AudioEar: Single-View Ear Reconstruction for Personalized Spatial Audio

Xiaoyang Huang[1], Yanjun Wang[1], Yang Liu[2], Bingbing Ni[1]*, Wenjun Zhang[1], Jinxian Liu[1], Teng Li[3]*

[1]Shanghai Jiao Tong University, Shanghai 200240, China,
[2]FocusMedia,
[3]Anhui University
{huangxiaoyang, nibingbing}@sjtu.edu.cn

Abstract

Spatial audio, which focuses on immersive 3D sound rendering, is widely applied in the acoustic industry. One of the key problems of current spatial audio rendering methods is the lack of personalization based on different anatomies of individuals, which is essential to produce accurate sound source positions. In this work, we address this problem from an interdisciplinary perspective. The rendering of spatial audio is strongly correlated with the 3D shape of human bodies, particularly ears. To this end, we propose to achieve personalized spatial audio by reconstructing 3D human ears with single-view images. First, to benchmark the ear reconstruction task, we introduce AudioEar3D, a high-quality 3D ear dataset consisting of 112 point cloud ear scans with RGB images. To self-supervisedly train a reconstruction model, we further collect a 2D ear dataset composed of 2,000 images, each one with manual annotation of occlusion and 55 landmarks, named AudioEar2D. To our knowledge, both datasets have the largest scale and best quality of their kinds for public use. Further, we propose AudioEarM, a reconstruction method guided by a depth estimation network that is trained on synthetic data, with two loss functions tailored for ear data. Lastly, to fill the gap between the vision and acoustics community, we develop a pipeline to integrate the reconstructed ear mesh with an off-the-shelf 3D human body and simulate a personalized Head-Related Transfer Function (HRTF), which is the core of spatial audio rendering. Code and data are publicly available in https://github.com/seanywang0408/AudioEar.

Introduction

Spatial audio is widely applied in virtual reality, gaming, and movie production (Begault and Trejo), for distinguishing sound source positions and generating immersive 3D sound. Without it, people would lose the spatial sense of sound. The rendering of spatial audio depends on the Head-Related Transfer Function (HRTF) (Elliott, Jung, and Cheer; Blauert, Allen, and Press). HRTF is adopted in mid-to-high-end audio equipment, such as stereos, headphones, Hi-Fi, and so on. HRTF varies from one person to another since it depends on the 3D structure of human bodies including ears, head, and torso. While a personalized HRTF could be approximated by acoustic simulation (Pierce) given a complete 3D human body structure, it is laborious to model human bodies using 3D scanners. Single-view reconstruction is an alternative solution. It is well studied to reconstruct personalized human heads and torsos within single-view images. However, without the modeling of ear structure, which is the central organ in the human hearing system, the simulated HRTF is still biased. In this work, we propose to obtain personalized HRTF by reconstructing the 3D mesh (Huang et al.) of human ears within a single-view image. [1]

To benchmark the task of ear reconstruction, we collect a high-quality 3D ear dataset with an advanced structured-light 3D scanner, named AudioEar3D. Compared to prior evaluation protocol that uses 2D ear landmark re-projection error as the metric, a 3D benchmark is more precise. AudioEar3D includes 112 ear point cloud scans with RGB images from 56 individuals, each scan with $100,000$ to $250,000$ points, which capture high-resolution shape characteristics of ears. To the best of our knowledge, this is the largest and most accurate 3D ear dataset that is publicly available, as analyzed in Section . We believe that except for spatial audio, this dataset could also contribute to other applications, including digital humans (Kappel et al.), morphology (Krishan, Kanchan, and Thakur) and medical surgery research (Varman et al.).

Prior reconstruction methods for other parts of human body train on large-scale 2D datasets self-supervisedly (Zhang et al.; Chen et al.). However, existing 2D ear datasets either lack semantic annotations or suffer from a small scale. This situation greatly limits the extension of CV studies on human ears. To this end, we build a large-scale 2D ear images dataset, including $2,000$ high-resolution ear images. They are selected from the FFHQ human face dataset, each of which accompanies manual annotations of occlusion and 55 ear landmarks. The landmarks annotations greatly enrich the semantics of the ear images, enabling their usage in extensive applications.

Based on the self-supervised reconstruction pipeline, we proposed AudioEarM, a depth-guided reconstruction method tailored for ears. Due to the lack of public ear tex-

*Corresponding Author: Teng Li, Bingbing Ni

[1]Strictly speaking, the ear structure include outer ear, middle ear, and inner ear. Our work only focuses on the visible structure (pinna) in the outer ear. Yet we still use the more understandable name *ear* in the paper, following prior works (Jin et al.).

ture, which hinders self-supervised training, we first build an ear texture space from DECA (Feng et al.) for the UHM Ear (Ploumpis et al.), the ear shape model we use in our work. We excavate UV coordinate correspondence between UHM Ear vertices and the texture of DECA via K-Nearest-Neighbors searching and distance-based weighting. Besides, motivated by the characteristics of ear data, we design two loss functions: (i) contour loss which better accommodates the annotation error than landmark loss; (ii) similarity loss which encourages the network to predict distinct shapes for different samples. Moreover, we introduce synthetic data to improve reconstruction quality. Instead of directly combining synthetic data with real data and using 3D supervision to train the model, we use the synthetic data to train a monocular depth model to extract depth feature from ear images. Then we inject the depth information into the main network to guide the whole reconstruction process.

Lastly, to fill the gap between ear reconstruction and acoustic applications, we develop a pipeline to integrate the reconstructed ear into a 3D mesh human torso for acoustic simulation. We propose *approximated Delaunay triangulation* on the 3D vertices to ensure the simulation validity of the integrated body mesh. We obtain the personalized HRTF with the reconstructed ear via simulation, and further demonstrate the effectiveness of the reconstruction model over the baseline by comparison on HRTF.

