Instant Multi-View Head Capture through Learnable Registration

Timo Bolkart[1] Tianye Li[2] Michael J. Black[1]
[1]MPI for Intelligent Systems, Tübingen [2]University of Southern California

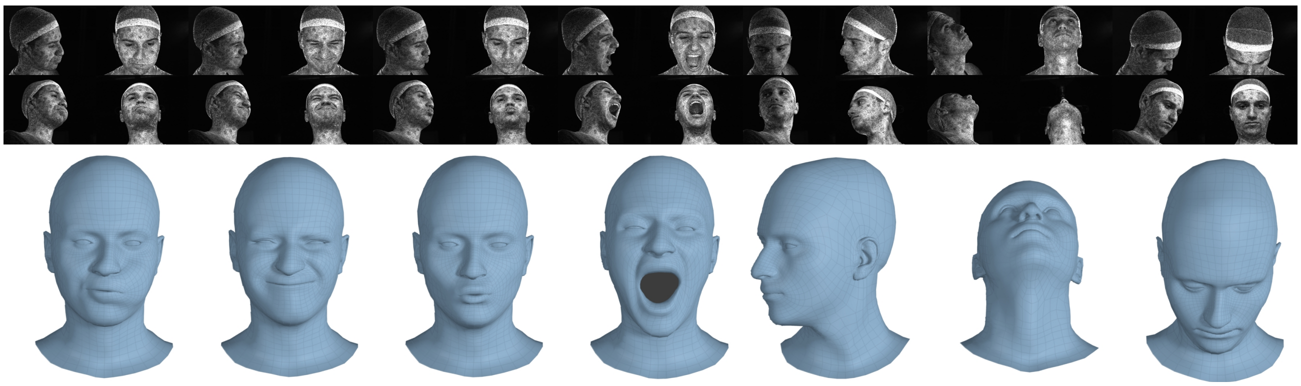

Figure 1. Given calibrated multi-view images (top: 4 of 16 views; contrast enhanced for visualization), TEMPEH directly infers 3D head meshes in dense semantic correspondence (bottom) in about 0.3 seconds. TEMPEH reconstructs heads with varying expressions (left) and head poses (right) for subjects unseen during training. Applied to multi-view video input, the frame-by-frame inferred meshes are temporally coherent, making them directly applicable to full-head performance-capture applications. See Sup. Mat. for the video output.

Abstract

Existing methods for capturing datasets of 3D heads in dense semantic correspondence are slow and commonly address the problem in two separate steps; multi-view stereo (MVS) reconstruction followed by non-rigid registration. To simplify this process, we introduce TEMPEH (Towards Estimation of 3D Meshes from Performances of Expressive Heads) to directly infer 3D heads in dense correspondence from calibrated multi-view images. Registering datasets of 3D scans typically requires manual parameter tuning to find the right balance between accurately fitting the scans' surfaces and being robust to scanning noise and outliers. Instead, we propose to jointly register a 3D head dataset while training TEMPEH. Specifically, during training, we minimize a geometric loss commonly used for surface registration, effectively leveraging TEMPEH as a regularizer. Our multi-view head inference builds on a volumetric feature representation that samples and fuses features from each view using camera calibration information. To account for partial occlusions and a large capture volume that enables head movements, we use view- and surface-aware feature fusion, and a spatial transformer-based head localization module, respectively. We use raw MVS scans as supervision during training, but, once trained, TEMPEH directly predicts 3D heads in dense correspondence without requiring scans. Predicting one head takes about 0.3 *seconds with a median reconstruction error of 0.26 mm, 64% lower than the current state-of-the-art. This enables the efficient capture of large datasets containing multiple people and diverse facial motions. Code, model, and data are publicly available at* https://tempeh.is.tue.mpg.de.

1. Introduction

Capturing large datasets containing 3D heads of multiple people with varying facial expressions and head poses is a key enabler for modeling and synthesizing realistic head avatars. Typically, building such datasets is done in two steps: unstructured 3D scans are captured with a calibrated multi-view stereo (MVS) system, followed by a non-rigid registration step to unify the mesh topology [23]. This two-stage process has major drawbacks. MVS reconstruction requires cameras with strongly overlapping views and the resulting scans frequently contain holes and noise. Registering a template mesh to these scans typically involves manual parameter tuning to balance the trade-off between accurately fitting the scan's surface and being robust to scan artifacts. Both stages are computationally expensive, each taking several minutes per scan. For professional captures, both steps are augmented with manual clean-up to enhance the quality of the output meshes [3, 67]. Such manual editing is infeasible for large-scale captures ($\gg$ 10K scans).

Instead, we advocate for a more practical setting that directly predicts 3D heads in dense correspondence from calibrated multi-view images, effectively bypassing the MVS step. We achieve this with TEMPEH (Towards Estimation of 3D Meshes from Performances of Expressive Heads), which quickly (~ 0.3 seconds per head on a NVIDIA A100-SXM GPU) infers accurate 3D heads (~ 0.26 mm median error) in correspondence, without manual user input.

While several methods exist that directly recover 3D faces in correspondence from calibrated multi-view images, they have high computational cost and require careful selection of optimization parameters per capture subject [9, 14, 29, 62]. These remain major obstacles for large-scale data captures. A few learning-based methods directly regress parameters of a 3D morphable model (3DMM) [80] or iteratively refine 3DMM meshes from multi-view images [6]. As shown by Li et al. [50], this 3DMM dependency constrains the quality and expressiveness of these methods.

The recent ToFu [50] method goes beyond these 3DMM-based approaches with a volumetric feature sampling framework to infer face meshes from calibrated multi-view images. While demonstrating high-quality predictions, ToFu has several limitations. (a) The training is fully-supervised with paired data of multi-view images and high-quality registered meshes; creating such data requires extensive manual input. (b) Only the face region is predicted; ears, neck, and the back of the head are manually completed in an additional fitting step. (c) Self-occlusions in scanner setups designed to capture the entire head result in mediocre predictions due to the naïve feature aggregation strategy that ignores the surface visibility. (d) Only a small capture volume is supported and increasing the size of the capture volume to cover head movements reduces the accuracy.

TEMPEH adapts ToFu's volumetric feature sampling framework but goes beyond it in several ways: (a) The training requires no manually curated data as we jointly optimize TEMPEH's weights and register the raw scans. Obtaining the clean, registered meshes required by ToFu is a key practical hurdle. TEMPEH learns from raw scans and is robust to their noise and missing data. This is done by directly minimizing the point-to-surface distance between scans and predicted meshes. (b) At run time the entire head is inferred from images alone and includes the ears, neck, and back of the head. (c) The feature aggregation accounts for surface visibility. (d) A spatial transformer module [42] localizes the head in the feature volume to only sample regions relevant for prediction, improving the accuracy.

In summary, TEMPEH is the first framework to accurately capture the entire head from multi-view images at near interactive rates. During training, TEMPEH jointly learns to predict 3D heads from multi-view images, and registers unstructured scans. Once trained, it only requires calibrated camera input and it generalizes to diverse extreme expressions and head poses for subjects unseen during training (see Fig. 1). TEMPEH is trained and evaluated on a dynamic 3D head dataset of 95 subjects, each performing 28 facial motions, totalling about 600K 3D head meshes. The registered dataset meshes, raw images, camera calibrations, trained model, and training code are publicly available.

