Supplementary Material

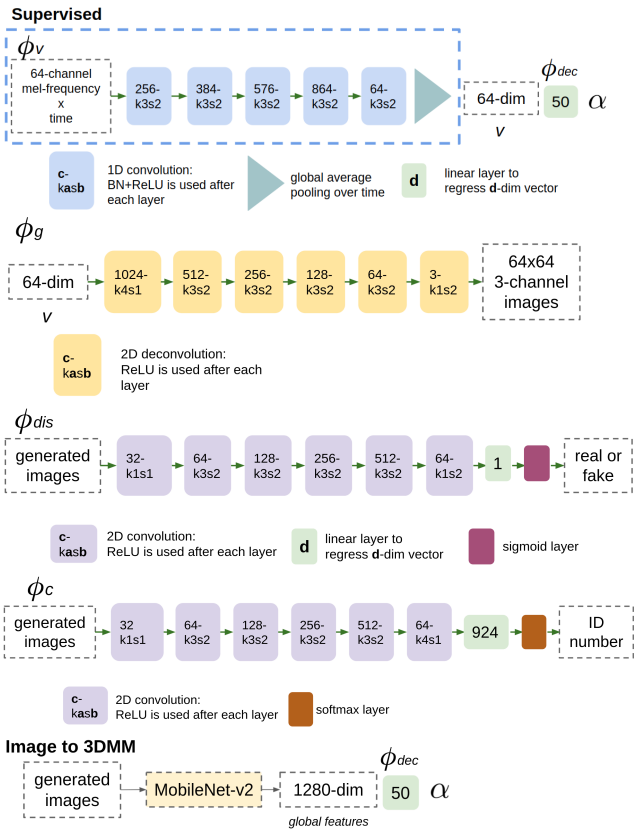


Figure S-1. **Illustrations of network architectures.** **c**-k**a**s**b** means a convolutional layer with a **c**-channel output using a kernel size **a** and stride **b**. **d** in the linear layer means to output a **d**-dimension vector. The first graphic contains our supervised learning framework with ϕ_v and ϕ_{dec}. The unsupervised setting contains ϕ_g, ϕ_{dis}, ϕ_c, ϕ_v (the same structure as in the supervised setting), and the image-to-3DMM module.

A. Overview

This supplementary document is organized as follows. In Sec. B, we show detailed network architecture for both supervised and unsupervised learning scenarios. In Sec. C, we describe details about our Voxceleb-3D training and evaluation split. In Sec. D, we respond to Q1-Q4 using our CMP- supervised learning setting. In Sec. E, we introduce both point-based and region-based metrics and compare results produced from our CMP with those from baselines. In Sec. F, we provide two simple non-network-based solutions as references using averaged face meshes in the training data as the predictions. In Sec. G, we show the robustness of head pose estimation of the expert network. In Sec. H, we present more results from our unsupervised setting. In Sec. I, we describe more on the applications of the cross-modal learning from voices to 3D faces. In Sec. J, we describe limitations of this work.

B. Network Architecture

We exhibit detailed network architectures for both supervised and unsupervised settings of our CMP in Fig. S-1.

C. Voxceleb-3D Training/Evaluation Split

We display details of the training/evaluation split in Table S-1. As described in our paper Sec. 3.2 and Sec. 4-Datasets, Voxceleb-3D inherited from Voxceleb and VGGFace contains 1225 people. Names starting with *A* to *E* are included in the evaluation set, and the others are in the training set. The training set contains as many utterances we can fetch from Voxceleb, and the evaluation set contains three utterances for each person, amounting to a total of 0.9K utterances. Face images are not included in the evaluation set because they cannot be used to calculate numerical 3D face errors, and thus we put a '-' mark in the table.

For 3D faces, we fit landmarks from images and obtain the optimized 3DMM and reconstructed 3D faces, as described in our paper Sec. 3.2 and Sec. 4-Datasets. There are several images associated with a person in VGGFace. We fit 3DMM parameters and reconstruct 3D meshes from these images. To create *reference face meshes* for a person to fulfill quantitative evaluation, we manually select one neutral 3D face from the pool that best fits 2D facial outlines on images. Therefore, there are 301 3D face meshes represented in 3DMM parameters for each person as the reference.

