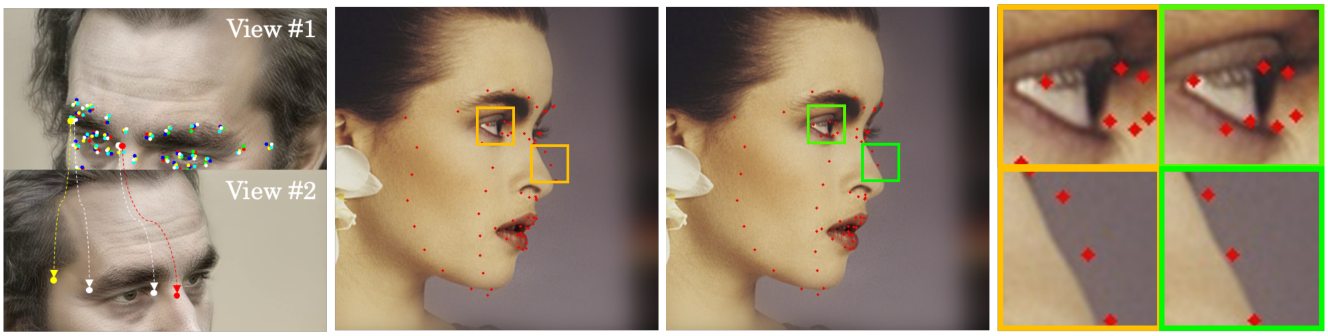
3D-aware Facial Landmark Detection via Multi-view Consistent Training on Synthetic Data

Libing Zeng [1*], Lele Chen[2], Wentao Bao[3*], Zhong Li[2], Yi Xu[2], Junsong Yuan[4], Nima K. Kalantari[1]
[1]Texas A&M University, [2]OPPO US Research Center, InnoPeak Technology, Inc,
[3]Michigan State University, [4]University at Buffalo



(a) Multi-view Inconsistency (b) DAD-3DNet (c) DAD-3DNet+ (Ours) (d)

Figure 1. We plot the landmark annotations labeled by different annotators with different colors in view #1 of (a). Accurate annotation of non-frontal faces with large angles like view #1 is challenging. This is a major problem since small differences between annotated landmarks in view #1, becomes substantially magnified when projected to view #2. Training a system on such datasets could lead to poor landmark detection accuracy, as shown in (b). We address this issue by proposing a 3D-aware optimization module that enforces multi-view consistency. We show the landmark detection improvement in (c). Magnified insets in (b) and (c) are shown in (d). After refined by the proposed 3D-aware learning, the detected facial landmark is better aligned with the identity.

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-

[*]This work was done when Libing Zeng and Wentao Bao were interns at OPPO US Research Center, InnoPeak Technology, Inc.

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.

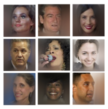
Dataset Type	Lab-Controlled	In-the-wild	Ours
Examples			
In-the-wild	✗	✓	✓
Large Scale	✗	✓	✓
Balanced	✓	✗	✓
Multiview Consistent	✓	✗	✓
Annotation Consistent	✓	✗	✓
Scalable	✗	✗	✓

Figure 2. The feature comparison of different type of datasets. For example, FaceScape [65] and MultiFace [64] are lab-controlled datasets, while 300W [47], AFLW2000 [68], and DAD-3DHeads [34] are in-the-wild datasets.

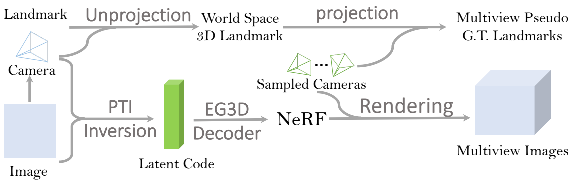


Figure 3. The proposed data simulation pipeline.

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 multi-view 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 [14]. 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 inconsistency and inaccuracy, we propose a novel 3D-Aware training module $\mathcal{R}$ to further improve the performance of baseline detection algorithm F.