2. Related work

Scan registration: Registering unstructured 3D scans to a common mesh topology has been extensively studied over the last two decades since the work of Blanz and Vetter [11]. For a comprehensive overview, we refer the reader to the survey of Egger et al. [23]. Most prior methods non-rigidly deform a template mesh [11, 13, 15, 43, 48, 49, 57, 65, 79, 81] with a generalization of the rigid iterative closest point (ICP) algorithm [10]. Existing methods mostly register 3D scans of faces in a neutral expression [11, 13, 57], a few static expressions [4, 65], or faces in motion [1]. Bolkart and Wuhrer [12] and Zhang et al. [83] jointly register static datasets of 3D faces while building a 3DMM. Only a few methods consider more than just the face by registering scans of entire heads [16, 17, 49, 81]. All these methods have in common that they are optimization-based, which makes them slow and sensitive to scans with noise or holes. Dealing with this to obtain accurate results requires manual parameter tuning. Few learning-based methods exist to directly go from a scan to a registered mesh [5, 52, 84]. While registering scans faster than previous optimization-based methods, these methods only register tightly cropped faces (i.e., no neck, ears, back of the head, etc.), they require facial landmarks, and scan noise can negatively impact the reconstructed meshes. TEMPEH takes inspiration from these scan registration methods by minimizing similar objective functions (namely, point-to-surface distance and an edge-based surface regularization [49]). However, in contrast to these registration methods, we jointly register the training scans and train TEMPEH. Once trained, TEMPEH requires no scans as input, but, instead, directly infers 3D heads in correspondence from multi-view images.

Image-based reconstruction: Undoubtedly, monocular images or videos have drawn the largest focus as input to 3D face reconstruction methods and we refer to recent surveys for a thorough overview [56, 88]. Most single-image-based 3D face reconstruction methods either fit some 3DMM to an image [2, 8, 11, 31, 59, 74, 78] in an analysis-by-synthesis fashion or regress the parameters of a 3DMM [20, 24, 47, 72, 73]. Reconstructing 3D from monocular images or videos is an ill-posed problem, as many 3D reconstructions give the same image when projected to 2D. To improve reconstruction accuracy, existing methods leverage multi-image constraints (e.g., same identity across images) [18, 25, 27, 66, 75] or multi-view constraints [68] during training. Others use paired image-3D data for training, obtained by fitting a 3DMM to images [26, 37, 41, 55, 77] or by sampling a 3DMM to generate synthetic data [22, 33, 61]. The state-of-the-art in single-image face reconstruction leverages a face recognition network trained on large amounts of image data, combined with supervised training from paired 2D-3D data obtained by registering 3D scans [87]. At test time, these methods reconstruct 3D faces from a single image without prior knowledge about the camera, image resolution, lighting, etc., making the problem ill-posed. The ambiguity between 3D face shape, camera, and distance to the camera limits the metric accuracy of their 3D reconstructions [7].

Instead, several methods focus on a more constrained scenario by reconstructing 3D faces from collections of images instead of just a single image. Such methods optimize the parameters of a parametric model [78] or non-rigidly deform a template mesh [6, 30, 45, 46, 64, 70] to fit multiple images of one subject. Learning-based methods independently regress 3DMM parameters from multiple images of the same subject and fuse the identity shape parameters to obtain coherent identity shape parameters per subject [21, 60, 71, 80]. As camera intrinsics are unknown for these image collections, the ambiguity between (unknown) focal length and object scale means that the reconstructed faces are not metrically accurate. Further, the approximate assumption about the camera (typically weak perspective projection) and the ambiguity of identity and expression-dependent shape result in reconstruction errors for each image. This limits the overall reconstruction accuracy when integrating the erroneous results across images (see Li et al. [50] for comparisons). TEMPEH instead leverages camera calibration information to reconstruct metrically accurate 3D heads, and it is designed for a multi-view setup (i.e., multiple time-synchronized images per expression) to disambiguate identity and expression shape variations.

Few methods directly reconstruct 3D faces from calibrated multi-view images. Among these, optimization-based approaches [9, 14, 29, 62] to date achieve the most impressive results with fine-scale geometric details, but at the cost of being computationally slow, and requiring carefully tuned parameters per subject. As these methods are tailored towards specific custom capture setups, they cannot be directly applied to our off-the-shelf active stereo system. The recent ToFu method [50] directly predicts 3D faces in correspondence, parameter-free at test time, and at near interactive rates. However, ToFu is trained fully-supervised from high-quality registered 3D faces, which are difficult to obtain and limit the applicability of ToFu to new capture setups. Further, its naïve multi-view feature integration neglects surface visibility, limiting generalization to 360° capture setups as, e.g., features from the back of the head contribute to reconstructing face vertices and vice versa. Additionally, it assumes the capture volume tightly encapsulates the face, leading to mediocre results for larger capture volumes, required for covering moving heads. ToFu is therefore only able to capture tightly cropped faces, while the rest of the head is completed in a manual post-processing step. We build on top of ToFu's design but to overcome its limitations, (1) we train our model directly from raw scans, which makes it easier to adapt to new capture setups, (2) our feature integration considers surface properties and visibility, and (3) we localize the head in the feature volume with a spatial transformer. These changes improve the reconstruction quality and enable us to infer entire heads.

Multi-view stereo: MVS capture systems are commonly used to reconstruct 3D faces [9, 34, 35, 54]. While reconstructing high-quality geometry, these methods are computationally expensive due to the pairwise feature matching across views. Recent learning-based methods [36, 39, 44, 69, 82] reduce this computational cost but lose accuracy. For an overview of learning-based MVS methods, see the survey of Wang et al. [76]. All these methods have in common that the reconstructed geometry is unstructured, while our goal is to reconstruct meshes in correspondence. However, we use a commercial MVS method to generate unstructured 3D head scans for our training data, and use these scans as a supervision signal. Once trained, TEMPEH directly predicts head meshes in correspondence from multi-view images without requiring MVS scans.

3. Method

Given sets of images $\{\mathcal{I}_i \in \mathbb{R}^{w \times h \times 3}\}_{i=1}^{k}$ and camera calibrations $\{C_i\}_{i=1}^{k}$ (i.e., camera intrinsics, extrinsics, and radial distortion parameters) of a multi-view capture system with k views, TEMPEH infers a head mesh $M_r := (\mathbf{V}^r, \mathbf{T})$ with vertices $\mathbf{V}^r \in \mathbb{R}^{n_v \times 3}$ and faces $\mathbf{T} \in \mathbb{R}^{n_f \times 3}$. Here, n_v and n_f are the number of vertices and faces, respectively. Note that n_v and $\mathbf{T}$ are fixed, as all meshes output for different sets of input images are in dense correspondence.

As outlined in Fig. 2, TEMPEH consists of two stages, a coarse head inference stage, which outputs an intermediate 3D head $M_c := (\mathbf{V}^c, \mathbf{T})$ with vertices $\mathbf{V}^c \in \mathbb{R}^{n_v \times 3}$, followed by a refinement stage that updates all vertex locations and outputs the final mesh M_r. This two stage process allows us to leverage the surface properties of M_c (i.e., vertex position, visibility and normals) in the second stage for multi-view feature aggregation and vertex refinement.

3.1. Coarse head prediction

Feature extraction: Each image $\mathcal{I}_i$ is processed by a shared convolutional U-Net $F_{\text{img}}(\mathcal{I}_i) \rightarrow \mathcal{F}_i \in \mathbb{R}^{w \times h \times d_f}$ to extract 2D feature maps $\mathcal{F}_i$ with d_f feature channels.