Related Work

Spatial Audio

Spatial audio, or spatial hearing, is a long-standing subject in acoustics (Culling and Akeroyd; Vorländer). It focuses on how humans localize the position of acoustic sources and how to reproduce the spatial sense with a sound system. Humans locate the sound sources by *biaural* and *monaural* cues. Biaural cues include interaural time difference (ITD) and interaural level difference (ILD), which means that the arriving sound signals are filtered by the diffraction and reflection of the human body, including ear (pinna), head, and torso, leading to a difference of time, and intensity of arrival. Monaural cue is the spectral distortion of sounds. Further, the 3D shape of the human body varies between individuals, resulting in different binaural and monaural cues between different people. To model different spatial hearing characteristics between people, the Head-Related Transfer Function (HRTF) (Li and Peissig; Xie, Zhong, and He) is proposed and widely used in spatial sound rendering. HRTF describes the Sound Pressure Level (SPL, dB) of a sound source in each direction for a specific individual. It is used to convert an arbitrary sound to a specific position as if the sound is originated from there. However, obtaining an accurate personalized HRTF is laborious since the measurement requires expensive equipment and special acoustic laboratories. Current implementations tend to deploy an average HRTF or choose one from an HRTF database (Guo et al.). Besides physical measurement, HRTF could also be numerically simulated (Conrad; Jensen) based on Boundary Element Method (BEM). Meshram et al. proposed to simulate a personalized HRTF by reconstructing the 3D human body using multi-view stereo (MVS). However, the reconstructed body is heavily blurred due to the imperfection of the reconstruction method, yielding large HRTF error.

3D Ear Dataset	Scale	with Image	Quality	Accessibility
UND-J2	$1,800$	✓	⋆	✓
York3DEar	500	✗	⋆	✓
SYMARE-1	20	✗	⋆⋆⋆	✓
SYMARE-2	102	✗	⋆⋆⋆	✗
Ploumpis et al.	234	✗	⋆⋆⋆	✗
AudioEar3D	112	✓	⋆⋆⋆⋆	✓

Table 1: AudioEar3D and its counterparts. Quality is assessed based on the precision of acquisition equipments.

2D Ear Dataset	Scale	Source	Annotations	
			Lmks.	Occ.
UND-E	464	Limited	✗	✗
AMI	700	Limited	✗	✗
IIT Delhi Ear	754	Limited	✗	✗
WPUTEDB	$3,348$	Limited	✗	✗
UBEAR	$4,410$	Limited	✗	✗
IBug-B	$2,058$	In-the-wild	✗	✗
AWE	$9,500$	In-the-wild	✗	✗
EarVN	$28,412$	In-the-wild	✗	✗
IBug-A	605	In-the-wild	✓	✗
AudioEar2D	$2,000$	In-the-wild	✓	✓

Table 2: Comparisons of AudioEar2D and its counterparts. *Lmks.* denotes landmark annotation. *Occ.* means annotations of whether the ear is partially occluded by hair or earrings.

Ear Datasets in 3D and 2D

Previous CV studies on human ears are mostly confined to biometric application (O'Sullivan and Zafeiriou), and do not prevail as other parts of human body, such as faces, hands and skeletons. This leads to the lack of a rich ear dataset in 3D and 2D, as stated in Table 1, 2.

Existing 3D ear datasets suffer from either non-accessibility, small scale, or low quality. Yan and Bowyer collect $1,800$ ear depth maps in a resolution of 640×480 using a depth sensor. Since the depth maps are single-view, they do not represent complete ear shapes. Dai, Pears, and Smith publish York3DEar, which is composed of 500 deformed 3D ear meshes. These meshes are not collected with instruments, but are estimated by a data-augmenting technique based on a 2D ear dataset, which introduces unforeseeable error. Jin et al. propose the SYMARE ear database consisting of the measured HRIR and the upper torso, head and ear mesh model collected by magnetic resonance imaging (MRI) from 61 individuals. However, only 10 (SYMARE-1) individuals' data are accessible, while the rest (SYMARE-2) are kept private. Ploumpis et al. collect 121 ear models from 64 adults and 133 ears models from children via CT scans. However, these data are not publicly available. Besides, most of the above 3D ear datasets lack the corresponding ear images (except for UND-J2), making it unfeasible to perform single-view reconstruction tasks upon them. Moreover, most of these data are collected by MRI, whose measurement error is generally about 1 millimeter (Nowogrodzki). This measurement error is not negligible for such an elaborately-structured organ, thereby introducing additional error in the simulation of HRTF.

Existing 2D ear datasets include UND-E (Chang et al.), AMI (Gonzalez, Alvarez, and Mazorra), IIT Delhi Ear (Kumar and Wu), WPUTEDB (Frejlichowski and Tyszkiewicz), UBEAR (Raposo et al.), IBug(-A/B) (Zhou and Zaferiou), AWE (Emeršič et al.) and EarVN (Hoang), most of which are for human identity recognition. The summarization of these datasets is shown in Table 2.

3D Morphable Models Reconstruction

A 3D Morphable Model (3DMM) is a parametric model that encodes the shape and texture of 3D meshes into a latent space. The most studied structure in 3DMM reconstruction are human faces (Ploumpis et al.; Ploumpis et al.), hands (Chen et al.; Wang et al.; Zhang et al.) and bodies (Choi, Moon, and Lee; Corona et al.). To achieve plausible performance, the 3DMM reconstruction algorithms require a large set of 2D images for self-supervised training, usually accompanied by semantic annotations, such as landmarks or poses. They use a feature extractor and a multi-layer perceptron (MLP) to regress the latent code of shape and texture. The latent codes are fed into the differentiable 3DMM models to obtain colored 3D meshes. Then the meshes are projected to images with a differentiable renderer (Kato, Ushiku, and Harada; Huang et al.). The photometric loss on images and the landmark re-projection loss are minimized to train the network. The popular face reconstruction algorithm DECA (Feng et al.) proposes to conduct robust detail reconstruction by regressing a UV displacement map. Sun, Pears, and Dai propose to reconstruct 3D ear meshes with the YEM model (Dai, Pears, and Smith). They derived a color model from images and minimized the photometric and landmark error to regress shape latent codes.