At test time, three utterances for each identity are used as inputs to reconstruct 3D faces. Those three predicted models are then used to compute quantitative results with the picked reference model for each identity.

Note that Voxceleb collects speech clips of interviews or talks for celebrities scraped from the web, and only gender labels are available in Voxceleb. Other features may require self-disclosure or are hard to trace, such as ages at the time of speaking, and thus are unavailable.

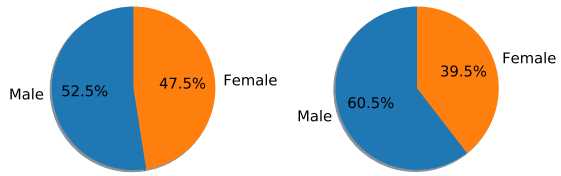
D. Response to Q1-Q4 using Supervised CMP

Following main paper Sec.4.1-Analysis, we present the counterpart of A1-A4 using our supervised framework.

A1-Face meshes from our supervised learning. In Fig. S-3, we present four types of face shapes – *skinny*, *wide*, *regular*, and *slim* – and show the reference images. The produced face meshes from our supervised learning setting exhibit the model's ability to produce various types of face shapes and are also consistent with the reference images, which are provided for shape identification purposes. This

Table S-1. **Voxceleb-3D training/evaluation split.** We provide data split details including number of utterances, number of face images, number of 3DMM parameters (equivalent to the number of 3D meshes), number of male and female, and number of people. Images for the evaluation set are not used for quantitative evaluation, and thus we mark the number '-'. We also display pie charts for gender below the table.

Dataset	Training	Evaluation
# of utterances	113K	0.9K
# of face images	107K	-
# of 3DMM param (face mesh)	107K	301
# of male/female	485/439	182/119
# of people	924	301



(a) Training split (b) Evaluation split

illustration also validates our findings in Table 1 of the paper: the lowest absolute ratio error is ear-to-ear ratio (ER) distance, which is associated with overall face shapes, indicating wider or thinner faces. We further investigate the proximity of the illustrated four face types, we calculate the parameter space Euclidean distances and show the confusion matrix in Fig. S-2.

A2-Mesh prediction coherence from our supervised learning. In Fig. S-4, we display coherence of test-time face mesh predictions from our supervised learning setting. We use different utterances from the same speaker at different time-step as inputs to produce the meshes. In Fig. S-4, one can observe from, for example, jaw widths that the output meshes are different for the two speakers; by contrast, meshes for the same speaker are highly coherent. These results demonstrate that our training strategy successfully predicts coherent geometry for the same speaker and can predict different topologies for different identities.

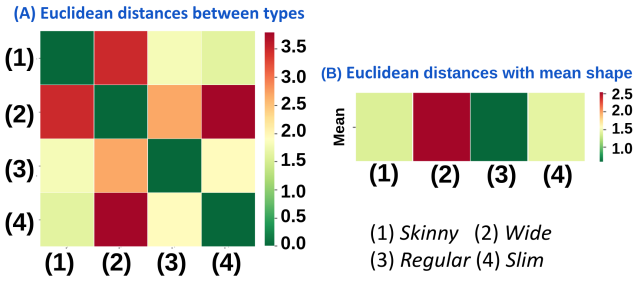


Figure S-2. **Proximity for four face types.** We show (A) a confusion matrix and (B) distances with mean face shape in 3DMM parameter space to help comprehension for the face type variation in Fig. S-3.

Finally, this coherence illustration also implies advantages over previous voice-to-face methods that work on the image domain [3, 12]. Their generation includes variations of background, hairstyles, and the rest. In contrast, Our results exclude these variations and focus on facial geometry to validate the correlation between face shapes and voices.