Volumetric feature sampling: Following Li et al. [50], all 2D feature maps are then unprojected and fused into

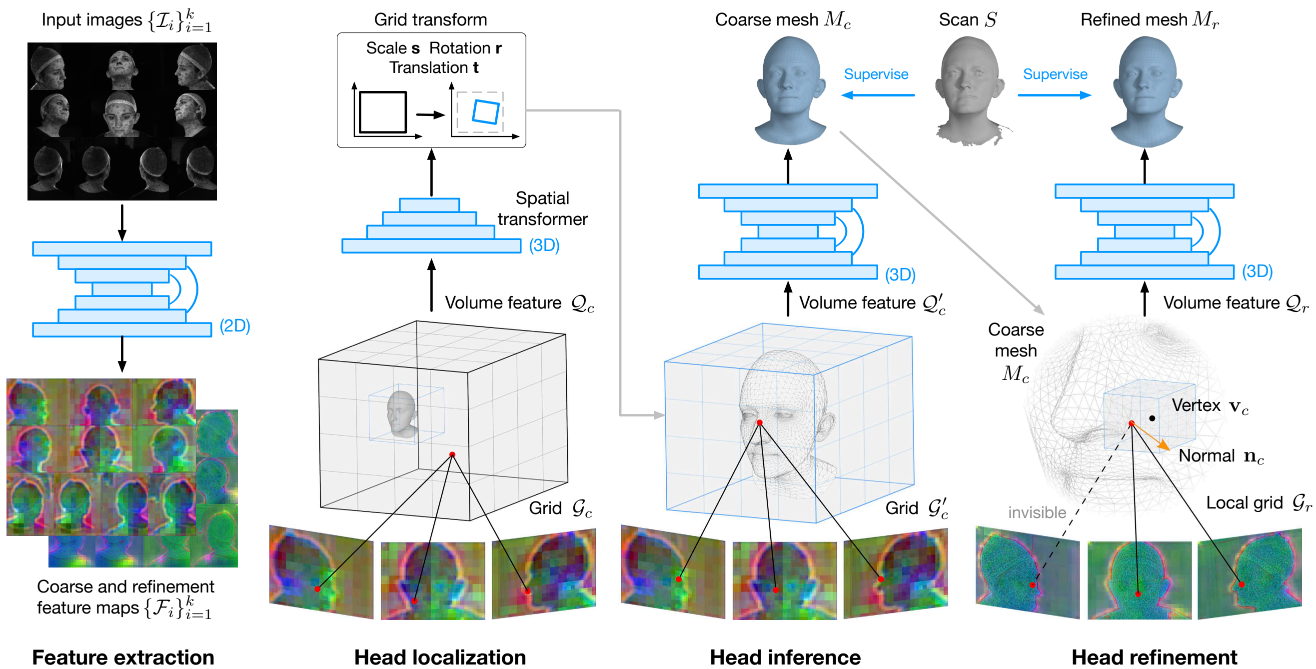

Figure 2. **Overview**. TEMPEH predicts a high-quality registered head mesh in two stages. The coarse stage builds a feature volume from the feature maps, localizes the moving head in this volume with a spatial transformer, and then infers an intermediate 3D head mesh from the localized feature volume. The refinement stage updates each vertex location by sampling features locally, fusing these features view- and surface-aware, and predicting the updated vertex location from the local feature volume. During training, raw MVS scans are used as supervision. Once trained, TEMPEH directly predicts head meshes from calibrated multi-view images without requiring scans as input.

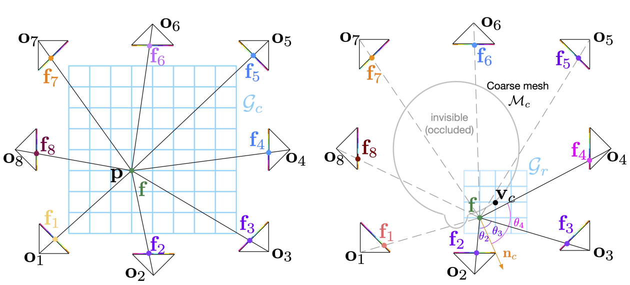

Figure 3. **Volumetric feature sampling illustrated in 2D**. While the coarse stage (left) directly fuses the feature vectors of all views, the refinement stage (right) performs a surface-aware feature fusion by leveraging the mesh estimated in the coarse stage. Specifically, image features for occluded views (dashed lines), and views with large angle between the surface normal and viewing direction are assigned a low importance weight, while non-occluded views with low angle receive a higher weight.

a feature cube (see Fig. 3 left). Specifically, for a point $\mathbf{p} \in \mathbb{R}^3$ and view i, a feature vector $\mathbf{f}_i \in \mathbb{R}^{d_f}$ is obtained by perspective projection $\Pi(.)$ and bilinear sampling as $\mathcal{F}_i(\Pi(\mathbf{p}, C_i)) \rightarrow \mathbf{f}_i$. The view feature vectors are then fused by computing the mean $\boldsymbol{\mu}_c = \frac{1}{k}\sum_{i=1}^{k}\mathbf{f}_i$ and variance $\boldsymbol{\sigma}_c^2 = \frac{1}{k}\sum_{i=1}^{k}\mathbf{f}_i^2 - \boldsymbol{\mu}_c^2$ (with a slight abuse of notation, where $\mathbf{f}_i^2$ and $\boldsymbol{\mu}_c^2$ are the element-wise squares) across all views, and concatenated to the feature vector $\mathbf{f} = (\boldsymbol{\mu}_c \oplus \boldsymbol{\sigma}_c^2) \in \mathbb{R}^{2d_f}$, where $\oplus$ denotes the vector concatenation. This procedure is performed for every point in a 3D sampling grid $\mathcal{G}_c \in \mathbb{R}^{d_c \times d_c \times d_c \times 3}$ that covers the entire capture volume with d_c samples per dimension, to obtain a feature cube $\mathcal{Q}_c \in \mathbb{R}^{d_c \times d_c \times d_c \times 2d_f}$.

Head localization: While the head to be captured only occupies a small subspace of the capture volume, the feature cube $\mathcal{Q}_c$ covers the entire space. This wastes learning capability due to the lower resolution of the feature cube in the relevant head region. Training a network to directly recover a 3D head from $\mathcal{Q}_c$, as done in ToFu, therefore reduces the reconstruction accuracy. However, $\mathcal{Q}_c$ contains information about the location and scale of the head. We leverage this with a trainable spatial transformer $F_{\text{loc}}(\mathcal{Q}_c) \rightarrow \{\mathbf{s}, \mathbf{r}, \mathbf{t}\}$ that predicts scale $\mathbf{s} \in \mathbb{R}^3$, rotation $\mathbf{r} \in \mathbb{R}^6$ [86], and translation $\mathbf{t} \in \mathbb{R}^3$ to localize the head in $\mathcal{Q}_c$. Transforming every point of $\mathcal{G}_c$ with $\{\mathbf{s}, \mathbf{r}, \mathbf{t}\}$ then gives a head localized feature grid $\mathcal{G}_c^{'}$. We then perform the volumetric feature sampling for $\mathcal{G}_c^{'}$ to obtain a localized feature cube $\mathcal{Q}_c^{'}$.

