DAD-3DHeads: A Large-scale Dense, Accurate and Diverse Dataset for 3D Head Alignment from a Single Image

Tetiana Martyniuk[1,2 *] Orest Kupyn[1,2 *] Yana Kurlyak[1,2] Igor Krashenyi[1,2]
Jiři Matas[3] Viktoriia Sharmanska[4, 5]
[1] Ukrainian Catholic University [2] Piñata Farms, Los Angeles, USA
[3] Visual Recognition Group, Center for Machine Perception, FEE, CTU in Prague
[4] University of Sussex [5] Imperial College London

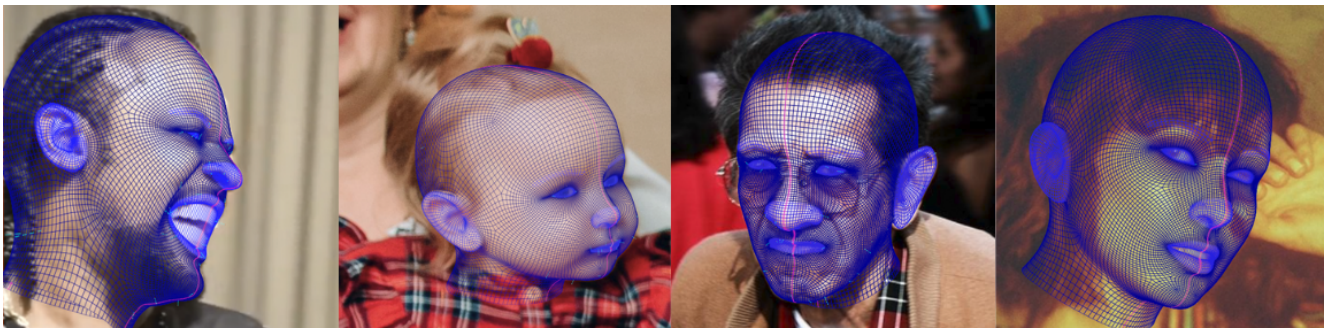

Figure 1. **DAD-3DHeads**, a **D**ense, **A**ccurate, and **D**iverse 3D Head dataset, labeled with over 3.5K verified accurate landmarks. A model trained on DAD-3DHeads achieves superior performance on diverse 3D head tasks. It is robust to domain shifts common in the wild, including head pose changes, occlusions, facial expressions, age groups, illumination conditions, and image quality. Best viewed in color.

Abstract

We present DAD-3DHeads, a dense and diverse large-scale dataset, and a robust model for 3D Dense Head Alignment in-the-wild. It contains annotations of over 3.5K landmarks that accurately represent 3D head shape compared to the ground-truth scans. The data-driven model, DAD-3DNet, trained on our dataset, learns shape, expression, and pose parameters, and performs 3D reconstruction of a FLAME mesh. The model also incorporates a landmark prediction branch to take advantage of rich supervision and co-training of multiple related tasks. Experimentally, DAD-3DNet outperforms or is comparable to the state-of-the-art models in (i) 3D Head Pose Estimation on AFLW2000-3D and BIWI, (ii) 3D Face Shape Reconstruction on NoW and Feng, and (iii) 3D Dense Head Alignment and 3D Landmarks Estimation on DAD-3DHeads dataset. Finally, diversity of DAD-3DHeads in camera angles, facial expressions, and occlusions enables a benchmark to study in-the-wild generalization and robustness to distribution shifts. The dataset webpage is https://p.farm/research/dad-3dheads.

[*]These authors contributed equally to this work.

1. Introduction

Tremendous progress in 3D face analysis has been made since the first 3D morphable model (3DMM) [4] from an image had been proposed [22]. The use cases for precise 3D face models are abundant: accurate face recognition and face detection [16], realistic 3D avatars and animation for VR and games [37], face re-enactment and synthesis for dubbing [59], virtual mirrors and try-on, statistical shape models for medical tasks such as segmentation and analysis of variations in anatomical structures [73].

These applications require not only accurate 3D face geometry but also (1) handling the diversity, e.g., ethnic, age, gender subgroups, and (2) generalizing to in-the-wild deployment conditions, i.e., beyond controlled capture and beyond the data they are trained on. The largest face models up-to-date [8, 45] have focused on the (1) aspect by collecting diverse 3D face and head scans, and building 3DMMs models for different age, gender and ethnicity. In-the-wild generalization has been identified as a pressing challenge of the next generation 3D face models [22]. This (2) aspect of in-the-wild generalization is the focus of our study.

The progress that we have witnessed with deep learn-

ing has impacted closely related facial analysis tasks such as Landmark Localisation [14, 20, 46, 51, 57], Facial Alignment in 2D and 3D [2, 10–12, 17, 31, 32, 48, 63, 70], and Face Detection [2, 17, 20, 25, 46, 70]. This has been driven by the community effort towards collecting and annotating large image datasets captured in unconstrained conditions, building enhanced models that can take advantage of such large datasets, and most importantly *openness*, i.e., making the models and datasets publicly available for research use.

Nevertheless, 3D face or head alignment from a single image in the wild remains an open challenge. The difficulty comes from (1) lack of 2D-3D ground-truth data and, as a result, (2) ambiguity of the task and reliance on 3D shape priors. Many methods have been developed to fill the gap of missing 2D-3D annotations (1), primarily using 2D landmarks datasets for fitting, or exploring extra knowledge such as identity invariance [53], or co-training with related face detection [20], [16] tasks to drive the recovery of 3D face geometry. Up until now, evaluation of the efficiency of these approaches has been problematic due to the lack of ground-truth data. Regarding (2), the state-of-the-art 3D face reconstruction methodologies such as non-linear 3DMMs and deep learning models [5, 6, 8, 38, 45] are based on learning a statistical 3D facial model and fitting it to the image as a shape (or shape and texture) prior. This direction has a long history tracing back to the seminal work of Blanz and Vetter [4]. It relies on a large and diverse dataset of 3D/4D scans to build the statistical 3D face model that can be decomposed into facial shape (identity and expression), and the camera parameters. This comes at the cost of laborious data collection with expensive 3D acquisition devices, and the fact that 3D acquisition devices cannot operate in arbitrary conditions. Hence, the current 3D facial databases have limited data sample size and have been captured not-quite-in-the-wild [53].

