FaceLift: Semi-supervised 3D Facial Landmark Localization

David Ferman Pablo Garrido Gaurav Bharaj

Flawless AI

Abstract

3D facial landmark localization has proven to be of particular use for applications, such as face tracking, 3D face modeling, and image-based 3D face reconstruction. In the supervised learning case, such methods usually rely on 3D landmark datasets derived from 3DMM-based registration that often lack spatial definition alignment, as compared with that chosen by hand-labeled human consensus, e.g.*, how are eyebrow landmarks defined? This creates a gap between landmark datasets generated via high-quality 2D human labels and 3DMMs, and it ultimately limits their effectiveness. To address this issue, we introduce a novel semi-supervised learning approach that learns 3D landmarks by directly lifting (visible) hand-labeled 2D landmarks and ensures better definition alignment, without the need for 3D landmark datasets. To lift 2D landmarks to 3D, we leverage 3D-aware GANs for better multi-view consistency learning and* in-the-wild *multi-frame videos for robust cross-generalization. Empirical experiments demonstrate that our method not only achieves better definition alignment between 2D-3D landmarks but also outperforms other supervised learning 3D landmark localization methods on both 3DMM labeled and photogrammetric ground truth evaluation datasets. Project Page: https://davidcferman.github.io/FaceLift*

1. Introduction

3D facial landmark localization plays a critical role in various applications, such as talking head generation [39], 3D face reconstruction [15, 29, 61], and learning 3D face models [56]. However, existing 3D facial landmark datasets based on 3D Morphable Model (3DMM) often lack alignment with 2D landmark definitions labeled by humans. This leads to a noticeable ambiguity between 2D and 3D datasets and limits their overall effectiveness, as shown in Fig. 1. We propose an algorithm to bridge this ambiguity by directly lifting hand-labeled 2D landmarks into 3D, without additional 3D landmark localization datasets.

Human-labeled 2D datasets are known to exhibit high-quality facial landmarks for visible facial regions, while self-occluded regions are labeled in a "landmark-marched" style [59], i.e., on the nearest visible boundary. On the other hand, current 3D datasets leave much to be desired in terms of accuracy and consistency w.r.t 2D landmark definitions. For example, human-labeled 2D facial landmark datasets focus on the apparent brow boundaries, whereas 3DMM-based models define the brow region structurally above the eyes, as fixed mesh vertices. However, the relationship between facial structure and brow appearance varies across identities, and hence, a 2D-3D inconsistency occurs, see Fig. 1.

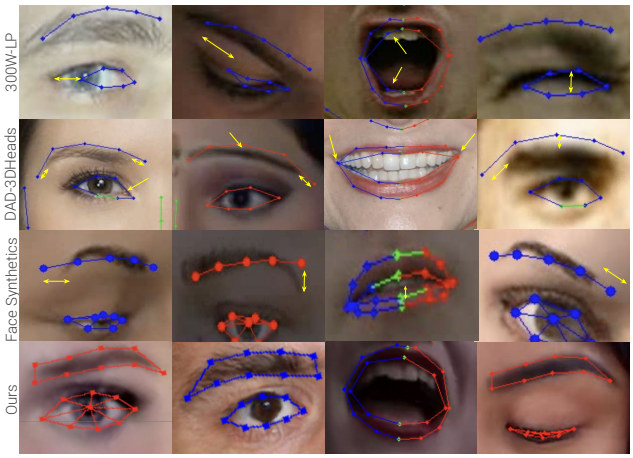

Figure 1. Comparison of our labels with 300W-LP [60], DAD3D-Heads [29], which are both labeled via 3DMM, and Microsoft's Face Synthetics [40] datasets.

Inconsistencies are particularly evident in fine-scale details not captured by the linear 3DMM fitting, often seen in the mouth and eyes where fine-scale details are crucial for accurate representation, as noted by [29], see Fig. 1. Additionally, unlike 2D methods, the SoTA models trained on such datasets tend to fail to capture blinks, as shown in Fig. 6. Finally, "hallucinated" self-occluded landmarks are prone to labeling errors due to the difficulty in labeling the non-visible regions [49]. We observe that the "visible" subset of 2D landmarks is fairly 3D consistent, see Fig. 3,

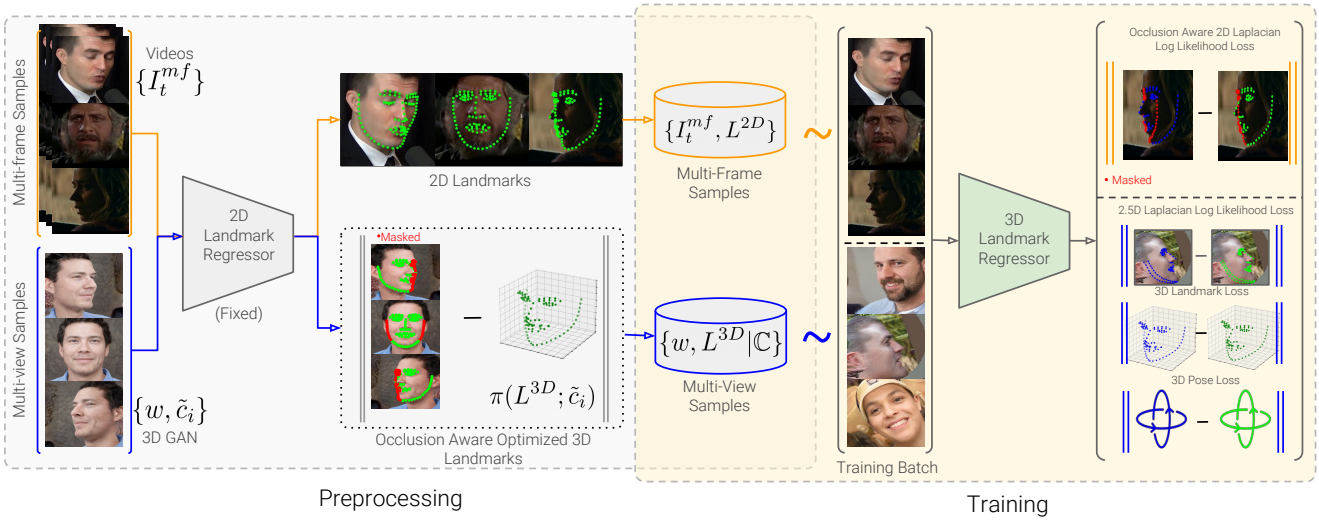

Figure 2. **System Pipeline:** We preprocess multi-frame videos, $\{I_t^{mf}\}_{t=1}^{T}$, and multi-view 3D-aware GAN samples, $\{I_i^{mf} = GAN(w; \tilde{c}_i)\}_{i=1}^{|\tilde{C}|}$, predicting 2D landmarks for each image. For each GAN latent w, we optimize a set of 3D landmarks to minimize a masked, occlusion-aware, reprojection error across views $\tilde{c}_i \in \tilde{C}$, to obtain 3D pseudo-labels. Next, we train a 3D landmark regressor on batches of 2D pseudo-labeled multi-frame samples and 3D pseudo-labeled multi-view 3D-aware GAN samples, supervising via a combination of 2D confidence-aware losses and 3D landmark and pose losses, masking the 2D pseudo-labels in an occlusion-aware manner.

i.e., they mimic what we refer to as "2.5D" (projected from 3D) landmarks. Inspired by these observations, we investigate whether it is feasible to lift visible 2D landmarks into 3D.