Head inference: The n_v vertices $\mathbf{v}_i^c$ of M_c are then predicted from $\mathcal{Q}_c^{'}$ following Li et al. [50]. For this, a convolutional 3D U-Net $F_{\text{rec}}(\mathcal{Q}_c^{'}) \rightarrow \mathcal{P}$ with a Softmax function applied to the output of the final layer's output across the first three dimensions predicts a probability volume $\mathcal{P} \in \mathbb{R}^{d_c \times d_c \times d_c \times n_v}$. In $\mathcal{P}$, each of the n_v channels encodes the probability distribution $\mathcal{P}_i$ of the 3D location for one vertex $\mathbf{v}_i^c$. Each $\mathbf{v}_i^c$ is then computed as the element-wise product of $\mathcal{P}_i$ and the grid $\mathcal{G}_c^{'}$, followed by a sum over the grid dimensions. Formally, $\mathbf{v}_i^c = \sum_{j,k,l=1}^{d_c} (\mathcal{P}_i \odot \mathcal{G}_c^{'})_{jkl}$, where $\odot$ denotes the Hadamard product, broadcasted across the fourth dimension of $\mathcal{G}_c^{'}$ (i.e., the xyz-channels), and $(.)_{jkl}$ is the jkl-th tensor element.

3.2. Geometry refinement

Volumetric surface-aware feature fusion: Following the coarse feature sampling stage, for a point $\mathbf{p} \in \mathbb{R}^3$ and view i, a feature vector $\mathbf{f}_i \in \mathbb{R}^{d_f}$ is obtained from the i-th view 2D feature map by perspective projection and bilinear sampling. Different from the coarse stage, which uses one big sampling grid, the refinement stage defines, for every vertex $\mathbf{v}^c$, a small 3D sampling grid $\mathcal{G}_r \in \mathbb{R}^{d_r \times d_r \times d_r \times 3}$, centered at $\mathbf{v}^c$. For simplicity, we omit the vertex index and describe the feature sampling for one particular vertex.

ToFu [50] naïvely fuses all $\mathbf{f}_i$ with equal importance across views, regardless of the visibility of point $\mathbf{p}$ in the i-th view. Instead, we use a surface-aware feature aggregation (see Fig. 3 right) in form of a weighted mean

$$\boldsymbol{\mu}_r = \frac{1}{\sum_{i=1}^{k} \eta_i} \sum_{i=1}^{k} \eta_i \mathbf{f}_i, \tag{1}$$

and weighted variance

$$\boldsymbol{\sigma}_r^2 = \frac{1}{\sum_{i=1}^{k} \eta_i} \sum_{i=1}^{k} \eta_i \mathbf{f}_i^2 - \boldsymbol{\mu}_r^2, \tag{2}$$

where η_i is a view- and surface-dependent weight that depends on the coarse mesh vertex $\mathbf{v}^c$, its corresponding vertex normal $\mathbf{n}^c$, and the camera location $\mathbf{o}_i$ (in C_i) of view i. The weight is computed as

$$\eta_i = \text{Softplus}(\delta_i \cdot \cos\theta_i), \tag{3}$$

where $\delta_i \in \{0, 1\}$ is the visibility of $\mathbf{v}^c$ from the camera centered at $\mathbf{o}_i$ (i.e., 0 invisible, 1 visible), and $\cos\theta_i = <\mathbf{n}^c, \mathbf{d}_i>$ measures the angle between the surface normal $\mathbf{n}^c$ and the negative viewing direction $\mathbf{d}_i$, where $\mathbf{d}_i = (\mathbf{o}_i - \mathbf{p})/\|\mathbf{o}_i - \mathbf{p}\|$. The Softplus enforces positivity of the weight to get non-zero gradients w.r.t. features of all views.

Finally, we compute the fused feature vector as $\mathbf{f} = (\boldsymbol{\mu}_r \oplus \boldsymbol{\sigma}_r^2 \oplus \mathbf{v}) \in \mathbb{R}^{2d_f+3}$, where $\mathbf{v} \in \mathbb{R}^3$ is the vertex corresponding to $\mathbf{v}^c$ of a fixed template mesh, and $\oplus$ denotes the concatenation operation. Adding $\mathbf{v}$ acts as identifier of the vertex to be processed to the refinement network. Performing the volumetric feature sampling for all points in $\mathcal{G}_r$ gives a local feature cube $\mathcal{Q}_r$ for every vertex $\mathbf{v}^c$.

Mesh refinement: Similar to the coarse face reconstruction, a convolutional 3D U-Net $F_{\text{ref}}(\mathcal{Q}_r) \rightarrow \mathcal{P}$, with a Softmax function applied to the final layer's output, predicts a probability volume $\mathcal{P} \in \mathbb{R}^{d_r \times d_r \times d_r}$. Note that different from the coarse stage, $\mathcal{P}$ encodes the probability distribution of the 3D location of one vertex $\mathbf{v}^c$. The final vertex location $\mathbf{v}^r$ is then reconstructed as $\sum_{j,k,l=1}^{d_r} (\mathcal{P} \odot \mathcal{G}_r)_{jkl}$.

3.3. Loss functions

Surface distance: TEMPEH reconstructs meshes $M \in \{M_c, M_r\}$ (i.e., either coarse mesh M_c or refined mesh M_r) in correspondence, which must closely resemble the raw training scans S. To enforce this, the training of TEMPEH minimizes the point-to-surface distance, given as

$$E_{\text{s2m}} = \lambda_{\text{s2m}} \frac{1}{|S|} \sum_{\mathbf{s} \in S} \rho \left(\min_{\mathbf{m} \in M} \|\mathbf{s} - \mathbf{m}\|_2^2 \right), \tag{4}$$

which measures the distances between points $\mathbf{s} \in \mathbb{R}^3$ on the surface of S and their closest points $\mathbf{m} \in \mathbb{R}^3$ on the surface of M. The Geman-McClure [32] robust penalty function $\rho(.)$ provides robustness to noise and outliers in the scans, the weight $\lambda_{\text{s2m}} \in \mathbb{R}_0^+$ controls the impact of the loss.
Surface regularization: Directly optimizing Eq. 4 results in poor registrations with overlapping and self-intersecting faces. To prevent this, we add a relative edge regularization

$$E_{\text{reg}} = \lambda_{\text{reg}} \frac{1}{n_e} \sum_{i=1}^{n_e} \gamma_i \left\| \mathbf{e}_i^m - \mathbf{e}_i^t \right\|_2^2, \tag{5}$$

which penalizes the difference between the 3D edge vectors $\mathbf{e}^m$ of $M \in \{M_c, M_r\}$ and the corresponding edge vectors $\mathbf{e}^t$ of the reference registration T. To account for varying scan quality in different face regions (e.g., eyes, lips, and hair regions are often more noisy), pre-defined edge weights $\gamma \in \mathbb{R}_0^+$ control the amount of regularization per face region, $\lambda_{\text{reg}} \in \mathbb{R}_0^+$ weights the overall regularization.
Registration error: The distance to the registration T is minimized as vertex-to-vertex distance, specifically as

$$E_{\text{v2v}} = \frac{1}{n_v} \sum_{i=1}^{n_v} \omega_i \left\| \mathbf{v}_i^m - \mathbf{v}_i^t \right\|_2^2, \tag{6}$$

where $\mathbf{v}^m$ and $\mathbf{v}^t$ are vertices of $M \in \{M_c, M_r\}$ and T, respectively. The binary weights $\omega_i \in \{0, 1\}$ control the impact of individual vertices. Note that E_{v2v} is only used to pre-train the coarse reconstruction stage for the full head (i.e., $\omega_i = 1 \; \forall i$). Afterwards, it only serves to regularize the eyeball location, with $\omega_i = 1$ for all eyeball vertices, and $\omega_i = 0$ otherwise.

4. Implementation details

Capture setup: We use a multi-camera active stereo capture system (3dMD LLC, Atlanta) with eight pairs of grayscale stereo cameras, and eight color cameras. The capture system is calibrated, providing parameters for camera extrinsics, intrinsics, and radial distortions. Each camera captures images of resolution 1600×1200 at 60 fps, where the color cameras are time synchronized with light LED panels, and the stereo cameras are synchronized with random speckle pattern projectors. The system uses a MVS method to reconstruct unstructured 3D scans at 60 fps, each with about 110K vertices, from the stereo images.