In this work, we show that *without expensive devices, like scanners, that are difficult to deploy in the wild, we can collect accurate annotations of 3D landmarks directly from images, which is labor-efficient and effective to push the state-of-the-art results for 3D head recovery from images*.

Our contributions are as follows:

- A new **D**ense, **A**ccurate and **D**iverse dataset for 3D Dense Head Alignment in-the-wild, **DAD-3DHeads**. It has over 3.5K verified accurate landmarks, the densest annotations for 3D dense head alignment in-the-wild currently available. DAD-3DHeads contains a variety of extreme poses, facial expressions, challenging illuminations, and severe occlusions cases. Accuracy and consistency of the annotations are compared to the ground truth 4D scans and head pose labels.
- A novel way to address the problems of shape reconstruction and pose estimation *simultaneously* during training via optimizing two loss components: (i) **Shape+Expression Loss** and (ii) **Reprojection Loss**. (i) is based on the normalized 3D vertices that enables disentangling the shape and expression information from the pose; (ii) is based on the full head dense 2D landmarks and assesses the pose accuracy. That makes the rich annotations fully utilized, which could not have been done previously due to the lack of GT annotations. Extensive ablation studies show the importance of both loss components.
- **DAD-3DNet** model that maps an input image to 3D mesh representation consistent with the FLAME topology. The model is trained end-to-end by regressing the 3DMM parameters and recovering the 3D head geometry with differential FLAME decoder. The proposed approach learns the head shape, pose, and expression simultaneously. DAD-3DNet outperforms state-of-the-art on a range of tasks, suggesting that dense supervision as provided in our dataset, enables a holistic framework for 3D Head Analysis from images.
- A novel benchmark with the evaluation protocol for quantitative assessment of 3D dense head fitting, i.e. 3D Head Estimation from dense annotations. Our evaluation protocol introduces two novel metrics: **Reprojection NME** computing the NME of the reprojected 3D vertices onto the image plane, and $\boldsymbol{Z_n}$ **Accuracy** evaluating the ordinal distance of the Z-coordinate and accuracy of the 3D fitting.

2. Related Work

This section provides an overview of the available 3D face datasets, followed by a survey of the methods targeting 3D head-related tasks.

3D Face Datasets. Existing 3D face datasets differ based on registration of a 3D face model. Model fitting datasets [5, 7, 30] fit the 3DMM to the images, which makes it suitable for large-scale datasets. The main limitation of such approach is shape detalization. To get a precise 3D facial shape, multi-view camera systems are applied [19, 72] or depth camera [18, 52, 54, 67, 68, 71], however, these sensors suffer from limited spatial resolution. The FaceScape dataset [65] contains textured 3D faces recorded using a dense camera array under controlled lighting, which retrieves the 3D facial model preserving low-level details such as small wrinkles and pores. The 3DFAW-Video dataset [34] lacks subjects diversity, and is not really "in-the-wild"; 300W-LP [50, 76] is synthetic and focuses only on faces. In contrast to our dataset, none of the datasets is diverse, accurate, dense, and in-the-wild at the same time.

3D Head Pose Estimation. *Classical methods* for head pose estimation are based on traditional techniques such as cascade detectors [60] or template matching [9]. Cascade detectors localize the head for each pose [35], while

a template matching approach compares query image with a set of pre-labeled templates and finds a corresponding pose [41, 55]. *Geometric methods* use facial landmarks retrieved from the input image and estimate the head pose empirically [13, 29]. *Regression and classification methods* include wide-ranging methods that fit a mathematical model to predict the head pose from labeled training data or discretized set of poses [3, 40, 49, 56, 64, 79]. *Multi-task* approach combines a head pose estimation learning with other facial analysis tasks, such as Face Detection [46, 47, 78], Face Recognition [47], Landmark Localization [46, 47, 78], Alignment [17, 47, 70]. Our approach is related to the latter one, where the study of 3D Head Reconstruction is coupled with learning parameters of a 3D head model and Landmark Localization.

3D Face Alignment. Early 3D Morphable Face Models (3DMM) [4, 44] were derived from a small amount of registered 3D scans, e.g., Basel Face Model (BFM) [44] has 200 human faces. More recent models such as FLAME [38] are learnt from a significantly larger amount of scans, i.e., FLAME uses 3,800 3D scans of human heads. Nevertheless, the diversity of the scans is limited.

RingNet [53] is trained to estimate the 3D face shape from a single image without direct 3D supervision to overcome this limitation. In contrast, we train our model to perform 3D head reconstruction from an image directly by the 2D-3D supervision as provided in our dataset. Similarly motivated is 3DDFA [77], a Cascaded CNN model which directly predicts a dense 3DMM from the facial image. This approach has been further extended and optimized in [28] with meta-joint optimization to facilitate parameters regression. Another approach called DECA [24] is trained to regress a parameterized face model. A recent FAN model [12] has been constricted by stacking four Hourglass models [42, 66] in which all bottleneck blocks were replaced with the hierarchical, multi-scale and parallel binary residual blocks [11]. Instead of using Landmark Localization, in [16] the authors propose to align human faces directly from an image using 6 degrees of freedom (6DoF 3D) – rotations and translations along X, Y, Z axis. [31] introduces a model with a lightweight attention mechanism for Face Alignment. In contrast, we collect a large-scale, diverse dataset with annotations directly in 3D and correspond with FLAME topology. This enables efficient training of the DAD-3DNet for a range of 3D head tasks.

3. DAD-3DHeads Dataset

To create a large-scale dataset of in-the-wild images, we repurpose a modern 3D modeling tool and introduce a novel annotation scheme that 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 for any head images. In this section we verify that obtained annotations are accurate compared to the GT 3D scans, and of high quality, i.e., reducing annotator's errors by half.

3.1. Data acquisition

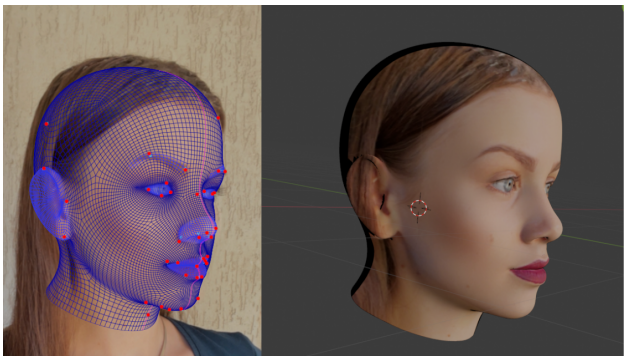

Figure 2. **A labeling tool example**. The annotator fits the 3D Head model to the image by anchoring pinpoints. The corresponding 3D textured render is available to ensure the visual plausibility of the head shape.