Thanks to recent advancements, volumetric 3D-aware GANs have enabled the generation of synthetic, yet photo-realistic, multi-view images with controllable ground-truth camera information. Despite remarkable progress, we observe that available methods are still imperfect w.r.t. multi-view appearance consistency [8, 36], while improved modeling is on-going [2]. In view of the present limitations, we hypothesize that we can exploit existing 3D-aware GANs as a 3D prior and 2D landmarks as 2D image constraints to reveal the 3D awareness of human faces while preserving the 2D-3D consistency.

In order to obtain 2D-3D consistent 3D landmarks, we propose a semi-supervised approach for 3D landmark detection, which leverages 1) a 3D-aware GAN prior for multi-view and multi-frame information from *in-the-wild* videos and 2) 2D landmark pseudo-labels[1] from a SoTA 2D detector. Our method trains jointly on multi-frame samples, pseudo-labeled by the 2D landmark detector, and on multi-view samples, with 3D pseudo-labels obtained via lifting 2D detections from multiple views. Training purely on multi-view 3D-aware GAN samples would introduce a bias and lack of sufficient variation in lighting, image quality, and facial expressions due to the limited diversity of the dataset, FFHQ [18] and its extrapolation, while *in-the-wild* videos contain such diversity.

As previously noted, 3D-aware GAN sampled data, in its current state, is still imperfect, as we observe certain fine-scale details can vary with pose, especially for features like eyelids and pupils, while large poses often lead to severe appearance degradation and background boundary artifacts. While multi-frame samples from videos contain rich diversity, we cannot rely on these samples exclusively as they lack the 3D constraints offered by the multi-view 3D-aware GAN samples, and only a subset of 2D landmarks are 2D-3D consistent, as previously noted. Additionally, while *in-the-wild* videos are biased toward frontal camera-facing head poses [58], sampling from a 3D-aware GAN offers full controllability over the 3D pose distribution, offering more balanced training. Thus, by combining the merits of multi-view and multi-frame samples we are, to the best of our knowledge, the first to achieve this 2D-3D consistency and thus enable 3D landmark localization consistent with 2D human-defined labels.

We evaluate our method on the *in-the-wild* DAD-3DHeads dataset and on high-quality ground-truth temporally consistent 3D mesh tracking dataset, Multiface [45]. On both datasets [29, 45], we achieve state-of-the-art accuracy when comparing to existing SoTA methods, despite being trained without a ground-truth 3D dataset. To summarize, our main contributions are as follows:

1. We introduce a semi-supervised approach that leverages

[1]Non-ground-truth labels

high-quality 2D landmarks along with a 3D-aware GAN prior to tackle the 2D to 3D lifting problem. The resulting pipeline is geometric prior free, enabling learning accurate 3D landmarks that align with 2D hand-labeled definitions, without any ground-truth 3D labels.
2. A novel 3D transformer formulation that leverages volumetric consistency (multi-view constraint) while training on real videos (multi-frame constraint) for *in-the-wild* generalization.
3. State-of-the-art accuracy on photogrammetric ground-truth Multiface [45] and human-labeled DAD-3DHeads [29] datasets, achieving cross-generalization.

2. Related Work

We review methods for 2D-to-3D pose and keypoint estimation and 3D facial landmark localization. In addition, we discuss existing facial landmark datasets.

2D-to-3D Uplifting for Pose and Landmark Estimation. Direct estimation of 3D pose and landmarks from images is an ill-posed problem [22], and ground truth 3D image annotations are often limited [30]. As such, methods in this category often require 3D priors [15, 22, 41] or depth supervision [30]. On the other hand, lifting methods leverage intermediate representations, such as 2D pose or landmark detectors [5, 28, 55], or temporal information e.g. via graph convolutional networks [25, 27, 54] to infer 3D information. Due to the excellent performance of 2D detectors [7, 14, 37], 2D-to-3D uplifting methods normally outperform direct 3D regression methods. Interestingly, while 2D-to-3D pose uplifting has been investigated more extensively, almost no work for face landmark estimation has been done [5, 50], mainly due to the wide availability of 3D face priors [12] and recent photo-realistic synthetic datasets [40, 60]. We note that the definitional gap between 2D and "2.5D", see bottom Fig. 3, presents an additional challenge for uplifting facial landmarks, as 2D labels cannot be modeled simply as projections from 3D, as in the case of human pose estimation. Despite their impressive performance, 2D-to-3D uplifting remains an inherently ill-posed problem, even when spatio-temporal modeling is adopted, since multiple solutions are available, especially when occlusions occur [24]. Recently, transformer-based methods have been introduced, which exploit attention to better reason over temporally neighboring 2D poses for temporal-aware lifting [24, 55]. While these methods attend to relevant temporal information for lifting to 3D, the problem of 3D facial localization lacks temporal ground-truth datasets and remains unexplored. To the best of our knowledge, no work for 2D-to-3D face landmark uplifting has been explored with 3D transformer architectures without ground truth 3D datasets.

3D Facial Landmark Detection on Images. Methods for 3D landmark estimation can be categorized as template-based, 3DMM-based, 3D aware, and 2D-to-3D uplifting. Template-based approaches exploit the template's underlying mesh topology for predicting spatial deformation maps in UV texture space [13, 32, 34] or dense 3D face deformations [4, 38]. 3DMM-based methods utilize a 3D face model, e.g., BFM [31] or FLAME [23], directly to estimate model parameters [41], often with surrogate 2D landmark supervision [15, 29], or as an intermediate representation for 3D landmark refinement [42, 60, 61]. While template- and model-based methods have demonstrated robustness for 3D landmark localization, the representation power is limited by the underlying 3D dense prior [33, 38]. 3D aware techniques leverage volumetric representations to embed 3D landmarks [51] or generate explicit multi-view image constraints for 3D consistent landmark prediction [49]. We note that both of these methods require 3D GAN inversion of monocular 2D images, either at inference or training, which is known to fail for large poses and occlusions [47]. Rather than inverting images, we lift 2D landmarks into 3D by exploiting the multi-view information of 3D-aware GAN samples, avoiding errors introduced by inversion. Unlike previous approaches, 2D-to-3D uplifting methods require no geometry prior and directly regress 3D landmarks [5] or 3D shape consistent landmarks with moving boundaries via heatmap-based regression [46, 59]. Alternatively, joint coordinate and adversarial voxel regression have been proposed [50]. As far as we are aware, the use of transformer-based 3D architectures without geometric priors for 3D sparse localization remains unexplored.

