

Synergy between 3DMM and 3D Landmarks for Accurate 3D Facial Geometry

Cho-Ying Wu Qiangeng Xu Ulrich Neumann

University of Southern California

Abstract

This work studies learning from a synergy process of 3D Morphable Models (3DMM) and 3D facial landmarks to predict complete 3D facial geometry, including 3D alignment, face orientation, and 3D face modeling. Our synergy process leverages a representation cycle for 3DMM parameters and 3D landmarks. 3D landmarks can be extracted and refined from face meshes built by 3DMM parameters. We next reverse the representation direction and show that predicting 3DMM parameters from sparse 3D landmarks improves the information flow. Together we create a synergy process that utilizes the relation between 3D landmarks and 3DMM parameters, and they collaboratively contribute to better performance. We extensively validate our contribution on full tasks of facial geometry prediction and show our superior and robust performance on these tasks for various scenarios. Particularly, we adopt only simple and widely-used network operations to attain fast and accurate facial geometry prediction. Codes and data: <https://choyingw.github.io/works/SynergyNet/>.

1. Introduction

Facial geometry prediction including 3D facial alignment, face orientation estimation, and 3D face modeling are fundamental tasks [8, 13, 27, 32, 50, 51, 61, 69, 73] and have applications on face recognition [11, 25, 45, 57, 58, 60], tracking [9, 12, 29, 30], and compression [54]. Recent works [17, 20, 49, 70, 71] predict facial geometry by estimating 3D Morphable Models (3DMM) parameters that include shape and expression variations. Yet, face orientation for these previous works is only a by-product without evaluation and discussion on relations with 3D landmarks and 3D face models. In contrast, we fully evaluate facial geometry, including 3D facial alignment, face orientation estimation, and 3D face modeling.

3D face meshes can be built from 3DMM parameters, and 3D landmarks can be extracted from vertices by querying associated indices. 3D landmarks are widely used to guide 3D facial geometry learning. Previous works

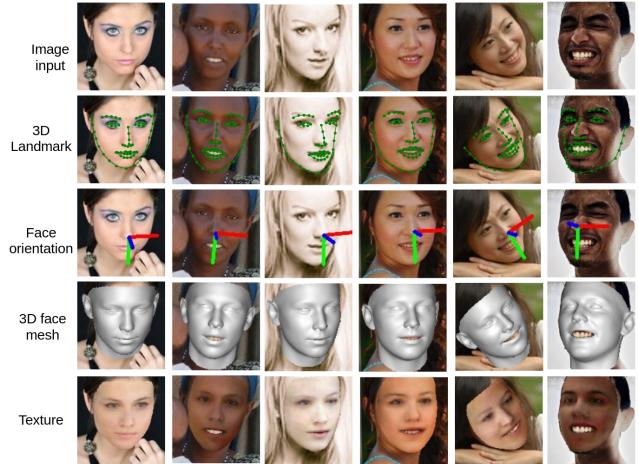


Figure 1. **Results from our SynergyNet with monocular image inputs.** Note that 3D landmarks can predict hidden face outlines in 3D rather than follow visible outlines on images.

[17, 20, 49, 70, 71] only directly extract coarse landmarks from fitted 3D faces and compute supervised alignment losses with groundtruth landmarks. These works utilize a representation direction, from 1D parameters to 3D landmarks. However, though 3D landmarks are very sparse (a 68-point definition is commonly used), they compactly and efficiently describe facial outlines in 3D space. We think 3D landmarks can be further exploited to predict underlying 3D faces as supportive information. Hence, in addition to only going from 1D parameters to 3D landmarks, we propose a further step to reversely regress 3DMM parameters from 3D landmarks and establish a representation cycle. The advantage is that *predicting 3D face using 3DMM from 2D images is naturally an ill-posed problem, but prediction from 3D landmarks can alleviate the intrinsic ill-posedness*. To our knowledge, we are the first to study this reverse representation direction, from 3D landmarks to 3DMM parameters. Together we build a representation cycle as a *synergy process* that adopts collaborative relation between 3DMM parameters and 3D landmarks to improve the information flow and attain better performance.

We propose **SynergyNet**, a synergy process network that includes two stages. The first stage contains a backbone net-

work to regress 3DMM parameters from images and construct 3D face meshes. After landmark extraction by querying associated indices, we propose a landmark refinement module that aggregates 3DMM semantics and incorporates them into point-based features to produce refined 3D landmarks. We closely validate how each information source contributes to 3D landmark refinement. From the representation perspective, the first stage goes from 1D parameters to 3D landmarks. Next, the second stage contains a landmark-to-3DMM module that predicts 3DMM parameters from 3D landmarks, which is a reverse representation direction compared with the first stage. We leverage this step to regress embedded facial geometry lying in sparse landmarks. The overall framework is in Fig. 2.

We will first review 3DMM as basics in Sec.3.1 used in the previous work [20, 49, 70, 71]. 3DMM regression contains pose, shape, and expression parameter estimation from a monocular face image through a backbone network. 3D faces are constructed as foundation models by 3DMM, and 3D landmarks are extracted from face meshes. Next, in Sec.3.2, we introduce the proposed multi-attribute feature aggregation (MAFA), including landmark features, image features, and shape and expression of 3DMM semantics. MAFA is then used to produce finer landmark structures. The advantage is that refinement based on only coarse landmarks is hard because the information is unitary. Joining different attributes can refine and correct raw structures.

In Sec.3.3, we introduce the reverse direction to regress 3DMM parameters from 3D landmarks, based on the assumption that 3D landmarks contain rough facial geometry. The advantage is that regressing 3DMM parameters from 3D landmarks can avoid inherent ambiguity in conventional 3DMM-based methods that predict facial geometry only from images. A self-constraining loss for both 3DMM parameters regressed from images and from 3D landmarks is used: since two 3DMM parameters describe the same identity, they should be numerically consistent.

Especially, our SynergyNet contains only simple and widely-used network operations in the whole synergy process. We quantitatively analyze performance gains introduced by each adopted information and each regression target with extensive experiments. We evaluate our SynergyNet on all tasks of facial alignment, face orientation estimation, and 3D face modeling using the standard datasets for each task. Our SynergyNet attains superior performance than other related work. Fig.1 demonstrates the ability of our SynergyNet.

In summary, we present the following contributions:

1. We propose SynergyNet to study a synergy process that leverages the collaborative relation between 3DMM parameters and 3D landmarks to learn better 3D facial geometry. This is the first study to include reverse representation direction, i.e., from 3D landmarks to 3DMM parameters.

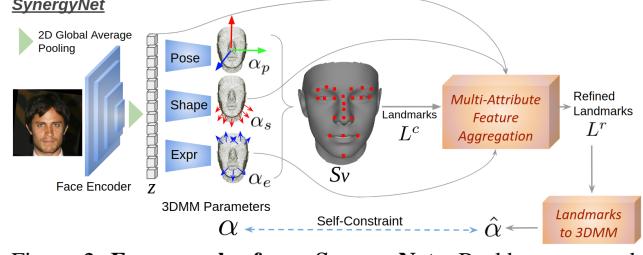


Figure 2. Framework of our SynergyNet. Backbone network learns to regress 3DMM parameters (α_p, α_s , and α_e) and reconstruct 3D face meshes from monocular face images. Multi-Attribute feature aggregation gathers underlying 3DMM semantics and the latent image code to refine landmarks further. The landmark-to-3DMM module regresses 3DMM from refined landmarks L^r to reveal the embedded facial geometry in 3D landmarks. A self-constraining consistency is applied to 3DMM parameters regressed from different sources. This synergy process includes a forward representation direction, from *3DMM parameters to refined 3D landmarks*, and a reverse direction, from *3D landmarks to regress 3DMM parameters*, to attain better performance. The red and blue arrows after shape and expression (expr) decoders show the main areas of deformation that each 3DMM semantics controls.

2. We propose multi-attribute feature aggregation for landmark refinement using multiple information sources and closely analyze performance gain for each information.

3. We conduct extensive and detailed benchmarking on 3D facial geometry, including facial alignment, face orientation estimation, and 3D face modeling, to validate our superior performance on these tasks.

2. Related Work

2.1. 3D Facial Alignment via 3D Face Modeling

3D facial alignment aims at predicting 3D landmarks on images. In contrast, 2D approaches [14, 15, 18, 26, 62] usually regress direct landmark coordinates or heatmaps based on *visible* facial parts. If input faces are self-occluded due to large face poses, their methods either only estimate landmarks along visible face outlines rather than hidden outlines or produce much larger errors at invisible parts that make the results unreliable.

3D approaches [4, 17, 20, 24, 43, 70, 71] predict aligned 3D faces with images. This way, occluded landmarks can be registered. 3DDFA [70, 71] adopts Basel Face Model (BFM) and use 3DMM fitting to reconstruct face meshes from monocular images. PRNet [17] predicts 2D UV-position maps that encode 3D points and uses BFM mesh connectivity to build face models. Compared with 3DDFA, PRNet might have higher mesh deformation ability since its 3D points are not from 3DMM parameterization. However, it is harder to obtain a smooth and reliable mesh for PRNet. 2DASL [49] based on 3DMM further adopts a differentiable renderer and a discriminator to produce high-

quality 3D face models. 3DDFA-V2 [20] based on 3DDFA further introduces a meta-joint optimization strategy and a short video synthesis to attain the current best result. 3D faces and 3D landmarks extracted by vertex indexing are outputs of these methods. However, their landmarks are raw without refinement. Our landmarks are refined with multi-attribute feature aggregation, and we further adopt 3DMM from 3D landmarks as another information source.

Another line of work adopts self-supervision from images and targets at more realistic 3D face synthesis [16, 42, 47, 48, 56]. Their self-supervised factors usually rely on visible facial areas to collect visual cues for prediction and may not be robust to large-pose cases.

2.2. Face Orientation Estimation

Face orientation estimation has applications on human-robot interaction [28, 36, 52]. Euler angles (yaw, pitch, roll) are used to represent the orientation. Deep Head Pose [35] uses networks to predict 2D landmarks and face orientation at the same time. HopeNet [40] uses bin-based angle regression and QuatNet [22] uses a multi-regression loss for head pose. FSA-Net [64] constructs a fine-grained structure mapping for features aggregation. TriNet [7] uses a vector-based representation for pose estimation. These works focus on face orientation as a standalone task. On the other hand, although 3DMM-based 3D alignment approaches estimate rotation matrices, previous works only focus on evaluation and discussion on landmarks and 3D faces [20, 49, 70, 71]. To gain an insight into full facial geometry, we benchmark both standalone orientation estimation methods and 3DMM-based approaches.

3. Method

Our method, illustrated in Fig. 2, aims at precise and accurate 3D facial alignment, face orientation estimation, and 3D face modeling by utilizing a synergy process of 3D landmarks and 3DMM parameters to guide 3D facial geometry learning better. The pipeline contains two stages. The first stage includes a preliminary 3DMM regression from images and a multi-attribute feature aggregation (MAFA) for landmark refinement. The second stage contains a landmark-to-3DMM regressor to reveal the embedded facial geometry in sparse landmarks.