We fit a 3D Morphable Model of a human head to a given photo with a simple interface. The annotators do not explicitly control or label either the 3DMM parameters or the blendshapes. The fitting is conditioned upon the visible part of the head and the prior FLAME model [38]. The annotators "pin" the points on the 3D mesh surface (see Fig. 2, left) to the specific pixels of the image. The mesh then undergoes the optimization of the 3DMM parameters, so that the "pin" reprojection error is minimized. During the labeling process, labelers can see the texture rendered onto the 3D mesh with respect to their fitting to verify that the results are visually plausible (Fig. 2, right). We use the 2D reprojection of the 3D mesh onto the image to ensure that the boundaries of the facial features and the skull are correct, and the relative depth information to confirm that the image provides realistic texture mapping onto the human head model. The details of the annotation procedure along with the visuals - images of the intermediate steps and the full video example - are provided in Supplementary. In total we receive 5,023 dense landmarks consistent with the FLAME topology, namely, FLAME mesh vertices.

3.2. Dataset Statistics

DAD-3DHeads dataset 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). For each image, we provide 5,023 vertices of the FLAME mesh, 3,669 of which are accurately labeled (we demonstrate it in Sec. 3.3), neck and eyeballs excluded. We refer to this subset of 3,669 landmarks as "head" (see Fig.5 in Supplementary). We also provide the model-view and frustum projection matrices that map the 3D mesh from model space

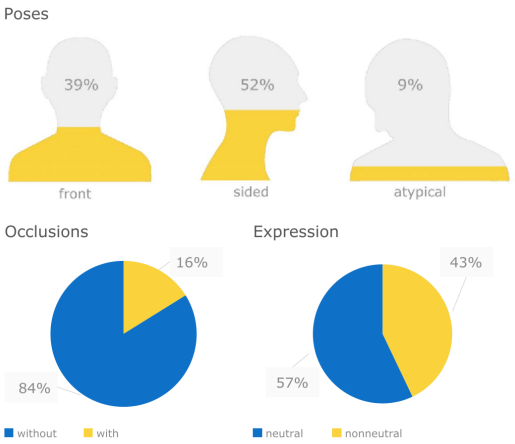

Figure 3. **Dataset Properties:** DAD-3DHeads is well balanced over a wide range of poses, face expressions, and occlusions. The attribute labels are a valuable signal for subgroup analysis and for generalization to in-the-wild deployment conditions.

onto the image for different training scenarios and evaluation purposes. In addition, we release rich attribute information, showing the variability and unbiasedness of the data. DAD-3DHeads attributes include head poses, presence of emotions, occlusions (see Fig. 3), as well as gender, age group, image quality, and illumination labels. The detailed dataset card can be found in the Supplementary.

3.3. Annotation accuracy

To check the accuracy of our annotations, we calculate the accuracy of the head shape reconstruction and head pose estimation compared to ground-truth 3D scans.

3D Head Shape Reconstruction. To validate that DAD-3DHeads annotations fit the head shape correctly, we compare the 3D meshes to the ground-truth scans provided in NoW [53] and Stirling [1] datasets, following the correspondent evaluation protocols (see Sec. 5.3).

As both benchmarks provide scans only of the frontal part of the face, the reconstruction of the whole skull can not be validated by those methods - that is where we resort to visual verification by our labelers as shown in Fig. 2 (right).

We explicitly validate the accuracy on neutral images only since the 3D scans do not capture emotions, see the quantitative results in Tab. 1a, Tab. 1b. For visual comparison, see Fig. 4. Note that the representation is coarse (same as FLAME topology [38]), and we do not aim to model wrinkles and other tiny details that scanners can capture.

3D Head Pose Estimation. To validate the goodness-of-fit of the head pose, we compare the rotation matrices from our annotations to the ground-truth matrices from the BIWI dataset [23]. They are captured by Kinect v2 sensors, the measurement error of which is 20mm [43].

(a) **NoW [53] Dataset,** "multiview_neutral" subset.

Model	Median(mm)	Mean(mm)	Std(mm)
3DDFA-V2 [27, 28]	1.360	1.762	1.621
RingNet [53]	1.316	1.659	1.392
DAD-3DHeads	**1.109**	**1.386**	**1.166**

(b) **Stirling [1] Database,** "Neutral expression, four views" subset.

Model	3DRMSE(mm)	Median(mm)	Mean(mm)	Std(mm)
RingNet [53]	2.793	1.633	2.112	1.828
3DDFA-V2 [27, 28]	2.550	1.508	1.927	1.670
DAD-3DHeads	**2.488**	**1.447**	**1.873**	**1.638**

Table 1. **DAD-3DHeads accuracy of 3D Face Shape Reconstruction** on NoW and Stirling DBs; SOTA methods as reference.

Method	$\|\|I - R_1 R_2^T\|\|_F$	Angle error (degrees)
Img2Pose [2]	0.228	9.336
DAD-3DHeads	**0.149**	**6.037**

Table 2. **DAD-3DHeads accuracy of 3D Head Pose estimation** on BIWI [23]; SOTA method as reference. The measure of $R_1 R_2^T$ deviation from identity matrix lies in the $(0, 2\sqrt{2})$ range [30].

Method	$\mathcal{F}_Q$ (avg NME)	Best sample NME
2D 68 keypoints	3.210	2.326
DAD-3DHeads 68 landm.	**1.737** ($\downarrow$ 45.8%)	**1.302**

Table 3. **Quality score.** Annotation in 3D reduces the global average NME by 45.8%, see e.g. Fig. 5.

To compare the matrices R_1 and R_2, we calculate the difference rotation $R_1 R_2^T$, and measure (i) Frobenius norm of the matrix $I - R_1 R_2^T$, as in [30], and (ii) the angle in axis-angle representation of $R_1 R_2^T$, see Tab. 2.