Data capture: We collect a multi-view 3D head dataset, referred to as FaMoS (Facial Motion across Subjects), from 93 subjects with our capture setup. Each subject performs 28 motion sequences, containing six prototypical expressions (i.e., Anger, Disgust, Fear, Happiness, Sadness, and Surprise), two head rotations (left/right and up/down), and diverse facial motions, including extreme and asymmetric expressions. All subjects wear a hair net. In total, FaMoS contains around 600K frames (i.e., $\sim$ 227 frames per sequence) of calibrated multi-view images and corresponding 3D head scans. See the Sup. Mat. for additional details.

Training data: We train TEMPEH on data of 78 FaMoS subjects (70 subjects for training, 8 for validation). For training, we randomly sample 40 frames per expression sequence, and 120 frames per head rotation sequence (88K frames in total). For validation, we sample 5 frames per sequence (1,118 frames in total). For each frame, we compute a reference registration in FLAME mesh topology with the fully-automatic registration method of Li et al. [49].

Test data: We qualitatively and quantitatively evaluate TEMPEH on all 28 sequences of 15 FaMoS subjects, disjoint from the training subjects. For each sequence, we randomly sample 20 frames; in total $8,350$ frames.

Implementation details: TEMPEH is implemented in PyTorch [58], and optimized with AdamW [53] with a learning rate of $1e-4$ for the parameters of the head localization networks, and $1e-3$ for all other parameters. The volumetric feature sampling is adapted from ToFu [50]. The differentiable distance between scans and reconstructed mesh is based on Kaolin [28], random points are sampled in each scan's surface during training with Trimesh [19].

TEMPEH takes the 16 gray-scale stereo images as input, as they are time synchronized with the 3D scans used as training supervision, and it outputs meshes in FLAME mesh topology [49] with $n_v = 5023$ vertices and $n_f = 9976$ faces. The feature network F_{img} is a U-Net [63] with a ResNet34 [38] backbone. The head localization network F_{loc} consists of two 3D convolutional layers and a fully-connected layer, each with ReLU activations, followed by a final linear layer. The reconstruction networks F_{rec} and F_{ref} are 3D U-Nets [40], with five and three down- and upsampling layers, respectively. See the Sup. Mat. for details on the model architecture and computational requirements.

Parameter settings: First, the coarse reconstruction stage is trained for 300K iterations on the reference registrations with $\omega_i = 1$ for all vertices, and without surface distance or regularization ($\lambda_{\text{s2m}} = \lambda_{\text{reg}} = 0$). Then, we train the coarse reconstruction stage with surface distance weight $\lambda_{\text{s2m}} = 10.0$ and surface regularization weight $\lambda_{\text{reg}} = 1.0$ for 300K iterations. Directly optimizing E_{s2m} and E_{reg} can result in dislocated eyeballs during training, as the eyeballs are parts separated from the head, and E_{reg} is translation invariant. To prevent this, we regularize the eyeball locations with $\omega_i = 1$ for all eyeball vertices, for all other vertices, we set $\omega_i = 0$. We train the refinement stage for 150K iterations with $\lambda_{\text{s2m}} = 10.0$ and $\lambda_{\text{reg}} = 0.3$. Following ToFu, the dimensions of the feature grids are $d_c = 32$ (coarse stage) and $d_r = 8$ (refinement stage), and the number of features is $d_f = 8$. To reduce the memory consumption, we downsample the input images by a factor of 8 for the coarse training (i.e., $w = 200, h = 150$), and by a factor of 4 for the refinement step (i.e., $w = 400, h = 300$). We use a batch size of 2 for the entire training.

5. Evaluation

Qualitative evaluation: We evaluate the quality of the predicted meshes on the FaMoS test data, and compare it with the predictions of the current state-of-the-art, ToFu [50]. As the publicly available ToFu model[1] does not generalize to our scanner setup and it only predicts faces, we train ToFu on our training data to predict complete heads. For this, we first predict a low-resolution head with 341 vertices in the global stage, followed by upsampling and refining the mesh to the final mesh resolution. Unlike ToFu, we use no hierarchical architecture. To factor out the potential impact of the intermediate mesh resolution, we additionally compare to ToFu without mesh hierarchy, where the global stage directly predicts meshes of the final resolution, followed by refining the mesh in the local stage. We refer to this option as ToFu+. To compare with a direct image-to-3DMM regressor, we extend the coarse model of DECA [25] to a multi-view setting by regressing FLAME [49] parameters from the concatenated view feature vectors, independently reconstructed for each view.

Figure 4 shows that TEMPEH reconstructs 3D heads with the lowest error in the neck region for extreme head poses (Row 1), and with head shapes closer to the reference scans (Rows 2 & 3). See the Sup. Mat. for additional results, including 3DMM regressor details and comparisons.

Quantitative evaluation: We quantify the accuracy of the predicted 3D heads on the FaMoS test data by computing the point-to-surface distances between the vertices of each reference scan, and their closest points in the predicted 3D head's surface. To analyze the accuracy in different head regions, we segment each scan into face, scalp, and neck regions (see Sup. Mat.), and report the reconstruction errors

[1] https://github.com/tianyeli/tofu

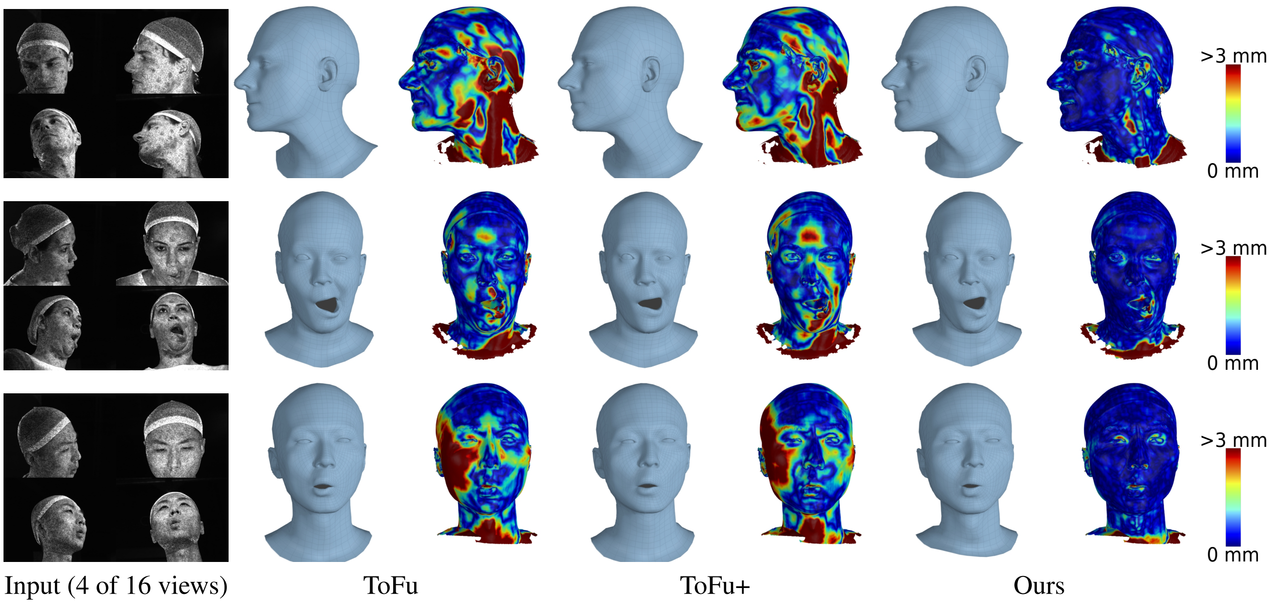

Figure 4. **Qualitative evaluation**. Comparison to ToFu [50] and ToFu+ on FaMoS test samples with varying expressions and head poses for subjects not present during training. For each method, we show the predicted mesh (left) and the color coded point-to-surface distance (right) between the reference scan and the predicted mesh as a heatmap on the scan's surface (red means $\geq$ 3 millimeter).

