Capturing the Unseen: Vision-Free Facial Motion Capture Using Inertial Measurement Units

Youjia Wang[1,2*], Yiwen Wu[1,2*], Hengan Zhou[1,2], Hongyang Lin[1,3], Xingyue Peng[1], Jingyan Zhang[1], Yingsheng Zhu[1], YingWenQi Jiang[1], Yatu Zhang[1], Lan Xu[1], Jingya Wang[1], Jingyi Yu[1†]

[1]ShanghaiTech University
[2]LumiAni Technology
[3]Deemos Technology
{wangyj2, wuyw2023, zhouha, linhy, pengxy2023, zhangjy7,
zhuysh, jiangywq, zhangyt2023, xulan1, wangjingya, yujingyi}@shanghaitech.edu.cn

Abstract

We present Capturing the Unseen (CAPUS), a novel facial motion capture (MoCap) technique that operates without visual signals. CAPUS leverages miniaturized Inertial Measurement Units (IMUs) as a new sensing modality for facial motion capture. While IMUs have become essential in full-body MoCap for their portability and independence from environmental conditions, their application in facial MoCap remains underexplored. We address this by customizing micro-IMUs, small enough to be placed on the face, and strategically positioning them in alignment with key facial muscles to capture expression dynamics. CAPUS introduces the first facial IMU dataset, encompassing both IMU and visual signals from participants engaged in diverse activities such as multilingual speech, facial expressions, and emotionally intoned auditions. We train a Transformer Diffusion-based neural network to infer Blendshape parameters directly from IMU data. Our experimental results demonstrate that CAPUS reliably captures facial motion in conditions where visual-based methods struggle, including facial occlusions, rapid movements, and low-light environments. Additionally, by eliminating the need for visual inputs, CAPUS offers enhanced privacy protection, making it a robust solution for vision-free facial MoCap.

Introduction

In the data-driven AI era, the efficacy of analytics tools is directly tied to their ability to adapt to the unique characteristics of different data modalities. For facial motion capture, the success of now widely adopted tools such as 3DDFA (Zhu et al. 2017; Guo et al. 2020), DECA (Feng et al. 2021) and Apple's ARKit (Apple 2023) are largely attributed to the availability of RGB and RGBD cameras by mobile devices. Using the captured images as the sole modality, these solutions offer a rapid means of acquiring facial geometry and expression to support various downstream tasks such as editing, relighting, animation, etc (Zhang et al. 2022). However, image as a modality has its own limitations. For example, mobile phone camera based solutions require the user facing the camera all the time, which is impractical in outdoor activities. Existing algorithms are also vulnerable to occlusions, motion blurs, and noise. In fact, for privacy protection, the use of images may even be deliberately avoided.

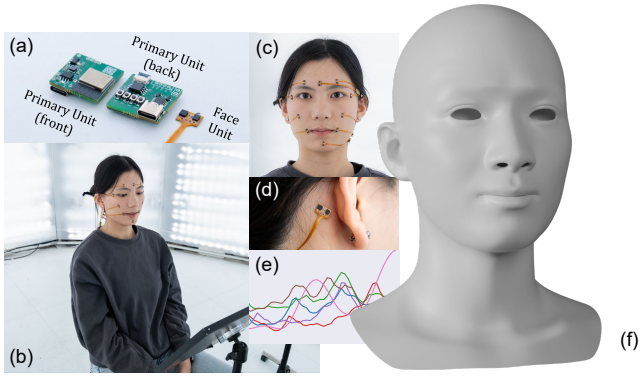

Figure 1: We introduce CAPUS, an innovative facial capture system based on IMUs. Using flexible electronic materials, we fabricate miniature IMUs that attach to the human face. Without relying on any visual signals, CAPUS can accurately reconstruct facial expressions.

In this work, we explore using a new type of data modality for facial motion capture. We observe full-body motion capture using visual signals encounters similar challenges but the latest successes unanimously resorted to the Inertial Measurement Units (IMUs) as the input signal Loper et al. (2015). By attaching the IMUs to various body joints, these solutions manage to capture essential acceleration and axis angle data for modeling body motion. Yi, Zhou, and Xu (2021) manages to achieve comprehensive body motion capture using as few as six IMUs whereas Li, Liu, and Wu (2023) leverages the stability and generative capabilities of Transformer Diffusion to further improve the robustness. It is not an exaggeration that IMUs have now become as integral as visual-based methods owing to their exceptional portability and minimal spatial demands. In particular, IMUs neither require using visual sensors nor rely on external envi-

*These authors contributed equally.
†Corresponding author.

ronmental conditions, offering unique advantages in outdoor activities and remote applications.

We introduce Capturing the Unseen (CAPUS), the first IMU-based facial motion capture solution that provides a camera-free alternative to traditional visual-based methods. CAPUS overcomes previous challenges related to the large size and lack of flexibility of IMUs, making them suitable for facial applications. To address this, we developed a new IMU design tailored specifically for facial use as shown in Fig. 1(a), with a strong focus on miniaturization. By separating the data acquisition and main control modules, CAPUS ensures that the face-attached device is both compact and lightweight. The acquisition module is designed using flexible materials to adhere comfortably to the face, ensuring accurate signal capture without compromising user comfort. This design minimizes interference with natural facial movements while enabling reliable data transmission and synchronization.

In terms of data processing, we observe that IMU signals tend to have much lower signal-to-noise ratios compared to visual input, leading to less reliable spatial features. Additionally, facial expressions are primarily driven by muscle movements, unlike body motion capture where spatial positions are closely linked to joint rotations. This poses a challenge in effectively interpreting IMU data for facial expressions. To address this, the proposed CAPUS adopts an anatomy-driven strategy by strategically placing IMUs in alignment with specific muscles that control facial expressions. Using CAPUS, we have created the first facial IMU dataset, which includes IMU signals, visual data, and ARKit parameters. This IMU-ARKit dataset records signals from participants performing various activities, such as speaking different languages, making facial expressions, and auditioning with emotional intonation. We then utilize this dataset to train a Transformer Diffusion-based neural network to infer Blendshape parameters directly from the IMU data. Our experiments validate the reliability of the dataset and the effectiveness of our approach.