A3+A4-Comparison against the baseline and the major improvement from our supervised learning. We further compare against Base-2 (See Sec.4-Baseline in the paper: the cascaded pretrained blocks). One reference face, one 3D face mesh produced by our method, and one by Base-2 are presented in Fig. S-5. For the example on the left, the person of interest has a wider jawbone, and the mesh produced from our method also shows a similar trait. On the right, the image shows a wider face shape and apparent cheeks. Our 3D model also displays a similar shape, but Base-2 shows a much thinner face. Our mesh can reflect the wideness of faces, which corresponds to our findings in paper Table 1 that the major improvement voice can hint is ER (ear-to-ear ratio). In summary, we use the above visual results to show that the supervised learning of the analysis framework is effective. Fig. S-3 shows the output face meshes have similar overall face shapes to the reference images, which shows the model's ability for various types of faces and is validated in Fig. S-2. Fig. S-4 shows our supervised method can predict coherent face shapes. Fig. S-5 shows the output face models are more similar to the reference than Base-2 in terms of overall face shapes, which again validates the ER improvements shown in the paper Table 1.

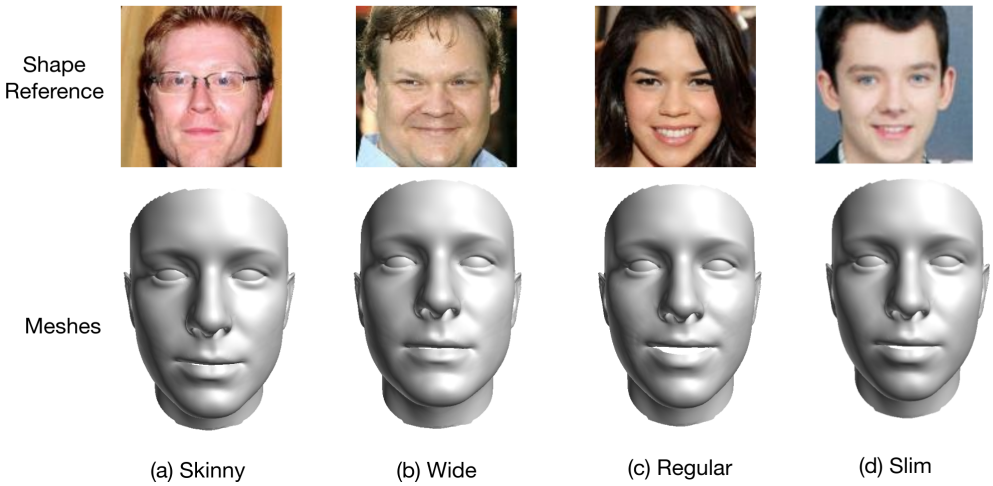


Figure S-3. **Visualization of predicted 3D face meshes from our supervised learning**. We display four face shapes, skinny, wide, regular, and slim, and their reference images to show the shape correspondence. References are provided to identify face shapes of the person of interest.

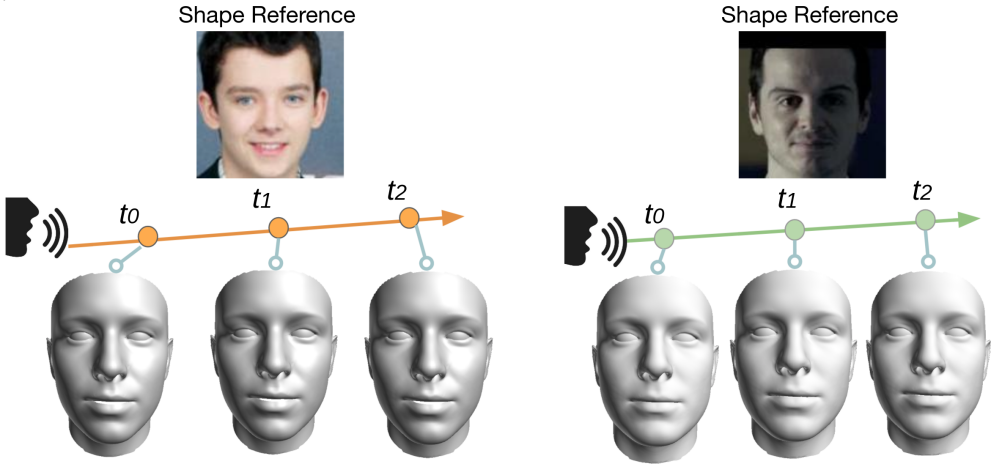


Figure S-4. **Inference coherence of meshes produced from our CMP- supervised learning.**