3.1. 3D Morphable Models (3DMM)

3DMM reconstructs face meshes using principal component analysis (PCA). Given a mean face $M \in \mathbb{R}^{3N_v}$ with N_v 3D vertices, 3DMM deforms M into a target face mesh by predicting the shape and expression variations. $U_s \in \mathbb{R}^{3N_v \times 40}$ is the basis for shape variation manifold that represents different identities, $U_e \in \mathbb{R}^{3N_v \times 10}$ is the basis for expression variation manifold, and $\alpha_s \in \mathbb{R}^{40}$ and $\alpha_e \in \mathbb{R}^{10}$ are the associated basis coefficients. The 3D face

reconstruction can be formulated in Eq.1.

$$S_f = \text{Mat}(M + U_s \alpha_s + U_e \alpha_e), \quad (1)$$

where $S_f \in \mathbb{R}^{3 \times N_v}$ represents a reconstructed frontal face model after the vector-to-matrix operation (Mat). To align S_f with input view, a 3×3 rotation matrix $R \in SO(3)$, a translation vector $t \in \mathbb{R}^3$, and a scale τ are predicted to transform S_f by Eq.2.

$$S_v = \tau R S_f + t, \quad (2)$$

where $S_v \in \mathbb{R}^{3 \times N_v}$ aligns with input view. τR and t are included as 3DMM parameters in most works [20, 61, 70], and thus we use $\alpha_p \in \mathbb{R}^{12}$ instead. We follow the current best work 3DDFA-V2 [20] to predict 62-dim 3DMM parameters α for pose, shape, and expression.

We follow 3DDFA-V2 to adopt MobileNet-V2 as the backbone network to encode input images and use fully-connected (FC) layers as decoders for predicting 3DMM parameters from the bottleneck image feature z . We separate the decoder into several heads by 3DMM semantics, which jointly predict the whole 62-dim parameters. The advantage of separate heads is that disentangling pose, shape, and expression controls better information flow. The illustration in Fig. 2 shows the encoder-decoder structure. The decoding is formulated as $\alpha_m = \text{Dec}_m(z)$, $m \in \{p, s, e\}$, showing pose, shape and expression. With groundtruth notation $*$ hereafter, the supervised 3DMM regression loss is shown as follows.

$$\mathbb{L}_{3DMM} = \sum_m \|\alpha_m - \alpha_m^*\|^2. \quad (3)$$

3.2. From 3DMM to Refined 3D Landmarks

After regressing 3DMM parameters $(\alpha_p, \alpha_s, \alpha_e)$, 3D face mesh for the input face can be constructed by Eq.1 and be aligned with input face by Eq.2. We adopt popular BFM [37], which includes about 53K vertices, as the mean face M in Eq.1. Then, 3D landmarks $L^c \in \mathbb{R}^{3 \times N_l}$ are extracted by landmark indices. $N_l = 68$ is used in 300W-LP [70] as our training dataset.

Previous studies [20, 70, 71] directly use extracted landmarks L^c to compute the alignment loss for learning 3D facial geometry. However, these extracted landmarks are raw without refinement. Instead, we adopt a refinement module that aggregates multi-attribute features to produce finer landmark structures. Landmarks can be seen as a sequence of 3D points. Weight-sharing multi-layer perceptrons (MLPs) are commonly used for extracting features from structured points. PointNet-based frameworks [33, 38, 39, 55, 59, 63] use an MLP-encoder to extract high-dimensional embeddings. At the bottleneck, global point max-pooling is applied to obtain global point features. Then an MLP-decoder is used to regress per-point attributes. An MLP-based refinement module takes sparse landmarks L^c

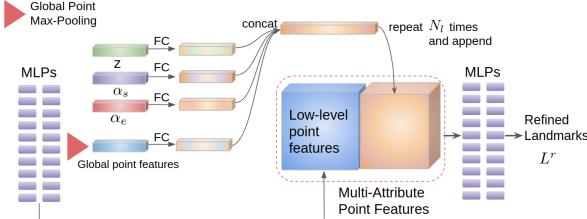


Figure 3. Structure of multi-attribute landmark refinement. The input is L^c from the foundation face model. The left MLPs extract global point features and fuse the global features with other attributes, including images features, shape, and expression parameters. The concatenation is appended to the low-level features to create multi-attribute point features, which are used to regress the refined landmarks.

as inputs and uses the MLP-encoder and MLP-decoder to produce finer landmarks.

Instead of using L^c alone for the refinement, our refinement module adopts multi-attribute feature aggregation (MAFA), including input images and 3DMM semantics that provide information from different domains. For example, shape contains information of thinner/thicker faces, and expression contains information for eyebrow or mouth movements. Therefore, these pieces of information can help regress finer landmark structures. Specifically, our MAFA fuses information of the image, using its bottleneck features z after global average pooling and shape and expression 3DMM parameters. These features and parameters are global information without spatial dimensions. We first use FC layers for domain adaption. Later we concatenate them into a multi-attribute feature vector and then repeat this vector N_l times to make multi-attribute features compatible with per-point features. We last append the repeated features to the low-level point features and feed them to an MLP-decoder to produce refined 3D landmarks. The overall design is shown in Fig. 3. Skip connection is used from the coarse to refined landmarks to facilitate training.

We use groundtruth landmarks to guide the training. The alignment loss function is formulated as follows.

$$\mathbb{L}_{lmk} = \sum_n \mathbb{L}_{smL1}(L_n^r - L_n^*), n \in [1, N_l], \quad (4)$$

where N_l is number of landmarks, $*$ denotes groundtruth, and \mathbb{L}_{smL1} is smooth L1 loss. So far, the operations of constructing 3D face meshes and landmark extraction and refinement transform 3DMM parameters to refined 3D landmarks.

3.3. From Refined Landmarks to 3DMM

We next describe the reverse direction of representation that goes from refined landmarks to 3DMM parameters.

Previous works only consider 3DMM parameter regression from images [13, 20, 24, 49, 61, 70, 71]. However, facial landmarks are sparse keypoints lying at eyes, nose, mouth,

and face outlines, which are principal areas that α_s and α_e control. We assume that approximate facial geometry is embedded in sparse landmarks. Thus, we further build a landmark-to-3DMM module to regress 3DMM parameters from the refined landmarks L^r using the holistic landmark features. To our knowledge, we are the first to study this reverse representation direction, from landmarks to 3DMM parameters.

The landmark-to-3DMM module also contains an MLP-encoder to extract high dimensional point features and use a global point max-pooling to obtain holistic landmark features. Later separate FC layers transform the holistic landmark features to 3DMM parameters to get $\hat{\alpha}$, including pose, shape, and expression. We refer $\hat{\alpha}$ to **landmark geometry**, since this 3DMM geometry is regressed from landmarks. We adopt a supervised loss with groundtruth α^* for $\hat{\alpha}$ as follows.

$$\mathbb{L}_{3DMM_{lmk}} = \sum_m \|\hat{\alpha}_m - \alpha_m^*\|^2, \quad (5)$$

where m contains pose, shape, and expression.

Furthermore, since $\hat{\alpha}$ regressed from the landmarks and α regressed from the face image describe the same identity, they should be numerically similar. We further add a novel self-supervision control as follows.

$$\mathbb{L}_g = \sum_m \|\alpha_m - \hat{\alpha}_m\|^2, \quad (6)$$

where $m \in \{p, s, e\}$. \mathbb{L}_g improves information flow that lets 3DMM regressed from images obtain support from landmark geometry.

The advantage of self-supervision control (Eq.6) is that since images and sparse landmarks are different data representations (2D grids and 3D points) using different network architectures and operations, more descriptive and richer features can be extracted and aggregated under this multi-representation strategy. Although conceptually sparse landmarks provide rough face outlines, our experiments show that this reverse representation direction further contributes to the performance gain and attains superior results than related work.

Overall, the total loss combination is shown as follows

$$\mathbb{L}_{total} = \lambda_1 \mathbb{L}_{3DMM} + \lambda_2 \mathbb{L}_{lmk} + \lambda_3 \mathbb{L}_{3DMM_{lmk}} + \lambda_4 \mathbb{L}_g, \quad (7)$$

where λ terms are loss weights.

3.4. Representation Cycle

Our overall framework creates a cycle of representations. First, the image encoder and separate decoders regress 1D parameters from a face image input. Then, we construct 3D meshes from parameters and refine extracted 3D landmarks—this is the forward direction that switches representations from 1D parameters to 3D points. Next, the reverse representation direction adopts a landmark-to-3DMM module to switch representations from 3D points back to 1D

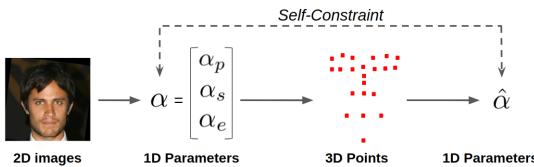


Figure 4. Illustration of representation cycle.

parameters. Therefore it forms a representation cycle (Fig. 4), and we minimize the consistency loss to facilitate the training. The forward and reverse representation direction between 3DMM parameters and refined 3D landmarks form a synergy process that collaboratively improves the learning of facial geometry. Landmarks are extracted and refined, and the refined landmarks and landmark geometry further supports better 3DMM parameter predictions using the self-supervised consistency loss (Eq.6).

Compared with a simple baseline using only the forward representation, i.e., going from image to 3DMM and directly extracting 3D points from built meshes to compute alignment loss, our proposed landmark refinement (MAFA) and the reverse representation (landmark-to-3DMM module) only bring about 5% more time in average for a single feed-forward pass. This is because landmarks are sparse and compact, and weight-sharing MLPs are lightweight.

We choose simple and widely-used network operations to show that without special operations, landmarks and 3DMM parameters can still guide the 3D facial geometry learning better. Through the following studies and experiments, we closely validate each module we introduce to the plain 3DMM regression from images, including MAFA for landmark refinement and landmark-to-3DMM module. Network details are described in the supplementary.

4. Experiments

Evaluation is conducted on the three focused tasks: facial alignment, face orientation estimation, and 3D face models. One of our main contributions is close studies that exhibit a detailed performance gain breakdown for each module we propose and each attribute we include in MAFA.

Procedure. We train on 300W-LP [70], which is a standard and widely-used training set on 3D face tasks. It collects in-the-wild face images and conducts face profiling [70] for producing 3DMM parameters with BFM [37] and FaceWarehouse texture [5]. The dataset also performs out-of-plane face rotation for augmentation, attaining more than 60K face images and fitted 3D models.

During training, we use a learning rate of 0.08, batch size of 1024, and momentum of 0.9 of the SGD [72] optimizer. We train our network with 80 epochs, and the learning rate decays to $\frac{1}{10}$ and $\frac{1}{100}$ of the initial after 48 and 64 epochs. We use random color jittering and random horizontal flip. We further adopt face-swapping augmentation that exchanges textures with others, overlays them on the

original mesh, and last renders the 3D model onto the original image. This is to increase the appearance variety that creates novel texture-geometry pairs. We train on 4x GTX 1080Ti GPUs, and the training takes about 8 hours.