Moreover, CAPUS supports reliable facial motion capture in traditionally challenging cases for visual-based solutions. In an era where digital privacy is a paramount concern, CAPUS offers a new reliable method of capturing facial expressions without visual input, thereby safeguarding portrait rights. In addition, by freeing a performer from holding a camera by hand toward the face, CAPUS supports facial motion capture while the performer is on the move, with normal body movements to convey body language. Finally, CAPUS can handle challenging scenarios when facial parts (e.g., the mouth) are severely occluded (e.g., during eating or drinking), where vision-based solutions would easily fail. Finally, some subtle changes, especially in the speed of muscle movements, are very challenging to visual sensors but are tractable using IMUs.

In conclusion, our contributions are as follows:

1. We introduce the first system capable of recovering human facial expressions using Inertial Measurement Units (IMUs), offering a novel approach to facial motion capture.
2. We design a new, lightweight IMU device that can be comfortably worn on the face, utilizing flexible electronic materials and weighing just 2.7% of a commercial Xsens IMU.
3. A new multi-modal dataset is proposed, which includes aligned IMU signals, visual data, audio signals, ARKit expression parameters, subject emotion labels, and the text of the subject's speech.
4. We introduce a Transformer Diffusion-based pipeline for inferring Blendshape parameters directly from IMU data, thereby enhancing the capabilities of facial motion capture systems.

Related Works

Facial Mocap Early works by (Ferrigno, Borghese, and Pedotti 1990; Bianchi et al. 1998; Guo, Xu, and Tsuji 1994) pioneered the realization of human motion capture. Subsequently, face motion capture systems based on multi-camera setups (Michoud et al. 2007; de Aguiar et al. 2004; Vlasic et al. 2008; Cao et al. 2017) became the mainstream solution. Over time, efforts such as (Yuan and Chen 2014; Von Marcard et al. 2017) reduced the number of cameras required for effective capture. During the same period, significant progress was made in 2D and 3D facial landmark detection(Cootes et al. 1995; Cootes, Edwards, and Taylor 2001; Cao et al. 2014; Zhou et al. 2005; Guo, Zhao, and Wang 2024; Li et al. 2024). More recently, specialized 3D reconstruction methods have emerged (Bao et al. 2021; Smith et al. 2020; Egger et al. 2020; Weise et al. 2011; Cao et al. 2013), with ARKit being a notable example (Apple 2023). However, vision-based approaches are often vulnerable to occlusion issues. The work of (Qammaz and Argyros 2023) addresses this challenge by predicting information about occluded regions. Additionally, some research has focused on reconstructing the movements of the human face from speech(Zhao et al. 2024; Stan, Haque, and Yumak 2023).

Sensor-based Mocap The advancements in inertial measurement units (IMUs), driven by works such as (Bachmann et al. 2001; Del Rosario et al. 2018; Foxlin 1996; Roetenberg et al. 2005; Vitali, McGinnis, and Perkins 2020; Liu et al. 2011; Vlasic et al. 2007; Ahmad et al. 2013), have significantly optimized their size and performance, establishing IMUs as a viable tool in the domain of human motion capture. Early works using IMUs (Schepers et al. 2018; Noitom) achieved full-body human motion capture by mapping IMU rotations to the angles of the human skeleton. Subsequent efforts (Huang et al. 2018; Riaz et al. 2015; Slyper and Hodgins 2008; Tautges et al. 2011; Von Marcard et al. 2017; Yi, Zhou, and Xu 2021; Li, Liu, and Wu 2023; Du et al. 2023), have gradually reduced the number of IMUs required for full-body mocap from 17 to as few as 6. These methods offer a broader capture range than vision-based methods and are not constrained by obstacles or lighting conditions.

Some studies (Makaussov et al. 2020; Mummadi et al. 2018) have utilized IMUs for hand motion capture, demonstrating the potential of IMUs for mocap on smaller body parts.

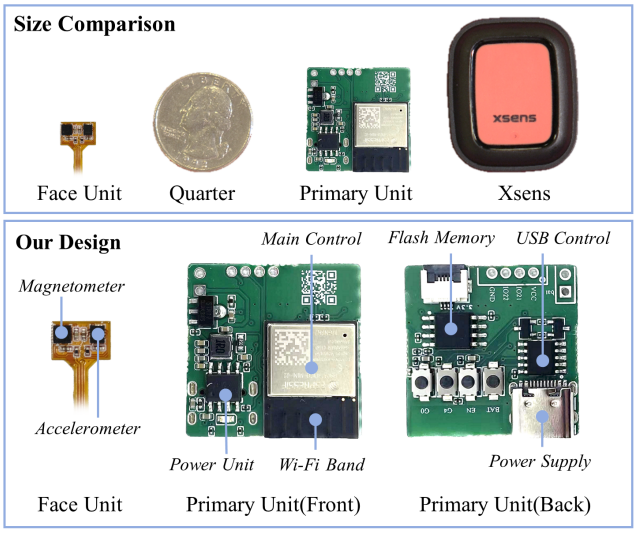

Figure 2: Our IMU has two main components: the face unit and the primary unit. Top: size comparison. Bottom: architecture design.

IMUs for Facial MoCap

Light-weight Facial IMU Sensor Design

Within the field of motion capture, IMU plays a critical role in reflecting the spatial movements of an object by measuring its orientation and acceleration. IMUs designed for full-body motion capture, such as Xsens, Sony Mocopi, and others, have been widely applied commercially. These units usually consist of various parts, including detecting sensors and data transmission modules, making them too hulking to be used for facial motion capture. Furthermore, employing multiple units of this model for facial capture can lead to severe occlusion, preventing observation of the participant's facial expressions. These necessitates the development of a custom-designed IMU, specifically tailored to meet the unique requirements and scale of facial motion capture.

Our design preserves the function of standard IMU while minimizing weight and size to cater to the requirements for facial capture. Fig. 2(top) compares the size of our IMU. We achieved significant miniaturization by separating the sensor module from the data transmission module. We designed the IMU's face module using flexible electronic materials to closely conform to the skin, ensuring that it does not impede natural facial movements. This design allowed our sensor module to be compact, measuring only 0.6 cm^2 and weighing merely 0.3 grams, a stark reduction to 5.4% the area and only 2.7% the weight of an Xsens module.