Facial Landmark Datasets. A variety of 2D, 2.5D, and 3D face datasets have been proposed to advance research in facial landmark localization. 2D datasets contain ground truth 2D landmarks annotated on real images [20, 26, 35, 43]. Here, visible landmarks are aligned to face image features, while object-occluded landmarks are hallucinated and self-occluded landmarks are snapped to image boundaries, thus destroying overall 3D face likeness. Kumar et al. [21] partially solve this problem by labeling landmarks with three visibility categories: visible, externally-occluded, and self-occluded. However, these categories, and especially the latter, are created based on human perception, not metrics, and thus they are error prone. 2.5D datasets are either synthetically generated from rendered 3D meshes that attempt to bridge the photorealism gap [40] or derived from real images by automatically fitting a 3DMM [60]. Pure 3D datasets are derived from coarse 3DMM-based renderings [1] or generated by densely fitting a 3DMM to real images with human supervision [29]. Both 2.5D and 3D synthetic datasets have landmarks registered to specific 3D face mesh locations, which are not always aligned to 2D

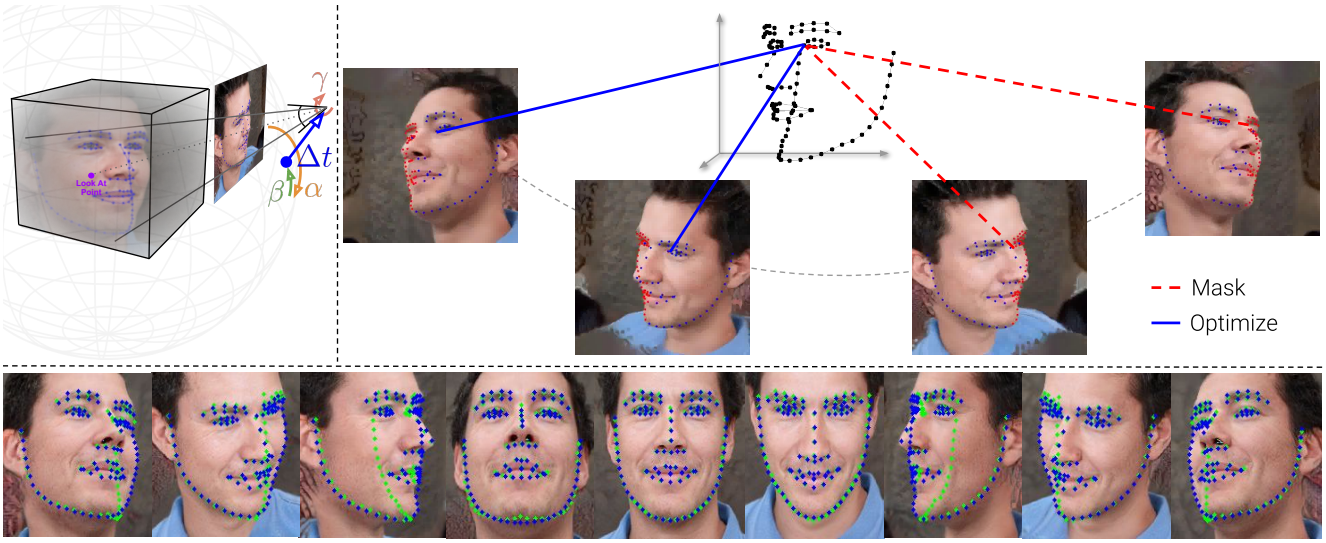

Figure 3. **Masked Multiview 3D Landmark Optimization:** Top Right: We define a fixed set of camera views and hand-design view-dependent landmark masks based on 3D landmark visibility and 2D landmark detector competency. For a given identity sampled from the 3D GAN, we render the set of views, predict landmarks via the 2D detector, and optimize 3D landmarks to minimize the view-dependent-masked reprojection error across all views. Top Left: Rendering of 3D-aware GAN with camera c parameterized by azimuth, elevation, and roll angles α, β, γ along a sphere, and translation Δt. Bottom: Illustration of 2D-3D consistency between optimized 2.5D projections of 3D pseudo-labels, green, and 2D landmark detections, blue.

facial image features, e.g. eyebrows. As 2.5D and 3D real datasets are derived from 3DMM-based fitting, which is an ill-posed problem, perfect annotations cannot always be achieved, as shown in Fig. 1. While there exist smaller-scale multiview face datasets captured in controlled studio setups with photogrammetry data [3, 6, 9, 45, 48, 52, 53], i.e., metrically accurate 3D reconstructions, these datasets are not applicable for in-the-wild facial landmark generalization. Our semi-supervised approach overcomes limitations of 2D, 2.5D, and 3D datasets by leveraging the accuracy of visible 2D landmarks on real images and lifting them via 3D prior supervision with our novel 3D transformer formulation. Thus, our method requires no large-scale annotated 3D datasets, which to date are non-existent and nearly impossible to generate.

3. Method

We introduce a semi-supervised approach for learning 3D facial landmarks from a 3D-aware GAN prior and high-quality 2D landmarks [14], without the use of 3D labels, see Fig. 2. Our method consists of a pre-processing stage and a training stage. We first pre-process our training data by predicting 2D landmarks on multiview 3D-aware GAN samples and *in-the-wild* videos. The multiview landmarks from GAN samples are lifted to 3D via an occlusion-aware masked optimization to obtain 3D landmark pseudo-labels for each GAN latent. In our second phase, we train jointly on multi-view GAN samples, supervised by ground-truth 3D pseudo-labels, and multi-frame *in-the-wild* videos, supervised via pose-dependently masked 2D pseudo-labels.

Pre-Processing We obtain data for training our method from multi-view 3D-aware GAN samples, along with multi-frame *in-the-wild* videos. For each video frame, I_t^{mf}, we predict N 2D landmarks, $L^{2D} \in \mathbb{R}^{N\times 2}$, using a high-quality 2D landmark detector [14], which was trained on the WFLW [43] and LaPa [26] 2D landmark datasets, concurrently. For each GAN sampled latent code, w, we render a set of views and fit 3D landmark pseudo-labels via an occlusion-aware objective on multi-view 2D detections. In the following, we introduce the camera model of the 3D-aware GAN, and describe our landmark pseudo-label optimization and 3D landmark localization model.