At test time, the refined landmarks L^r , fitted 3D face S_v with its orientation from α_p are the outputs for evaluation. The processing of landmark geometry is saved at test time since its information is auxiliary, and we leave the discussion and evaluation of landmark geometry in the supplementary. Our inference attains 2600fps inference speed on average for the 3D landmark prediction and about 2300fps for the dense 3D face prediction on a single GPU with the MobileNet backbone. The speed calculation includes inference time batching and communication overhead of data loading to GPU-memory. This satisfies the real-world applications for fast inference.

In addition to facial geometry, we also study texture synthesis in the supplementary for more realistic 3D faces.

Test sets for facial alignment. Facial alignment is standardly evaluated on AFLW2000-3D [70], which contains the first 2000 images of AFLW [34] with a *68-point* landmark annotation. Two landmark sets of AFLW2000-3D are present, original and reannotated by LS3D-W [4]). The reannotated one carries better quality. We separately report performances on the two versions to fairly compare with related work. We also follow [20] to evaluate on the full AFLW set, which contains 21K images with a *21-point* landmark annotation. The two datasets are used for showing evaluation on different numbers of facial landmarks.

Test sets for 3D face modeling. We evaluate 3D face modeling on AFLW2000-3D [70], MICC Florence [2], 300VW [44], and Artistic-Faces [66] for both quantitative and qualitative analysis. AFLW2000-3D contains 2000 fitted 3D faces. Florence contains high-resolution real face scans of 53 individuals. 300VW collects talks or interviews from the web, and Artistic-Faces gathers artistic style faces.

Test sets for face orientation estimation. Most previous 3DMM-based works [4, 17, 20, 20, 70] focus on the facial alignment and 3D face modeling. To fully evaluate facial geometry, we introduce face orientation estimation for evaluation. AFLW2000-3D contains large-pose faces with orientation annotation that make it suitable for evaluation.

4.1. Facial Alignment Evaluation

Metrics. Normalized mean error (NME) in Eq.8 for sparse facial landmarks with Euclidean distance is reported.

$$\text{NME} = \frac{1}{T} \sum_{t=1}^T \frac{\|u_t - v_t\|_2}{B}, \quad (8)$$

where u_t and v_t are landmark prediction/groundtruth that are both registered with face images, T is the number of samples, and B is bounding box size, square root of box areas, as the normalization term for each face.

Table 1. Ablative for facial alignment. The first table is for AFLW2000-3D using original groundtruth annotation. The second table is for the reannotated version. ‘-’ means the module is not used, and the corresponding loss terms are not introduced. The first row setting without all the introduced modules contains only a simple baseline of a backbone network to regress 3DMM parameters from only images.

AFLW2000-3D Original	Multi-Attribute Feature Aggregation for Refinement	Landmark-to- 3DMM	0 to 30	30 to 60	60 to 90	All
-	-	-	2.99	3.80	4.86	3.88
	✓	-	2.68	3.32	4.35	3.49
	-	✓	2.69	3.57	4.69	3.65
	✓	✓	2.66	3.30	4.27	3.41
AFLW2000-3D Reannotated	Multi-Attribute Feature Aggregation for Refinement	Landmark-to- 3DMM	0 to 30	30 to 60	60 to 90	All
-	-	-	2.34	2.99	4.27	3.20
	✓	-	2.24	2.67	3.76	2.89
	-	✓	2.23	2.69	3.90	2.94
	✓	✓	2.16	2.61	3.66	2.81

Ablation study. We show three different ablation studies in this section to thoroughly examine the contribution of each part in SynergyNet. The aim is to validate our introduced multi-attribute feature fusion and analyze how each attribute contributes to the final performance.

(1) We first conduct ablation studies of SynergyNet on AFLW2000-3D. Compared with conventional frameworks that only use a backbone network to regress 3DMM parameters from images, the proposed multi-attribute feature aggregation for landmark refinement and the landmark-to-3DMM module in the synergy process are examined. Following [17, 20, 71], we report NME under three yaw angle ranges. The results are shown in Table 1.

The table shows that both landmark refinement and landmark geometry support stages contribute to the final performance for better facial landmark estimation. MAFA adopts the advantage of multi-attribute information fusion to refine landmarks. The landmark-to-3DMM reveals the geometric information embedded in the 3D landmarks and helps 3DMM prediction with the representation cycle. From both tables’ third and fourth rows, directly predicting landmark geometry from raw landmarks without refinement only obtains limited performance gains. By contrast, based on finer landmarks, the reverse representation stage further improves the results. This validates our SynergyNet design that both MAFA and the landmark-to-3DMM are required to attain the best performance.

(2) We then study the performance contribution of *each attribute at MAFA* and *each 3DMM regression target* at the reverse representation direction stage. For the former, we experiment with different feature aggregation and fusion: only point feature, point+image feature, and all attributes in Fig. 3 (point, image, and 3DMM semantics). For the latter, we examine the performance gain of each regression target at the landmark-to-3DMM, including pose, shape, and expression from 3D landmarks. Results are shown in Table 2. Row 1- 3 show that the gain of using image and 3DMM semantics mainly comes from small or medium pose ranges because images and the derived 3DMM parameters capture more descriptive features on frontal faces. Large poses

Table 2. Study on different attributes used at landmark refinement and different regression targets at landmark-to-3DMM ($L \rightarrow 3D$). AFLW2000-3D Original is adopted for facial alignment evaluation. The first three rows exploit different attributes for feature aggregation. Row 3 uses all attributes shown in Fig.3. Row 4 to 6 further study performance gains of different regression targets at the reverse representation direction.

Structures	0 to 30	30 to 60	60 to 90	All
Point feature only	2.73	3.51	4.51	3.58
Point + image feature	2.68	3.40	4.61	3.56
MAFA	2.67	3.34	4.51	3.51
MAFA+ $L \rightarrow 3D$ (pose)	2.68	3.31	4.55	3.51
MAFA+ $L \rightarrow 3D$ (shape,expr)	2.67	3.32	4.47	3.48
MAFA+ $L \rightarrow 3D$ (all)	2.66	3.30	4.27	3.41

Table 3. Landmark-to-3DMM network structure study. Facial alignment on AFLW2000-3D Original is evaluated. Refer to Section 4.1 for the two comparison settings. The last row is the adopted setting introduced in Sec.3.3.

Settings	0 to 30	30 to 60	60 to 90	All
Comparison 1	2.63	3.35	4.50	3.49
Comparison 2	2.62	3.31	4.31	3.41
Adopted setting	2.66	3.30	4.27	3.41

may cause *self-occlusion* on images and make prediction unreliable. Row 4 to 6 exhibit effects of regressing pose only (Row 4), shape and expression (Row 5), and all (Row 6). The improvements mainly come from large-pose cases. The reason is that the reverse direction regresses parameters from 3D landmarks that provide features from the 3D space, which naturally avoids self-occlusion compared with 2D. Such a strategy benefits alignment for large poses.

(3) We further investigate the other two possible network designs of landmark-to-3DMM. Comparison 1 refines landmarks and regresses $\hat{\alpha}$ at the same step, and thus $\hat{\alpha}$ is regressed from L^c in this setting. *This is to study whether the reverse representation direction is benefited from refined landmarks L^r .* Comparison 2 further includes z , α_s , and α_e in the landmark-to-3DMM regression in Fig. 2, forming another multi-attribute feature aggregation to regress the $\hat{\alpha}$. *This setting analyzes whether aggregation can also assist $\hat{\alpha}$ prediction.*

From Table 3, results of Comparison 1 show that regress-

Table 4. **Benchmark on AFLW2000-3D for facial alignment.** The original annotation version is used. Our performance is the best with a gap over others on large poses.

AFLW2000-3D Original	0 to 30	30 to 60	60 to 90	All
ESR [6]	4.60	6.70	12.67	7.99
3DDFA [70]	3.43	4.24	7.17	4.94
Dense Corr [67]	3.62	6.06	9.56	6.41
3DSTN [3]	3.15	4.33	5.98	4.49
3D-FAN [4]	3.16	3.53	4.60	3.76
3DDFA-PAMI [71]	2.84	3.57	4.96	3.79
PRNet [17]	2.75	3.51	4.61	3.62
2DASL [49]	2.75	3.46	4.45	3.55
3DDFA-V2 (MR) [20]	2.75	3.49	4.53	3.59
3DDFA-V2 (MRS) [20]	2.63	3.42	4.48	3.51
SynergyNet (our)	2.65	3.30	4.27	3.41

Table 5. **Quantitative facial alignment comparison on AFLW with 21-point landmark definition.**

AFLW	0 to 30	30 to 60	60 to 90	All
ESR [6]	5.66	7.12	11.94	8.24
3DDFA [70]	4.75	4.83	6.39	5.32
3D-FAN [4]	4.40	4.52	5.17	4.69
3DSTN [3]	3.55	3.92	5.21	4.23
3DDFA-PAMI [71]	4.11	4.38	5.16	4.55
PRNet [17]	4.19	4.69	5.45	4.77
3DDFA-V2 [20]	3.98	4.31	4.99	4.43
SynergyNet (our)	3.76	3.92	4.48	4.06

ing $\hat{\alpha}$ from L^c does not perform better than L^r due to the finer and more accurate structure of L^r . Results of Comparison 2 show that multi-attribute aggregation used at the landmark-to-3DMM does not bring better performance. We assume this is because the information has been joined at the landmark refinement phase.

Comparison to related work. We benchmark performance on the widely-used AFLW2000-3D. The two versions of annotations (original and reannotated) are used. To have a fair comparison, we show results on the two different sets separately and compare them with other reported performances. Table 4 shows the comparison on the original, and the reannotated version is in the supplementary. Our SynergyNet holds the best performance among all the related work on this standard dataset in Table 4. From the breakdown, our performance gain mainly comes from medium and large pose cases. We find that prior works encounter performance bottlenecks since they only regress 3DMM from images. However, referring to Table 1, MAFA has already shown the best performance compared with prior arts due to its ability to fuse multi-attribute features. Then the landmark-to-3DMM further shows lower errors to break through the performance bottleneck.

Following 3DDFA-V2 [20], we next use the AFLW full set for evaluation (21K testing images with 21-point landmarks). We show the comparison in Table 5. Our work has the best performance, especially with a performance gap over others on large-pose cases.

Table 6. **Ablative for face orientation estimation.** The same modules are studied in Table 1 for facial alignment. MAE of Euler angles in degree is reported.

Multi-Attribute Feature Aggregation	Landmark-to-3DMM	Yaw	Pitch	Roll	Mean
-	-	3.97	4.93	3.28	4.06
✓	-	3.72	4.37	2.88	3.65
-	✓	3.67	4.48	2.95	3.70
✓	✓	3.42	4.09	2.55	3.35

Table 7. **Study on different attributes and different regression targets for face orientation estimation.** The same structures are also studied in Table 2 for the alignment task. The first three rows study different attributes for aggregation. Based on the landmark refinement, Row 4 to 6 further study performance gains of different regression targets at the reverse representation stage.