	Complete head			Face			Scalp			Neck		
Method	Median ↓	Mean ↓	Std ↓	Median ↓	Mean ↓	Std ↓	Median ↓	Mean ↓	Std ↓	Median ↓	Mean ↓	Std ↓
3DMM regressor	9.42	12.06	10.11	8.74	10.38	7.91	9.19	11.80	9.74	13.02	16.45	13.29
ToFu	0.72	1.44	2.72	0.63	0.93	1.50	0.62	1.46	3.27	1.30	2.39	3.24
ToFu+	0.82	1.59	2.84	0.68	1.00	1.52	0.73	1.59	3.16	1.50	2.77	3.82
Ours	**0.26**	**0.51**	**1.22**	**0.21**	**0.34**	**1.22**	**0.26**	**0.41**	**0.66**	**0.38**	**0.95**	**1.91**

Table 1. **Quantitative evaluation**. Reconstruction error for varied head regions (FaMoS test set). We compare to a 3DMM regressor, ToFu [50] and ToFu without hierarchical architecture (ToFu+), all trained to predict entire heads on the TEMPEH training data. Errors in mm.

for the entire head and the individual segments. Table 1 shows that TEMPEH outputs 3D heads with 64% lower reconstruction error compared to ToFu, and 68% lower than ToFu+. See the Sup. Mat. for cumulative error plots and additional qualitative comparisons.

Training TEMPEH minimizes the distance between the predicted heads and MVS scans, hence it effectively registers the scans. TEMPEH closely fits the training scans, with a median error of 0.17 mm (coarse stage: 0.80 mm).

Ablation experiments: To quantify the impact of individual design choices, we train following model variants: (1) *Coarse w/o s2m loss*: fully supervised training with v2v loss only (Eq. 6) (i.e., no scan supervision). (2) *Coarse w/o head localization*: direct prediction of the head mesh from the coarse feature volume without head localization. (3) *Refinement w/ naïve feature fusion*: feature aggregation as mean and variance across views, without surface-aware feature fusion. (4) *Ours w/o s2m loss*: training of coarse and refinement stage with v2v loss only. (5) *Ours w/o head localization*: coarse model w/o head localization with refinement stage. (6) *Ours color images input*: use of color images as input, instead of gray-scale stereo images. (7) *Ours hierarchical*: predicting a mesh with 1000 vertices in the coarse stage, followed by upsampling and refinement in the second stage. Fig. 5 and Tab. 2 compare different model variants qualitatively and quantitatively. We find that both, surface distance loss and head localization are essential for the coarse head inference, as ablating either of them leads to worse performance. Further, our model with surface-aware feature fusion and surface distance loss predicts heads with lowest error. While the models without head localization (5) or with hierarchical architecture (7) infer heads with comparably low errors, they reconstruct e.g., the lips region with lower fidelity due to the worse expression initialization from the coarse stage (see Fig. 5 and Sup. Mat.).

6. Discussion

Reference registrations: TEMPEH uses reference registrations for pre-training and regularization. While these reg-

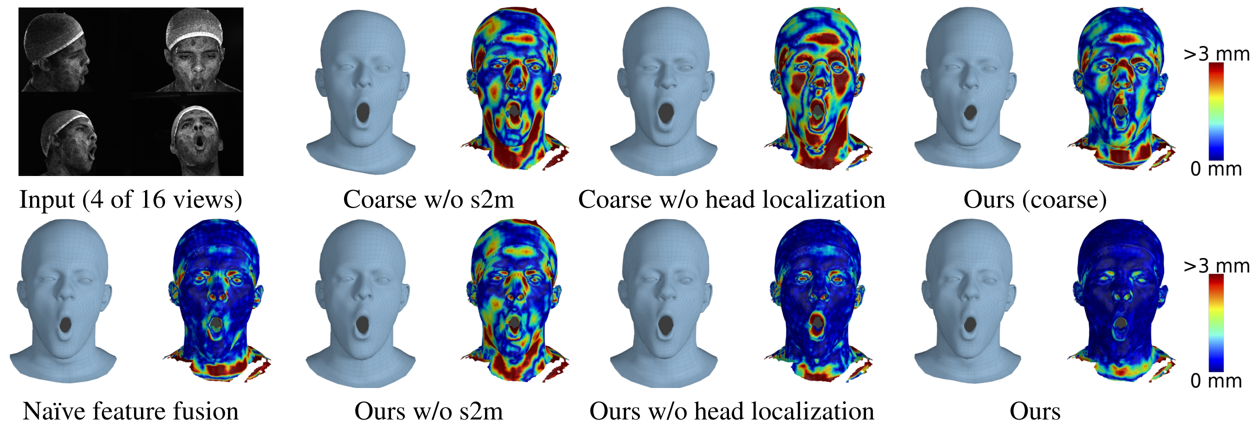

Figure 5. **Ablation experiments**. For each model variant, we show the reconstructed mesh (left) and the color coded point-to-surface distance (right) between the reference scan and the reconstructed mesh as a heatmap on the scan's surface (red means $\geq$ 3 millimeter).

	Complete head			Face			Scalp			Neck		
Method	Median ↓	Mean ↓	Std ↓	Median ↓	Mean ↓	Std ↓	Median ↓	Mean ↓	Std ↓	Median ↓	Mean ↓	Std ↓
Coarse w/o s2m loss	1.15	1.85	2.76	0.91	1.21	1.56	1.22	2.13	3.44	1.75	2.60	2.91
Coarse w/o head localization	1.16	1.60	1.79	1.12	1.43	1.61	1.00	1.33	1.26	1.69	2.48	2.65
Ours coarse	**0.71**	**1.11**	**1.56**	**0.68**	**0.93**	**1.40**	**0.61**	**0.92**	**1.18**	**1.09**	**1.81**	**2.28**
Refinement w/ naïve feature fusion	0.35	0.70	1.36	0.27	0.45	1.26	0.36	0.63	0.90	0.54	1.23	2.06
Ours w/o s2m loss	0.78	1.44	2.59	0.64	0.89	1.43	0.76	1.62	3.22	1.27	2.18	2.81
Ours w/o head localization	0.28	0.58	1.32	0.22	0.39	1.24	0.28	0.44	**0.62**	0.40	1.09	2.16
Ours color images input	0.44	0.77	1.33	0.34	0.54	1.27	0.46	0.71	0.93	0.65	1.28	1.94
Ours hierarchical	0.27	0.61	1.55	**0.21**	0.36	1.24	0.27	0.62	1.54	**0.37**	0.97	2.04
Ours	**0.26**	**0.51**	**1.22**	**0.21**	**0.34**	**1.22**	**0.26**	**0.41**	0.66	0.38	**0.95**	**1.91**

Table 2. **Ablation experiments**. Effects of training from registered meshes instead of scans (w/o s2m loss), reconstructing heads from the entire feature volume (w/o head localization), aggregating features without leveraging surface and visibility information (w/ naïve feature fusion), using the capture system's 8 color images as input, and using a hierarchical architecture (ours hierarchical). Errors in mm.

istrations are obtained fully automatically [49], their computation is slow and computationally expensive. Instead, directly regularizing to a statistical model [23] during training by jointly optimizing the statistical model's parameters (i.e., as done for coupled registrations [49]) could mitigate the need for registrations. This, however, adds an additional level of complexity, which goes beyond our current scope.