Augmented Camera Space We share a perspective camera model [62] between the volumetric rendering of the 3D-aware GAN and projecting 3D landmarks to screen space. Typically, volumetric face GANs use a camera with extrinsics $\mathsf{M} = \left(\begin{smallmatrix}\mathsf{R} & \mathsf{t}\\ \mathbf{0} & 1\end{smallmatrix}\right) \in \mathbb{R}^{4\times 4}$ parameterized by an azimuth angle, α, and elevation angle, β, such that the camera is situated on a sphere pointing at the look-at point. We augment M with camera roll γ and Δt applied to R and t, respectively, see top-left Fig. 3. 3D-aware face GAN's also define camera intrinsics, $\mathsf{K} \in \mathbb{R}^{3\times 3}$, with a fixed focal length, as described in [36]. Let $\mathbb{C}$ be the space of cameras projections s.t. $c = (\mathsf{K}, \mathsf{M}) \in \mathbb{C}$ iff $\alpha \in [-A, A], \beta \in [-B, B], \gamma \in$

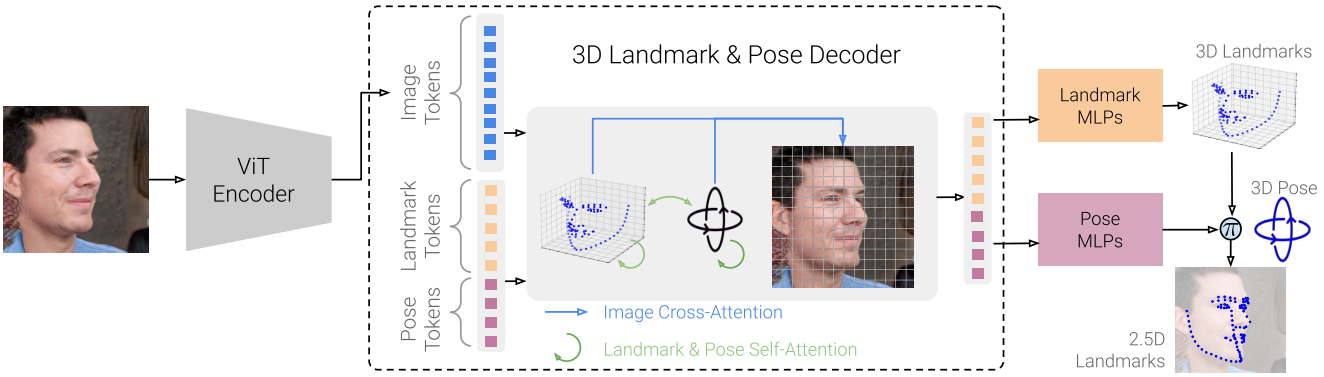

Figure 4. **3D Landmark Regressor Architecture:** Face images are embedded via a ViT encoder to obtain image tokens. Landmark and pose tokens are initialized from a learned embedding and passed through a 3D landmark and pose decoder, in which landmark and pose tokens cross-attend to the image tokens and perform self-attention over the "sequence" of landmark and pose tokens. Each landmark and pose token are routed to an MLP head to predict 3D landmarks and 3D pose, respectively. Finally, the 3D landmarks are projected to 2.5D landmarks via the predicted 3D pose.

$[-\Gamma, \Gamma]$, and Δt such that the bounding box of projected facial landmarks is contained within the image and has minimum dimension greater than half the image dimension. A 3D landmark, $l^{3D} \in \mathbb{R}^3$, is projected from the GAN's canonical space to screen space via the perspective projection function $l^{2D} = \pi(l^{3D}; c)$:

$$\pi(l^{3D}; c) = [l_x^{2D}, l_y^{2D}, w]^\top / w; \quad (1)$$

$$[l_x^{2D}, l_y^{2D}, w] = \mathsf{K} \cdot (\mathsf{R} \cdot l^{3D} + \mathsf{t}). \quad (2)$$

Model Architecture We employ a transformer encoder-decoder model for predicting 3D landmarks, as shown in Fig. 4. We use a ViT encoder [11], known as FaRL [57], pre-trained for human face perception tasks, which we show yields slightly better performance than Resnet152 [16]. We design a transformer decoder with a token per landmark and pose tokens for rotation, Txy and Tz. These tokens pass through three blocks, each containing an image-cross-attention layer, landmark-pose self-attention layer, and MLP, with layer-normalization prior to each. Finally, we pass the landmark and pose tokens individually through MLP heads, which predict the 3D landmarks, Cholesky factorization of the 2D covariances of projected 2.5D landmarks, and the 3D rotation and translation. We apply the 3D landmark predictions as offsets to a template, defined as the landmark-wise mean of our 3D pseudo-labels obtained during pre-processing, to obtain 3D landmark predictions, $\hat{L}^{3D} \in \mathbb{R}^{N \times 3}$. The pose is predicted via a 6D rotation representation, akin to [17] from which a rotation matrix, $\hat{\mathsf{R}}$, is extracted. From $\hat{\mathsf{R}}$, we compute the 3D translation to the camera sphere, $\bar{\mathsf{t}}_{\hat{\mathsf{R}}}$, and predict $\hat{\Delta}$t to obtain $\hat{\mathsf{t}} = \bar{\mathsf{t}}_{\hat{\mathsf{R}}} + \hat{\Delta}\mathsf{t}$. Finally, we form our predicted camera, $\hat{c} = [\mathsf{K}, \hat{\mathsf{M}}]$, with fixed intrinsics, K, and obtain 2.5D landmarks $\hat{L}^{2.5D} = \pi(\hat{L}^{3D}; \hat{c})$.

Training Methodology We train our 3D landmark detector jointly on multi-view GAN sampled images, and multi-frame *in-the-wild* video frames. For multi-view image, $I^{mv} = GAN(w; c)$, we sample a latent code w from a set of pre-processed latents, along with a random camera $c \in C$. For each multi-frame sample, we sample a random video from our pre-processed video dataset followed by a random frame I^{mf} from the video. Each batch of training consists of 4 multi-view samples, (I^{mv}, L^{3D*}, c^*), and 4 multi-frame samples, (I_t^{mf}, L^{2D*}). We formulate our loss function as a combination of multi-frame and multi-view losses. The multi-view loss consists of a 2.5D uncertainty-aware landmark loss, 3D landmark loss, and 3D pose loss. We employ a Laplacian Log Likelihood (LLL) objective parametrized by predicted Cholesky factorization of landmark covariances, akin to [14, 21]. Such parameterization enables the energy landscape to adapt to noise caused by rendering artifacts and allows the model to weigh the loss for each landmark prediction based on its 2D anisotropic confidences. For 3D landmark loss, along with the translation loss, on Δt, we adopt mean-squared-error, while for 3D rotation, we follow head pose estimation work [17] and use geodesic loss. Thus, our multi-view loss is defined as:

$$\mathcal{L}_{\text{mv}} = \mathcal{L}_{\text{MSE\_}L^{3D}} + \mathcal{L}_{\text{MSE\_}\Delta t} + \mathcal{L}_{\text{LLL\_}L^{2.5D}} + \mathcal{L}_{\text{Geo\_R}}, \quad (3)$$