Algorithm 1 3D-Aware Plug-in Module.

```
Input: pretrained detector $F$ with weights $\theta$, $M$ single-view images $I_{1,\dots,M} \in \mathcal{D}$ along with ground truth landmark $\mathrm{L}_{1,\dots,M}$ , paired $N$ multi-view images $V_{1,\dots,N} \in \hat{\mathcal{D}}$ along with ground truth landmark $\mathrm{L}_{1,\dots,N}$.
Output: detector $F$ with updated weights $\theta^*$
Initialization: set $\theta$ to pre-trained weights
Unfreeze $\theta$
for number of iterations do
    Output predicted landmarks $\hat{\mathrm{L}}_{1,\dots,N}$ for each view.
    Randomly sample $P$ landmarks from them, $(1 < P \leq N)$.
    Cast the landmarks into world space and estimate the approximate 3D landmark $\dot{\mathrm{L}}$ using Eq. 2, 3, 4, 5
    Project $\dot{\mathrm{L}}$ onto the image planes of remaining $Q$ views ($Q = N - P$) using Eq. 6, 7
    Calculate Total Loss $\mathcal{L}$ using Eq. 11
    $\theta^* \leftarrow Adam\{\mathcal{L}\}$
```

Given a facial landmark detection network $F_\theta(\cdot)$ pretrained on dataset $\mathcal{D}$, the proposed module $\mathcal{R}$ further refines the network parameters θ by leveraging our simulated DAD-3DHeads-Syn dataset $\hat{\mathcal{D}}$ in addition to the original dataset $\mathcal{D}$. Our module $\mathcal{R}$ can be formulated as:

$$F_{\theta^*} \leftarrow \mathcal{R}(F_\theta, X, V_{1,\dots,N}), X \in \mathcal{D}, V_{1,\dots,N} \in \hat{\mathcal{D}}, \tag{1}$$

where X is the image batch sampled from $\mathcal{D}$ and $V_{1,\dots,N}$ are N multi-view images sampled from $\hat{\mathcal{D}}$. We refine the network parameters θ through exploring 3D information among multi-view images and applying a novel projection consistency during the fine-tuning process. Our module $\mathcal{R}$ does not result in any new network parameters and can be plugged into any learning-based network. We show the training protocol in Alg. 1.

4.2. Multi-view Consistency Supervision

We propose a novel multi-view supervision to force the baseline network to learn to be 3D consistent. To simplify notation, we ignore the batch dimension and fixed camera intrinsic matrix. For every training iteration, we randomly sample N image and landmark pairs $\{V, \mathrm{L}\}_{1,\dots,N}$ from $\hat{\mathcal{D}}$ and M image and landmark pairs $\{I, \mathrm{L}\}_{1,\dots,M}$ from initial dataset $\mathcal{D}$[*].

We pass $V_{1,\dots,N}$ to the baseline network F to obtain predicted landmarks $\hat{\mathrm{L}}_{1,\dots,N}$ which are shown with green

[*] $\mathcal{D}$ is DAD-3DHeads dataset when training DAD-3DNet and is AFLW2000-3D when training 3DDFA.

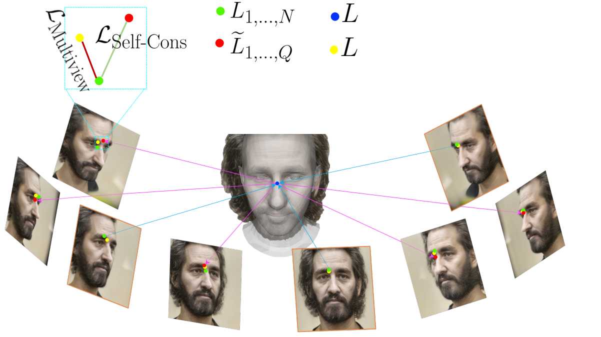

$\mathcal{L}_{\text{Multiview}}$
$\mathcal{L}_{\text{Self-Cons}}$
$\hat{L}_{1,\ldots,N}$
$\tilde{L}_{1,\ldots,Q}$
$\dot{L}$
L