Fig. 2(down) provides a detailed overview of the specific hardware components utilized in our study. The sensor module incorporates a total of nine-axis sensing sub-units, which include the QMC5883P (QST Inc.) from Silicon Power, a three-axis magnetic field sensor with a measurement range of ± 30 gauss, and the QMI8658 (QST Inc.) integrated chip, which combines a three-axis gyroscope and accelerometer. These sensors are capable of accurately recording spatial orientations and accelerations at a rate of 60fps. The data transmission module is primarily based on the ESP32 controller. It employs the UDP protocol to collect and correct data detected by the sensor module. Additionally, we use a Wi-Fi module to transmit the computed data to the host computer. The data includes time stamps, quaternion representations, and acceleration values at each recorded instance.

The data transmission module of our face IMU sensor system requires only a 5V battery supply. This setup provides the essential conditions for the portability and wearability of the face IMU sensor system. Furthermore, the connection to the host computer via Wi-Fi allows users to move freely within the Wi-Fi signal range while wearing the Face IMU, enabling high degrees of mobility.

We further delved deeply into the essential technology for capturing facial information in synchrony using multiple IMUs. To achieve this, it is imperative to address two fundamental challenges: synchronization and calibration. We designated one ESP32 as the auxiliary ESP32, employing it as a benchmark for synchronizing and calibrating the others. We integrated a calibration program into this ESP32 within the data transfer module during hardware design and used the data module of the auxiliary ESP32's clock as a reference point. We transmitted pulse signals through the DuPont line to each IMU's ESP32 for calibration purposes. Upon receiving this pulse signal, each ESP32 aligns its internal clock with the external reference, synchronizing the timestamps across all IMUs.

Next, acknowledging the variability in facial structures and the potential for slight discrepancies in IMU placement each time, we adopted the concept of a Neutral facial performance, similar to the approach used by (Yi, Zhou, and Xu 2021; Egger et al. 2020) in body mocap. After wearing the IMUs for the participants, we had each participant relax the facial muscles, presenting a Neutral state, and recorded the orientation of each IMU. In subsequent calculations, we used the orientation relative to this pose as a baseline. To eliminate the interference with expression prediction caused by head rotation, we strategically place an auxiliary IMU behind the ear, as shown in Fig. 1(d). We use this IMU to record the overall rotation of the head. We provide a detailed description in the supplementary materials.

Capturing IMU-ARKit Dataset

To accurately capture facial movements, it is imperative to attach IMUs to distinct regions on the surface of the face. The layout in Fig. 1(c) is informed by a detailed analysis of the distribution of facial muscles(Uldis 2017). We demarcated distinct facial zones, the zygomaticus area, the buccinator and mentalis area, the orbicularis oculi area, and the frontalis area. In every designated region, we meticulously placed at least one IMU to ensure comprehensive monitoring of the key muscle groups and facial zones. Acknowledging the sensitivity of certain facial regions, we intentionally avoided placing IMUs on the eyelids and lips, and used the surrounding IMUs to accurately predict their movement.

Our goal is to recover the 3D geometry of the face from the captured IMU data $\mathcal{S}$. A common approach for this task is to use blendshapes as the 3D representation. Blend-

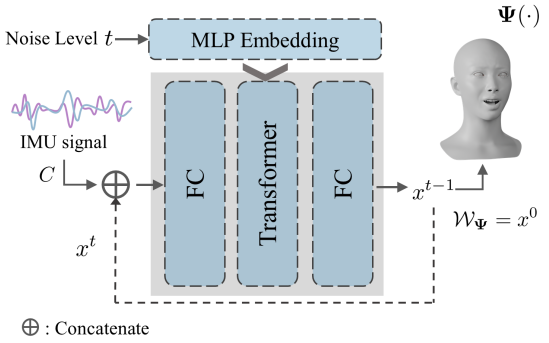

Figure 3: Our transformer diffusion network architecture. We use IMU signal C as a condition input to the network. In each iteration, the network denoises x^t, and finally outputs the predicted blendshape parameters x^0.

shape technology, widely adopted in facial animation and motion capture, operates on the principle of parametric modeling, enabling the generation of highly realistic and nuanced expressions. Specifically, a blendshape model is defined by a collection of blendshape weights, denoted as $\mathcal{W} = \{w_1, w_2, \cdots, w_m\}$, a facial expression blendshape model can be represented as:

$$M(\mathcal{W}) = B_0 + \sum_k^m w_k B_k. \tag{1}$$

where B_0 represents the neutral face, B_k is the blendshape basis vector, and m is the number of blendshapes. By linearly interpolating between different blendshapes, this approach allows the creation of multiple facial expressions.

Given that the IMU is capable of capturing acceleration and orientation, we propose a method for mapping these physical measurements to blendshape weights $\mathcal{W}$. This requires the development of an algorithm that converts IMU readings into meaningful hybrid shape parameters.

In order to realize a data-driven solution for predicting facial blendshape weights using IMU, we set out to create a facial IMU dataset aligned with ARKit parameters, as demonstrated in Fig. 1. This dataset was carefully compiled to contain paired data of IMU signals and ARKit parameters to ensure a comprehensive base for model training.

Our dataset contains records from 20 different participants. These individuals are all within the 18-40 age range and proficient in English, providing richly varied and vivid facial expressions. Fig. 1(b) shows an example of the data collection setup. Each participant wore a set of 11 IMUs and sat in the acquisition seat, with the teleprompter screen placed directly in front of the participant, next to an iPhone that captured the visual information. We used LiveLinkface(UnrealEngine 2023) to capture the visual information, which is divided into two parts: the RGB video sequence and the ARKit Parameters.