Since the set of views encountered when training on *in-the-wild* videos is not fixed, such as in the 3D pseudo-labeling optimization, we employ a simple heuristic for obtaining masks, $m \in \{0, 1\}^N$. We define a template of normal vectors for each landmark, apply the estimated rotation to each normal, and threshold the dot product with the forward vector to obtain the mask. Thus, we supervise the multi-frame

video samples via their 2D pseudo-labels, L^{2D^*} as:

$$\mathcal{L}_{\text{mf}} = \sum_{i=1}^{N} m_n \cdot \mathcal{L}_{lll}(l_n^{2D^*}, \hat{l}_n^{2.5D}; \Sigma_n), \tag{4}$$

where Σ_n refers to the covariance matrix obtained via predicted Cholesky factorization, and L_{lll} denotes the 2D laplacian-log-likelihood. Refer to [21] for details. For video training, this anisotropic confidence weighted loss enables the energy landscape to adapt to systematic noise from the 2D detector, such as extreme pose samples where the 2D detector may fail, while GAN-based extreme pose samples are constrained via fixed 3D losses. Thus, our complete objective is: $\mathcal{L} = \mathcal{L}_{\text{mf}} + \mathcal{L}_{\text{mv}}$.

4. Results & Analysis

Training Implementation Details We train our method jointly on *in-the-wild* videos obtained from the CelebV-HQ [58] dataset, selecting the first 10K videos, along with GAN samples obtained from IDE-3D [36], sampling latents from the first 10K random seeds. Our pre-processing stage takes roughly 3 days on a single A10G GPU with 24GB RAM to obtain pseudo-labels. Since geometric augmentations (e.g., scale and translation) would break our 3D ground truth under perspective projection, we render GAN-generated images on the fly during training, sampling cameras uniformly from augmented camera space, $\mathbb{C}$. To obtain a sensible pose distribution, we softly decrease extreme rotation angle combinations by accepting sampled rotations with probability $e^{-((\frac{\alpha}{A})^2+(\frac{\beta}{B})^2+(\frac{\gamma}{\Gamma})^2)}$, with $A = 110, B = 60, \Gamma = 90$. Our IDE-3D renders and video frame crops are 224x224, to match the input dimension of our FaRL [57] backbone. We train with a learning rate of 1e-5 for 225 epochs, with the Adam [19] optimizer, decaying the learning rate exponentially by a factor of 0.9 every 3 epochs, taking roughly 4 days on a single GPU machine. We overcome an IDE-3D artifact, where a large pose causes the background to occlude the face, by exploiting IDE-3D's semantic field to set the density of background points in the near half of the viewing frustum to $-\infty$, prior to rendering. We ignore GAN-rendered pupil landmarks during training, as we observe a bias where pupils tend to follow the camera, breaking multi-view consistency.

Normalized Mean Local Consistency Metric Previous works [5, 15, 42, 49, 60] train and evaluate NME on datasets where selected indices of face mesh vertices define the landmarks. Across datasets, we observe that landmark definitions are globally aligned, i.e., same general semantic position, but suffer local definition bias, see Fig. 5, due to differences in vertex selections and mesh topology, e.g., 300WLP [60] uses BFM [31] and DAD-3DHeads [29] uses FLAME [23], while ours does not use a mesh. We report cross-dataset

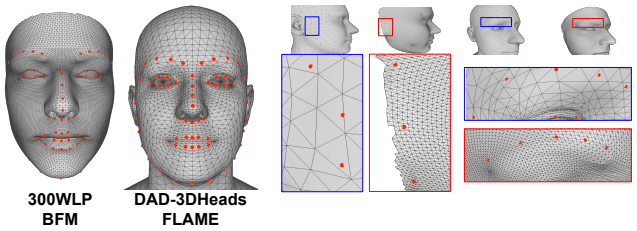

Figure 5. Global alignment of landmarks with local definition bias.

evaluations in the supplementary document, showing that traditional NME leads to unfair comparisons due to the local definition bias while still capturing a ballpark notion of global alignment. As such, we need a landmark definition agnostic metric for meaningful local comparison, which we introduce as an extension of standard NME. We first define NME as parametrized by the vertex indices. For a test set of M images, with predicted landmarks $\hat{L}_m^{2.5D} \in \mathbb{R}^{N\times 2}$, projected vertex labels, $V_m^{2.5D} \in \mathbb{R}^{|V|\times 2}$, vertex indices, $K \in \{1, .., |V|\}^N$:

$$\mathcal{M}(\hat{L}, V; K) = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} z_m ||\hat{l}_{m,n}^{2.5D} - v_{m,K_n}^{2.5D}||_2 \tag{5}$$

where $z = (h_{box} \times w_{box})^{-\frac{1}{2}}$, the diagonal of the face bounding box. Given the dataset-specific landmark definition, $\tilde{K}$, NME $= \mathcal{M}(\hat{L}, V; \tilde{K})$. Our normalized mean local consistency metric (NMLC) replaces the dataset-specific landmark definition with a model-specific one:

$$\text{NMLC} = \min_{K} \mathcal{M}(\hat{L}, V; K), \tag{6}$$

enabling fair cross-dataset comparison. Unlike NME, consistent local bias w.r.t. a desired landmark definition, $\tilde{K}$, will not be penalized. Trivially, NMLC $\leq$ NME. Non-triviality of NMLC is ensured by a large test set with pose, identity, and expression variations. Our NMLC comparisons correlate with qualitative results, see Fig. 6, as our method's leading performance appears to be reflected.