3.4. Annotation quality

To verify the quality of our annotations, we have selected a subset of $N = 30$ images from different categories in the dataset. Each image was manually labeled with 68 facial landmarks, in the traditional configuration of [26], by $m = 10$ different annotators. The same pictures were labeled following our annotation scheme (Sec. 3.1) in the 3D labeling tool. The 68 reprojected landmarks were computed from the 3D annotations to be comparable with the manual 2D-point labels. We compute the *quality score* $\mathcal{F}_Q$ for each approach (see Table 3), averaging across the images, as a normalized mean error between each pair of labels:

$$\mathcal{F}_Q = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{d_n}\cdot\frac{2}{m(m-1)}\sum_{i=1}^{m}\sum_{j>i}\left|\left|\overrightarrow{x_n^i} - \overrightarrow{x_n^j}\right|\right|_2, \quad (1)$$

where d_n is the head bounding box size, as used in [32, 69], $\overrightarrow{x}$ is an array of 68 labeled landmarks. As our data is mainly non-frontal, we do not use eye landmark distances as a normalization factor.

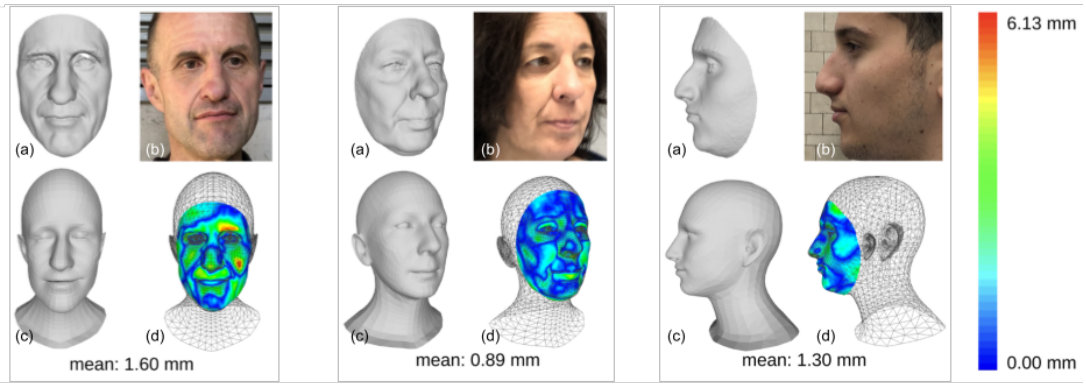

Figure 4. **DAD-3DHeads accuracy** on selected samples from the NoW dataset. **(a)** GT scan; **(b)** input image; **(c)** the result of our annotation; **(d)** alignment of the mesh (wireframe) and the GT scan (with color-coded errors overlayed). The scale of the errors relates to the real-world size of the scans. Note that the resulting meshes accurately capture the coarse shape of the frontal part of the head, the regions of higher error heavily overlap with finer facial structures. We provide more examples, in high resolution, in the Supplementary visualising this phenomenon. Best viewed zoomed in and in color.

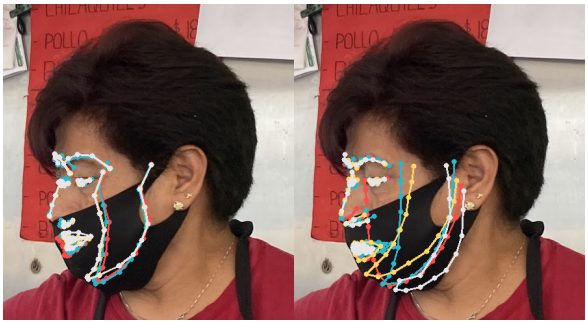

Figure 5. **Annotation Consistency.** Images labeled with our 3D annotation scheme (**left**) and with 68 2D points (**right**). Different colors correspond to different labelers. The annotators are consistent due to the conditioning by a 3D head model prior, which ensures high quality of the DAD-3DHeads dataset even under extremely diverse conditions. Labeling of invisible landmarks is highly inconsistent with the traditional approach, while using the 3D mesh fitting ensures high consistency even on occluded parts.

Limitations. Such labelling scheme provides only partial control over depth. To mitigate this issue, we (i) provide the annotators with the ability to see the rendered texture onto the mesh in 3D, so they can inspect visually whether the lack of depth information corrupted the skull shape, and if the image provides realistic texture; (ii) propose Z_n metric (see Sec. 5.1) that assesses the depth quality.

4. Method

Our goal is to estimate a compact 3D Head representation from a single image. Given an image, we assume the head is detected, loosely cropped, and approximately centered. We introduce a novel architecture, DAD-3DNet, that predicts a vector of 3DMM parameters disentangled into shape, expression and pose, and a dense set of 2D landmarks. The landmarks serve as additional supervision and regularization and extend the range of applications that could benefit from the DAD-3DNet model. The DAD-3DNet architecture is illustrated in Fig. 6.

4.1. DAD-3DNet Architecture

Our architecture consists of (i) a CNN Encoder to extract features from the image, (ii) a Landmark Heatmap Estimator based on the BiFPN [58] to predict coarse locations of 2D landmarks, (iii) a Fusion Module that fuses the heatmap predictions with the encoder features, and (iv) a Regression Module that predicts finer facial landmarks locations and 3DMM parameters. We also use (v) a differential FLAME Layer that maps the 3DMM vector to the 3D mesh vertices.

A pre-trained CNN Encoder extracts features from the first four stages of a backbone network. The Landmark Heatmap Estimator takes second to fourth stage feature maps as an input and predicts coarse Gaussian heatmaps using BiFPN, allowing easy and fast multi-scale feature fusion. The Gaussian heatmaps with $1/4$ of the original spatial resolution are then interpolated to the size of the fourth stage feature maps. The Fusion Layer incorporates the interpolated Gaussian heatmaps, the original feature map, and the BiFPN feature maps to encode a multi-scale representation with an Inception Module. A linear layer follows encoder representation to extract 2D landmarks locations.