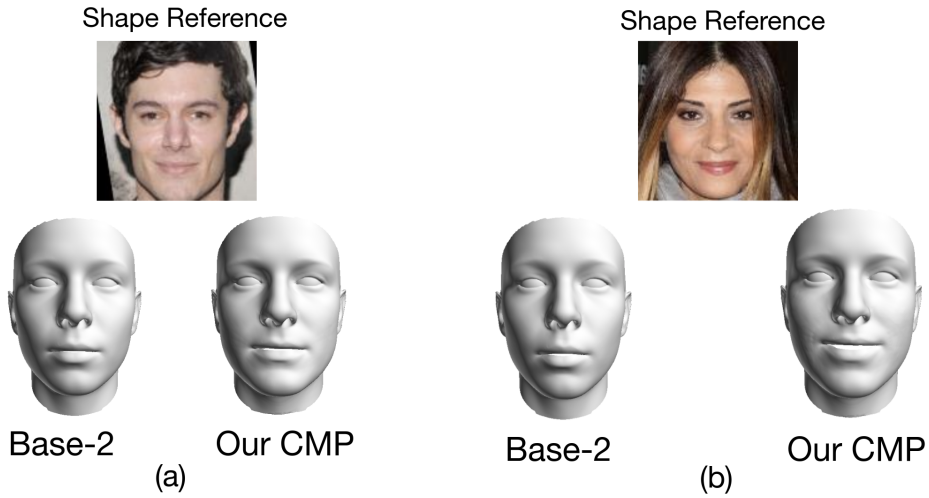


Figure S-5. **Comparison of output face meshes from our CMP of the supervised learning and Base-2**. In case (a), our mesh shows a more squared face with a wider jawbone, but Base-2 only shows a slim face. Reference face in (b) is wider and bears apparent cheeks, and our result is much more similar to the reference.

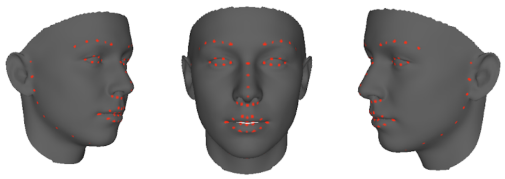


Figure S-6. **Illustration of commonly-used 68-point 3D facial landmarks.**

E. Point-based and Region-based Metrics

Normalized Mean Error (**NME**, point-based) of facial landmarks. BFM Face annotates 68 points 3D facial landmark points that lie on the eyes, nose, mouth, and face outlines (shown in Fig. S-6). We calculate the NME of the landmark point set between the predicted and reference 3D face meshes, i.e., first calculate the Euclidean distance of two landmark sets and then normalize the distance by the face size (square root of face width $\times$ length).

Results in Table S-2 show NME for 3D facial landmark alignment. Quantitative results under this metric show improvements, but the gains are smaller. It is because most facial landmarks concentrate on eyes, nose, and mouth parts that naturally bear more minor deformations across people. For example, the nose tip and mid-dorsum usually lie on the centerline of faces, and alar base and columella are located around them closely. (See Fig. S-7 for the definition of these physiology terms.)

Point-to-Plane Root Mean Square Error (**Point-to-Plane RMSE**, region-based). We follow the surface registration for 3D models using the popular iterative closest point (ICP) [9] algorithm to align the predicted and reference meshes. We then calculate point-to-plane RMSE. Registration for the holistic face and facial parts (illustrated in Fig. S-8) are considered and shown in Table S-3 and Table S-4. Both supervised and unsupervised CMP outperforms the baselines in either holistic or part-based registrations. These evaluations indicate the capability of cross-modal learning, from voice inputs to 3D face outputs.

F. Simple Oracles

We provide numerical results of simple oracles as other baselines. Oracle (1): We take average 3D faces in the Voxceleb-3D training set and use the mean shape directly as predictions for testing data; Oracle (2): We take the mean 3D face for the male/female group in the training set and use the mean shape for male/female as predictions at test time. The following Table S-5 shows results using line-based (Mean ARE), point-based (NME), and region-based (RMSE) metrics, compared with Base-2 described in paper paper Sec.4-Baseline. Simple oracles perform worse than Base-2, which means directly taking average faces is naive and weaker than the network-based solutions. This validates our baseline construction, and our CMP framework can further predict more accurate face geometry from voices for each person of interest.