Comparisons We evaluate our method on the studio-captured photogrammetric ground-truth Multiface [45] dataset, along with 3DMM-labeled *in-the-wild* images from DAD-3DHeads [29]. As the Multiface dataset is extremely large (65TB), we select 6 sequences, which cover a range of facial expressions, including asymmetric facial deformations. See supplemental document for curation details. The DAD3D-Heads training set offers category labels for pose, expression, occlusion, quality, lighting, gender, and age. So, we chose the above for our evaluations to obtain fine-grained analysis (quality, lighting, gender, and age reported in supplementary). Since our detector outputs 98 landmarks, we generate the corresponding 68 landmark subset to compare

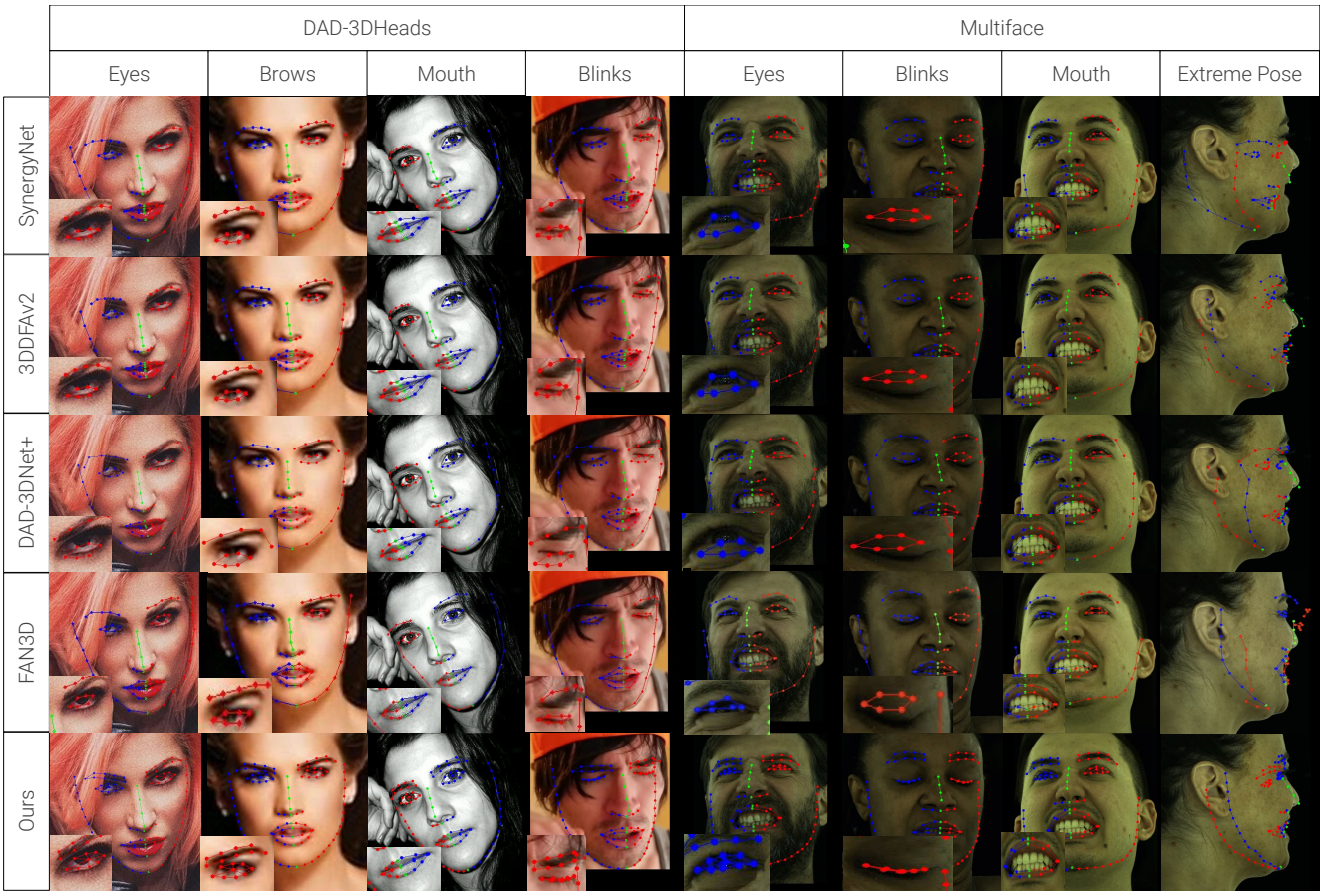

Figure 6. **Visual Results:** Ours, SynergyNet [42], 3DDFAv2 [15], DAD-3DNet+ [49], FAN3D [5] on DAD-3DHeads [29] and Multiface [45] samples.

	Face Regions						Pose			Expression		Occlusions	
Model	full	contours	brows	nose	eyes	mouth	front	sided	atypical	True	False	True	False
SynergyNet [42]	2.81	4.17	2.87	1.96	2.19	2.39	2.59	2.91	3.27	2.52	2.99	3.65	2.68
3DDFA [60]	3.45	4.79	3.51	2.76	2.78	2.99	3.21	3.48	4.33	2.98	3.73	4.62	3.26
3DDFA+ [49]	3.21	4.50	3.26	2.60	2.56	2.75	2.98	3.26	3.95	2.86	3.42	4.06	3.07
3DDFAv2 [15]	2.45	3.76	2.37	1.77	1.92	2.01	2.33	2.49	2.77	2.22	2.59	2.81	2.39
DAD-3DNet★ [29]	2.08	2.91	2.27	1.54	1.68	1.77	1.87	2.19	2.45	1.95	2.17	2.21	2.06
DAD-3DNet+★ [49]	2.06	2.86	2.26	1.54	1.68	1.75	1.87	2.16	2.37	1.93	2.14	2.19	2.04
FAN3D [5]	2.46	4.24	2.34	1.58	1.77	1.82	2.35	2.52	2.64	2.26	2.58	2.77	2.41
Ours	1.91	3.13	2.02	1.28	1.33	1.45	1.76	2.00	2.07	1.77	1.99	2.07	1.88
Ours (Resnet50)	2.23	3.71	2.23	1.59	1.51	1.71	2.05	2.36	2.38	2.05	2.35	2.68	2.16
Ours (Resnet152)	2.06	3.46	2.11	1.36	1.45	1.53	1.89	2.17	2.18	1.88	2.17	2.51	1.99
Ours (MF only)	2.06	3.62	2.10	1.31	1.40	1.45	1.76	2.26	2.26	1.88	2.17	2.17	2.04
Ours (MV only)	2.33	3.46	2.31	1.70	1.83	1.95	2.14	2.43	2.56	2.13	2.44	2.85	2.24
Ours (100)	2.12	3.27	2.21	1.47	1.59	1.71	1.96	2.22	2.26	1.96	2.22	2.50	2.06
Ours (1k)	2.00	3.17	2.09	1.40	1.47	1.57	1.85	2.09	2.13	1.86	2.09	2.24	1.96

Table 1. SoTA evaluation (top) and ablations (bottom) on DAD-3DHeads [29]. We report the NMLC for each model, when averaging across various facial regions and categories. {Model}★ denotes the model was trained on the data samples used for our evaluation, and thus not included in our statements relating accuracy.

with SoTA methods. We observe that, despite training without ground-truth 3D labels, our method outperforms previous SoTA on each dataset by 22% and 19%, as shown in Tab. 1 and Tab. 2, respectively. Fig. 6 shows that our method cap-

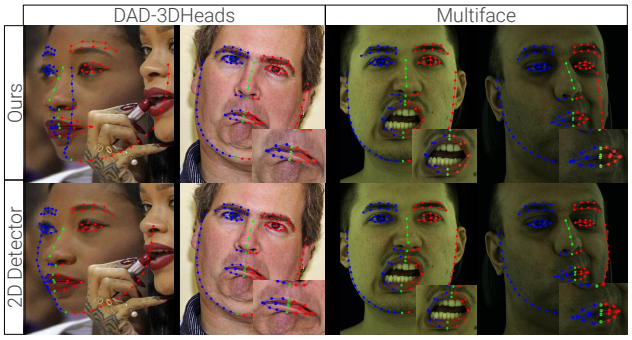

Figure 7. We observe failure cases for our model, including extreme asymmetric expressions and overlapping faces with occlusion, and visualize our model outputs alongside the outputs of the 2D detector used in pseudo-labeling.