Before the formal data collection process began, participants were given time to adapt to the sensation of wearing IMUs, ensuring captured facial movements were natural and unrestricted. Participants were instructed to tap the IMU located on mentalis at the start of each recording, as reference frames for synchronizing the IMU signals with the visual signals. The data was divided into three parts by intentionally designed content that disentangles facial expressions into plain facial movements and emotions. In the first part, participants read aloud the provided content in a calm tone, with a split between native language and English. This was done to capture the natural facial movements associated with the language. In the second part, participants were asked to sequentially make a series of facial expressions that were based on specific classifications, ensuring a full range of emotions and movements. Finally, participants were asked to perform lines of one specific emotion from a set of emotions, joining plain facial movements with emotions.

Our IMU-ARKit dataset provides aligned data pairs of synchronized IMU signals from 11 IMUs, RGB frames, audio signals, and ARKit parameter sequences, with the emotion and content of each sequence annotated. We showcased samples of our dataset in the supplementary video.

IMU-Based Facial Tracker

Considering the IMU signals provide information not as plain as visual inputs, we chose a lightweight Transformer Diffusion-based network, to interpret the IMU signals meaningfully.

As shown in Fig. 3, our network $\mathbf{\Psi}(\cdot)$ comprises two parts, an MLP embedding network $\mathbf{em}(\cdot)$ and a denoising network $\psi(\cdot)$. The denoising network has an initial Fully Connected (FC) layer, a concluding FC layer, and a transformer-based core.

A single frame IMU signal $c_j^i = [a_j^i, q_j^i] \in \mathbb{R}^7$ contains acceleration a_j^i and spatial orientation q_j^i, where q_j^i is represented as quaternion. We concatenate signals of each frame from all 11 IMUs as $C^i \in \mathbb{R}^{77}$, and stack signals of T consecutive frames to produce the input IMU signal $C \in \mathbb{R}^{T\times 77}$.

To reconstruct blendshape parameters from IMU signals, we utilized the denoising process of the diffusion model with IMU signal C as the condition. Specifically, for the denoising process at noise level t, we concatenate the noised blendshape parameters $x^t \in \mathbb{R}^{T\times m}$ with condition C, combined with noise embedding $\mathbf{em}(t)$ as input to ψ, and estimate $x^{t-1} \in \mathbb{R}^{T\times m}$. Here m is the number of blendshapes.

$$x^{t-1} = \psi(\mathbf{em}(t), x^t, C). \tag{2}$$

We repeat such process until obtaining the predicted blendshape parameters $\mathcal{W}_{\mathbf{\Psi}} = x^0 \in \mathbb{R}^{T\times d}$ as final output.

The paired sequence of blendshape parameters and IMU signals are provided in training. Given ground truth blendshape parameters $\mathcal{W}$, the training loss is defined as follows:

$$\mathcal{L} = ||\mathcal{W}_{\mathbf{\Psi}} - \mathcal{W}||_1 . \tag{3}$$

We trained our network in a supervised manner on our IMU-ARKit dataset, with $T = 120$ for both training and testing. To avoid jittering at inference, we set an overlap of 60 frames to ensure the network has sufficient prior information to accurately determine the initial state of the face within the time window.

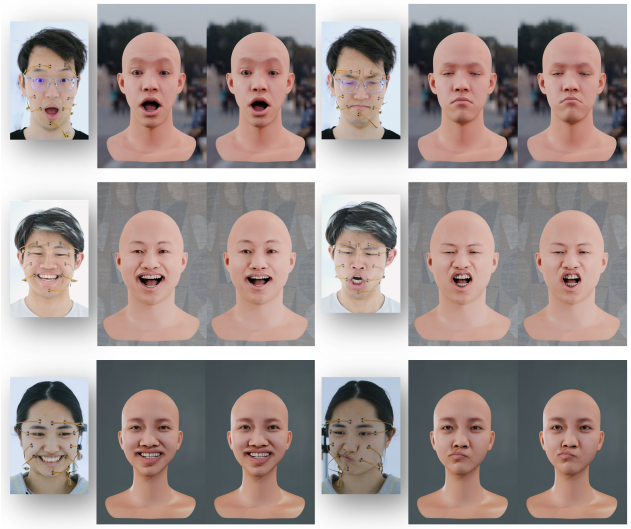

Figure 4: Gallery. We present three subjects, with each row corresponding to two different expressions of a single participant. For each subfigure, Left: Image reference. Middle: Facial motion reconstructed by our pipeline. Right: Recorded result by ARKit(Apple 2023). Our method achieves results that are comparable to those obtained using ARKit.

We adopted ICT Face Model (Li et al. 2020) as our blendshapes, and the number of blendshapes $m = 53$. The ARkit parameters are mapped into ICT blendshape parameters.

Experimental Evaluations

In Fig. 4, we use CAPUS to recover a variety of facial expressions. We include sequences of facial expressions that represent signature emotions as well as sequences of a performer speaking. The video results can be found in the supplementary video.

Following the similar network architecture as Li, Liu, and Wu (2023), CAPUS uses the noise as inputs, imposes the IMU data as transformer conditions, and outputs the inferred blendshape weights to control facial motions. We use Adam as the optimizer with a learning rate $2 \times 10^{-4}, \alpha = 0.9, \beta = 0.999$. We train and evaluate CAPUS on a single NVIDIA RTX3090 GPU. The training process takes ≈ 1 hours on all identities with paired data. For the generation and rendering of facial assets, we leverage the off-the-shelf technique DreamFace (Zhang et al. 2023) to maintain high fidelity and realistic results.

Evaluations on IMU Locations

We qualitatively evaluate how IMU placements across different facial regions affect final facial expression estimation, as shown in Fig. 5. The far left image compares various IMU position schemes, with white dots on the face representing the final locations CAPUS adopts. The images, arranged from left to right, depict the experimental positioning of test points in the Frontalis Area, Zygomaticus Area, and Buccinator and Mentalis Area, respectively. The top row shows the locations we have experimented with for placing the IMUs, with the red and purple ones as the final positions we chose to use.

In our studies, we strategically select the candidates for placing the IMUs to best reduce interference and align with the underlying muscles. The middle row demonstrates the specific facial movements performed by participants. We collect the acceleration data from respective IMUs during specific facial movements, shown in the bottom row of the images. Our selected IMU locations unanimously produce strong signals that correspond to higher sensitivity under motion. Such placements result in signals with a high SNR suitable for recovering accurate and reliable facial motions. Table 2 further shows the quantitative results.