4.2. Objective Function

We introduce a multi-component loss function for the end-to-end training of DAD-3DNet to provide supervision for different branches of the network. The loss function consists of four different parts: **Shape+Expression Loss** measuring goodness-of-fit of the 3D Head Shape (L_{3D}),

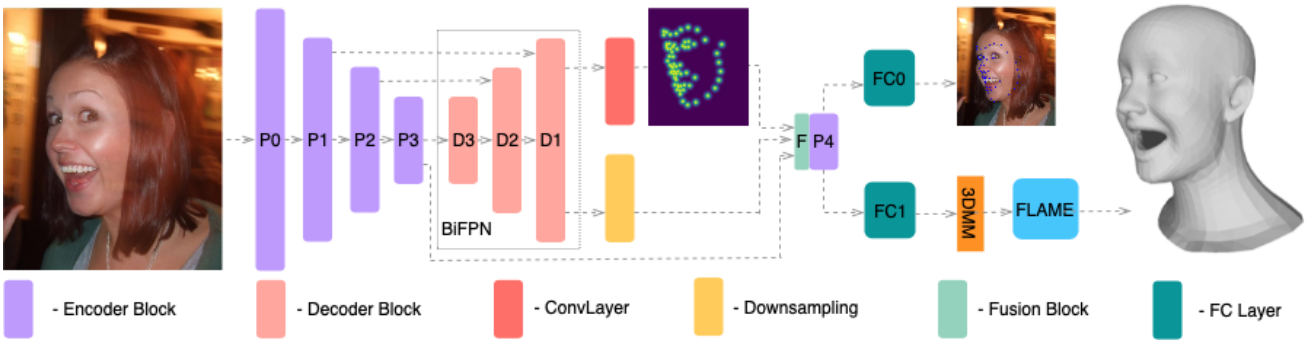

Figure 6. **DAD-3DNet architecture design** and model training benefit from the rich annotations in a multi-branch setup. The Gaussian heatmap estimator predicts coarse locations of the head landmarks. The fusion block combines the coarse heatmap, BiFPN feature map, and CNN encoder output to regress a set of 3D head model parameters and finer locations of head landmarks.

Reprojection Loss (L_{proj}) that incorporates pose information, Landmark Regression (L_1) and Gaussian Heatmap Loss (L_{AWing} [61]) to provide the supervision for the 2D Facial Landmarks prediction branch. The detailed ablation studies (Sec. 5.4) show the importance of each component.

Shape+Expression Loss: Following the notations used in [53] , we denote the 3DMM coefficients as follows: shape coefficients $\overrightarrow{\beta} \in \mathbb{R}^{|\overrightarrow{\beta}|}$, expression coefficients $\overrightarrow{\psi} \in \mathbb{R}^{|\overrightarrow{\psi}|}$. The global rotation pose is modeled by $\overrightarrow{\theta_r} \in \mathbb{R}^6$ for continuity of representation [74], and is separated from the jaw rotation pose vector $\overrightarrow{\theta_j} \in \mathbb{R}^3$. In our approach we assume that neck $\overrightarrow{\theta_n} \in \mathbb{R}^3$ and eyeballs $\overrightarrow{\theta_e} \in \mathbb{R}^6$ rotation coefficients are equal to zero. The global rotation predictions are set to zero to evaluate the discrepancy between our predictions and the ground truth in 3D. The 3D vertices are computed from the 3DMM parameters using a differentiable FLAME layer. As FLAME model [38] contains both the head and the neck, but our task is narrowed down to the head mesh estimation, we subsample the vertices vector $\overrightarrow{v} = \overrightarrow{v}(\overrightarrow{\beta}, \overrightarrow{\psi}, \overrightarrow{\theta_j})$ over the set of "head" vertex indices I: $\overrightarrow{v}|_I$.

The ground truth and the predicted mesh can differ in scale and location, so we normalize φ both to fit into the unit cube after subsampling.

The final loss term measures discrepancy between normalized subsampled vertices:

$$L_{3D}\left(\overrightarrow{\beta}, \overrightarrow{\psi}, \overrightarrow{\theta_j}\right) = \left|\varphi\left(\overrightarrow{v_{pred}|_I}\right) - \varphi\left(\overrightarrow{v_{GT}|_I}\right)\right|_2. \quad (2)$$

Reprojection Loss is computed by projecting the 3D vertices of the posed mesh onto the image. The posed mesh is a "zero-pose" mesh described above, to which we apply the similarity transform (rotation $R(\overrightarrow{\theta_r})$, uniform scaling s, and translation $\overrightarrow{t}$). The reprojection then is a simple orthographic projection onto the image plane. Here, as well, only the "head" vertices are included in the loss computation. The L_1 criterion is used as a discrepancy measure between the reprojected subsampled vertices.

The overall loss is a combination of the four terms:

$$L = \lambda_1 L_{3D} + \lambda_2 L_1 + \lambda_3 L_{proj} + \lambda_4 L_{AWing}.$$

We use 50.0, 1.0, 0.05, 1.0 as λ_1, λ_2, λ_3, λ_4 respectively.

4.3. Implementation details

We implemented all of our models using PyTorch. The backbone network is initialized using the pre-trained weights on ImageNet. The differentiable FLAME layer is kept fixed during the training. The number of learnable head shape and expression parameters are set to 300 and 100, respectively. All the models are trained using **1 RTX A6000** GPU, with a batch size of 256. We use an ADAM optimizer with a learning rate = $1 * 10^{-4}$ and a plateau learning rate reducer with a reduce factor = 0.5 every six epochs when the validation loss stops decreasing. The training takes one day to converge. To preserve the scale ratio and shape of the head, images are padded to the square size and then resized to 256x256. We trained all models without any image augmentations.

5. Experimental Evaluation

We propose DAD-3DHeads Benchmark for evaluating (i) the task of 3D Dense Head Alignment from an image, (ii) in-the-wild model generalization (when trained on our data) to a range of 3D Head Learning tasks, and (iii) robustness to extreme poses. To address (i), we provide a comprehensive analysis of DAD-3DNet and several existing methods on our benchmark, and report the findings in Tab. 4. To test generalization (ii), we analyze the performance of DAD-3DNet on the established benchmarks for 3D Face Shape Reconstruction and 3D Head Pose Estimation, detailed in Sec. 5.2, Sec. 5.3. To test robustness (iii), we evaluate DAD-3DNet under train/test distribution shift in camera poses and report our findings in Supplementary.

Table 4. **Comparison with state-of-the-art 3D Dense Head Alignment models on DAD-3DHeads Benchmark:** We compute the metrics on full test dataset as well as on challenging atypical poses (Pose), compound expressions (Expr.) and heavy occlusions (Occl.) subsets. DAD-3DNet shows superior performance on all subsets. Note: Z_n is computed only for methods that use FLAME mesh topology.