Structures	Yaw	Pitch	Roll	Mean
Point-feature only	3.81	4.42	2.89	3.71
Point + image feature	3.72	4.39	2.85	3.66
MAFA	3.72	4.37	2.88	3.65
MAFA+ L \rightarrow 3D(pose)	3.58	4.06	2.57	3.40
MAFA+ L \rightarrow 3D(shape,expr)	3.47	4.23	2.59	3.43
MAFA+ L \rightarrow 3D(all)	3.42	4.09	2.55	3.35

4.2. Face Orientation Estimation Evaluation

Metrics and studies. Following the evaluation protocol in [7, 10, 40, 64], we calculate the mean absolute error (MAE) of predicted Euler angles in degrees. Groundtruth angles for each face in AFLW2000-3D are used, except for 31 samples whose yaw angles are outside the range [-99°, 99°]. We first study different combinations of the proposed modules in Table 6. From the results, finer landmark structures from MAFA lead to better orientation estimation. The landmark-to-3DMM further contributes to better performance since pose parameter regression from 3D points leads to more robust poses than 2D images.

We further study information fusion of using different attributes at MAFA and examine different parameter regression targets at the landmark-to-3DMM in Table 7. This analysis shows that more accurate orientation estimation is mainly benefited by pose regression at the reverse representation direction stage because Euler angles estimated from 3D representation are more robust than from 2D images. This breakdown also explains the performance gain of the landmark-to-3DMM in Table 6.

Benchmark comparison. We collect works that focus only on face orientation estimation [7, 10, 22, 40, 64] and 3DMM-based methods [20, 49, 70, 71]. The 3DMM-based works do not include evaluation of this task. To show the full benchmark list, we evaluate their methods using the released pretrained models. Table 8 shows that our method is the best and holds a performance gap over others. We display visual comparison in Fig. 5 and more studies and discussion in the supplementary.

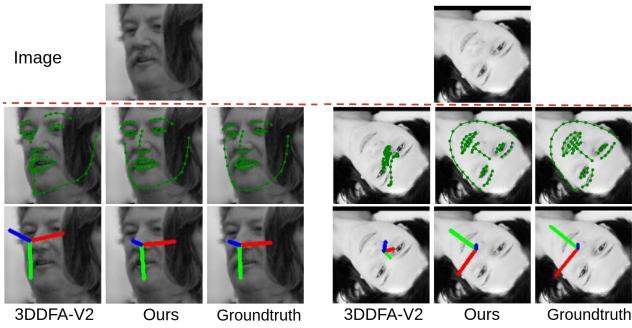


Figure 5. **Qualitative comparison of facial alignment and orientation estimation.** The case on the left is low-resolution, blurry, and thus challenging. The case on the right is of rare and extreme roll rotation. Our results show more robustness over 3DDFA-V2.

Table 8. **Face orientation estimation benchmark comparison on AFLW2000-3D.** PnP shows solving perspective-n-point problems using groundtruth landmarks. We do not include PRNet here since it does not infer face orientation directly and also obtains poses by PnP with predicted landmarks.

AFLW2000-3D	Yaw	Pitch	Roll	Mean
PnP-landmark	5.92	11.76	8.27	8.65
FAN-12 point [4]	6.36	12.30	8.71	9.12
HopeNet [40]	6.47	6.56	5.44	6.16
SSRNet-MD [65]	5.14	7.09	5.89	6.01
FSANet [64]	4.50	6.08	4.64	5.07
QuatNet [22]	3.97	5.62	3.92	4.15
TriNet [7]	4.20	5.77	4.04	3.97
RankPose [10]	2.99	4.75	3.25	3.66
3DDFA-TPAMI [71]	4.33	5.98	4.30	4.87
2DASL [49]	3.85	5.06	3.50	4.13
3DDFA-V2 [20]	4.06	5.26	3.48	4.27
SynergyNet (our)	3.42	4.09	2.55	3.35

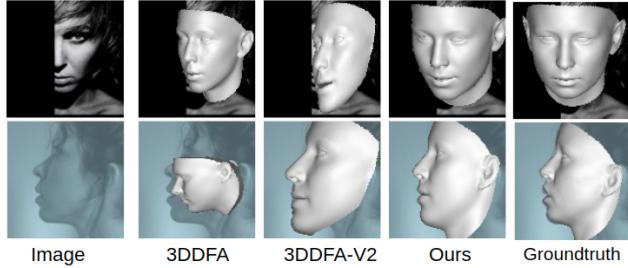


Figure 6. **Qualitative comparison of 3D face models.** Our results are robust to rare and out-of-domain face examples.

4.3. 3D Face Modeling Evaluation

Metrics and Comparison. Following [17, 20, 49], we first evaluate 3D face modeling on AFLW2000-3D. Two protocols are used. Protocol 1 suggested by [17, 49, 70] uses the iterative closet point (ICP) algorithm to register groundtruth 3D models and predicted models. NME of per-point error normalized by interocular distances is calculated. Protocol 2, suggested by [20] and also called dense alignment, calculates the per-point error normalized by bounding box sizes with groundtruth models aligned with images. Since ICP is not used, pose estimation would

Table 9. **3D face modeling comparison on AFLW2000-3D.** Refer to Sec. 4.3 for the protocol details.

Protocol-1	3DDFA [70]	DeFA [31]	PRNet [17]	2DASL [49]	SynergyNet (our)
NME	5.37	5.55	3.96	2.10	1.97
Protocol-2	3DDFA [20]	DeFA [31]	3DDFA-V2 [20]	SynergyNet (our)	
NME	6.56	6.04	4.18		4.06

Table 10. **3D face modeling comparison on Florence.** Point-to-plane RMSE is calculated for evaluation.

Florence	PRNet [17]	2DASL [49]	3DDFA-V2 [20]	SynergyNet (our)
RMSE	2.25	2.05	2.04	1.87

affect the performance under this protocol, and the NME would be higher. We illustrate numerical comparison in Table 9. The results show the ability of SynergyNet to recover 3D face models from monocular inputs and attain the best performance. In addition, we further exhibit visual comparison in Fig. 6. Our SynergyNet is capable of recovering 3D faces under rare and out-of-domain scenarios, such as heavily cropped or underwater cases.

Next, we evaluate the performance of 3D face modeling on Florence [2] with real scanned 3D faces. We follow the protocol from [17, 20], which renders 3D face models on different views with pitch of -15, 20, and 25 degrees and yaw of -80, -40, 0, 40, and 80 degrees. The rendered images are used as the test inputs. After reconstruction, face models are cropped to 95mm from the nose tip, and ICP is performed to calculate point-to-plane root mean square error (RMSE) with cropped groundtruth. Numerical results are shown in Table 10. An error curve that shows our robustness to yaw angle changes and a qualitative comparison on Florence are displayed in the supplementary.

5. Conclusion

This work proposes a synergy process that utilizes the relation between 3D landmarks and 3DMM parameters, and they collaboratively contribute to better performance. We establish a representation cycle, including forward direction, from 3DMM to 3D landmarks, and reverse representation direction, from 3D landmarks to 3DMM. Specifically, We propose two modules, multi-attribute feature aggregation for landmark refinement and the landmark-to-3DMM module. Extensive experiments validate our network design, and we show a detailed performance breakdown for each included attribute and regression target. Our SynergyNet only adopts simple network operations and attains superior performance, making it a fast, accurate, and easy-to-implement method.

Acknowledgement

We sincerely thank Jingjing Zheng, Jim Thomas, and Cheng-Hao Kuo for their detailed feedback on this paper.

References

- [1] Nvidia maxine cloud-ai video-streaming platform. https://developer.nvidia.com/maxine?ncid=so-yout-26905#cid=d113_so-yout_en-us. 3
- [2] Andrew D. Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2d/3d hybrid face dataset. In *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*, J-HGBU '11. ACM, 2011. 5, 8
- [3] Chandrasekhar Bhagavatula, Chenchen Zhu, Khoa Luu, and Marios Savvides. Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In *ICCV*, pages 3980–3989, 2017. 7
- [4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017. 2, 5, 7, 8, 1
- [5] Chen Cao, Yanlin Weng, Shun Zhou, Yiyi Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 20(3):413–425, 2013. 5
- [6] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision (IJCV)*, 107(2):177–190, 2014. 7
- [7] Zhiwen Cao, Zongcheng Chu, Dongfang Liu, and Yingjie Chen. A vector-based representation to enhance head pose estimation. In *WACV*, 2021. 3, 7, 8
- [8] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gérard Medioni. Deep, landmark-free fame: Face alignment, modeling, and expression estimation. *International Journal of Computer Vision (IJCV)*. 1
- [9] Grigoris G Chrysos, Epameinondas Antonakos, Patrick Snape, Akshay Asthana, and Stefanos Zafeiriou. A comprehensive performance evaluation of deformable face tracking “in-the-wild”. *International Journal of Computer Vision (IJCV)*, 2018. 1
- [10] Donggen Dai, Wangkit Wong, and Zhuojun Chen. Rankpose: Learning generalised feature with rank supervision for head pose estimation. In *BMVC*, 2020. 7, 8
- [11] Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou. Masked face recognition challenge: The insight-face track report. In *ICCV Workshops*, 2021. 1
- [12] Jiankang Deng, Anastasios Roussos, Grigoris Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *International Journal of Computer Vision (IJCV)*, 127(6-7):599–624, 2019. 1
- [13] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPR Workshops*, pages 0–0, 2019. 1, 4, 3
- [14] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *CVPR*, pages 379–388, 2018. 2
- [15] Xuanyi Dong and Yi Yang. Teacher supervises students how to learn from partially labeled images for facial landmark detection. In *CVPR*, pages 783–792, 2019. 2
- [16] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 2021. 3
- [17] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, pages 534–551, 2018. 1, 2, 5, 6, 7, 8, 3, 4
- [18] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *CVPR*, pages 2235–2245, 2018. 2
- [19] James D Foley, Foley Dan Van, Andries Van Dam, Steven K Feiner, John F Hughes, Edward Angel, and J Hughes. *Computer graphics: principles and practice*, volume 12110. Addison-Wesley Professional, 1996. 4
- [20] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *ECCV*, 2020. 1, 2, 3, 4, 5, 6, 7, 8
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1
- [22] Heng-Wei Hsu, Tung-Yu Wu, Sheng Wan, Wing Hung Wong, and Chen-Yi Lee. Quatnet: Quaternion-based head pose estimation with multiregression loss. *IEEE Transactions on Multimedia (TMM)*, 2018. 3, 7, 8
- [23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 4
- [24] Aaron S. Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *ICCV*, Oct 2017. 2, 4
- [25] Aniwat Juhong and C Pintavirooj. Face recognition based on facial landmark detection. In *2017 10th Biomedical Engineering International Conference (BMEiCON)*, pages 1–4. IEEE, 2017. 1
- [26] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, pages 1867–1874, 2014. 2
- [27] Hyeyoung Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 1
- [28] Séverin Lemaignan, Fernando Garcia, Alexis Jacq, and Pierre Dillenbourg. From real-time attention assessment to “with-me-ness” in human-robot interaction. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 157–164. Ieee, 2016. 3
- [29] Chung-Ching Lin and Ying Hung. A prior-less method for multi-face tracking in unconstrained videos. In *CVPR*, 2018. 1