Method	PVE [mm]↓	PVE_LMK [mm]↓	MSE↓
Ours	**0.075 ± 0.055**	**0.126 ± 0.091**	**0.0075**
Small Dataset	0.089 ± 0.063	0.150 ± 0.102	0.0093
Fewer IMU	0.078 ± 0.055	0.129 ± 0.089	0.0082

Table 1: Quantitative ablation study of our method.

Area	#1	#2	#3	#4	#5
Frontalis	**0.71**	0.57	0.52	0.58	-
Zygomaticus	**0.64**	**1.20**	0.57	0.32	-
Buccinator / Mentalis	0.84	**1.38**	0.54	0.88	**1.77**

Table 2: Quantitative evaluations on IMU placements. The table shows the variations times 10^{-3} where higher value essentially corresponds to higher sensitivity. Numbers in **Bold** fonts correspond to the placements that CAPUS uses.

Evaluations on Facial Capture

Next, we compare CAPUS with the state-of-the-art vision-based techniques DECA (Feng et al. 2021) and 3DDFA_V2 (Guo et al. 2020). Specifically, we experiment on a new IMU-ARKit dataset that takes the image captured by iPhone as the input of DECA and 3DDFA_V2 along with CAPUS. The results are shown in Fig. 6. Columns 3, 7, and 8 correspond to the results from CAPUS vs. 3DDFA_V2 and DECA. Visual quality wise, CAPUS estimations are comparable to the SOTA visual-based methods. Compared with DECA, CAPUS performs better near the eye region. Compared with 3DDFA_V2, CAPUS better recovers eyebrow movements induced by facial expressions.

We then demonstrate the necessity of our dataset. Much work in the field of human motion capture uses simulated datasets for training and tests on a small number of IMU datasets (Yi, Zhou, and Xu 2021; Li, Liu, and Wu 2023). However, our experiments show that the same approach does not work in facial capture. We used the approach of (Yi, Zhou, and Xu 2021) to generate a set of simulated datasets using the ARKit parameters of the training set. The performance on the test set after we trained on these simulated data is shown in Figure 6. Unlike the performance of hu-

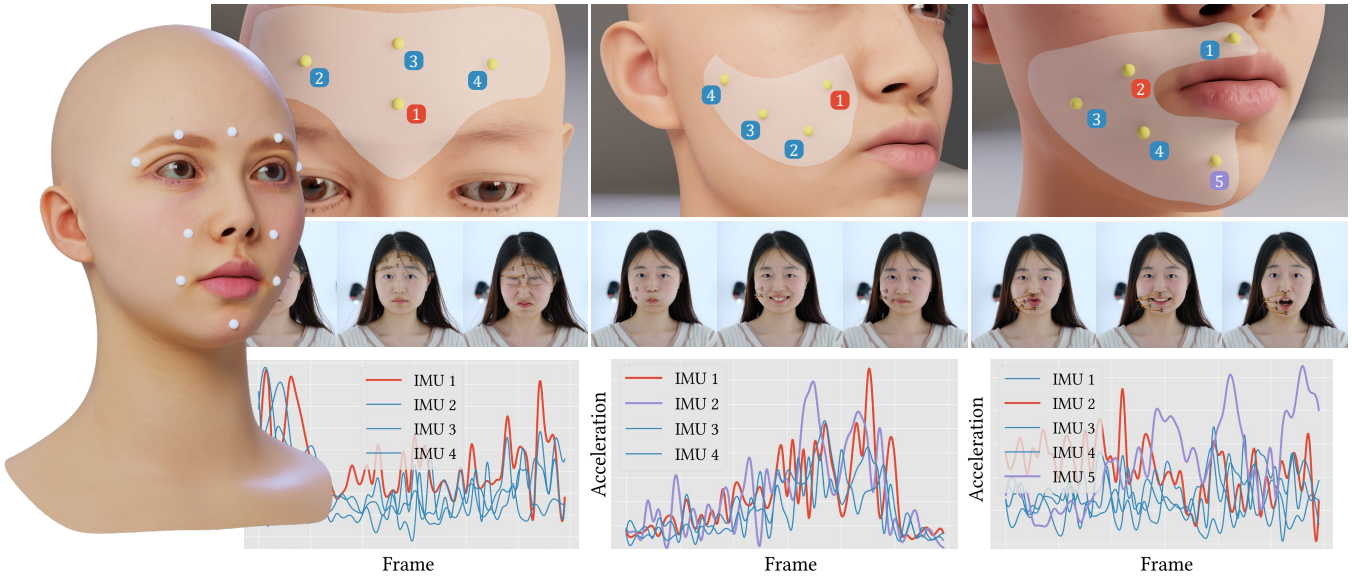

Figure 5: Experiment on IMU placement. This figure presents our anatomically-based facial partitioning, highlighting the selected points and the corresponding experiments for each facial region. The left image shows our chosen points on the face, while the other images elaborate on the individual experiments conducted for each specific area.

man motion capture, the simulated data in face capture does not yield correct results for the network, which fails to make correct predictions in the vast majority of movements.

We further conduct two ablation experiments to evaluate our dataset and the IMU placements: (1) *Fewer IMU*: we train the network using only a fraction of IMUs, i.e., the ones placed on the eyebrows (2 IMUs), jaw (1 IMU), and cheeks (2 IMUs). (2) *Small Dataset*: we train the network using 1/3 of the dataset.

The variations are illustrated in columns 4, and 5 of Fig. 6 sequentially. The results in column 4 show some examples that CAPUS fails to faithfully predict the motion, e.g., closed eyes. This is largely attributed to the locations where we place the IMUs. The results in column 5 manage to recover challenging facial distortions under extreme expressions. This indicates that our training dataset is sufficiently rich to cover these movements and the trained network is robust enough to generalize to reproduce these distortions.