Method

Data Collection

We collect two ear datasets for personalized spatial audio. One is AudioEar3D, a 3D ear scan dataset, for benchmarking the ear reconstruction task. The other is AudioEar2D, for the training of ear reconstruction models.

AudioEar3D Among various 3D scanning equipments, we choose a structured-light 3D scanner to collect our 3D ear data, for the following reasons: (i) Compared to previous 3D ear scanning methods, such as MRI and CT, a structured-light scanner is much more convenient and accessible (it can be integrated into a small portable device), which makes it easy to cover a larger population; (ii) The operation and post-processing are simpler; (iii) The resolution of structured-light scanners is higher compared to others, under the same price. Specifically, we use MantisVision® F6-SR [2], a portable high-resolution 3d scanner, which is designed to be able to scan detailed organ models in medical applications. Its plane resolution and depth resolution are 0.1 mm and 0.4 mm respectively (compared to a resolution of 1 mm in most MRI scanners), with a frame rate of 8 FPS.

To alleviate the self-occlusion issue brought by the complex ear structure and obtain a clear and complete 3D ear scan, we scan each individual with the above device for about 90 seconds from different directions. We remove the other parts of the body and only keep the ear part. This yields approximately $100,000$ to $250,000$ points for each ear scan, which captures high-resolution shape characteristics of ears. Besides, we take frontal pictures of the ears from a distance similar to the scanner. At present we have obtained data of 112 ears from 56 subjects. We plan to further extend the scale to cover a larger population. We show several samples in Figure 1. The out-coming dataset is completely *anonymized* for privacy issues.

AudioEar2D We obtain our 2D ear dataset from a high-quality public human face dataset, Flickr-Faces-HQ Dataset (FFHQ) (Karras, Laine, and Aila), which consists of $70,000$ face images in the resolution of $1,024 \times 1,024$ that are various in terms of pose, age, and ethnicity. To find images that contain clear ears, we first train an ear detection model on the IBug Collection-B dataset (Zhou and Zaferiou), which contains the bounding box of ears, from a pre-trained Yolov4 (Bochkovskiy, Wang, and Liao). Then we adopt the trained detection model on the FFHQ dataset to roughly sift out high-quality ear images according to detection confidence and bounding box area. Besides, we leverage WHENet (Zhou and Gregson), a pre-trained head pose estimation model, to further remove those images that have bad view angles. We cut the remaining images into squares based on the bounding box. We preserve the original resolution of the cut ear image to avoid information loss induced by interpolation resizing. Thanks to the bounding box area filtering, the image resolution is mostly above 300×300, as illustrated in Figure 1 bottom (a). Next, we manually filter the remaining images and annotate the 55 landmarks, following a landmark protocol in IBug Ear (Zhou and Zaferiou). Moreover, we annotate whether the ear is clean or is occluded by hair or by earrings, which could act as noise in some applications. The distribution of occlusion is shown in Figure 1 bottom (b). We obtain $2,000$ high-resolution ear images, each with 55 annotated landmarks in the end.

AudioEarM: Depth-Guided Ear Reconstruction

We develop our method upon a common pipeline of parametric shape reconstruction, which regresses the latent code of the 3D shape, texture, camera, and lighting via self-supervised training. We first adapt this pipeline to ears with several effective modifications, which are motivated by the specific characteristic of ears. Besides, to further boost the performance, we propose to leverage synthetic ear images to train a depth estimation model and fuse the depth information into the reconstruction model via multi-scale feature alignment. We name our method AudioEarM.

Texture Acquisition We use the UHM Ear (Ploumpis et al.), the best-quality ear 3DMM to transform the shape