Pose	NME↓				Z_5 Accuracy↑				Chamfer Distance↓				Pose Error↓			
	Overall	Pose	Expr.	Occl.	Overall	Pose	Expr.	Occl.	Overall	Pose	Expr.	Occl.	Overall	Pose	Expr.	Occl.
3DDFA-V2 [27, 28]	3.580	7.630	3.168	3.195	-	-	-	-	6.17	8.878	6.410	6.400	0.527	0.790	0.455	0.542
RingNet [53]	8.757	26.732	5.010	12.660	0.880	0.743	0.913	0.860	5.166	5.704	5.792	5.993	0.438	1.076	0.294	0.551
DAD-3DNet	**2.302**	**6.049**	**1.748**	**2.036**	**0.954**	**0.916**	**0.958**	**0.943**	**3.178**	**4.094**	**3.375**	**3.774**	**0.138**	**0.343**	**0.112**	**0.203**

5.1. Metrics

Given a ground-truth mesh M on a particular frame, and post-processed model output - predicted 3D vertices V, we calculate how well V fit M. The goodness-of-fit measures the pose fitting, and both face and head shape matching. We propose two new metrics for the evaluation protocol: Reprojection NME and Z_n accuracy, in addition to Chamfer Distance and Pose Error reported previously for the 3D Head Learning tasks.

Reprojection NME: we compute the normalized mean error of the reprojected 3D vertices onto the image plane, taking X and Y coordinates into account. Similar to Eq. (1), we use head bounding box size for normalization. The metric is computed on 68 landmarks [26].

Z_n Accuracy: as our annotation scheme is conditioned only upon model prior and the reprojection onto the image, we do not guarantee the absolute depth values to be as accurate as sensor data. We address this issue by measuring the *relative depth* as an ordinal value of the Z-coordinate. For each of K vertices v_i of the GT mesh, we choose K closest vertices $\{v_i^1, ..., v_i^K\}$, and calculate which of them are closer to (or further from) the camera. Then, we compare if for every predicted vertex w_i this configuration is the same: $Z_n = \frac{1}{K}\frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n}\Big((v_i \succeq_z v_i^j) == (w_i \succeq_z w_i^j)\Big)$. We do so on the "head" subset of the vertices only.

Chamfer Distance: as the Z_n metric is valid only for predictions that follow FLAME mesh topology, we add Chamfer distance to measure the accuracy of fit. To ensure generalization to any number of predicted vertices, we measure a one-sided Chamfer distance from our ground-truth mesh to the predicted one. We align them by seven key points correspondences [53], and compute the distances from the "face" subset of the vertices only (see Fig.5 in Supplementary), following the traditional approach [39, 53].

Pose Error: measuring the accuracy of pose prediction, we want to overcome the issues observed in AFLW2000-3D [36] Dataset. Creators of AFLW2000-3D measure the head pose resorting to Euler angles. Such representation is highly dependent on the order in which the rotations are applied. Whenever the second rotation reaches over $\frac{\pi}{2}$ in any direction, i.e., extreme head poses, other rotation axes become linearly dependent, yielding an infinite number of representations for the same transformation [15]. One can observe inconsistencies caused by this in the AFLW2000-3D [36] benchmark in Fig. 7.

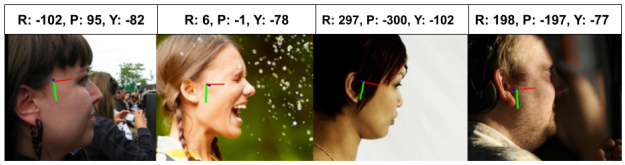

Figure 7. **AFLW2000-3D label inconsistencies.** Some labels of side or extreme atypical poses are inconsistent as the Euler angle representation used is ambiguous due to gimbal lock.

To avoid that, we measure accuracy of pose prediction based on rotation matrices [30] (see Sec. 3.3):

$$Error_{pose} = ||I - R_1 R_2^T||_F$$

5.2. 3D Head Pose Estimation

We evaluate DAD-3DNet on AFLW2000-3D and BIWI datasets for the task of 3D Head Pose Estimation.

BIWI Dataset [23] is gathered in a laboratory setting by recording RGB-D video of different subjects across various head poses using a Kinect v2 device. It contains frames with the rotations ±75° for yaw, ±60° for pitch, and ±50° for roll. A 3D model was fit to each individual's point cloud, and the head rotations were tracked to produce the pose annotations.

AFLW2000-3D Dataset [76] consists of the first 2,000 subjects of the in-the-wild AFLW dataset, which has been re-annotated with image-level 68 3D landmarks and consequently, contain fine-grained pose annotations.

Results: We report the results in Tab. 5a, Tab. 5b. The proposed model outperforms all other 3DMM estimation methods by a large margin, and shows comparable performance to other state-of-the-art methods for head pose estimation.

5.3. 3D Face Shape Reconstruction

For the task of 3D Face Shape reconstruction, we compare the performance of DAD-3DNet with two state-of-the-art publicly available methods: 3DDFA-V2 [27, 28], and RingNet [53] on two 3D Face Shape reconstruction benchmarks: NoW [53] and Feng et al. [39].

NoW Face Challenge: NoW benchmark is designed for the task of 3D face reconstruction from single monocular

Table 5. **3D head pose estimation** results.

Model	MAE ↓	Pitch MAE ↓	Roll MAE ↓	Yaw MAE ↓
3DDFA [27, 77]	19.07	12.25	8.78	36.18
Fan (12 points) [16]	7.88	7.48	7.63	8.53
Dlib (68 points) [33]	12.25	13.80	6.19	16.76
HopeNet [21]	4.90	6.61	3.27	4.81
Img2Pose [2]	**3.79**	**3.55**	3.24	4.57
3DDFA-V2 [27, 28]	8.81	12.08	7.54	6.80
RingNet [53]	7.34	5.37	7.82	8.82
WHENet [75]	3.81	4.39	3.06	3.99
DAD-3DNet	3.98	5.24	**2.92**	**3.79**

(a) **BIWI** [23]

Model	MAE ↓	Pitch MAE ↓	Roll MAE ↓	Yaw MAE ↓
3DDFA [27, 77]	7.39	8.53	7.39	5.40
Fan (12 points) [16]	9.12	12.28	8.71	6.36
Dlib (68 points) [33]	13.29	12.60	9.00	18.27
HopeNet [21]	6.16	6.56	5.44	6.47
RetinaNet [20]	6.22	9.64	3.92	5.10
Img2Pose [2]	3.91	5.03	3.28	3.43
SynergyNet [62]	**3.35**	**4.09**	**2.55**	3.42
3DDFA-V2 [27, 28]	7.56	8.48	9.89	4.30
RingNet [53]	8.27	4.39	13.51	6.92
DAD-3DNet	3.66	4.76	3.15	**3.08**

(b) **AFLW2000-3D** [76]

Table 6. **3D face shape reconstruction** results.