Registration quality: While TEMPEH's reconstructions well resemble the reference scans, expressions like eye blinks are not well captured. This is due to the poor quality of the scans in the eye region, the fast motion of the eyelids, and the absence of a clear signal in the optimized point-to-surface distance (Eq. 4). We plan to add an additional eyelid landmark error to improve the eyelid tracking in the future.

Representation: Several methods exist to learn deep implicit functions with dense correspondence from scans [51, 84, 85]. Replacing TEMPEH's mesh representation with such implicit functions is an interesting future direction.

Camera calibrations: TEMPEH is designed for lab environments with a carefully calibrated capture system. Adapting TEMPEH to less constrained scenarios with noisy or unknown camera calibration goes beyond the current scope.

7. Conclusion

We have presented TEMPEH, a framework to predict entire 3D heads in dense correspondence from calibrated multi-view images. TEMPEH infers 3D heads with reconstruction accuracy that is 64% lower than the previous state-of-the-art. We achieve this by training TEMPEH directly from scans, using a spatial transformer head localization module, and surface-aware feature fusion. Intuitively, the training from scans overcomes ambiguous correspondence across subjects and imperfect correspondence across expressions. The head localization enables the coarse stage to handle a large capture volume by focusing on the region of interest, and it provides a better initialization for the geometry refinement. The surface-aware feature fusion accounts for self-occlusions. Due to the inferred geometry accuracy and inference speed, TEMPEH is useful for applications like multi-view head performance capture.

Acknowledgement: We thank T. Alexiadis, M. Höschle, Y. Fincan, B. Pellkofer for data capture and IT support, T. McConnell for voice over, and R. Daněček, W. Zielonka, P. Patel, and P. Kulits for proofreading. **Disclosure:** https://files.is.tue.mpg.de/tbolkart/disclosure.txt

References

[1] Victoria Fernández Abrevaya, Stefanie Wuhrer, and Edmond Boyer. Spatiotemporal modeling for efficient registration of dynamic 3D faces. In *International Conference on 3D Vision (3DV)*, pages 371–380, 2018. 2

[2] Oswald Aldrian and William AP Smith. Inverse rendering of faces with a 3D morphable model. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(5):1080–1093, 2013. 2

[3] Oleg Alexander, Mike Rogers, William Lambeth, Matt Jen-Yuan Chiang, and Paul E. Debevec. The digital Emily project: photoreal facial modeling and animation. In *SIGGRAPH Courses*, pages 12:1–12:15, 2009. 2

[4] Brian Amberg, Reinhard Knothe, and Thomas Vetter. Expression invariant 3D face recognition with a morphable model. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 1–6, 2008. 2

[5] Mehdi Bahri, Eimear O' Sullivan, Shunwang Gong, Feng Liu, Xiaoming Liu, Michael M. Bronstein, and Stefanos Zafeiriou. Shape my face: Registering 3D face scans by surface-to-surface translation. *International Journal of Computer Vision (IJCV)*, 129(9):2680–2713, 2021. 2

[6] Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep facial non-rigid multi-view stereo. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5849–5859, 2020. 2, 3

[7] Anil Bas and William A. P. Smith. What does 2D geometric information really tell us about 3D face shape? *International Journal of Computer Vision (IJCV)*, 127(10):1455–1473, 2019. 3

[8] Anil Bas, William A. P. Smith, Timo Bolkart, and Stefanie Wuhrer. Fitting a 3D morphable model to edges: A comparison between hard and soft correspondences. In *Asian Conference on Computer Vision Workshops*, pages 377–391, 2017. 2

[9] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul A. Beardsley, Craig Gotsman, Robert W. Sumner, and Markus H. Gross. High-quality passive facial performance capture using anchor frames. *Transactions on Graphics, (Proc. SIGGRAPH)*, 30(4):75, 2011. 2, 3

[10] Paul J. Besl and Neil D. McKay. A method for registration of 3-D shapes. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 14(2):239–256, 1992. 2

[11] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pages 187–194, 1999. 2

[12] Timo Bolkart and Stefanie Wuhrer. A groupwise multilinear correspondence optimization for 3D faces. In *International Conference on Computer Vision (ICCV)*, pages 3604–3612, 2015. 2

[13] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David J. Dunaway. A 3D morphable model learnt from 10,000 faces. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5543–5552, 2016. 2

[14] Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. High resolution passive facial performance capture. *Transactions on Graphics (TOG)*, 29(4), 2010. 2, 3

[15] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. FaceWarehouse: A 3D facial expression database for visual computing. *Transactions on Visualization and Computer Graphics (TVCG)*, 20(3):413–425, 2014. 2

[16] Hang Dai, Nick E. Pears, William A. P. Smith, and Christian Duncan. A 3D morphable model of craniofacial shape and texture variation. In *International Conference on Computer Vision (ICCV)*, pages 3104–3112, 2017. 2

[17] Hang Dai, Nick E. Pears, William A. P. Smith, and Christian Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision (IJCV)*, 128(2):547–571, 2020. 2

[18] Radek Daněček, Michael J. Black, and Timo Bolkart. Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022. 3

[19] Dawson-Haggerty et al. Trimesh. `https://trimsh.org/`, 2019. 6

[20] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, pages 285–295, 2019. 2

[21] Pengfei Dou and Ioannis A. Kakadiaris. Multi-view 3D face reconstruction with deep recurrent neural networks. *Image and Vision Computing (IVC)*, 80:80–91, 2018. 3

[22] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. End-to-end 3D face reconstruction with deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5908–5917, 2017. 3

[23] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, 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. *Transactions on Graphics (TOG)*, 39(5), 2020. 1, 2, 8

[24] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J. Black, and Victoria Fernández Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation. In *European Conference on Computer Vision (ECCV)*, volume 13673, pages 72–90, 2022. 2

[25] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *Transactions on Graphics, (Proc. SIGGRAPH)*, 40(4):88:1–88:13, 2021. 3, 6

[26] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3D face reconstruction and dense alignment with position map regression network. In *European Conference on Computer Vision (ECCV)*, volume 11218, pages 557–574, 2018. 3

[27] Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3D facial expression reconstruction from videos, 2022. 3