[2]Product introduction: https://mantis-vision.com/handheld-3d-scanners/

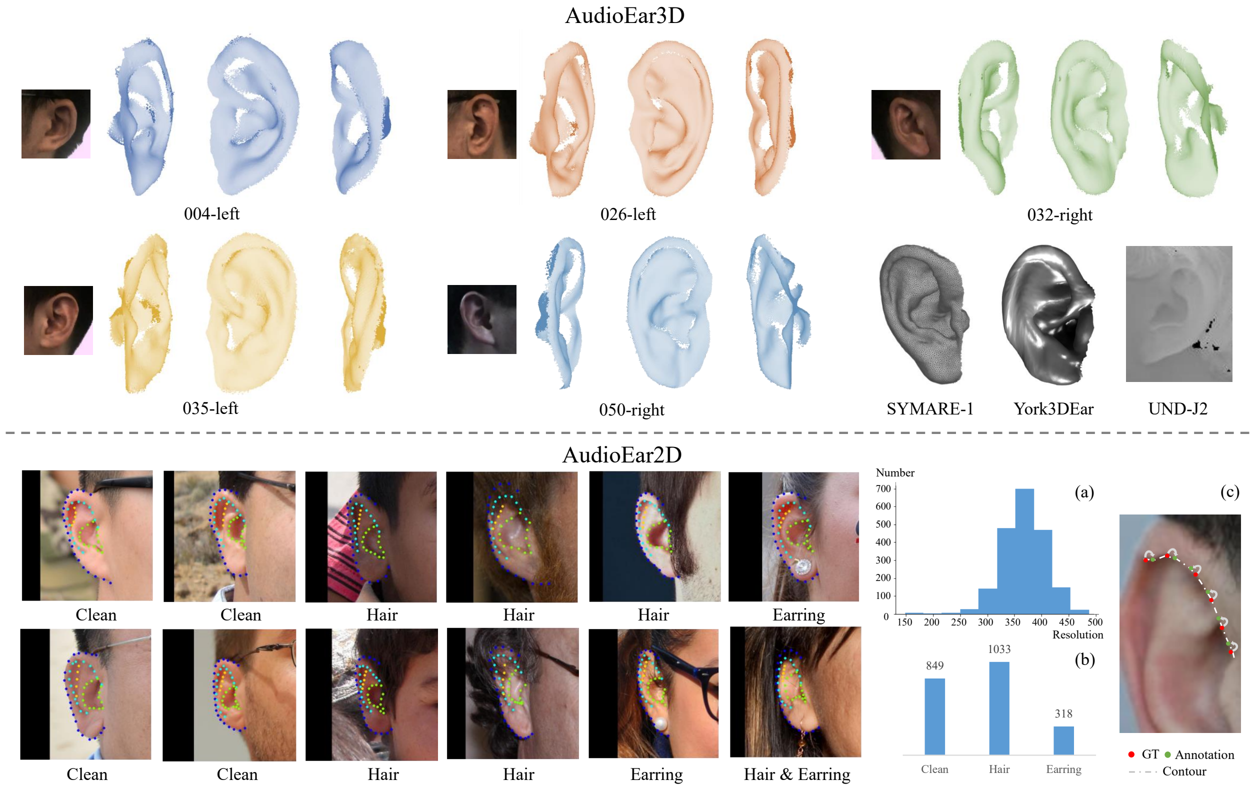

Figure 1: Visualization of the two collected datasets. AudioEar3D (top): We show five samples with their three-view rendered point clouds and RGB images. Compared to existing 3D ear dataset, AudioEar3D is high-quality, large-scale, and also publicly available. AudioEar2D (bottom): We illustrated several samples with their occlusion annotations and the 55 landmarks. The four colors of the landmarks indicate four main contours of the ears. We flip the left ears to the right for better visualization.

latent code into a 3D mesh. The UHM Ear reduces the N vertices (2800) of an ear mesh to a lower dimension (236) using PCA. Let $\overline{\mathbf{S}} \in \mathbb{R}^{3N}$ denotes the mean shape, $\vec{v} \in \mathbb{R}^{236}$ and $\mathcal{U} \in \mathbb{R}^{3N\times 236}$ denote the eigenvalues and eigenvectors of the 3DMM. The shape variations are modeled by eigenvalue-weighted linear blendshapes: $B_{\mathcal{U}}(\vec{\beta};\mathcal{U}) = \sum_{n=1}^{236} v_n\beta_n\mathcal{U}_n$, where $\vec{\beta} \in \mathbb{R}^{236}$ is the predicted shape latent code. The final ear mesh is obtained by:

$$M(\vec{\beta}) = \sum_{n=1}^{236} v_n\beta_n\mathcal{U}_n + \overline{\mathbf{S}} \tag{1}$$

The UHM Ear does not provide texture, which is indispensable to minimize the photometric loss in a self-supervised reconstruction pipeline. We extract the ear texture space from DECA (Feng et al.) to enable colored rendering. In DECA, given a texture latent code $\vec{\theta} \in \mathbb{R}^{|\vec{\theta}|}$, a texture map $T(\boldsymbol{\theta}) \in \mathbb{R}^{h\times w\times 3}$ ($h \times w$ is the resolution) could be calculated. Each vertex in the DECA mesh corresponds to a UV coordinate p_{deca} in the texture map. However, the vertices of the UHM ear do not have correspondence with DECA. To this end, we first visually align the mean UHM ear with the ear in DECA as close as possible. For each vertex $v_{uhm} \in \mathbb{R}^3$ in UHM model, we then search its K-Nearest-Neighbors $v^i_{deca} \in \mathbb{R}^3, i = 1, 2 \ldots, k$ in DECA vertices (k is 3 in our implementation). We assign the KNN-distance-weighted UV coordinates to the UHM vertices based on Euclidean distance between v_{uhm} and v^i_{deca}. Formally, the UV coordinates of v_{uhm} are computed by:

$$p_{uhm} = \sum_{i=1}^{k} \frac{D(v_{uhm}, v^i_{deca})}{\sum_{j=1}^{k} D(v_{uhm}, v^j_{deca})} p^i_{deca} \tag{2}$$

Contour Loss. Contour loss is concerned with the four contours formed by the 2D landmarks. (The four contours are indicated by four different landmark colors in AudioEar2D in Figure 1.) As illustrated in Figure 1 bottom (c), the landmark annotation have bias *along* the contour, while the contour is well represented by the annotations. The reason is that the contour is more visually salient for human eyes. Hence measuring the error of contour prediction in training is superior to landmark error. Practically, we connect the landmarks one by one to form four polylines and uniformly re-sample dense points on these polylines. We measure the chamfer distance of the re-sampled points $\mathcal{P}$ between the ground truth and prediction as contour loss:

$$L_{contour} = \frac{1}{4}\sum_{i=1}^{4} ChamferDistance(\mathcal{P}^i_{gt}, \mathcal{P}^i_{pred}) \tag{3}$$

where the superscript i is the index of the four contours.

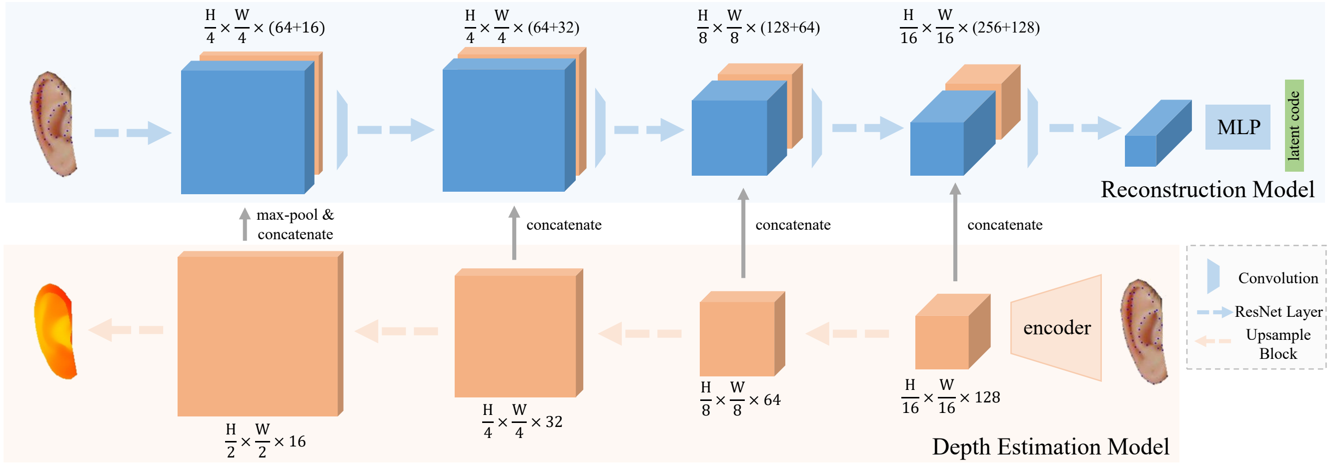

Figure 2: The architecture of our depth-guided reconstruction model, AudioEarM. A depth estimation model, which is trained on a synthetic dataset, guides the main reconstruction network with multi-scale feature alignment.