Model	full	contours	brows	nose	eyes	mouth
SynergyNet [42]	3.59	4.59	3.16	3.05	2.66	3.74
3DDFA [60]	5.06	7.31	4.29	4.36	3.70	4.66
3DDFA+ [49]	4.92	6.88	4.19	4.39	3.56	4.69
3DDFAv2 [15]	3.25	4.46	2.63	2.57	2.34	3.39
DAD-3DNet [29]	3.30	4.33	3.21	2.80	2.54	3.16
DAD-3DNet+ [49]	3.28	4.29	3.18	2.85	2.53	3.10
FAN3D [5]	3.13	5.04	2.58	2.14	2.18	2.80
Ours	2.52	3.49	2.27	2.11	1.57	2.56
Ours (Resnet50)	2.87	4.13	2.61	1.79	2.13	2.86
Ours (Resnet152)	2.62	3.75	2.38	2.14	1.60	2.60
Ours (MF only)	3.51	6.21	2.94	2.44	2.24	2.75
Ours (MV only)	2.98	3.70	2.81	2.46	2.07	3.22
Ours (100)	2.84	3.82	2.49	1.84	2.37	3.0
Ours (1k)	2.73	3.69	2.49	2.31	1.72	2.84

Table 2. SoTA evaluation (top) and ablations (bottom) on Multiface [45]. We report the NMLC for each model, when averaging across various facial regions.

tures more fine-scale details in the eye, mouth, and brow regions, and it can properly handle blinks while other methods fail. We remark, however, that our method still fails for extreme expressions such as "puckers" and asymmetric deformations. We hypothesize that improving the 2D detector for such cases will propagate through our pipeline toward improving 3D results, as suggested by observations in Fig. 7. Note that the model used for 2D pseudo-labels fails similarly for mouth deformations.

Ablation Studies We conduct ablation studies to observe the effects of the two data sources independently, our choice of encoder backbone, and the impact of sample size. We train our method with multi-view GAN samples and multi-frame video samples independently, referred to as Ours (MV only) and Ours (MF only), respectively, to observe the strengths and weaknesses of each in isolation. We also train our method with a standard Resnet152 [16] backbone, pre-trained on ImageNet [10], of similar parameter count (60M) to our FaRL [57] backbone (87M), and refer to this model as Ours (Resnet152). Additionally, we include Ours (Resnet50) to compare with the same backbone used by [29, 49]. Finally, we decrease the number of videos/GAN latents from 10k to 1k and 100 samples, referred to as Ours (1K) and Ours (100). Tab. 1 and Tab. 2 show that, when training without multi-view samples (MF only), the model failures for the Multiface dataset are quite pronounced for the contours, as the method struggles to capture large pose variation without the 3D constraints of the multi-view training, which is accentuated for contour landmarks. When training without multi-frame samples (MV only), we observe a sizable performance decrease for both mouth and occlusions, as its training distribution is limited by the FFHQ-trained GAN. Replacing our ViT backbone with a Resnet152 of similar parameter count yielded a slight drop in performance. Interestingly, occlusions yield a more significant drop. We hypothesize that it is a result of the ViT's global reasoning capacity, ability to selectively ignore occluded regions, and the backbone encoder's (FaRL) face image embedding prior. We observe a trend that performance improves with the number of training pseudo-labels we generate.

5. Conclusion

In this paper, we have introduced a semi-supervised method for geometric prior-free localization with accurate 3D facial landmarks, aligned with 2D human labels, by exploiting multi-view 3D-aware GANs and using 2D landmarks with no ground-truth 3D dataset. We have shown, for the first time, that SoTA 3D landmarks can be learned without 3D labels, paving the way toward improving 3D facial landmarks beyond the limitations of current 3D labeling techniques.

Limitations & Future Directions Despite the promising results demonstrated by our 3D facial landmark localization method, there are still some limitations. We are heavily dependent on the quality of both a 2D landmark detector and a 3D-aware facial GAN. Improvements to either of these dependencies should result in improvements when training with our approach. Currently, we observe limitations in the 2D landmark detector for facial expressions, such as puckers and asymmetric deformations, constraining the performance of 3D uplifting. However, correcting this may be as simple as labeling such examples when training the 2D detector. As has been noted by previous approaches [44], 3D-aware GANs have limited pose and expression distributions, limiting their downstream application for multi-view consistency. Noting the failure case observed where the method fails due to an occlusion Fig. 7, future work may include investigating GAN sample augmentation via volumetrically generated occlusions. Finally, as we observe a strong trend in increasing performance improvement with the number of pseudo-labels generated, future work may explore the asymptotic limits of such improvement.

References

[1] Synthesis AI. Face api dataset. https://github.com/Synthesis-AI-Dev/face_api_dataset, 2022. [Online; accessed 21-Sep-2022]. 3

[2] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Ümit Y. Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360°. *CoRR*, abs/2303.13071, 2023. 2

[3] Andrew D. Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2d/3d hybrid face dataset. In *ACM HGBU*, pages 79–80. ACM, 2011. 4

[4] Chandrasekhar Bhagavatula, Chenchen Zhu, Khoa Luu, and Marios Savvides. Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In *ICCV*, pages 3980–3989. IEEE CS, 2017. 3

[5] 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 *ICCV*, pages 1021–1030. IEEE CS, 2017. 3, 6, 7, 8

[6] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE TVCG*, 20(3):413–425, 2013. 4

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE TPAMI*, 43(1): 172–186, 2021. 3

[8] 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, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, pages 16102–16112. IEEE, 2022. 2

[9] Darren Cosker, Eva Krumhuber, and Adrian Hilton. A FACS valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *ICCV*, pages 2296–2303. IEEE CS, 2011. 4

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE CS, 2009. 8

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021. 5

[12] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhöfer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models - past, present, and future. *ACM TOG*, 39(5): 157:1–157:38, 2020. 3