- [30] Qingshan Liu, Jing Yang, Jiankang Deng, and Kaihua Zhang. Robust facial landmark tracking via cascade regression. *Pattern Recognition (PR)*, 66:53–62, 2017. 1
- [31] Yaojie Liu, Amin Jourabloo, William Ren, and Xiaoming Liu. Dense face alignment. In *ICCV*, pages 1619–1628, 2017. 8
- [32] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *CVPR*, pages 3317–3326, 2017. 1
- [33] Xinzhu Ma, Shinan Liu, Zhiyi Xia, Hongwen Zhang, Xingyu Zeng, and Wanli Ouyang. Rethinking pseudo-lidar representation. *ECCV*, 2020. 3
- [34] Peter M. Roth Martin Koestinger, Paul Wohlhart and Horst Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011. 5
- [35] Sankha S Mukherjee and Neil Martin Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia (TMM)*, 17(11):2094–2107, 2015. 3
- [36] Oskar Palinko, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *IROS*, pages 5048–5054. IEEE, 2016. 3
- [37] 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. 3, 5
- [38] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 3, 2
- [39] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, pages 5099–5108, 2017. 3, 2
- [40] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *CVPR Workshops*, pages 2074–2083, 2018. 3, 7, 8
- [41] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018. 1
- [42] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *CVPR*, pages 7763–7772, 2019. 3
- [43] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Ming-min Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *ECCV*, 2020. 2, 1, 3
- [44] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kos-saifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *CVPR Workshops*, pages 50–58, 2015. 5, 4
- [45] Jiazheng Shi, Ashok Samal, and David Marx. How effective are landmarks and their geometry for face recognition? *Computer vision and image understanding (CVIU)*, 102(2):117–133, 2006. 1, 3
- [46] Bin Sun, Ming Shao, Siyu Xia, and Yun Fu. Deep evolutionary 3d diffusion heat maps for large-pose face alignment. In *BMVC*, page 256, 2018. 1, 3
- [47] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgarib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Fml: Face model learning from videos. In *CVPR*, pages 10812–10822, 2019. 3
- [48] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *CVPR*, pages 2549–2559, 2018. 3
- [49] Xiaoguang Tu, Jian Zhao, Mei Xie, Zihang Jiang, Akshaya Balamurugan, Yao Luo, Yang Zhao, Lingxiao He, Zheng Ma, and Jiashi Feng. 3d face reconstruction from a single image assisted by 2d face images in the wild. *IEEE Transactions on Multimedia (TMM)*, 2020. 1, 2, 3, 4, 7, 8
- [50] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *CVPR*, pages 5163–5172, 2017. 1
- [51] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *CVPR*, pages 3935–3944, 2018. 1
- [52] Kang Wang, Rui Zhao, and Qiang Ji. Human computer interaction with head pose, eye gaze and body gestures. In *FG*, pages 789–789. IEEE, 2018. 3
- [53] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 4
- [54] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. *CVPR*, 2021. 1, 3
- [55] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, pages 8445–8453, 2019. 3
- [56] Yandong Wen, Weiyang Liu, Bhiksha Raj, and Rita Singh. Self-supervised 3d face reconstruction via conditional estimation. In *CVPR*, 2021. 3
- [57] Cho-Ying Wu and Jian-Jiun Ding. Occlusion pattern-based dictionary for robust face recognition. In *ICME*, 2016. 1
- [58] Cho-Ying Wu and Jian Jiun Ding. Occluded face recognition using low-rank regression with generalized gradient direction. *Pattern Recognition (PR)*, 2018. 1
- [59] Cho-Ying Wu, Xiaoyan Hu, Michael Happold, Qiang-geng Xu, and Ulrich Neumann. Geometry-aware instance segmentation with disparity maps. *arXiv preprint arXiv:2006.07802*, 2020. 3

- [60] Cho-Ying Wu and Ulrich Neumann. Efficient multi-domain dictionary learning with gans. In *GlobalSIP*, 2019. [1](#)
- [61] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *CVPR*, pages 959–968, 2019. [1](#), [3](#), [4](#)
- [62] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, pages 2129–2138, 2018. [2](#)
- [63] Qiangeng Xu, Xudong Sun, Cho-Ying Wu, Panqu Wang, and Ulrich Neumann. Grid-gcn for fast and scalable point cloud learning. In *CVPR*, pages 5661–5670, 2020. [3](#), [2](#)
- [64] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *CVPR*, pages 1087–1096, 2019. [3](#), [7](#), [8](#)
- [65] Tsun-Yi Yang, Yi-Husan Hunag, Yen-Yu Lin, Pi-Cheng Hsiu, and Yung-Yu Chuang. Ssr-net: A compact soft stage-wise regression network for age estimation. In *IJCAI*, 2018. [8](#)
- [66] Jordan Yaniv, Yael Newman, and Ariel Shamir. The face of art: landmark detection and geometric style in portraits. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019. [5](#)
- [67] Ronald Yu, Shunsuke Saito, Haoxiang Li, Duygu Ceylan, and Hao Li. Learning dense facial correspondences in unconstrained images. In *ICCV*, pages 4723–4732, 2017. [7](#)
- [68] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020. [1](#)
- [69] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters (SPL)*, 23(10):1499–1503, 2016. [1](#)
- [70] 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. [1](#), [2](#), [3](#), [4](#), [5](#), [7](#), [8](#)
- [71] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2019. [1](#), [2](#), [3](#), [4](#), [6](#), [7](#), [8](#)
- [72] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In *NeurIPS*, pages 2595–2603, 2010. [5](#)
- [73] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer Graphics Forum*, volume 37, pages 523–550. Wiley Online Library, 2018. [1](#)

Supplemental Materials:

A. Overview

We document this supplementary into the following sections. In Section **B**, we provide details of our network architectures and loss weights for training. We further present a study for network backbone choices and a study for showing that our performance gain is not simply from using more network parameters. In Section **C**, evaluation of facial alignment on AFLW2000-3D reannotation version is exhibited. In Section **D**, we present analysis and discussion on the reverse representation direction. In Section **E**, we dig into performance comparison using L^r and L^c . In Section **F**, an error curve and visual comparison of 3D face modeling on Florence are illustrated. In Section **G**, we describe texture synthesis using introduced UV-texture GAN and further compare with textures from 3DMM fitting. In Section **H**, we add more qualitative results from our face geometry prediction using the 300VW video dataset and Artistic Faces.

B. Network Architecture, Hyper-Parameters, and Network Parameter Studies

Details of architecture. Based on Fig. 2 in the main paper (pipeline graph of our SynergyNet), detailed network architecture is described here. Following [20], we use MobileNet-V2 as the backbone for 3DMM regression from images. The latent image features z after the global max-pooling is 1280-dim. The pose, shape, and expression decoders are fully-connected (FC) layers with input z and output 3DMM parameters of 12 (α_p), 40 (α_s), and 10 (α_e) dimensions for pose, shape, and expression.

Fig. S1 shows the network architecture of multi-attribute feature aggregation. Aggregation of the latent image features, α_s , and α_e and global point features forms a 2354-dim feature vector. We repeat this vector and append it to the low-level point features to obtain multi-attribute point features, whose size is 68×2418 . Later we use another MLP-block to obtain refined landmarks L^r .

Fig. S2 illustrates the architecture of the landmark-to-3DMM module. With L^r as the module input, this module reverses the representation direction and regresses 3DMM parameters $\hat{\alpha}$, also referred to as landmark geometry in the paper.

Loss weights. For weights of loss terms (Eq.7 in paper), we choose $\lambda_1 = 0.02$, $\lambda_2 = 0.03$, $\lambda_3 = 0.02$, and $\lambda_4 = 0.001$ for training.

Study on the backbone for 3DMM regression from images. We next conduct a study on the backbone choices for facial alignment using the AFLW full set. We select MobileNet-V2 [41], ResNet50 [21], ResNet101 [21], ResNeSt50 [68], and ResNeSt101 [68] for comparison.

Table S2. **Backbone study on the AFLW full set.** We compare MobileNet [41], ResNet [21], and ResNeSt [68], a ResNet variant with split-attention.

Backbone	0 to 30	30 to 60	60 to 90	All
MobileNet	3.86	4.13	4.61	4.20
ResNet50	3.76	3.92	4.48	4.06
ResNeSt50	3.76	3.92	4.52	4.07
ResNet101	3.90	4.14	5.08	4.38
ResNeSt101	3.78	4.04	4.62	4.15

From Table S2, one could see that the 101-residual layer network is too deep, and thus the performance drops compared with the 50-residual layer network. ResNeSt, a split-attention variant of ResNet, can improve the performance for the 101-residual layer case, but the performance of ResNeSt50 is on par with ResNet50. We think this is because the split-attention scales better and remedies the undesirable effects of deeper networks, which are described in their work [68].

Study on the number of network parameters. We further conduct a study to verify the effectiveness of the introduced multi-attribute feature aggregation (MAFA) for landmark refinement and the landmark-to-3DMM modules. The following experiment shows that our performance gain comes from designing the two proposed modules in our synergy process, rather than simply using more network parameters.

By using MobileNet-V2 as the face image encoder backbone, network parameters of our SynergyNet amount to 3.8M (3.0M for the backbone) and 0.8M for the MAFA and landmark-to-3DMM modules). We build another baseline model, Image \rightarrow 3DMM (larger), that only contains 3DMM parameter regression from images using more network parameters. We add additional MLP layers with ReLU and BN, which amount to 0.8M parameters, after the image bottleneck feature z for regressing α_p , α_s , and α_e . Thus, this baseline model and our SynergyNet have approximately the same number of network parameters.

In Table S2, we show experiments with the baseline model on facial alignment and face orientation estimation. More network parameter adoption for regressing 3DMM from images only leads to minor improvements. Especially for facial alignment, using extra 0.8M parameters of MLPs only gives 0.02 overall performance gain. The results validate the designed synergy process. Without the proposed modules, using more parameters only brings minor improvements.