Figure 4. Multi-view Consistency Supervision. Predicted landmarks $\hat{\mathrm{L}}_{1,\ldots,N}$, estimated 3D landmark $\dot{\mathrm{L}}$, projected landmarks $\tilde{\mathrm{L}}_{1,\ldots,Q}$, and ground truth landmarks L are denoted as green, blue, red, and yellow points respectively. The processes of calculating 3D landmark $\dot{\mathrm{L}}$ 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 $\hat{\mathrm{L}}_{1,\ldots,P} \in \mathbb{R}^{P\times 68\times 2}$ from $\hat{\mathrm{L}}_{1,\ldots,N}$ to calculate the "canonical" 3D landmark $\dot{\mathrm{L}} \in \mathbb{R}^{68\times 3}$, as shown by the blue point in Fig. 4. We calculate each keypoint of the "canonical" 3D landmark $\dot{\mathrm{L}}^{(k)} \in \mathbb{R}^3, 1 \le k \le 68$ through Direct Linear Transformation (DLT) [16, 25], as follows:

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

$$\upsilon_p = \mathbb{M}_p[1,:] - \mathbb{M}_p[2,:] \cdot \hat{\mathrm{L}}_p^k[1] \in \mathbb{R}^4, \tag{3}$$

$$\mathbf{A} = [\mu_1 \mid \mu_2 \mid \ldots \mid \mu_p \mid \upsilon_1 \mid \upsilon_2 \mid \ldots \mid \upsilon_p]^T \in \mathbb{R}^{2P\times 4}, \tag{4}$$

$$\dot{\mathrm{L}}^{(k)} = \Big(\mathbf{A}[:,:3]^T \quad \mathbf{A}[:,:3]\Big)^{-1}\mathbf{A}[:,:3]^T(-\mathbf{A}[:,3]), \tag{5}$$

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

After obtaining the "canonical" 3D landmark $\dot{\mathrm{L}}$, we project it onto the image planes of rest of $Q = N - P$ views to obtain the projected landmark $\tilde{\mathrm{L}}_{1,\ldots,Q}$, shown as red points in Fig. 4, by the following equations:

$$s = \mathbb{M}_q[:,:3]\dot{\mathrm{L}}^{(k)} + \mathbb{M}_q[:,3] \in \mathbb{R}^{3\times 1}, \tag{6}$$

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

where, in our case, $1 \le q \le Q$. Eq. 6 transfroms 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 $\hat{\mathrm{L}}_{1,\ldots,Q}$ and the projected landmarks $\tilde{\mathrm{L}}_{1,\ldots,Q}$ should be consistent. Therefore, we propose to minimize the error between the predicted and projected landmarks as follows:

$$\mathcal{L}_{\text{Self-Cons}} = \sum_{q=1}^{Q} \parallel \hat{\mathrm{L}}_q - \tilde{\mathrm{L}}_q \parallel_1 . \tag{8}$$

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:

$$\mathcal{L}_{\text{Mesh-Cons}} = \sum_{n=1}^{N} \parallel \hat{\mathrm{M}}_n - \dot{\mathrm{M}} \parallel_2, \tag{9}$$

where $\hat{\mathrm{M}}_n$ is the predicted mesh of view n in the canonical space, and $\dot{\mathrm{M}}$ 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:

$$\mathcal{L}_{\text{Multiview}} = \sum_{q=1}^{N} \parallel \hat{\mathrm{L}}_q - \mathrm{L}_q \parallel_1 . \tag{10}$$

We also incorporate the original loss of the baseline method computed with the image and landmark pairs $\{I, L\}_{1,..,M}$ from dataset $\mathcal{D}$ 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.

Table 1. Facial landmark detection result (NME) on DAD-3DHeads [34], FaceScape [65], and MultiFace [64]. Lower values mean better results.

Method	DAD-3DHeads	FaceScape	MultiFace
FAN [6]	7.141	16.74	16.143
Dlib [31]	10.841	29.431	18.205
3DDFA-V2 [23]	2.926	6.853	5.942
3DDFA [22]	4.082	7.988	8.121
3DDFA+	3.784	7.425	7.305
DAD-3DNet [34]	2.599	6.681	5.786
DAD-3DNet+	2.503	6.050	5.480

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 1×10^{-4} on each baseline network. As to computational cost, fine-tuning DAD-3DNet take about and 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, respetively.