Similarity Loss. In our early experiments, we find that the predicted shape latent codes are too similar across all instances. This similarity is undesirable since the shape should be distinct across different individuals. We encourage the network to predict distinct shapes by measuring the mean cosine similarity between the shape latent codes in a batch and penalizing high similarity:

$$L_{sim} = \frac{1}{bs \times (bs-1)} \sum_{i,j=0,i\neq j}^{bs-1} \frac{\beta_i \beta_j}{|\beta_i||\beta_j|} \tag{4}$$

where bs denotes the batch size and β_i, β_j denote the ith, jth sample in a batch.

Besides, we prevent extreme rotation angles and scale parameters in camera with $L1$-penalization, since the angle and scale of ears in images are generally within a certain scope. We combine these three loss functions with the commonly used losses: landmark, photometric loss, mesh smooth loss and latent code regularization. The overall loss is the weighted sum of the losses above.

Depth-Guided Reconstruction Zhang et al. demonstrates that introducing synthetic data with 3D supervision into trainset could improve the quality of single-view hand reconstruction. Yet we empirically find that directly migrating this approach to our task brings negative effect (see ablation study). Alternatively, we leverage the power of pretraining (Yang et al.), that is to use synthetic data to pretrain a monocular depth estimation (MDE) model. Then we leverage the depth feature extracted by the MDE model to facilitate the whole reconstruction process, as shown in Figure 2.

To generate a synthetic ear dataset with 3D supervision, we randomly sample shape and texture latent codes from a Gaussian distribution, and render corresponding RGB images and depth maps. We generate $10,000$ samples in this way. Then we train an MDE model on this synthetic dataset, using the depth map as direct supervision. The MDE model (Laina et al.) consists of an encoder and a decoder (Figure 2). The ResNet-34 (He et al.) encoder extracts high-level feature from images, while the decoder predicts a depth map from the feature. The extracted feature is upsampled by the decoder layer by layer until the original resolution is restored. We aim to leverage the depth information obtained by decoder to guide the reconstruction process. To this end, we concatenate the multi-scale upsampled feature with the feature before each layer in the main network. We insert an additional convolution layer each time after concatenation to transform the increased channel numbers into original ones, so as to match the channel of the next layer. Since the last-layer resolution of the MDE model is twice as big as the first-layer resolution in the main network, we conduct max-pooling on the last-layer MDE feature before concatenation. In this way, we explicitly fuse the depth information into the main network to guide the reconstruction process. Ablation study shows that this method brings notable improvement.

Acoustic Simulation

Spatial audio rendering depends on the HRTF, which is defined as the Sound Pressure Level (SPL) measured at the eardrum (or the ear canal entrance in practice). We simulate a personalized HRTF with COMSOL$^{\text{TM}}$, a Multiphysics software that uses the Boundary Element Method (BEM) for simulation. However, the simulation requires a complete human upper body, not just an ear. To this end, we develop a pipeline to integrate our reconstructed ears into an off-the-shelf human body 3D model (Braren and Fels; Braren and Fels). The pipeline is illustrated in Figure 3.

We manually remove the original ears of the given 3D body and place the reconstructed ear to the consequent hole. To combine the ear and body mesh together, we propose a method named *approximated Delaunay triangulation*. Delaunay triangulation is a triangulation method for 2D point sets, which maximize the minimal angle of all the angles of the triangles in the triangulation process, such that sliver

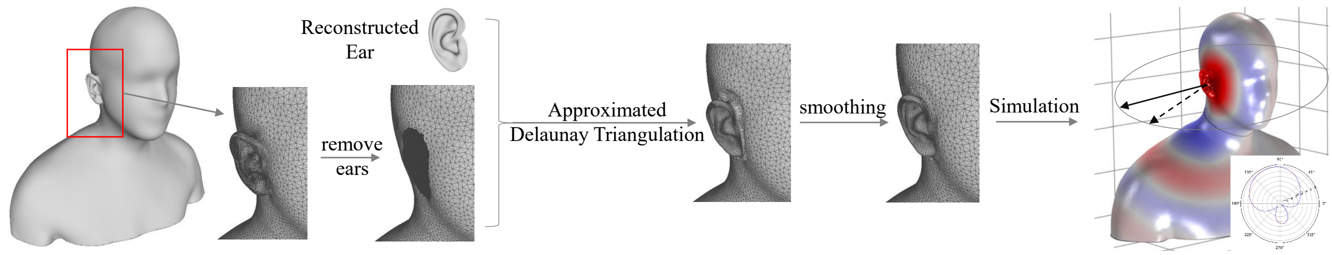

Figure 3: The proposed integration and simulation pipeline. We conduct *approximated Delaunay triangulation* to combine the reconstructed ear with the body and simulate the personalized HRTF in all azimuth angles (361 angles) on the horizontal plane.

triangles are avoided. Sliver triangles are inappropriate for BEM simulation. Since both the edges of the ear and body hole are not on a 2D plane (but close to a plane), we project the edges to a plane which is found by the averaged normals of the mesh faces around the edges. In this way, we can conduct *approximated Delaunay triangulation* on the vertices of the two edges and fix the crack between ears and heads. Since the geometry of the edges is not greatly changed by the projections, the sliver triangles could still be avoided as much as possible. We leverage the Detri2 program [3] for Delaunay triangulation. At last, we smooth the generated mesh and use it for simulation. We simulate the HRTF in all azimuth angles (361 angles in total) on the horizontal plane for three frequencies (f = 1033.6 Hz, 2067.5 Hz, and 3962.1 Hz), following a general protocol. Note that the simulation pipeline involves manual efforts and is not part of the evaluation protocol of ear reconstruction, but for application use. We evaluate reconstruction performance on AudioEar3D.