We further conduct quantitative evaluations in Fig. 7 and Table. 1. Same as (Feng et al. 2021; Guo et al. 2020), we calculate the 3D per vertex error (PVE)(Shimada et al. 2023) on the deformed mesh as an indicator of the similarity between ARKit vs. IMSUE predictions. Specifically, we use the 3D landmark vertex error (PVE_LMK) to demonstrate the fidelity of CAPUS estimations on visually significant areas. We further calculate the MSE using the predicted blendshape weights with ARKit as the ground truth. The red curve represents the metrics for each frame using CAPUS whereas the purple and blue curves represent the metrics for the two ablation experiments.

In the supplementary material, we further compare our network with other architectures, and demonstrate that our network is not overfitted to the training set.

Applications

Camera-Free Facial Capture In traditional facial capture systems, users need to always face the camera, which limits the head and body movements. For example, while on the move, users have to hold their phones by hand, making it difficult to perform normal body movements and convey body language. We demonstrate using CAPUS as a portable facial capture solution, as shown in the supplementary video. Due to the modular design of the IMUs, the user's facial skin experiences minimal weight. All IMUs are powered by a portable power bank, using Wi-Fi module to communicate with the computer. As a result, CAPUS allows for accurate facial capture while a person is walking, preserving complete facial information and freeing the user's hands.

Occluded Facial Capture In some scenarios, facial capture encounters unavoidable occlusions, such as during eating or drinking. Professional actors commonly resort to 'mimicking' eating to avoid this issue, which can result in a lack of authenticity. We demonstrate using CAPUS to conduct robust motion capture in such scenarios. We showcase CAPUS's accurate and stable motion capture capabilities in the heavily occluded 'eating an apple' situation, as shown in Fig. 8. While eating, the user's hands and the food largely occlude the face, particularly the mouth regions, rendering vision-based methods ineffective. CAPUS, instead, does not rely on any video signals, allowing for accurate capture of the mouth movements. The supplementary video includes several dynamic sequences.

Conclusion and Discussion

We have presented CAPUS, a novel vision-free facial motion capture technique that takes only IMU signals as input.

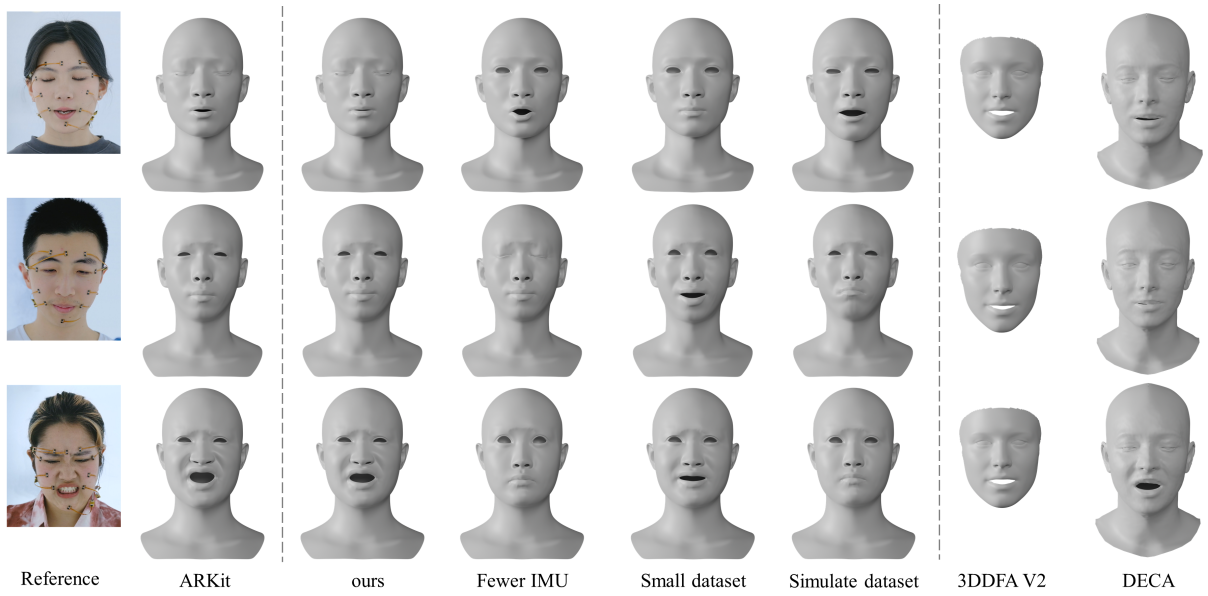

Figure 6: Qualitative comparison and ablation study. The first column displays the reference image. The second column illustrates the record result by ARKit (Apple 2023). The third column shows the reconstruction results of our pipeline. Columns 4, 5, and 6 illustrate the result of our ablation experiment *Fewer IMU*, *Small Dataset* and *Simulate Dataset* respectively. Columns 7 and 8 illustrate the results of 3DDFA V2 (Guo et al. 2020) and DECA (Feng et al. 2021) respectively.

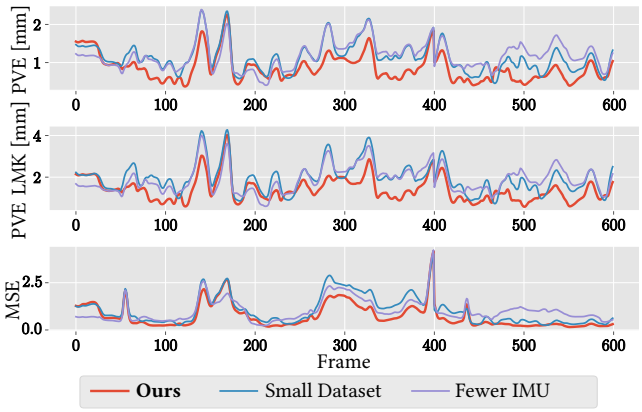

Figure 7: Quantitative result of our method on test data. We plot the PVE, PVE_LMK and MSE calculated per frame with ARKit as ground truth on a sequence.