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-

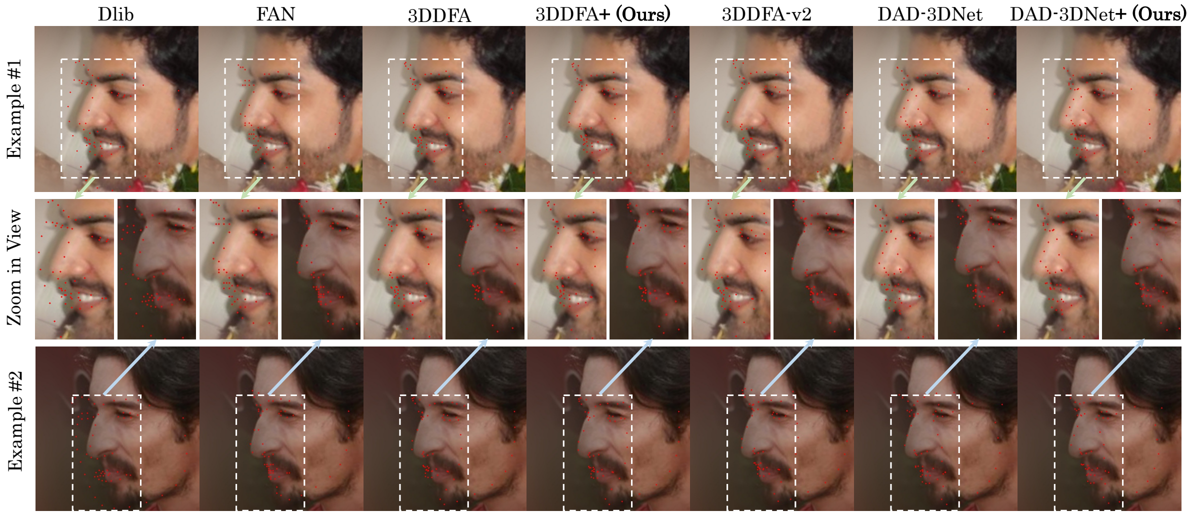


Figure 5. The visual results of Dlib [31], FAN [5], 3DDFA [22], our refined 3DDFA+, 3DDFA-V2, DAD-3DNet [34], and our refined DAD-3DNet+ on images randomly sampled from DAD-3DHeads [34] testing set. We show the enlarged error region (while box) in the middle row.

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

	DAD-3DHeads				FaceScape			
	Pitch	Roll	Yaw	Overall	Pitch	Roll	Yaw	Overall
FAN [5]	9.765	5.376	6.390	7.177	8.774	4.895	6.556	6.742
Dlib [31]	13.352	11.799	14.654	13.268	17.861	12.663	19.548	16.691
3DDFA-V2 [23]	7.901	4.989	6.088	6.326	13.741	9.718	11.353	11.604
3DDFA [22]	9.895	7.977	8.996	8.956	20.789	18.145	19.692	19.752
3DDFA+	9.195	6.792	8.692	8.226	20.996	16.426	19.054	18.826
DAD-3DNet [34]	8.274	4.666	9.206	7.382	15.851	9.676	18.346	14.624
DAD-3DNet+	7.700	4.274	7.528	6.500	14.466	7.247	13.876	11.863

ity to our 3D-aware multi-view consistency training, which lets the refined model gain the better sense in 3D space, and therefore, improve the landmark detection results.

To further validate the improvement gained by the proposed 3D-aware multi-view consistency training, we show the visual results (Fig. 6) of 3DDFA [22], our refined 3DDFA+, DAD-3DNet [34], and our refined DAD-3DNet+ on images sampled from four different test sets. We can find that our proposed refinement improves the landmark detection results in the eye, mouth, and face contour regions, which usually contain more appearance dynamics than the other areas.

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.