Experiments

Ear Reconstruction Benchmark

Evaluation We leverage the full AudioEar3D dataset to evaluate the performance of ear reconstruction models. We compute the distance from each point in the ground-truth ear scan to the predicted ear mesh. (The distance from a point to a mesh is defined as the distance from the point to the closest triangular face in the mesh.) We average the distance across all points in the scan as the evaluation metric, named scan2mesh (S2M). However, since the scale and position are not aligned between the scans and the predicted meshes, we first register the mesh with the scan in a gradient-based method. We iteratively optimize scale and position parameters for registration. To reduce the evaluation time and avoid convergence in a local minimum, we design a three-stage registration scheme. First, we minimize four manually chosen key points to obtain a coarse registration. Next, we randomly sample points on the predicted mesh surfaces and minimize the chamfer distance between the sampled points and the scans. Last, we directly minimize the S2M distance between the scan and mesh. Since the dense points clouds in AudioEar3D would cause intense computation, we randomly down-sample the scan to $1,000$ points for the S2M and only compute the S2M distance on the whole scan in the last iteration. We use Adam (Kingma and Ba) with an initial learning rate of 0.45 for optimization. Each stage lasts for 166 iterations. The learning rate is multiplied by 0.1 when entering the next stage. After numerous tests, we empirically find that this setting can make most registrations converge to a global minimum. We conduct the registration process on all the 112 samples in AudioEar3D, which takes about half an hour, and average the S2M across all samples as the final evaluation metric. The left ear scans are transformed to the right ones by reflection in the sagittal plane.

[3]http://www.wias-berlin.de/people/si/detri2.html

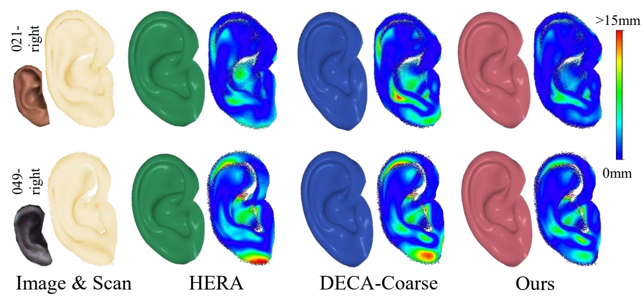

Figure 4: We visualize the S2M error map of our method and two baselines on testset (AudioEar3D). Our method has lower error compared to baselines.

Setting To compare, we send the average ear of UHM into the registration process, as a baseline without any personalization. We also compare our method with a prior ear reconstruction method named HERA (Sun, Pears, and Dai). Since HERA did not publish their code and ear model, we implement it using our ear model. Besides, we would like to compare AudioEarM with the SOTA face reconstruction algorithm, DECA (Feng et al.). Since a detailed UV displacement map in DECA is not available for ears, we only implement the coarse branch in DECA, denoted as DECA-coarse.

Results As listed in Table 3 #6-9, our method surpasses the average ear baseline, HERA and DECA-coarse, with a final S2M distance of 1.28 mm. The average ear yields an S2M of 1.78 mm, while HERA and DECA-coarse yields an S2M of 1.70 and 1.46 mm. We visualize the **S2M error map** of our method and two baselines on testset (AudioEar3D). Our method has lower error compared to baselines.

#	Model	S2M ↓
1	Naive	1.73mm
2	#1 + texture	1.65mm
3	#2 + similarity loss	1.63mm
4	#2 + contour loss	1.52mm
5	#2 + contour & similarity	1.47mm
6	**Ours** (#5 + depth-guided)	**1.28mm**
7	Avg ear	1.78mm
8	HERA	1.70mm
9	DECA-coarse	1.46mm

Table 3: Comparisons of reconstruction methods. {#1 + texture} means texture is added on top of Method#1. Similar meanings for other $+,-$ notations.

Comparisons of Reconstruction Models

We examine the impact of each part in our method (Table 3 #1-6). All the models are trained with the same setting described in Section . We first implement a baseline model (Method#1) in which a common self-supervised reconstruction pipeline is adopted to our data without modifications. A single skin color is assigned to the texture instead of our texture space. Method#1 yields an S2M of 1.73 mm, which has little improvement over the average ear. Then we replace the single color with our texture space (Method#2), which reduces the S2M error to 1.65 mm. On top of it, we add the similarity loss and contour loss into the training process (Method#3-5). Adding the two losses together yields a result of 1.47 mm. The effect of contour loss is more notable than the similarity loss. Finally, we use the depth estimation model to perform depth-guide reconstruction, and obtain an S2M of 1.28 mm, which is our best result. The reason behind the superiority is that predicting a regular depth map is less ill-posed than predicting a shape latent code. Besides, depth could inform the object's geometry. Similar motivation is enlightening in pseudo-lidar approaches for 3D object detection (Qian et al.).

HRTF Simulation

Evaluation. In this part, we conduct the HRTF simulation using our predicted ears. We compare the HRTF of the predicted ears against the ground-truth ears. The HRTF of the ground-truth ears are simulated using the ear meshes registered from the raw ear scans in AudioEar3D. Specifically, we send the mean UHM Ear into the same registration process described in Section , except that the optimized parameters in the second and third stages include the shape latent code of UHM besides the original ones. The registration yields an average S2M of 0.11 mm among all scans, which is fairly low. The low S2M error indicates the optimized meshes are rather close to the true ear shape. We measure the mean SPL error in absolute value across all angles between *ground-truth* and predicted ones. As comparisons, we replace the predicted ears with the mean UHM Ear (denoted as *Avg* in Table 4). The SPL error of the mean UHM Ear represents the spatial audio effect without personalization.