C. Evaluation on Reannotated AFLW2000-3D

Reannotation of AFLW2000-3D is provided in LS3DW [4]. Few works, Deng *et al* [13], DHM [46] and MCG-Net [43], report their performance on the annotated version. To aggregate more results for this study, we also in-

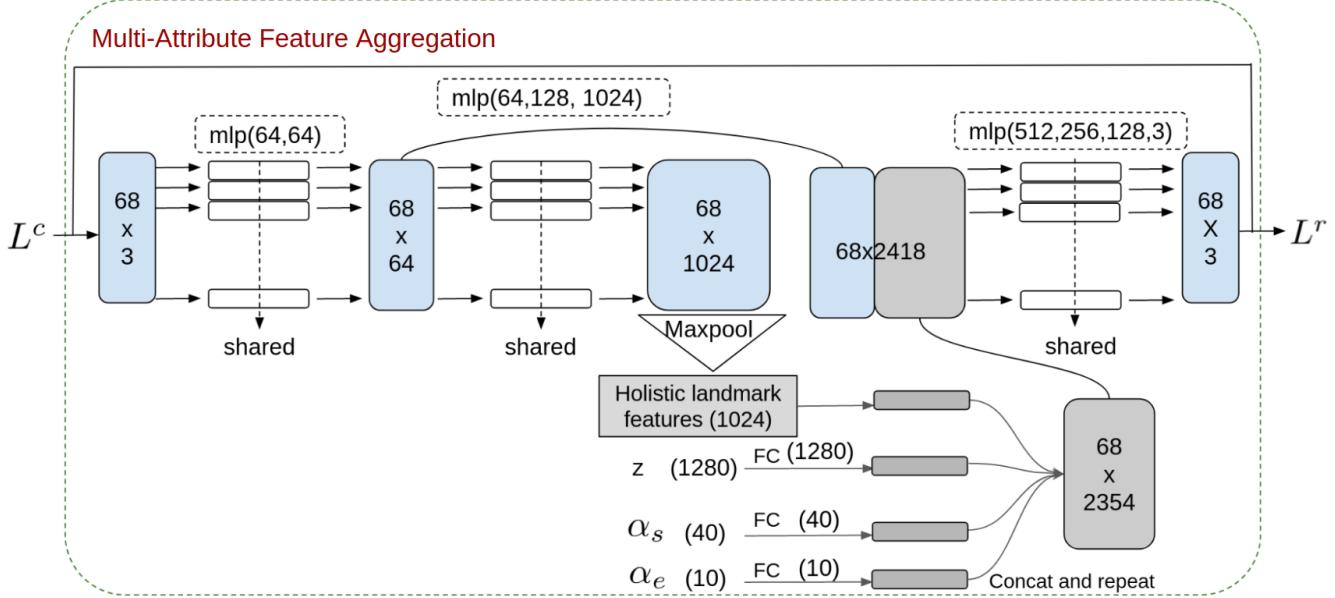


Figure S1. **Detailed structure of MAFA.** $\text{mlp}(64,64)$ means two MLP layers with output channel sizes 64 and 64. ReLU and batch normalization are used for each layer. The notations correspond to those in the main paper. (z is latent image feature, α_s and α_e are 3DMM shape and expression parameters regressed from images, L^c and L^r are 3D landmarks before and after the landmark refinement.)

Table S2. **Comparison with the baseline that simply uses more network parameters.** The first table shows results on AFLW2000-3D Original for facial alignment, and the second table shows results also on AFLW2000-3D for face orientation estimation. # of params means the number of network parameters. More network parameter use for the baseline 3DMM regression from images only results in a limited performance gain. The experiment validates that the performance gain of our SynergyNet is not simply from more parameter adoption.

Structures	# of params	0 to 30	30 to 60	60 to 90	All
Image → 3DMM	3.0M	2.99	3.80	4.86	3.88
Image → 3DMM (larger)	3.8M	2.98	3.82	4.77	3.86
SynergyNet	3.8M	2.66	3.30	4.27	3.41
Structures	# of params	Yaw	Pitch	Roll	Mean
Image → 3DMM	3.0M	3.97	4.93	3.28	4.06
Image → 3DMM (larger)	3.8M	3.80	4.62	2.84	3.75
SynergyNet	3.8M	3.42	4.09	2.55	3.35

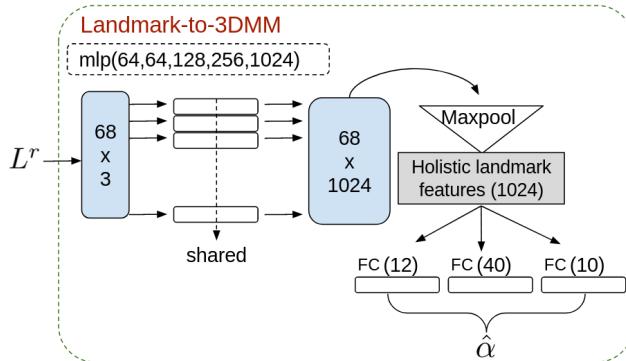


Figure S2. **Detailed structure of the landmark-to-3DMM module.** The notations correspond to those in the main paper. L^r is refined 3D landmarks, and $\hat{\alpha}$ is regressed landmark geometry.

clude evaluation of 3DDFA [70], PRNet [17], and 3DDFA-V2 [20] using their pretrained models. From Table S2, nor-

malized mean errors (NMEs) are generally lower than using the original annotation. This shows the higher quality of the reannotation. Among the methods for comparison, our result is the best and holds a performance gap over others. Compared with PRNet, the second-best method in the table, our improvements are derived from large pose cases.

D. Discussion on Reverse Representation Direction

Why use sparse landmarks rather than full vertices?

Landmark geometry $\hat{\alpha}$ in Sec.3.3 of the main paper describes revealing facial geometry underlying in sparse 3D landmarks. In contrast to sparse landmarks (68 points in our work), mesh from BFM Face includes 53.5K vertices (45K if excluding the neck and ears). When surveying on point processing, research usually adopts only 1024 or 2048 points [38, 39, 63]. Much denser points are inefficient for

Table S2. Comparison on AFLW2000-3D Reannotation. Our method has the best alignment result and holds a performance gap over others.

AFLW2000-3D Reannotated	0 to 30	30 to 60	60 to 90	All
DHM [46]	2.28	3.10	6.95	4.11
3DDFA [70]	2.84	3.52	5.15	3.83
PRNet [17]	2.35	2.78	4.22	3.11
MGCNet [43]	2.72	3.12	3.76	3.20
Deng <i>et al</i> [13]	2.56	3.11	4.45	3.37
3DDFA-V2 [20]	2.84	3.03	4.13	3.33
SynergyNet (our)	2.05	2.49	3.52	2.65

point processing, and the accommodation is also limited by GPU memory. On the other hand, because facial alignment is considered as an upstream task for the downstream application such as face recognition [45] or recent streaming video compression [1,54], high efficiency is more desirable.

Although a point-sampling strategy could be used for downsizing, 3D landmarks are very efficient and compact for expressing facial traits and outlines. Therefore, 3D landmarks are desirable for predicting facial geometry, and our focus of this work is to exert the synergy between facial landmarks and 3DMM parameters, which collaboratively contribute to better performance.

Advantage of the reverse representation direction. Compared with 2D images, 3D sparse landmarks describe facial traits and approximate face outlines. Although landmarks are sparse, the representation provides another view to learn facial geometry and complements with 3DMM regressed from images. For example, facial geometry for large pose cases is hard to estimate from the 2D due to self-occlusion. Further, the face orientation is defined in the 3D space; thus, it is more advantageous to estimate face orientation from 3D points, whose learning paradigm provides less ambiguity. From Table 2 and 7 in the main paper that study contributions for each regression target at the $L \rightarrow 3D$ stage, MAFA+ $L \rightarrow 3D$ (all) improves the performance on facial alignment and face orientation estimation. The results show the ability of the reverse representation direction.

We also conduct evaluations using $\hat{\alpha}$ as the output on facial alignment and face orientation estimation using AFLW2000-3D. The 3D landmarks reconstructed by $\hat{\alpha}$ and the face orientation converted from its pose parameter $\hat{\alpha}_p$ attain an NME of 6.48 on the alignment and an MAE of 5.76 on the orientation estimation. The results are reasonable since the direct input to the landmark-to-3DMM module is L^r , 68-point sparse landmarks that present only approximate facial traits and outlines. However, these numerical results are comparable with some methods in Table 4 and 8 of the main paper. The results validate our training strategy so that *3DMM estimation directly from sparse landmarks achieves on par performance with some studies for 3DMM*

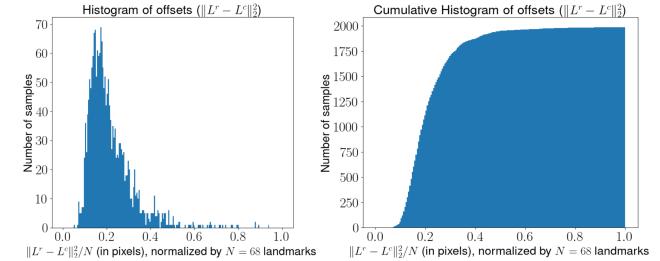


Figure S3. Histogram and cumulative histogram of the offset term: $\|L^r - L^c\|_2^2$. These plots show the difference of the landmark set before and after refinement.

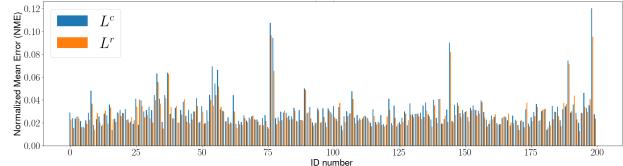


Figure S4. NME of random 200 people: The lower the better. Zoom in for the best view.

regression from images, whose information lies on dense grids.

E. Evaluation on L^r v.s. L^c

From Table 1 and 6 in the paper (1st row: without refinement and use L^c for evaluation; 2nd row: with refinement and use L^r for evaluation), one can observe from these two tables that with the refinement branch, the performance is significantly improved. Specifically, Table 1-facial alignment error: 3.88 (L^c) to 3.49 (L^r); Table 6-face orientation error: 4.06 (L^c) to 3.65 (L^r).

We further show the histograms of offsets $\|L^r - L^c\|_2^2$ in Fig. S3. We use AFLW2000-3D for evaluation. One can see that the difference of L^r and L^c peaks at 0.22 pixels for each landmark. This matches the purpose of refinement that by predicting each landmark more precisely, the total improvements can break through the performance bottleneck in the benchmark list (paper Table 4). In addition, we plot the normalized mean error (NME) for random 200 people in Fig. S4 (zoom in for the best view). One can find that the blue bars are higher overall, meaning the error using L^c is higher than L^r .

F. 3D Faces on Florence

Based on the Florence experiment in Section 4.3 and Table 10 of the main paper, we further show an error curve comparison in Fig. S5 for 3D face modeling. Our method is robust to pose changes and attains a *nearly flat* error curve. Although the results of 2DASL for low and medium cases are close to ours, they are not robust for large pose cases.

We visualize the reconstructed meshes and compare with the current best-performing 3DMM-based method

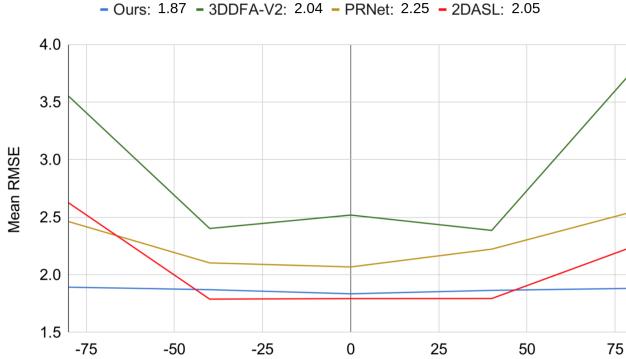


Figure S5. Error curve for 3D face modeling by yaw angle on the Florence dataset. The numbers at the top are mean RMSE over all testing data. Our method is rather robust to pose changes and attains the lowest overall point-to-plane RMSE.