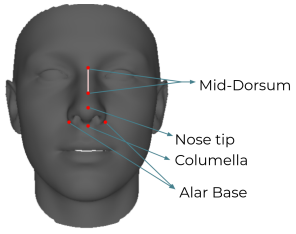


Figure S-7. **Illustration of physiology terms in Sec. E.**

Table S-2. **NME for point-based metric study.** 68 facial landmarks annotated in BFM Face [7] are used for measurements.

Landmark Alignment	Base-1	Base-2	CMP-supervised	CMP- unsupervised
NME	0.2979	0.2969	0.2723	0.2904

Table S-3. **Point-to-Plane RMSE study.** ICP is used to align the predicted and reference meshes. We calculate point-to-plane RMSE after ICP.

Holistic Registration	Base-1	Base-2	CMP-supervised	CMP- unsupervised
RMSE	1.357	1.348	1.210	1.312

Table S-4. **Part-based point-to-plane RMSE study.**

Part Registration	Base-1	Base-2	CMP- supervised	CMP- unsupervised
Left Eye	0.3961	0.3945	**0.3517**	0.3779
Right Eye	0.3667	0.3656	**0.3349**	0.3488
Nose	0.5258	0.5250	**0.5141**	0.5177
Mouth	0.3466	0.3435	**0.2958**	0.3149
Left Cheek	0.4748	0.4735	**0.4654**	0.4711
Right Cheek	0.5078	0.5061	**0.4916**	0.4919

Figure S-8. **Illustration of regions.** We show regions of the left eye, right eye, nose, mouth, left cheek, and right cheek that are used in the part-based point-to-plane evaluation using ICP in our paper Table 3.

G. Pose from the Pretrained Expert

Here we study the robustness of head poses estimated from the expert, which helps our visualization (in Fig.7-

Table S-5. **Oracle results**. We show quantitative evaluations of simple oracles explained in Sec. F.

Metrics	Type	Oracle(1)	Oracle(2)	Base-2
Mean ARE	line	0.0319	0.0311	0.0302
NME	point	0.3058	0.3021	0.2969
RMSE	region	1.540	1.529	1.348

10 of the paper) of laying 3D face meshes onto images to show the fitness. SynergyNet [13] as an expert used in the unsupervised framework predicts 3DMM parameters (α_s and α_e) as pseudo-groundtruth based on images synthesized from GAN. Here we verify the robustness of pose estimation from the expert. As illustrated in our paper Fig. 8, synthesized faces from GAN are almost frontal because face images in VGGFace [5] for GAN-training are with small poses. We adopt widely-used AFLW2000-3D [15] including 2K in-the-wild face images to examine the performance of head pose estimation. Then, we calculate the mean absolute error (MAE) for three estimated Euler angles (yaw, pitch, roll). MAE is 1.49 degrees for faces whose yaw angle (left/right turns) lies in [-15°, 15°]. This result justifies the robustness of head pose estimation from the pretrained expert.

H. Ablation Study on Knowledge Distillation

We conduct an extensive survey for the performance of various recent KD strategies on our unsupervised framework. We include vanilla KD [2], Attention [14], SP [11], Correlation [8], RKD [4], CRD [10], VID [1], PKT [6], and train our unsupervised framework with different $\mathcal{L}_{KD}$. Here we show the results in Table S-6. We find that more recent and advanced KD methods attain similar results. For example, RKD, CRD, and PKT have very close performance, compared with earlier methods such vanilla version or using attention map similarity. Therefore, the study validates our adoption of conditional probability in our paper Eq.(4)[1], introduced in PKT [6].

Further, we exhibit more qualitative comparisons in Fig. S-9 extending Fig. 10 of the main paper.

I. Applications of Voice-to-3D-Face Task

We focus on the analysis that purposes to validate the correlation between voice and 3D face geometry. Here we describe more on application sides, where our work has potential at:

1. Our work can be used for public security, such as recovering the face shape of the unheard speech of a suspect or a masked robber.

[1]The scaled and shifted cosine similarity is $K_{cosine}(z_i, z_j) = \frac{1}{2}\left((z_i^E z_j / \|z_i^E\|_2 \|z_j\|_2) + 1\right)$.