Sample	S2M ↓ (mm)	SPL Error ↓ (dB× 10) f=1kHz	f=2kHz	f=4kHz
033-right	1.49	1.12±0.75	3.59±5.97	4.13±5.44
	0.94	0.33±0.19	0.59±0.89	2.32±2.84
021-left	1.66	0.58±0.27	1.20±2.19	4.25±3.43
	0.88	0.30±0.12	1.30±1.47	2.39±1.85
036-left	2.23	1.23±1.17	3.52±9.16	6.97±11.97
	1.85	1.13±0.89	2.69±4.58	4.75±5.89
008-left	1.23	1.69±1.31	4.50±6.71	5.29±5.40
	0.74	1.35±0.69	3.532±5.97	6.75±6.49
048-left	1.35	0.40±0.30	1.27±1.64	5.00±4.76
	0.94	0.23±0.14	1.18±1.23	3.31±2.88
040-right	1.27	1.00±0.66	2.83±5.65	5.54±10.23
	1.58	1.17±0.81	3.57±6.54	5.99±10.43
029-left	1.33	2.11±0.98	5.95±9.76	8.06±9.47
	1.46	1.12±0.76	3.44±5.98	10.09±10.74

Table 4: Results of HRTF simulation measured in SPL error. The first and second row in each sample denote *Avg* and *Pred* results respectively. We see that better reconstruction results bring more realistic HRTF. The predicted ears in the first 5 samples have lower S2M and also lower SPL error than average, vice versa andfor the last 2 samples.

Larger errors indicate worse spatial audio. Since the simulation experiment is labour-intense and time-consuming, we only show the results of several samples for demonstration.

Results. Table 4 compares the results between *Avg* and *Pred* of each sample. The first and second row in each sample denote *Avg* and *Pred* results respectively. We see that better reconstruction results (lower S2M) bring more realistic HRTF (lower SPL error). For those samples whose reconstructed S2M distances are lower than the mean UHM Ear (the first 5 samples), both the mean value and variation of SPL error are also lower, meaning that they obtain a closer HRTF to the ground truth than the mean UHM Ear. For the last two samples, the S2M of predicted ears is higher and the SPL error is somewhat larger, indicating that the obtained HRTF is highly related to the reconstruction results. It is necessary to develop advanced ear reconstruction algorithms for more realistic spatial sound effects.

Conclusion

This work considers 3D ear reconstruction from single-view images for personalized spatial audio. For this purpose, we collect a 3D ear dataset for benchmarking and a 2D dataset for the training of ear reconstruction models. Besides, we propose a reconstruction method guided by a depth estimation network that is trained on synthetic data, with two loss functions tailored for ear data. Lastly, we develop a pipeline to integrate the reconstructed ear mesh with a human body for acoustic simulation to obtain personalized spatial audio.

Acknowledgements

This work was supported by National Science Foundation of China (U20B2072, 61976137). This work was also partially supported by Grant YG2021ZD18 from Shanghai Jiao Tong University Medical Engineering Cross Research.

References

Begault, D. R.; and Trejo, L. J. 2000. 3-D sound for virtual reality and multimedia. *No. NASA/TM-2000-209606*.

Blauert, J.; Allen, J.; and Press, M. 1997. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press. ISBN 9780262024136.

Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Braren, H.; and Fels, J. 2020a. A High-Resolution Individual 3D Adult Head and Torso Model for HRTF Simulation and Validation: 3D Data. *Medical Acoustics Group, Institute of Technical Acoustics, RWTH Aachen*.

Braren, H.; and Fels, J. 2020b. A High-Resolution Individual 3D Adult Head and Torso Model for HRTF Simulation and Validation: HRTF Measurement. *Medical Acoustics Group, Institute of Technical Acoustics, RWTH Aachen*.

Chang, K.; Bowyer, K. W.; Sarkar, S.; and Victor, B. 2003. Comparison and combination of ear and face images in appearance-based biometrics. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9): 1160–1165.

Chen, W.; Ling, H.; Gao, J.; Smith, E.; Lehtinen, J.; Jacobson, A.; and Fidler, S. 2019. Learning to predict 3d objects with an interpolation-based differentiable renderer. *Advances in Neural Information Processing Systems*, 32: 9609–9619.

Chen, Y.; Tu, Z.; Kang, D.; Bao, L.; Zhang, Y.; Zhe, X.; Chen, R.; and Yuan, J. 2021. Model-based 3D hand reconstruction via self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10451–10460.

Choi, H.; Moon, G.; and Lee, K. M. 2020. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *European Conference on Computer Vision*, 769–787. Springer.

Conrad, Z. 2018. Hats Off to the Boundary Element Method. *IEEE Spectrum Multiphysics Simulation*, 30.

Corona, E.; Pumarola, A.; Alenya, G.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. SMPLicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11875–11885.

Culling, J.; and Akeroyd, M. 2012. Spatial hearing. *Oxford Handbook of Auditory Science Hearing*.

Dai, H.; Pears, N.; and Smith, W. 2018. A data-augmented 3D morphable model of the ear. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 404–408. IEEE.

Elliott, S. J.; Jung, W.; and Cheer, J. 2018. Head tracking extends local active control of broadband sound to higher frequencies. *Scientific reports*, 8(1): 1–7.

Emeršič, Ž.; Štepec, D.; Štruc, V.; Peer, P.; George, A.; Ahmad, A.; Omar, E.; Boult, T. E.; Safdaii, R.; Zhou, Y.; et al. 2017. The unconstrained ear recognition challenge. In *2017 IEEE international joint conference on biometrics (IJCB)*, 715–724. IEEE.

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.

Frejlichowski, D.; and Tyszkiewicz, N. 2010. The west pomeranian university of technology ear database–a tool for testing biometric algorithms. In *International Conference Image Analysis and Recognition*, 227–234. Springer.

Gonzalez, E.; Alvarez, L.; and Mazorra, L. 2022. AMI Ear Database. https://ctim.ulpgc.es/research_works/ami_ear_database/. Accessed: 2022-07-24.

Guo, X.; Bai, Y.; Liu, Q.; Xiong, D.; Wang, Y.; and Du, J. 2020. 3D Head Anthropometry for Head-Related Transfer Function of Chinese Pilots. In *International Conference on Man-Machine-Environment System Engineering*, 171–178. Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hoang, V. T. 2019. EarVN1. 0: A new large-scale ear images dataset in the wild. *Data in brief*, 27.