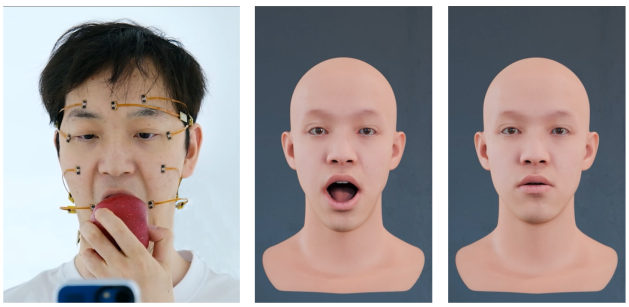

Figure 8: Facial capture during occlusions.

Our tailored micro-IMUs, strategically attached to facial regions aligned with facial anatomy, enable us to capture facial movements from nuanced to dramatic. We have collected the first-ever IMU-ARKit dataset with synchronized IMU and visual signals of diverse expressions from various performers. We further developed a framework for reliable motion inference. Both the dataset and the code are released to the community for comprehensive evaluations.

We believe our IMU-based facial motion capture is an innovative and potentially advantageous solution. In full-body motion capture, due to the exceptional portability and minimal spatial demands of IMUs, we have witnessed a transition from vision-only to vision-IMU hybrid and most recently to IMU-only solutions. Similarly, we believe that IMU-based methods will also become mainstream in the field of facial motion capture. CAPUS illustrates this possibility by achieving results comparable to visual methods while allowing users to move their heads freely. This is particularly advantageous in scenarios where visual signals are unavailable or intentionally avoided, such as industrial facial capture solutions that require subjects to wear helmets, or smartphone-based solutions that require users to always face the camera. As a prototype, CAPUS is far from perfect and still has many issues (comfort, sensor sizes, wiring, etc). Yet, we believe using IMU as a new modality in facial motion capture may stimulate significant future developments in facial animation, capture, and beyond.

Acknowledgements

This work is supported in part by Natural Science Foundation of China (61977047, 61976138), Science and Technology Commission of Shanghai Municipality (21010502400), Shanghai Local College Capacity Building Program (23010503100), Shanghai Clinical Research and Trial Center and National Key R&D Program of China (2022YFF0902301). The authors appreciate the anonymous reviewers for their valuable comments. They also acknowledge support from Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI), MoE Key Lab of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University) and Design Interaction Visual Lab (ShanghaiTech University).

References

Ahmad, N.; Ghazilla, R. A. R.; Khairi, N. M.; and Kasi, V. 2013. Reviews on various inertial measurement unit (IMU) sensor applications. *International Journal of Signal Processing Systems*, 1(2): 256–262.

Apple. 2023. ARKit. https://developer.apple.com/arkit/.

Bachmann, E. R.; McGhee, R. B.; Yun, X.; and Zyda, M. J. 2001. Inertial and magnetic posture tracking for inserting humans into networked virtual environments. In *Proceedings of the ACM symposium on Virtual reality software and technology*, 9–16.

Bao, L.; Lin, X.; Chen, Y.; Zhang, H.; Wang, S.; Zhe, X.; Kang, D.; Huang, H.; Jiang, X.; Wang, J.; et al. 2021. High-fidelity 3D digital human head creation from RGB-D selfies. *ACM Transactions on Graphics (TOG)*, 41(1): 1–21.

Bianchi, L.; Angelini, D.; Orani, G.; and Lacquaniti, F. 1998. Kinematic coordination in human gait: relation to mechanical energy cost. *Journal of neurophysiology*, 79(4): 2155–2170.

Cao, C.; Weng, Y.; Zhou, S.; Tong, Y.; and Zhou, K. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3): 413–425.

Cao, X.; Wei, Y.; Wen, F.; and Sun, J. 2014. Face alignment by explicit shape regression. *International journal of computer vision*, 107: 177–190.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.

Cootes, T. F.; Edwards, G. J.; and Taylor, C. J. 2001. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6): 681–685.

Cootes, T. F.; Taylor, C. J.; Cooper, D. H.; and Graham, J. 1995. Active shape models-their training and application. *Computer vision and image understanding*, 61(1): 38–59.

de Aguiar, E.; Theobalt, C.; Magnor, M.; Theisel, H.; and Seidel, H.-P. 2004. M/sup 3: marker-free model reconstruction and motion tracking from 3D voxel data. In *12th Pacific Conference on Computer Graphics and Applications, 2004. PG 2004. Proceedings.*, 101–110. IEEE.

Del Rosario, M. B.; Khamis, H.; Ngo, P.; Lovell, N. H.; and Redmond, S. J. 2018. Computationally efficient adaptive error-state Kalman filter for attitude estimation. *IEEE Sensors Journal*, 18(22): 9332–9342.

Du, Y.; Kips, R.; Pumarola, A.; Starke, S.; Thabet, A.; and Sanakoyeu, A. 2023. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 481–490.

Egger, B.; Smith, W. A.; Tewari, A.; Wuhrer, S.; Zollhoefer, M.; Beeler, T.; Bernard, F.; Bolkart, T.; Kortylewski, A.; Romdhani, S.; et al. 2020. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG)*, 39(5): 1–38.

Feng, Y.; Feng, H.; Black, M. J.; and Bolkart, T. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4): 1–13.

Ferrigno, G.; Borghese, N.; and Pedotti, A. 1990. Pattern recognition in 3D automatic human motion analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 45(4): 227–246.

Foxlin, E. 1996. Inertial head-tracker sensor fusion by a complementary separate-bias Kalman filter. In *Proceedings of the IEEE 1996 Virtual Reality Annual International Symposium*, 185–194. IEEE.

Guo, J.; Zhu, X.; Yang, Y.; Yang, F.; Lei, Z.; and Li, S. Z. 2020. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, 152–168. Springer.

Guo, K.; Zhao, C.; and Wang, J. 2024. A fast mask synthesis method for face recognition. *Visual Intelligence*, 2(1): 25.

Guo, Y.; Xu, G.; and Tsuji, S. 1994. Understanding human motion patterns. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5)*, volume 2, 325–329. IEEE.

Huang, Y.; Kaufmann, M.; Aksan, E.; Black, M. J.; Hilliges, O.; and Pons-Moll, G. 2018. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6): 1–15.

Li, J.; Liu, K.; and Wu, J. 2023. Ego-Body Pose Estimation via Ego-Head Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17142–17151.