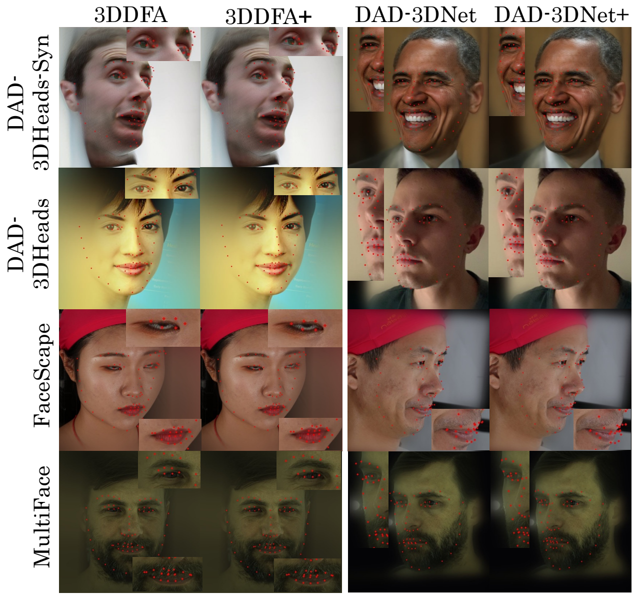


Figure 6. The visual comparisons between baseline methods and the refined methods on four testing sets. The left column and upper row list the dataset and method names, respectively. '+' denotes the model that has been refined by our 3D-aware training.

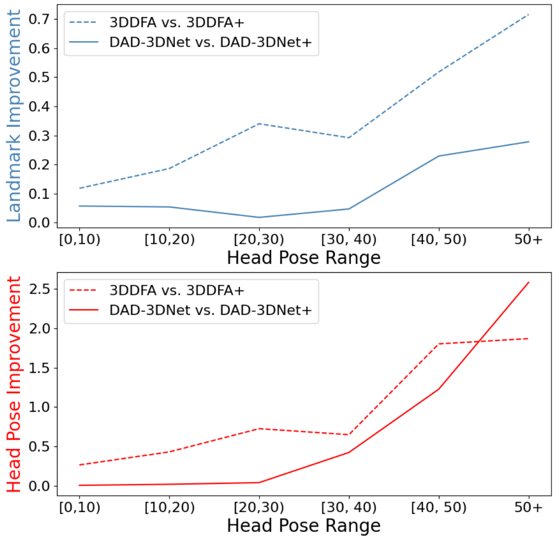


Figure 7. The landmark (top) and head pose (bottom) error improvement over DAD-3DNet [34] and 3DDFA [22] on images from different head pose ranges. The solid and dotted lines indicate DAD-3DNet [34] vs. DAD-3DNet+ (ours) and 3DDFA [22] vs. 3DDFA+ (ours).

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.

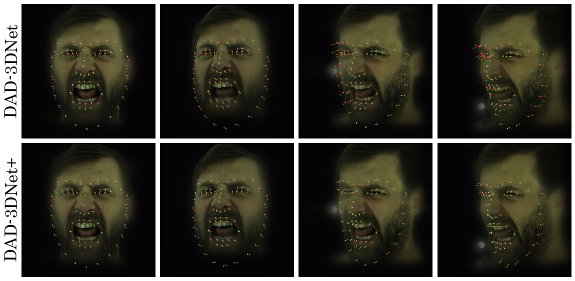


Figure 8. The error visualization of DAD-3DNet [34] and our DAD-3DNet+ on MultiFace [64] dataset. The white and green dots are the ground truth and predicted landmarks, respectively. We use the red line to show the error distance. From left to right, the head pose increases gradually.

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

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

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 pretrianed 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

[1] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7617–7627, 2021.

[2] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 189–205, 2018.

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

[4] Adrian Bulat and Georgios Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *European Conference on Computer Vision*, pages 616–624. Springer, 2016.

[5] Adrian Bulat and Georgios Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3706–3714, 2017.

[6] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017.

[7] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE international conference on computer vision*, pages 1513–1520, 2013.

[8] Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013.

[9] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.

[10] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *European conference on computer vision*, pages 109–122. Springer, 2014.

[11] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7832–7841, 2019.

[12] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. In *European conference on computer vision*, pages 484–498. Springer, 1998.

[13] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5203–5212, 2020.

[14] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[15] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, June 2018.

[16] Xuanyi Dong, Yi Yang, Shih-En Wei, Xinshuo Weng, Yaser Sheikh, and Shoou-I Yu. Supervision by registration and triangulation for landmark detection. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3681–3694, 2020.

[17] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5908–5917, 2017.