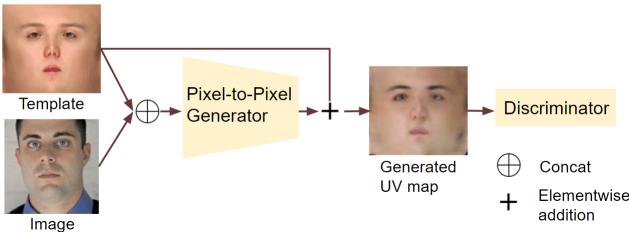


Figure S6. UV-texture GAN. The generator produces a UV map from a fixed template and an input image. The generated UV map combines structures of the template and the skin color of the image.

(3DDFA-V2 [20]) and UV-position-based method (PRNet [17]) in Fig. S9. We mark point-to-plane RMSEs beside each face model. Our reconstructed faces show narrower eye-to-side distances, higher cheeks, and pointed chins from the upper example. These features are consistent with the groundtruth model. On the other hand, 3DDFA-V2 shows wider faces, unapparent cheeks, and non-pointed chins; thus, their errors are higher. Besides, for a cropping range of 95mm from the nose tip, 3DDFA-V2 shows more forehead areas than the groundtruth model, which means the geometry prediction is inaccurate. PRNet is not 3DMM-based. Although PRNet has higher flexibility to predict per-vertex deformation due to its non-parametric nature, it is also harder to estimate a precise 3D face via vertex regression on a UV-position map. From the lower example, our faces are wider and consistent with the groundtruth. In contrast, 3DDFA-V2 shows more elongated shapes for large pose cases. PRNet also shows skewed faces under large pose scenarios.

G. Texture Synthesis

Most previous works for 3D facial alignment via 3D face modeling mainly focus on the geometry [?, 17, 20, 49, 70, 71]. To get more realistic 3D face models, here we also conduct

a smaller study on texture synthesis based on our predicted 3D face models.

Similar to 3DMM fitting for 3D faces, as illustrated in the main paper Eq.(1), textures can also be synthesized by adding a mean texture term and a multiplication term of texture basis and parameters. For example, BFM Face contains texture parameters with a 199-dim texture basis. However, 3DMM texture fitting usually produces over-smooth textures that lack reality. (See examples in Fig. S8).

Here we introduce a simple but effective UV-texture Generative Adversarial Network (UV-texture GAN) for texture synthesis. The model structure is illustrated in Fig. S6. UV mapping [19] involves per-vertex color mappings from UV-texture maps. Each vertex is associated with its (u, v) -coordinate for querying vertex color from the three-channel UV-texture maps.

The introduced UV-texture GAN adopts a pixel-to-pixel image translator that transforms unstructured in-the-wild images to a canonical UV space to generate the UV-texture maps from images. However, pixel-to-pixel style transfer [23, 53] retains input image structures, such as salient object outlines, and produces a different style artifact. It is hard to map unconstrained face images onto the canonical UV space by direct pixel-to-pixel translators. To resolve the issue, we further feed a template UV map together with a face image as inputs to the generator (Fig. S6). The template is projected from the mean texture of BFM Face. Further, we shortcut the template to the output for facilitating the training procedure, where the generator learns a mapping from the six-channel input to the residual UV space. We display the ability of the template in Fig. S7.

To form our training set, we collect about 2K in-the-wild frontal face images and warp the faces onto the UV space with the aids of facial landmarks. The generator and discriminator architectures are the same as pix2pix model [23]. Least-square GAN (LSGAN) [?] is used as the loss for training. We train the network with 300 epochs. Adam is adopted as the optimizer with an initial learning rate of 0.0002. After 100 epochs, the learning rate starts to drop linearly to 0.

We show a comparison in Fig. S8 for synthesized textures from the introduced UV-texture GAN and conventional 3DMM texture fitting. Results of UV-texture GAN are more realistic and are not over-smooth compared with textures from 3DMM fitting. Skin colors are more similar to images since the introduced UV-texture GAN combines hues from images and structures from the template to produce more realistic textures.

H. More Qualitative Results

Here we further show more qualitative results on the 300VW dataset for talks or interview videos [44] in Fig. S10, S11, S12, S13 and Artistic Faces (AF) for different

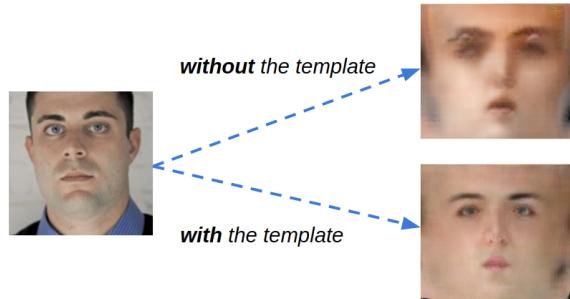


Figure S7. Effects of using the template in the UV-texture GAN.
Without the template shown in Fig. S6, facial traits such as eyes and mouth are blurry and inaccurate.

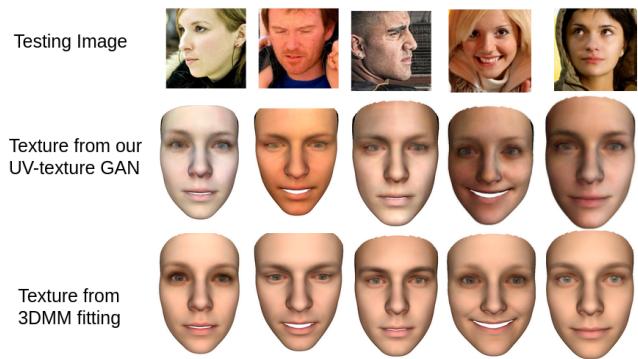


Figure S8. Synthesized texture comparison. Textures synthesized by the introduced UV-texture GAN are more realistic than textures from 3DMM fitting.

artistic style faces [66] in Fig. S14, S15.

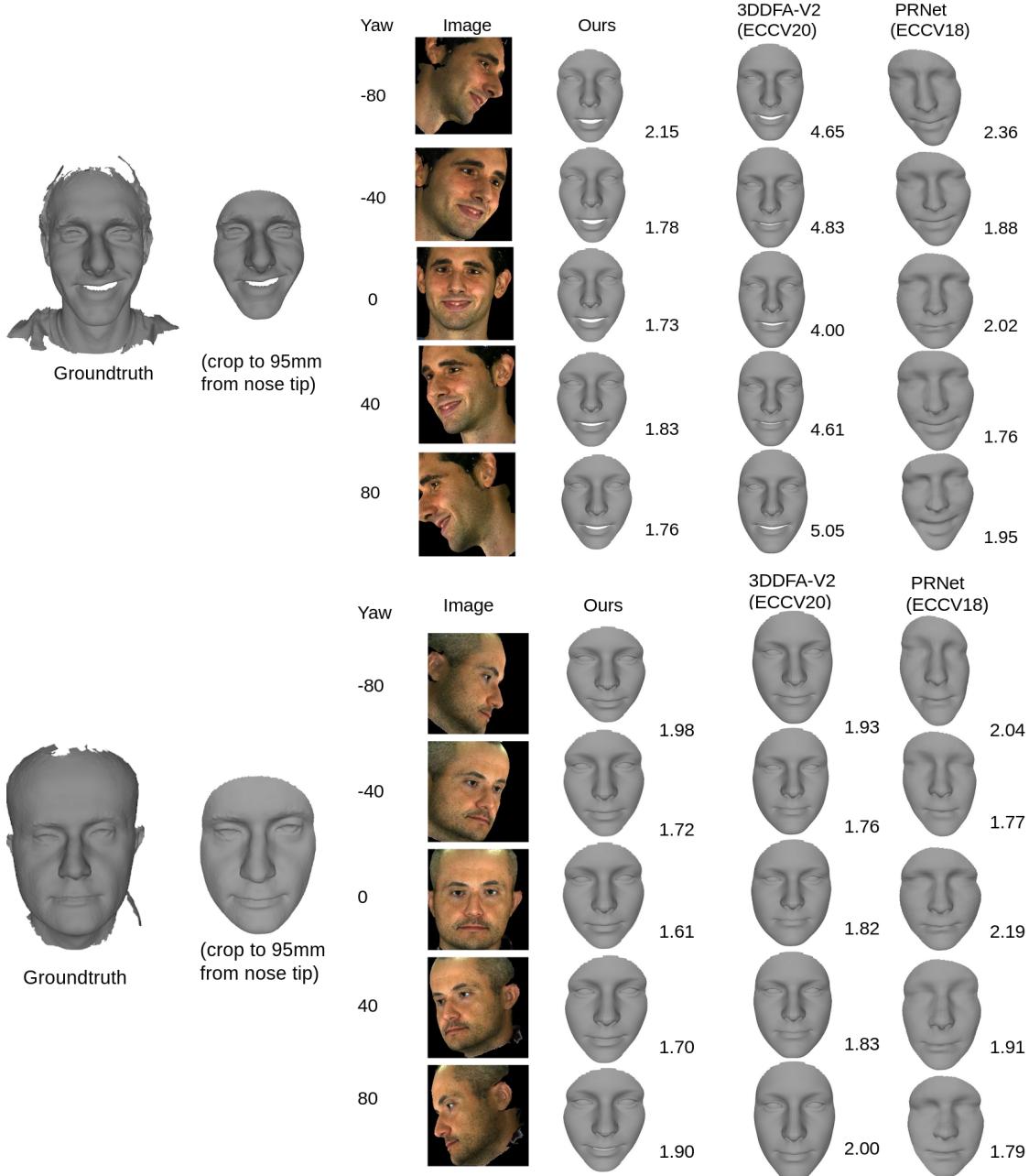


Figure S9. Reconstructed face comparison by yaw angle on examples from the Florence dataset. Numbers beside the reconstructed models are their normalized point-to-plane RMSEs. Our results are robust to pose changes. For the upper example, 3DDFA-V2 shows wider faces, unapparent cheeks, non-pointed chins, and larger forehead areas with a cropping range of 95mm from the nose tip; therefore, their results hold higher errors. PRNet shows imprecise facial structures. In addition, their faces are twisted for large pose cases. For the lower example, the groundtruth face is wider, but 3DDFA-V2 shows more elongated faces, and PRNet predictions are unreliable. Our results are more similar to the groundtruth shape.



Figure S10. **Results of 3D geometry prediction on 300VW from our method.** Row 1-4: images, 3D landmarks, face orientation, 3D faces.



Figure S11. (Continued) **Results of 3D geometry prediction on 300VW from our method.** Our result is robust to motion blur for the right-hand-side case.