Model	Median(mm) ↓	Mean(mm) ↓	Std(mm) ↓
3DDFA-V2 [27, 28]	1.234	1.566	1.391
RingNet [53]	**1.207**	**1.535**	1.306
DAD-3DNet	1.236	1.541	**1.285**

(a) **NoW [53]**

Model	3DRMSE↓	Median(mm)↓		Mean(mm)↓		Std(mm)↓	
		HQ	LQ	HQ	LQ	HQ	LQ
3DDFA-V2 [27, 28]	2.998	**1.500**	1.779	1.942	2.350	1.704	2.149
RingNet [53]	2.809	1.698	1.634	2.161	2.113	1.832	1.831
DAD-3DNet	**2.749**	1.558	**1.624**	**1.940**	**2.082**	**1.581**	**1.795**

(b) **Feng et al. [39]**

images. The dataset contains 2054 2D images of 100 subjects. Following the evaluation protocol, we predict the meshes that are then rigidly aligned with corresponding ground truth scans based on seven landmark points. The scan-to-mesh distance is computed between them. The calculated mean, median, and standard deviation errors are reported in Table 6a.

Feng et al. Benchmark: [39] provides a subset of Stirling/ESRC 3D face database as the test dataset for their challenge. The test dataset consists of 2,000 2D various expression facial images, including 656 high-quality (HQ) images taken in controlled scenarios and 1,344 low-quality (LQ) images extracted from video frames [1]. Following [39] protocol that rehearses [53] we perform the aforementioned steps and compute the scan-to-mesh distance between the predicted meshes and the ground truth scans. These distances are used to compute the 3DRMSE. We also compute mean, median and standard deviation errors for HQ and LQ images separately for in-depth analysis. The results of evaluation are provided in the Table 6b.

Results: DAD-3DNet shows superior performance to other methods for coarse 3D dense head alignment without using explicit Shape and Expression disentanglement loss.

5.4. Ablation study

In this section, we verify the efficiency of the separate loss components and demonstrate the impact of the training data. We report the results of an ablation study in Table 7.

	Component	NME↓	Z_5 Acc.↑	Pose↓
1	*baseline*	2.576	0.880	0.267
2	+ full face reprojection loss	2.395	0.873	0.263
3	+ full head reprojection loss	2.500	0.943	0.172
4	+ shape+expression loss	2.471	0.951	0.139
5	+ landmark prediction head	**2.302**	**0.954**	**0.138**

Table 7. **DAD-3DNet ablation study on DAD-3DHeads:** The loss terms have significant impact on the fitting accuracy, and the multi-head architecture improves the model generalization.

Reprojection Loss: Supervision based on reprojected landmarks is a core part of the training algorithms. Compared to the models that use supervision based on 68 keypoints, we have only added the reprojection loss based on all available face and head points. Incorporating information about other facial landmarks improves the accuracy of reprojected 68 landmarks but does not impact the other metrics and does not improve the 3D fitting; adding points of the whole head improved all the metrics by a large margin. Additional full head supervision improves the model stability by enforcing to learn the entire head shape.

Shape+Expression Loss: Rich supervision of the normalized 3D vertices locations enables the model to encode more nuanced information about the 3D head pose. As shown in Table 7 this component improves all of the metrics and reduces the 3D head pose error significantly.

Landmarks Head: Multi-task training improves the model stability and enforces the model to prefer more general representations. With the landmarks regression and coarse heatmap estimation modules, the model achieves a significant boost in performance on all metrics yet again.

6. Conclusions

We introduce DAD-3DHeads, a dense, accurate, and diverse 3D Head dataset in the wild. We demonstrate the efficiency and accuracy of the data and novel loss components by training a data-driven DAD-3DNet model. DAD-3DNet achieves superior performance on diverse 3D head tasks and successfully generalizes to in-the-wild conditions.

Acknowledgements. We thank the Armed Forces of Ukraine for providing security to complete this work.

References

[1] Stirling ESRC 3D Face database. http://pics.stir.ac.uk/ESRC/. 4, 8

[2] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7617–7627, 2021. 2, 4, 8

[3] Ben Benfold and Ian Reid. Colour invariant head pose classification in low resolution video. In *British Machine Vision Conference (BMVC)*, pages 1–10. Citeseer, 2008. 3

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 1, 2, 3

[5] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models "in-the-wild". In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 48–57, 2017. 2

[6] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2):233–254, 2018. 2

[7] James Booth, Anastasios Roussos, Evangelos Ververas, Epameinondas Antonakos, Stylianos Ploumpis, Yannis Panagakis, and Stefanos Zafeiriou. 3d reconstruction of "in-the-wild" faces in images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2638–2652, 2018. 2

[8] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5543–5552, 2016. 1, 2

[9] Roberto Brunelli. *Template matching techniques in computer vision: theory and practice*. John Wiley & Sons, 2009. 2

[10] 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 (ECCV)*, pages 616–624. Springer, 2016. 2

[11] Adrian Bulat and Georgios Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 3

[12] 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 *International Conference on Computer Vision (ICCV)*, pages 1021–1030, 2017. 2, 3

[13] Patrick Burger, Martin Rothbucher, and Klaus Diepold. Self-initializing head pose estimation with a 2d monocular usb camera. Technical report, Lehrstuhl für Datenverarbeitung, 2014. 3

[14] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *International Conference on Computer Vision (ICCV)*, pages 1513–1520, 2013. 2

[15] Zhiwen Cao, Zongcheng Chu, Dongfang Liu, and Yingjie Chen. A vector-based representation to enhance head pose estimation. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1188–1197, 2021. 7

[16] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. Faceposenet: Making a case for landmark-free face alignment. In *International Conference on Computer Vision (ICCV) Workshops*, Oct 2017. 1, 2, 3, 8

[17] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *European Conference on Computer Vision (ECCV)*, pages 109–122. Springer, 2014. 2, 3

[18] Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 4dfab: A large scale 4d facial expression database for biometric applications. *arXiv preprint arXiv:1712.01443*, 2017. 2

[19] Darren Cosker, Eva Krumhuber, and Adrian Hilton. A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *2011 International Conference on Computer Vision*, pages 2296–2303, 2011. 2