Li, R.; Bladin, K.; Zhao, Y.; Chinara, C.; Ingraham, O.; Xiang, P.; Ren, X.; Prasad, P.; Kishore, B.; Xing, J.; et al. 2020. Learning formation of physically-based face attributes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3410–3419.

Li, Z.; Lv, X.; Yu, W.; Liu, Q.; Lin, J.; and Zhang, S. 2024. Face shape transfer via semantic warping. *Visual Intelligence*, 2(1): 26.

Liu, H.; Wei, X.; Chai, J.; Ha, I.; and Rhee, T. 2011. Realtime human motion control with a small number of inertial sensors. In *Symposium on interactive 3D graphics and games*, 133–140.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics*, 34(6).

Makaussov, O.; Krassavin, M.; Zhabinets, M.; and Fazli, S. 2020. A low-cost, IMU-based real-time on device gesture recognition glove. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 3346–3351. IEEE.

Michoud, B.; Guillou, E.; Briceno, H.; and Bouakaz, S. 2007. Real-time marker-free motion capture from multiple cameras. In *2007 IEEE 11th International Conference on Computer Vision*, 1–7. IEEE.

Mummadi, C. K.; Philips Peter Leo, F.; Deep Verma, K.; Kasireddy, S.; Scholl, P. M.; Kempfle, J.; and Van Laerhoven, K. 2018. Real-time and embedded detection of hand gestures with an IMU-based glove. In *Informatics*, volume 5, 28. MDPI.

Noitom. 2015. Noitom Motion Capture Systems. https://www.noitom.com/.

Qammaz, A.; and Argyros, A. A. 2023. A Unified Approach for Occlusion Tolerant 3D Facial Pose Capture and Gaze Estimation using MocapNETs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3178–3188.

QST Inc. 2012. QST Corporation Limited. https://www.qstcorp.com/.

Riaz, Q.; Tao, G.; Krüger, B.; and Weber, A. 2015. Motion reconstruction using very few accelerometers and ground contacts. *Graphical Models*, 79: 23–38.

Roetenberg, D.; Luinge, H. J.; Baten, C. T.; and Veltink, P. H. 2005. Compensation of magnetic disturbances improves inertial and magnetic sensing of human body segment orientation. *IEEE Transactions on neural systems and rehabilitation engineering*, 13(3): 395–405.

Schepers, M.; Giuberti, M.; Bellusci, G.; et al. 2018. Xsens MVN: Consistent tracking of human motion using inertial sensing. *Xsens Technol*, 1(8): 1–8.

Shimada, S.; Golyanik, V.; Pérez, P.; and Theobalt, C. 2023. Decaf: Monocular Deformation Capture for Face and Hand Interactions. arXiv:2309.16670.

Slyper, R.; and Hodgins, J. K. 2008. Action capture with accelerometers. In *Proceedings of the 2008 ACM SIGGRAPH/Eurographics symposium on computer animation*, 193–199.

Smith, W. A.; Seck, A.; Dee, H.; Tiddeman, B.; Tenenbaum, J. B.; and Egger, B. 2020. A morphable face albedo model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5011–5020.

Stan, S.; Haque, K. I.; and Yumak, Z. 2023. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, 1–11.

Tautges, J.; Zinke, A.; Krüger, B.; Baumann, J.; Weber, A.; Helten, T.; Müller, M.; Seidel, H.-P.; and Eberhardt, B. 2011. Motion reconstruction using sparse accelerometer data. *ACM Transactions on Graphics (ToG)*, 30(3): 1–12.

Uldis, Z. 2017. *Anatomy of Facial Expressions*. Anatomy Next, Inc.

UnrealEngine. 2023. Live Link Face. https://apps.apple.com/us/app/live-link-face/id1495370836.

Vitali, R. V.; McGinnis, R. S.; and Perkins, N. C. 2020. Robust error-state Kalman filter for estimating IMU orientation. *IEEE Sensors Journal*, 21(3): 3561–3569.

Vlasic, D.; Adelsberger, R.; Vannucci, G.; Barnwell, J.; Gross, M.; Matusik, W.; and Popović, J. 2007. Practical motion capture in everyday surroundings. *ACM transactions on graphics (TOG)*, 26(3): 35–es.

Vlasic, D.; Baran, I.; Matusik, W.; and Popović, J. 2008. Articulated mesh animation from multi-view silhouettes. In *Acm Siggraph 2008 papers*, 1–9.

Von Marcard, T.; Rosenhahn, B.; Black, M. J.; and Pons-Moll, G. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, volume 36, 349–360. Wiley Online Library.

Weise, T.; Bouaziz, S.; Li, H.; and Pauly, M. 2011. Realtime performance-based facial animation. *ACM transactions on graphics (TOG)*, 30(4): 1–10.

Yi, X.; Zhou, Y.; and Xu, F. 2021. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4): 1–13.

Yuan, Q.; and Chen, I.-M. 2014. Localization and velocity tracking of human via 3 IMU sensors. *Sensors and Actuators A: Physical*, 212: 25–33.

Zhang, L.; Qiu, Q.; Lin, H.; Zhang, Q.; Shi, C.; Yang, W.; Shi, Y.; Yang, S.; Xu, L.; and Yu, J. 2023. DreamFace: Progressive Generation of Animatable 3D Faces under Text Guidance. arXiv:2304.03117.

Zhang, L.; Zeng, C.; Zhang, Q.; Lin, H.; Cao, R.; Yang, W.; Xu, L.; and Yu, J. 2022. Video-driven Neural Physically-based Facial Asset for Production. arXiv:2202.05592.

Zhao, Q.; Long, P.; Zhang, Q.; Qin, D.; Liang, H.; Zhang, L.; Zhang, Y.; Yu, J.; and Xu, L. 2024. Media2face: Co-speech facial animation generation with multi-modality guidance. In *ACM SIGGRAPH 2024 Conference Papers*, 1–13.

Zhou, Y.; Zhang, W.; Tang, X.; and Shum, H. 2005. A bayesian mixture model for multi-view face alignment. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, 741–746. IEEE.

Zhu, X.; Liu, X.; Lei, Z.; and Li, S. Z. 2017. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 41(1): 78–92.