[18] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 534–551, 2018.

[19] Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv preprint arXiv:1506.08347*, 2015.

[20] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010.

[21] Kuangxiao Gu, Yuqian Zhou, and Thomas Huang. Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10861–10868, 2020.

[22] Jianzhu Guo, Xiangyu Zhu, and Zhen Lei. 3ddfa. https://github.com/cleardusk/3DDFA, 2018.

[23] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, pages 152–168. Springer, 2020.

[24] Xiaojie Guo, Siyuan Li, Jinke Yu, Jiawan Zhang, Jiayi Ma, Lin Ma, Wei Liu, and Haibin Ling. Pfld: A practical facial landmark detector. *arXiv preprint arXiv:1902.10859*, 2019.

[25] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[26] Stefan Hinterstoisser, Vincent Lepetit, Paul Wohlhart, and Kurt Konolige. On pre-trained image features and synthetic images for deep learning. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[27] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, 2022.

[28] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[29] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4188–4196, 2016.

[30] Ira Kemelmacher-Shlizerman and Ronen Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE transactions on pattern analysis and machine intelligence*, 33(2):394–405, 2010.

[31] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[32] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2144–2151. IEEE, 2011.

[33] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017.

[34] Tetiana Martyniuk, Orest Kupyn, Yana Kurlyak, Igor Krashenyi, Jiři Matas, and Viktoriia Sharmanska. Dad-3dheads: A large-scale dense, accurate and diverse dataset for 3d head alignment from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 20942–20952, 2022.

[35] Nikolaus Mayer, Eddy Ilg, Philipp Fischer, Caner Hazirbas, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. What makes good synthetic training data for learning disparity and optical flow estimation? *International Journal of Computer Vision*, 126(9):942–960, 2018.

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

[37] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[38] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–626, 2008.

[39] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 947–954. IEEE, 2005.

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

[41] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):121–135, 2017.

[42] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1259–1268, 2017.

[43] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022.

[44] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[45] Nataniel Ruiz, Samuel Schulter, and Manmohan Chandraker. Learning to simulate. In *International Conference on Learning Representations*, 2019.

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

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

[48] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 896–903, 2013.

[49] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 50–58, 2015.

[50] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017.

[51] Linsen Song, Wayne Wu, Chaoyou Fu, Chen Change Loy, and Ran He. Audio-driven dubbing for user generated contents via style-aware semi-parametric synthesis. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[52] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 17:585–598, 2022.

[53] Yang Song, Jingwen Zhu, Dawei Li, Andy Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 919–925. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[54] Luuk Spreeuwers, Maikel Schils, and Raymond Veldhuis. Towards robust evaluation of face morphing detection. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1027–1031. IEEE, 2018.

[55] Keqiang Sun, Wayne Wu, Tinghao Liu, Shuo Yang, Quan Wang, Qiang Zhou, Zuochang Ye, and Chen Qian. Fab: A robust facial landmark detection framework for motion-blurred videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5462–5471, 2019.

[56] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[57] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020.

[58] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[59] Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233, 2021.

[60] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[61] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021.

[62] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 131–138, 2016.

[63] Yue Wu, Zuoguan Wang, and Qiang Ji. Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3452–3459, 2013.

[64] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. In *arXiv*, 2022.

[65] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 601–610, 2020.

[66] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2019.

[67] Ran Yi, Zipeng Ye, Ruoyu Fan, Yezhi Shu, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Animating portrait line drawings from a single face photo and a speech signal. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–8, 2022.

[68] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *In Proceeding of International Conference on Computer Vision*, Venice, Italy, October 2017.

[69] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019.

[70] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2315–2324, 2019.

[71] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[72] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–6. IEEE, 2013.

[73] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021.

[74] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer, 2014.

[75] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2015.

[76] Aihua Zheng, Feixia Zhu, Hao Zhu, Mandi Luo, and Ran He. Talking face generation via learning semantic and temporal synchronous landmarks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3682–3689. IEEE, 2021.

[77] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 386–391, 2013.

[78] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021.

[79] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020.

[80] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pages 592–608. Springer, 2020.

[81] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[82] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.

[83] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.

[84] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):78–92, 2017.

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