the improvement to our proposed 3D aware multi-view training. One interesting phenomenon is that all the methods perform better on DAD-3DHeads dataset than the other two lab-captured datasets. We attribute this to the extreme head pose and challenging facial expressions in the other two datasets. We plot the head pose distribution of DAD-3DHeads (see supplementary materials) and find that distribution of head pose is not as uniform as the other two lab-controlled datasets.

Head Pose Estimation Results. Tab. 2 shows the head pose estimation error on DAD-3DHeads [34] and FaceScape [65]. Our DAD-3DNet+ achieves best performance in most metrics. Similar to the landmark results, we can also conclude that head pose detection accuracy of the baseline methods (3DDFA and DAD-3DNet) is improved by our 3D aware multi-view consistency (3DDFA+ and DAD-3DNet+). For example, after refinement, DAD-3DNet+ achieves 11.9% and 18.8% performance boosts in overall head pose error on DAD-3DHeads and FaceScape dataset, respectively.

5.3. Qualitative Evaluation

We fist show visual comparisons on images randomly sampled from DAD-3DHeads test set [34] in Fig. 5. The landmark predicted by our DAD-3DNet+ model fits the individual’s face tighter than the other predictions. Furthermore, by comparing the third (3DDFA [22]) and forth columns (ours), we can see that refining model (3DDFA+) improves the landmark accuracy dramatically. Similar visual improvements can be found in sixth (DAD-3DNet) and seventh (DAD-3DNet+) columns as well. Comparing the sixth and seventh column, we can see that the refinement training drags and rotates the landmark in 3D space to better fit it to the individual’s face surface. We attribute this abil-

<table>
<thead>
<tr>
<th>Method</th>
<th>DAD-3DHeads</th>
<th>FaceScape</th>
<th>MultiFace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dlib [31]</td>
<td>10.841</td>
<td>29.431</td>
<td>18.205</td>
</tr>
<tr>
<td>3DDFA-V2 [23]</td>
<td>2.926</td>
<td>6.853</td>
<td>5.942</td>
</tr>
<tr>
<td>3DDFA [22]</td>
<td>4.082</td>
<td>7.988</td>
<td>8.121</td>
</tr>
<tr>
<td>3DDFA+</td>
<td>3.784</td>
<td>7.425</td>
<td>7.305</td>
</tr>
<tr>
<td>DAD-3DNet [34]</td>
<td>2.599</td>
<td>6.681</td>
<td>5.786</td>
</tr>
<tr>
<td>DAD-3DNet+</td>
<td>2.503</td>
<td>6.050</td>
<td>5.480</td>
</tr>
</tbody>
</table>
Figure 5. The visual results of Dlib [31], FAN [5], 3DDFA [22], our refined 3DDFA+, 3DDFA-V2, DAD-3DNet [34], and our refined DAD-3DNet+ on images randomly sampled from DAD-3DHeads [34] testing set. We show the enlarged error region (while box) in the middle row.

Table 2. Head pose estimation results (head pose error) on DAD-3DHeads [34], FaceScape [65]. Lower values mean better results.

<table>
<thead>
<tr>
<th></th>
<th>DAD-3DHeads</th>
<th>FaceScape</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pitch</td>
<td>Roll</td>
</tr>
</tbody>
</table>

5.4. Performance Improvement Analysis

To systematically understand the source of improvement after refining the baseline methods (DAD-3DNet [34] and 3DDFA [22]) with our proposed 3D-aware multi-view consistency training, we further calculate and plot the landmark and head pose error improvements on DAD-3DHeads [34] (see Fig. 7). Instead of calculating the overall improved error score, we split all the testing images into different groups according to their head pose value and calculate the improved error score within each group. We can find that the improvement by our training gets more obvious as the head pose gets more challenging. For example, the landmark error improvement (Fig. 7 upper section) using our method built on top of 3DDFA [22] increases from 0.12 to 0.71. Similarly, the head pose estimation error (Fig. 7 lower section) improvement using our method built on top of DAD-3DNet [34] increases from 0.02 to 2.7. We also show the detection result visualization in Fig. 8. We can see that from left to right, as the head pose increases, the error of the DAD-3DNet+ (second row) is more stable than the error (first row) of the DAD-3DNet. Base on this trend, we conclude that our proposed 3D-aware multi-view consistency training provides a more significant improvement over the baselines on images with larger head pose. This verifies our hypothesis that multi-view consistency training enables the network to learn 3D-aware information, which benefits the detection results on images with large head pose.
Figure 6. The visual comparisons between baseline methods and the refined methods on four testing sets. The left column and upper row list the dataset and method names, respectively. ‘+’ denotes the model that has been refined by our 3D-aware training.

Figure 7. The landmark (top) and head pose (bottom) error improvement over DAD-3DNet [34] and 3DDFA [22] on images from different head pose ranges. The solid and dotted lines indicate DAD-3DNet [34] vs. DAD-3DNet+ (ours) and 3DDFA [22] vs. 3DDFA+ (ours).

5.5. Ablation Study

We conduct ablation study on FaceScape [65] to verify the importance of main components of our novel design. As shown in Tab. 3, we calculate NME of landmark and MAE of pose estimation in these ablation experiments. Based on these numbers, we can see the performance degrades drastically when we remove $\mathcal{L}_{\text{Self-Cons}}$. Moreover, removing $\mathcal{L}_{\text{Mesh-Cons}}$ negatively impacts the results, demonstrating its importance. Moreover, estimating the 3D landmarks in the world space using fewer views leads to better results. This is a significant advantage as it makes our fine-tuning process more efficient.

Table 3. Ablation Study on FaceScape [65]. The top 2 numbers are shown in bold.

<table>
<thead>
<tr>
<th>Component</th>
<th>NME ↓</th>
<th>Pose ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. full model (P=4)</td>
<td>6.050</td>
<td>11.863</td>
</tr>
<tr>
<td>2. w/o $\mathcal{L}_{\text{Mesh-Cons}}$</td>
<td>6.168</td>
<td>12.327</td>
</tr>
<tr>
<td>3. w/o $\mathcal{L}_{\text{Self-Cons}}$</td>
<td>6.541</td>
<td>13.623</td>
</tr>
<tr>
<td>4. full model (P=8)</td>
<td>6.048</td>
<td>11.923</td>
</tr>
<tr>
<td>5. full model (P=16)</td>
<td>6.098</td>
<td>11.902</td>
</tr>
<tr>
<td>6. full model (P=32)</td>
<td>6.139</td>
<td>11.912</td>
</tr>
</tbody>
</table>

6. Conclusion

We propose 3D-aware multi-view consistency training, a new framework for improving deep-learning base landmark detection algorithms. Through a set of novel loss functions, we force the network to produce landmarks that are 3D consistent. We additionally introduce a novel dataset simulation pipeline to combine the merits of lab-controlled captures and in-the-wild collected images. The model refined by our method outperforms previous approaches in terms of landmark detection accuracy and head pose estimation accuracy. Admittedly, our work has some limitations. For example, our proposed training relies on the performance of the baseline method. If the pretrained baseline yield poor initial predictions, our DLT would fail to estimate reasonable canonical 3D landmark, affecting the performance of the proposed self-projection consistency loss. Investigating ways to reduce the reliance on the accuracy of the baseline methods would be an interesting future research.
References