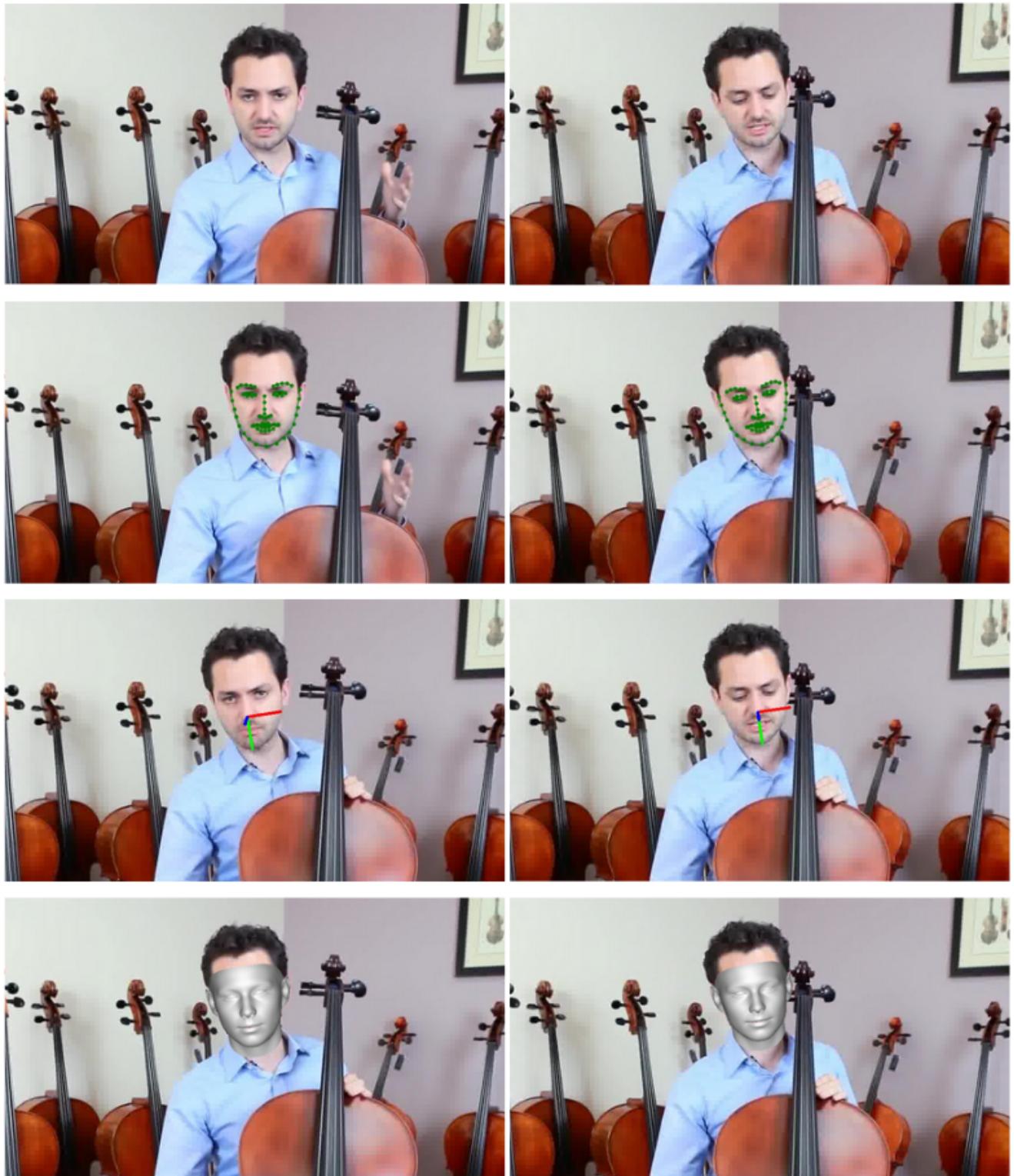


Figure S12. (Continued) Results of 3D geometry prediction on 300VW from our method.

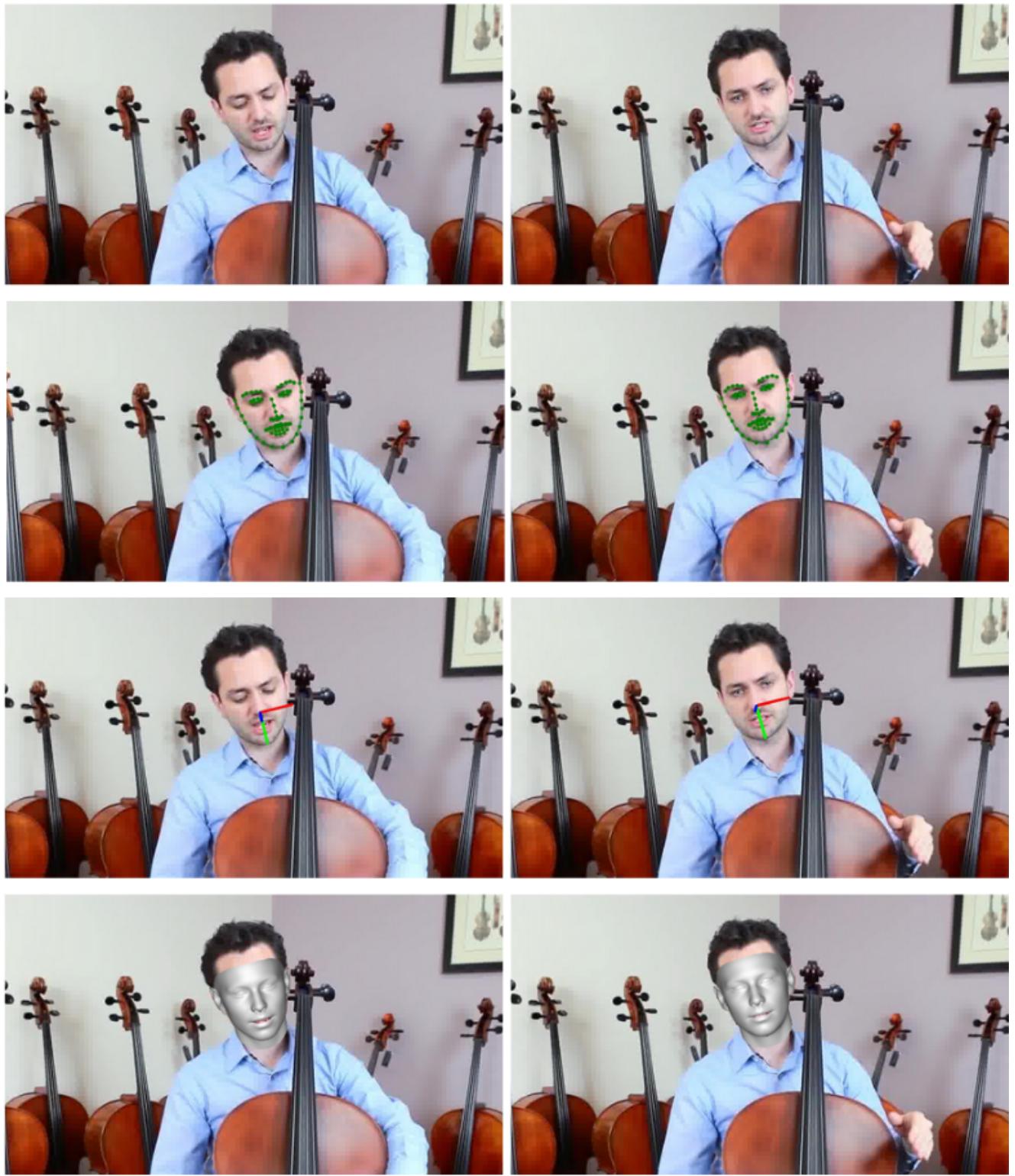


Figure S13. (Continued) Results of 3D geometry prediction on 300VW from our method.

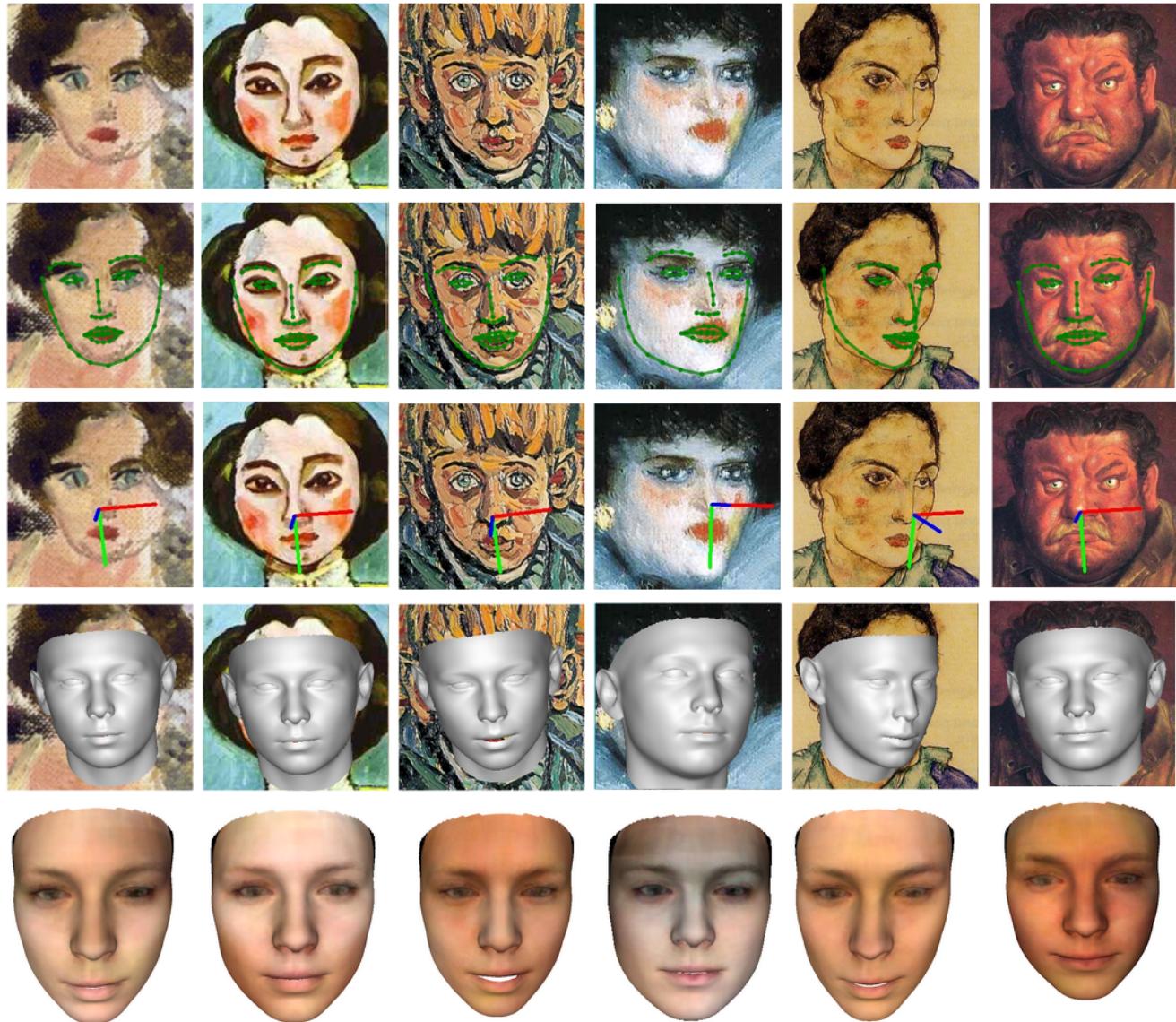


Figure S14. **Results of 3D geometry prediction on Artistic Faces from our method.** Row 1-5: images, 3D landmarks, face orientation, 3D faces, textures.



Figure S15. (Continued) Results of 3D geometry prediction on Artistic Faces from our method.