[20] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 8

[21] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J. Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 8

[22] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020. 1

[23] Gabriele Fanelli, Thibaut Weise, Juergen Gall, and Luc Van Gool. Real time head pose estimation from consumer depth cameras. In *Joint pattern recognition symposium*, pages 101–110. Springer, 2011. 4, 7, 8

[24] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 3

[25] Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv preprint arXiv:1506.08347*, 2015. 2

[26] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010. 4, 7

[27] Jianzhu Guo, Xiangyu Zhu, and Zhen Lei. 3ddfa. https://github.com/cleardusk/3DDFA, 2018. 4, 7, 8

[28] 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 (ECCV)*, 2020. 3, 4, 7, 8

[29] Rainer Herpers, Markus Michaelis, K-H Lichtenauer, and Gerald Sommer. Edge and keypoint detection in facial re-

gions. In *International Conference on Automatic Face and Gesture Recognition*, pages 212–217. IEEE, 1996. 3

[30] Du Q Huynh. Metrics for 3d rotations: Comparison and analysis. *Journal of Mathematical Imaging and Vision*, 35(2):155–164, 2009. 2, 4, 7

[31] Lei Jiang, Xiao-Jun Wu, and Josef Kittler. Dual attention mobdensenet(damdnet) for robust 3d face alignment. In *International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 2, 3

[32] Amin Jourabloo and Xiaoming Liu. Pose-invariant 3d face alignment. In *International Conference on Computer Vision (ICCV)*, pages 3694–3702, 2015. 2, 4

[33] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 8

[34] Rohith Krishnan Pillai, Laszlo Attila Jeni, Huiyuan Yang, Zheng Zhang, Lijun Yin, and Jeffrey F. Cohn. The 2nd 3d face alignment in the wild challenge (3dfaw-video): Dense reconstruction from video. In *International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 2

[35] Jeffrey Ng Sing Kwong and Shaogang Gong. Composite support vector machines for detection of faces across views and pose estimation. *Image Vis. Comput.*, 20:359–368, 2002. 2

[36] Martin Köstinger, 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 *IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011. 7

[37] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2755–2764, June 2021. 1

[38] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6), 2017. 2, 3, 4, 6

[39] Feng Liu, Luan Tran, and Xiaoming Liu. 3d face modeling from diverse raw scan data. In *International Conference on Computer Vision (ICCV)*, Seoul, South Korea, October 2019. 7, 8

[40] Sankha Mukherjee and Neil Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia*, 17(11):2094–2107, 2015. 3

[41] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31(4):607–626, 2009. 3

[42] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499. Springer, 2016. 3

[43] Diana Pagliari and Livio Pinto. Calibration of kinect for xbox one and comparison between the two generations of microsoft sensors. *Sensors*, 15(11):27569–27589, 2015. 4

[44] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009. 3

[45] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William A. P. Smith, and Stefanos Zafeiriou. Combining 3d morphable models: A large scale face-and-head model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2

[46] 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 (TPAMI)*, 41(1):121–135, 2017. 2, 3

[47] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 17–24. IEEE, 2017. 3

[48] Zeyu Ruan, Changqing Zou, Longhai Wu, Gangshan Wu, and Limin Wang. Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction. *IEEE Transactions on Image Processing*, 2021. 2

[49] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2074–2083, 2018. 3

[50] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016. 2

[51] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *International Conference on Computer Vision (ICCV) Workshops*, pages 397–403. IEEE, 2013. 2

[52] Wojciech Sankowski, Piotr Stefan Nowak, and Paweł Krotewicz. Multimodal biometric database dmcsv1 of 3d face and hand scans. In *International Conference Mixed Design of Integrated Circuits & Systems (MIXDES)*, pages 93–97. IEEE, 2015. 2

[53] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7763–7772, 2019. 2, 3, 4, 6, 7, 8

[54] Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. Bosphorus database for 3d face analysis. In *European Workshop on Biometrics and Identity Management*, pages 47–56. Springer, 2008. 2

[55] Jamie Sherrah, Shaogang Gong, and Eng-Jon Ong. Understanding pose discrimination in similarity space. In *British Machine Vision Conference (BMVC)*, pages 1–10. Citeseer, 1999. 3

[56] Sujith Srinivasan and Kim L Boyer. Head pose estimation using view based eigenspaces. In *Object recognition supported*

by user interaction for service robots, volume 4, pages 302–305. IEEE, 2002. 3

[57] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 2

[58] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10781–10790, 2020. 5

[59] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision (ECCV)*, pages 716–731. Springer, 2020. 1

[60] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2001. 2

[61] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *International Conference on Computer Vision (ICCV)*, pages 6971–6981, 2019. 6

[62] Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. *arXiv preprint arXiv:2110.09772*, 2021. 8

[63] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2129–2138, 2018. 2

[64] Yan Yan, Elisa Ricci, Ramanathan Subramanian, Oswald Lanz, and Nicu Sebe. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *International Conference on Computer Vision (ICCV)*, pages 1177–1184, 2013. 3

[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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 601–610, 2020. 2

[66] Jing Yang, Qingshan Liu, and Kaihua Zhang. Stacked hourglass network for robust facial landmark localisation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2025–2033, 2017. 3

[67] Baocai Yin, Yanfeng Sun, Chengzhang Wang, and Yun Ge. Bjut-3d large scale 3d face database and information processing. *Journal of Computer Research and Development*, 46(6):1009, 2009. 2

[68] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 211–216. IEEE, 2006. 2

[69] Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, and Dimitris N Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *International Conference on Computer Vision (ICCV)*, pages 1944–1951, 2013. 4

[70] 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. 2, 3

[71] 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 *IEEE International Conference and workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013. 2

[72] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 2

[73] Guoyan Zheng, Shuo Li, and Gabor Szekely. *Statistical shape and deformation analysis: methods, implementation and applications*. Academic Press, 2017. 1

[74] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 6

[75] Yijun Zhou and James Gregson. WHENet: Real-time fine-grained estimation for wide range head pose. *arXiv preprint arXiv:2005.10353*, 2020. 8

[76] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, 2016. 2, 7, 8

[77] 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. 3, 8

[78] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886. IEEE, 2012. 3

[79] Youding Zhu and Kikuo Fujimura. Head pose estimation for driver monitoring. In *IEEE Intelligent Vehicles Symposium, 2004*, pages 501–506. IEEE, 2004. 3