[28] Clement Fuji Tsang, Maria Shugrina, Jean Francois Lafleche, Towaki Takikawa, Jiehan Wang, Charles Loop, Wenzheng Chen, Krishna Murthy Jatavallabhula, Edward Smith, Artem Rozantsev, Or Perel, Tianchang Shen, Jun Gao, Sanja Fidler, Gavriel State, Jason Gorski, Tommy Xiang, Jianing Li, Michael Li, and Rev Lebaredian. Kaolin: A pytorch library for accelerating 3d deep learning re-

search. https://github.com/NVIDIAGameWorks/kaolin, 2022. 6
[29] Graham Fyffe, Koki Nagano, Loc Huynh, Shunsuke Saito, Jay Busch, Andrew Jones, Hao Li, and Paul E. Debevec. Multi-view stereo on consistent face topology. *Computer Graphics Forum (CGF)*, 36(2):295–309, 2017. 2, 3
[30] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3D face rigs from monocular video. *Transactions on Graphics (TOG)*, 35(3):28, 2016. 3
[31] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. GANFIT: generative adversarial network fitting for high fidelity 3D face reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1164, 2019. 2
[32] S. Geman and D. E. McClure. Statistical methods for tomographic image reconstruction. *Proceedings of the 46th Session of the International Statistical Institute, Bulletin of the ISI*, 52, 1987. 5
[33] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3D morphable model regression. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8377–8386, 2018. 3
[34] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul E. Debevec. Multiview face capture using polarized spherical gradient illumination. *Transactions on Graphics (TOG)*, 30(6):129, 2011. 3
[35] Michael Goesele, Brian Curless, and Steven M. Seitz. Multi-view stereo revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2402–2409, 2006. 3
[36] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2492–2501, 2020. 3
[37] Riza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. DenseReg: Fully convolutional dense shape regression in-the-wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6799–6808, 2017. 3
[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6
[39] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. DPSNet: End-to-end deep plane sweep stereo. In *International Conference on Learning Representations (ICLR)*, 2019. 3
[40] Karim Iskakov, Egor Burkov, Victor S. Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *International Conference on Computer Vision (ICCV)*, pages 7717–7726, 2019. 6
[41] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *International Conference on Computer Vision (ICCV)*, pages 1031–1039, 2017. 3
[42] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2017–2025, 2015. 2
[43] Penglei Ji, Hanchao Li, Luyan Jiang, and Xinguo Liu. Lightweight multi-view topology consistent facial geometry and reflectance capture. In *Advances in Computer Graphics (CGI)*, volume 13002, pages 139–150, 2021. 2
[44] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 365–376, 2017. 3
[45] Ira Kemelmacher-Shlizerman. Internet based morphable model. In *International Conference on Computer Vision (ICCV)*, pages 3256–3263, 2013. 3
[46] Ira Kemelmacher-Shlizerman and Steven M. Seitz. Face reconstruction in the wild. In *International Conference on Computer Vision (ICCV)*, pages 1746–1753, 2011. 3
[47] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. InverseFaceNet: Deep monocular inverse face rendering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4625–4634, 2018. 2
[48] Hao Li, Bart Adams, Leonidas J. Guibas, and Mark Pauly. Robust single-view geometry and motion reconstruction. *Transactions on Graphics (TOG)*, 28(5):175, 2009. 2
[49] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2, 6, 8
[50] Tianye Li, Shichen Liu, Timo Bolkart, Jiayi Liu, Hao Li, and Yajie Zhao. Topologically consistent multi-view face inference using volumetric sampling. In *International Conference on Computer Vision (ICCV)*, pages 3824–3834, 2021. 2, 3, 4, 5, 6, 7
[51] Feng Liu and Xiaoming Liu. Learning implicit functions for topology-varying dense 3D shape correspondence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 8
[52] Feng Liu, Luan Tran, and Xiaoming Liu. 3D face modeling from diverse raw scan data. In *International Conference on Computer Vision (ICCV)*, pages 9407–9417, 2019. 2
[53] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 6
[54] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Félix Chabert, Malte Weiss, and Paul E. Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the Eurographics Symposium on Rendering Techniques*, pages 183–194, 2007. 3
[55] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20910–20920, 2022. 3
[56] Araceli Morales, Gemma Piella, and Federico M. Sukno. Survey on 3D face reconstruction from uncalibrated images. *Computer Science Review*, 40:1–35, 2021. 2

[57] Georgios Passalis, Panagiotis Perakis, Theoharis Theoharis, and Ioannis A. Kakadiaris. Using facial symmetry to handle pose variations in real-world 3D face recognition. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(10):1938–1951, 2011. 2

[58] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 6

[59] Stylianos Ploumpis, Evangelos Ververas, Eimear O' Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick E. Pears, William A. P. Smith, Baris Gecer, and Stefanos Zafeiriou. Towards a complete 3D morphable model of the human head. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(11):4142–4160, 2021. 2

[60] Eduard Ramon, Janna Escur, and Xavier Giró-i-Nieto. Multi-view 3D face reconstruction in the wild using siamese networks. In *International Conference on Computer Vision Workshops (ICCV-W)*, pages 3096–3100, 2019. 3

[61] Elad Richardson, Matan Sela, and Ron Kimmel. 3D face reconstruction by learning from synthetic data. In *International Conference on 3D Vision (3DV)*, pages 460–469, 2016. 3

[62] Jérémy Riviere, Paulo F. U. Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Single-shot high-quality facial geometry and skin appearance capture. *Transactions on Graphics (TOG)*, 39(4):81, 2020. 2, 3

[63] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015. 6

[64] Joseph Roth, Yiying Tong, and Xiaoming Liu. Unconstrained 3D face reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2606–2615, 2015. 3

[65] Augusto Salazar, Stefanie Wuhrer, Chang Shu, and Flavio Prieto. Fully automatic expression-invariant face correspondence. *Machine Vision and Applications*, 25(4):859–879, 2014. 2

[66] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7763–7772, 2019. 3

[67] Mike Seymour, Chris Evans, and Kim Libreri. Meet Mike: Epic avatars. In *SIGGRAPH*, 2017. 2

[68] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3D face reconstruction by occlusion-aware multi-view geometry consistency. In *European Conference on Computer Vision (ECCV)*, volume 12360, pages 53–70, 2020. 3

[69] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. DeepVoxels: Learning persistent 3D feature embeddings. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2437–2446, 2019. 3

[70] Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M Seitz. Total moving face reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 796–812, 2014. 3

[71] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. FML: Face model learning from videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10812–10822, 2019. 3

[72] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 Hz. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2549–2559, 2018. 2

[73] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. MoFA: model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 1274–1283, 2017. 2

[74] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-time face capture and reenactment of RGB videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016. 2

[75] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1599–1608, 2017. 3

[76] Xiang Wang, Chen Wang, Bing Liu, Xiaoqing Zhou, Liang Zhang, Jin Zheng, and Xiao Bai. Multi-view stereo in the deep learning era: A comprehensive review. *Displays*, 70:102102, 2021. 3

[77] Huawei Wei, Shuang Liang, and Yichen Wei. 3D dense face alignment via graph convolution networks. *arXiv preprint arXiv:1904.05562*, 2019. 3

[78] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljevic, Daniel Wilde, Stephan Garbin, Chirag Raman, Jamie Shotton, Toby Sharp, Ivan Stojiljkovic, Tom Cashman, and Julien Valentin. 3D face reconstruction with dense landmarks. In *European Conference on Computer Vision (ECCV)*, pages 160–177, 2022. 2, 3

[79] Chenglei Wu, Takaaki Shiratori, and Yaser Sheikh. Deep incremental learning for efficient high-fidelity face tracking. *Transactions on Graphics (TOG)*, 37(6):234, 2018. 2

[80] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. MVF-Net: Multi-view 3D face morphable model regression. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 959–968, 2019. 2, 3

[81] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 598–607, 2020. 2

[82] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, volume 11212, pages 785–801, 2018. 3

[83] Chao Zhang, William A. P. Smith, Arnaud Dessein, Nick E. Pears, and Hang Dai. Functional faces: Groupwise dense correspondence using functional maps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5033–5041, 2016. 2

[84] Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. ImFace: A nonlinear 3D morphable face model with implicit neural representations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20320, 2022. 2, 8

[85] Zerong Zheng, Tao Yu, Qionghai Dai, and Yebin Liu. Deep implicit templates for 3D shape representation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1429–1439, 2021. 8

[86] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019. 4

[87] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)*, pages 250–269, 2022. 3

[88] 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. *Computer Graphics Forum (CGF)*, 37(2):523–550, 2018. 2