[13] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, pages 557–574. Springer, 2018. 3

[14] David Ferman and Gaurav Bharaj. Multi-domain multi-definition landmark localization for small datasets. In *ECCV*, pages 646–663. Springer, 2022. 3, 4, 5

[15] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *ECCV*, pages 152–168. Springer, 2020. 1, 3, 6, 7, 8

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE CS, 2016. 5, 8

[17] Thorsten Hempel, Ahmed A. Abdelrahman, and Ayoub Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *ICIP*, pages 2496–2500, 2022. 5

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE TPAMI*, 43(12):4217–4228, 2021. 2

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[20] 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 *ICCVW*, pages 2144–2151. IEEE CS, 2011. 3

[21] Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *CVPR*, pages 8233–8243. CVF / IEEE, 2020. 3, 5, 6

[22] Sijin Li and Antoni B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, pages 332–347. Springer, 2014. 3

[23] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM TOG*, 36(6):194:1–194:17, 2017. 3, 6

[24] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *CVPR*, pages 13137–13146. IEEE, 2022. 3

[25] Kenkun Liu, Rongqi Ding, Zhiming Zou, Le Wang, and Wei Tang. A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In *ECCV*, pages 318–334. Springer, 2020. 3

[26] Yinglu Liu, Hailin Shi, Yue Si, Hao Shen, Xiaobo Wang, and Tao Mei. A high-efficiency framework for constructing large-scale face parsing benchmark. *CoRR*, abs/1905.04830, 2019. 3, 4

[27] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Hai Ci, and Yizhou Wang. Context modeling in 3d human pose estimation: A unified perspective. In *CVPR*, pages 6238–6247. CVF / IEEE, 2021. 3

[28] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, pages 2659–2668. IEEE CS, 2017. 3

[29] Tetiana Martyniuk, Orest Kupyn, Yana Kurlyak, Igor Krashenyi, Jiri Matas, and Viktoriia Sharmanska. Dad-3dheads: A large-scale dense, accurate and diverse dataset for 3d head alignment from a single image. In *CVPR*, pages 20910–20920. IEEE, 2022. 1, 2, 3, 6, 7, 8

[30] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *CVPR*, pages 7307–7316. CVF / IEEE, 2018. 3

[31] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *IEEE AVSS*, pages 296–301. IEEE CS, 2009. 3, 6

[32] Jingtan Piao, Chen Qian, and Hongsheng Li. Semi-supervised monocular 3d face reconstruction with end-to-end shape-preserved domain transfer. In *ICCV*, pages 9397–9406. IEEE, 2019. 3

[33] Mallikarjun B. R., Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt. Learning complete 3d morphable face models from images and videos. In *CVPR*, pages 3361–3371. CVF / IEEE, 2021. 3

[34] 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 TIP*, 30:5793–5806, 2021. 3

[35] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV*, pages 397–403. IEEE CS, 2013. 3

[36] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. IDE-3D: interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM TOG*, 41(6):270:1–270:10, 2022. 2, 4, 6

[37] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703. CVF / IEEE, 2019. 3

[38] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. FML: Face model learning from videos. In *CVPR*, pages 10812–10822. CVF / IEEE, 2019. 3

[39] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, pages 10039–10049. CVF / IEEE, 2021. 1

[40] Erroll Wood, Tadas Baltrusaitis, 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 *ICCV*, pages 3661–3671. IEEE, 2021. 1, 3

[41] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljevic, Daniel Wilde, Stephan J. Garbin, Toby Sharp, Ivan Stojiljkovic, Tom Cashman, and Julien P. C. Valentin. 3d face reconstruction with dense landmarks. In *ECCV*, pages 160–177. Springer, 2022. 3

[42] Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In *3DV*, pages 453–463. IEEE, 2021. 3, 6, 7, 8

[43] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, pages 2129–2138. CVF / IEEE, 2018. 3, 4

[44] Yiqian Wu, Jing Zhang, Hongbo Fu, and Xiaogang Jin. Lpff: A portrait dataset for face generators across large poses. *ArXiv*, abs/2303.14407, 2023. 8

[45] 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 M. 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. *CoRR*, abs/2207.11243, 2022. 2, 3, 4, 6, 7, 8

[46] Shengtao Xiao, Jiashi Feng, Luoqi Liu, Xuecheng Nie, Wei Wang, Shuicheng Yan, and Ashraf A. Kassim. Recurrent 3d-2d dual learning for large-pose facial landmark detection. In *ICCV*, pages 1642–1651. IEEE CS, 2017. 3

[47] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. *CVPR*, 2023. 3

[48] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. A 3d facial expression database for facial behavior research. In *IEEE FGR*, pages 211–216. IEEE CS, 2006. 4

[49] Libing Zeng, Lele Chen, Wentao Bao, Zhong Li, Yi Xu, Junsong Yuan, and NimaKalantari. 3d-aware facial landmark detection via multiview consistent training on synthetic data. In *CVPR*. IEEE, 2023. 1, 3, 6, 7, 8

[50] Hongwen Zhang, Qi Li, and Zhenan Sun. Adversarial learning semantic volume for 2d/3d face shape regression in the wild. *IEEE Trans. Image Process.*, 28(9):4526–4540, 2019. 3

[51] Hao Zhang, Tianyuan Dai, Yu-Wing Tai, and Chi-Keung Tang. Flnerf: 3d facial landmarks estimation in neural radiance fields. *CoRR*, abs/2211.11202, 2022. 3

[52] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun J. Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image Vis. Comput.*, 32 (10):692–706, 2014. 4

[53] Zheng Zhang, Jeffrey M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur A. Ciftci, Shaun J. Canavan, Michael Reale, Andrew Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *CVPR*, pages 3438–3446. IEEE CS, 2016. 4

[54] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *CVPR*, pages 3425–3435. CVF / IEEE, 2019. 3

[55] Ce Zheng, Sijie Zhu, Matías Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *ICCV*, pages 11636–11645. IEEE, 2021. 3

[56] Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. Imface: A nonlinear 3d morphable face model with implicit neural representations. In *CVPR*, pages 20311–20320. IEEE, 2022. 1

[57] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *CVPR*, pages 18676–18688. IEEE, 2022. 5, 6, 8

[58] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A

large-scale video facial attributes dataset. In *ECCV*, pages 650–667. Springer, 2022. 2, 6

[59] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, pages 787–796. IEEE CS, 2015. 1, 3

[60] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *CVPR*, pages 146–155. IEEE CS, 2016. 1, 3, 6, 7, 8

[61] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z. Li. Face alignment in full pose range: A 3d total solution. *IEEE TPAMI*, 41(1):78–92, 2019. 1, 3

[62] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. *Comput. Graph. Forum*, 37(2):523–550, 2018. 4