Huang, X.; Yang, J.; Wang, Y.; Chen, Z.; Li, L.; Li, T.; Ni, B.; and Zhang, W. 2022a. Representation-Agnostic Shape Fields. In *International Conference on Learning Representations*.

Huang, X.; Zhang, Y.; Ni, B.; Li, T.; Chen, K.; and Zhang, W. 2022b. Boosting Point Clouds Rendering via Radiance Mapping. *arXiv preprint arXiv:2210.15107*.

Jensen, M. H. 2009. Improving the Performance of Hearing Aids Using Acoustic Simulations. *COMSOL Conference*.

Jin, C. T.; Guillon, P.; Epain, N.; Zolfaghari, R.; Van Schaik, A.; Tew, A. I.; Hetherington, C.; and Thorpe, J. 2013. Creating the Sydney York morphological and acoustic recordings of ears database. *IEEE Transactions on Multimedia*, 16(1): 37–46.

Kappel, M.; Golyanik, V.; Elgharib, M.; Henningson, J.-O.; Seidel, H.-P.; Castillo, S.; Theobalt, C.; and Magnor, M. 2021. High-Fidelity Neural Human Motion Transfer from Monocular Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1541–1550.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Kato, H.; Ushiku, Y.; and Harada, T. 2018. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3907–3916.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. In *ICLR*.

Krishan, K.; Kanchan, T.; and Thakur, S. 2019. A study of morphological variations of the human ear for its applications in personal identification. *Egyptian Journal of Forensic Sciences*, 9(1): 1–11.

Kumar, A.; and Wu, C. 2012. Automated human identification using ear imaging. *Pattern Recognition*, 45(3): 956–968.

Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; and Navab, N. 2016. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, 239–248. IEEE.

Li, S.; and Peissig, J. 2020. Measurement of head-related transfer functions: A review. *Applied Sciences*, 10(14): 5014.

Meshram, A.; Mehra, R.; Yang, H.; Dunn, E.; Franm, J.-M.; and Manocha, D. 2014. P-HRTF: Efficient personalized HRTF computation for high-fidelity spatial sound. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 53–61. IEEE.

Nowogrodzki, A. 2018. The world's strongest MRI machines are pushing human imaging to new limits. *Nature News Feature*, 563(7729): 24–26.

O'Sullivan, E.; and Zafeiriou, S. 2020. 3D Landmark Localization in Point Clouds for the Human Ear. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 402–406. IEEE.

Pierce, A. D. 2019. *Acoustics: an introduction to its physical principles and applications*. Springer.

Ploumpis, S.; Ververas, E.; O'Sullivan, E.; Moschoglou, S.; Wang, H.; Pears, N.; Smith, W.; Gecer, B.; and Zafeiriou, S. P. 2020. Towards a complete 3D morphable model of the human head. *IEEE transactions on pattern analysis and machine intelligence*.

Ploumpis, S.; Wang, H.; Pears, N.; Smith, W. A.; and Zafeiriou, S. 2019. Combining 3D Morphable Models: A Large scale Face-and-Head Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10934–10943.

Qian, R.; Garg, D.; Wang, Y.; You, Y.; Belongie, S.; Hariharan, B.; Campbell, M.; Weinberger, K. Q.; and Chao, W.-L. 2020. End-to-end pseudo-lidar for image-based 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5881–5890.

Raposo, R.; Hoyle, E.; Peixinho, A.; and Proença, H. 2011. Ubear: A dataset of ear images captured on-the-move in uncontrolled conditions. In *2011 IEEE workshop on computational intelligence in biometrics and identity management (CIBIM)*, 84–90. IEEE.

Sun, H.; Pears, N.; and Dai, H. 2021. A Human Ear Reconstruction Autoencoder. In *2021 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021)*, 136–145.

Varman, R. M.; Schrader, K.; Fort, C.; Daniel, H.; Al Mekdash, M. H.; and Demke, J. 2021. Ability to Identify External Ear Deformities and Normal External Ear Anatomy Based on Year and Specialty of Medical Training. *The American Journal of Cosmetic Surgery*, 0748806821993681.

Vorländer, M. 2020. *Auralization*. Springer.

Wang, J.; Mueller, F.; Bernard, F.; Sorli, S.; Sotnychenko, O.; Qian, N.; Otaduy, M. A.; Casas, D.; and Theobalt, C. 2020. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *ACM Transactions on Graphics (TOG)*, 39(6): 1–16.

Xie, B.; Zhong, X.; and He, N. 2015. Typical data and cluster analysis on head-related transfer functions from Chinese subjects. *Applied Acoustics*, 94: 1–13.

Yan, P.; and Bowyer, K. W. 2007. Biometric recognition using 3D ear shape. *IEEE Transactions on pattern analysis and machine intelligence*, 29(8): 1297–1308.

Yang, J.; Huang, X.; He, Y.; Xu, J.; Yang, C.; Xu, G.; and Ni, B. 2021. Reinventing 2d convolutions for 3d images. *IEEE Journal of Biomedical and Health Informatics*, 25(8): 3009–3018.

Zhang, F.; Bazarevsky, V.; Vakunov, A.; Tkachenka, A.; Sung, G.; Chang, C.-L.; and Grundmann, M. 2020a. Mediapipe hands: On-device real-time hand tracking. *Fourth Workshop on Computer Vision for AR/VR*.

Zhang, Y.; Chen, W.; Ling, H.; Gao, J.; Zhang, Y.; Torralba, A.; and Fidler, S. 2020b. Image GANs meet Differentiable Rendering for Inverse Graphics and Interpretable 3D Neural Rendering. In *International Conference on Learning Representations*.

Zhou, Y.; and Gregson, J. 2020. WHENet: Real-time Fine-Grained Estimation for Wide Range Head Pose. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Zhou, Y.; and Zaferiou, S. 2017. Deformable models of ears in-the-wild for alignment and recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 626–633. IEEE.