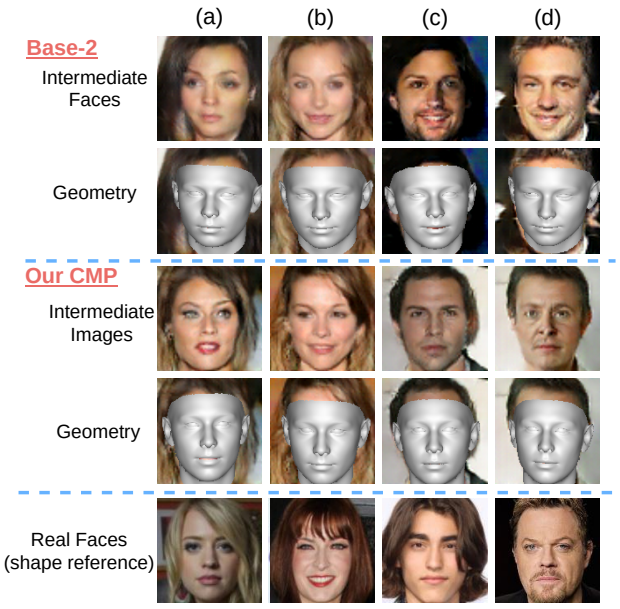


Figure S-9. **Qualitative comparison between our unsupervised CMP and the baseline**. This figure presents more results that extend Fig. 10 of the main paper.

2. Our work can generate personal avatars in gaming or virtual reality systems: it is helpful to create a rough 3D face model from voices as the initialization, and users can refine its shape based on one's preference.

3. 3D faces from voice can potentially provide another verification mode for person identification other than speech and face image verification.

J. Limitations

Our work focus on the analysis between voices and 3D faces, and generating high-quality meshes is not our aim. In fact, using voices as inputs to produce face meshes has its inherent limitations since our face wideness might be gleaned from voices from our intuition, but more subtle details, such as bumps or wrinkles, cannot be hinted at from this modality. We target at analysis between one's normal voices and face shapes and utilize Voxceleb as the speech source, which is primarily interviews or talks. As pointed out in main paper Sec.5- Ethical statement, more implicit factors like talking after drinking or abnormal health conditions may affect the analysis, but this requires large data corpse from a medical or physiological view to further validate these effects.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information dis-

Table S-6. **Study on performances of different KD strategies.** ARE is used as the metric for comparison.

ARE	Vanilla KD	Attention	SP	Correlation	RKD	CRD	VID	PKT
ER	0.0306	0.0318	0.0230	0.0227	0.0172	0.0198	0.0213	0.0184
FR	0.0173	0.0172	0.0169	0.0173	0.0171	0.0172	0.0172	0.0172
MR	0.0173	0.0173	0.0179	0.0179	0.0195	0.0177	0.0178	0.0176
CR	0.0540	0.0551	0.0471	0.0471	0.0474	0.0481	0.0471	0.0484
Mean	0.0298	0.0304	0.0262	0.0263	**0.0254**	0.0255	0.0259	**0.0254**

tillation for knowledge transfer. In *CVPR*, pages 9163–9171, 2019. 5

[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 5

[3] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. Speech2face: Learning the face behind a voice. In *CVPR*, pages 7539–7548, 2019. 2

[4] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, pages 3967–3976, 2019. 5

[5] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, 2015. 5

[6] Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Probabilistic knowledge transfer for lightweight deep representation learning. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2020. 5

[7] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. IEEE, 2009. 4

[8] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *CVPR*, pages 5007–5016, 2019. 5

[9] François Pomerleau, Francis Colas, and Roland Siegwart. A review of point cloud registration algorithms for mobile robotics. *Foundations and Trends in Robotics*, 2015. 4

[10] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020. 5

[11] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *CVPR*, pages 1365–1374, 2019. 5

[12] Yandong Wen, Bhiksha Raj, and Rita Singh. Face reconstruction from voice using generative adversarial networks. In *NeurIPS*, volume 32, 2019. 2

[13] Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. *3DV*, 2021. 5

[14] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. 2017. 5

[15] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *CVPR*, pages 146–155, 2016. 5