6DoF Head Pose Estimation through Explicit Bidirectional Interaction with Face Geometry

Sungho Chun and Ju Yong Chang

Department of ECE, Kwangwoon University, Korea
{asw9161,jychang}@kw.ac.kr

Abstract. This study addresses the nuanced challenge of estimating head translations within the context of six-degrees-of-freedom (6DoF) head pose estimation, placing emphasis on this aspect over the more commonly studied head rotations. Identifying a gap in existing methodologies, we recognized the underutilized potential synergy between facial geometry and head translation. To bridge this gap, we propose a novel approach called the head **T**ranslation, **R**otation, and face **G**eometry network (TRG), which stands out for its explicit bidirectional interaction structure. This structure has been carefully designed to leverage the complementary relationship between face geometry and head translation, marking a significant advancement in the field of head pose estimation. Our contributions also include the development of a strategy for estimating bounding box correction parameters and a technique for aligning landmarks to image. Both of these innovations demonstrate superior performance in 6DoF head pose estimation tasks. Extensive experiments conducted on ARKitFace and BIWI datasets confirm that the proposed method outperforms current state-of-the-art techniques. Codes are released at `https://github.com/asw91666/TRG-Release`.

Keywords: 6DoF head pose estimation · bidirectional interaction · landmark-based approach

1 Introduction

Six-degrees-of-freedom (6DoF) head pose estimation is a crucial concern in both computer vision and graphics communities owing to its broad applications in augmented/virtual reality, vehicular monitoring systems, and sports analytics. Despite its prominence, existing studies [3, 4, 21, 27, 38, 41, 45–47] have primarily focused on estimating head orientation, whereas research on head translation estimation has not received as much attention. Some studies [1, 44] have estimated pseudo-depth calculated from fitted data [55] without exploring methods to estimate the actual distance between the camera and head.

Estimating head translation from a single image using learning-based methods poses significant challenges, which can be attributed to roughly two reasons. First, head translation estimation depends on real-scale face geometry. However,

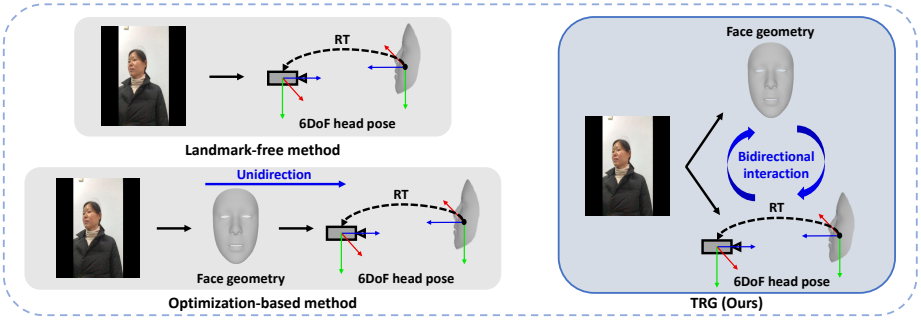

Fig. 1: Methods of inferring 6DoF head pose. The landmark-free approach [1] directly calculates the head pose from the image. Optimization-based methods [18, 25, 56] first predict face geometry, and then calculate the head pose. In contrast, TRG simultaneously estimates both face geometry and head pose to leverage the synergy between them.

the estimation of real-scale face geometry suffers from head translation ambiguities. In other words, the estimation of head translation and the estimation of actual size face geometry are strongly correlated, and there exists ambiguity due to their mutual absence. Second, learning-based head translation estimation encounters severe generalization issues with out-of-distribution data. Unlike head rotation, the range of head translation is infinite, necessitating a generalization strategy to address it.

However, existing works [1, 18, 25, 56] do not address the aforementioned issues. Fig. 1 provides an overview of the 6DoF head pose estimation methods used by existing models. In [18, 25, 56], face geometry is first inferred from an image, followed by the calculation of the 6DoF head pose using an optimization-based method. In other words, these methods [18,25,56] do not model the transfer of information from head pose to face geometry. This unidirectional information transfer method may face difficulties in predicting the actual size face geometry due to the absence of depth information. Consequently, the resulting face prior could create a vicious cycle, further reducing the accuracy of head translation prediction.

Landmark-free approach [1] estimates head translation directly from an image using a learning-based method; however, it does not utilize face geometry information during the inference process. Directly estimating head depth from an image is highly non-linear, making the landmark-free approach challenging for estimating head translation.

To overcome the limitations of existing models [1,18,25,56], we propose a head **T**ranslation, **R**otation, and face **G**eometry network (*TRG*), which is a landmark-based method for estimating a 6DoF head pose. The TRG is designed with an explicit bidirectional interaction structure that leverages the complementary characteristics between the 6DoF head pose and face geometry. Specifically, we

propose a method that simultaneously estimates the head pose and dense 3D landmarks, using each other's information to iteratively improve one another.

To achieve generalizable head translation estimation, TRG does not directly estimate depth, but utilizes the position and size information of the bounding box. The center coordinates of the bounding box are typically well-aligned with the coordinates of the head center, and the size of the bounding box inversely reflects the head's depth. These relationships make the bounding box a useful tool for estimating head translation in 3D space. However, reliance on the bounding box alone is insufficient. This is due to potential misalignments between the bounding box center and the head center, and the bounding box size being influenced by factors beyond depth, such as face size and head rotation. To address these discrepancies, we propose to estimate bounding box correction parameters and calculate head translation using these parameters and bounding box information. The proposed method has been found to achieve high accuracy and to be robust even for out-of-distribution data.

Additionally, TRG aligns the estimated 3D landmarks with the image through perspective projection. By iterating this process, TRG not only enhances the performance of head translation estimation but also improves head rotation accuracy. This landmark-to-image alignment framework is inspired by the architecture of PyMAF [49, 50], which is a model used to reconstruct a human mesh. However, PyMAF is not designed to estimate the camera-to-human distance and fundamentally differs from TRG as it does not leverage the synergy between real-scale human geometry and depth.

Furthermore, we discovered that TRG can accurately predict 3D face landmarks from a single image, even when strongly affected by perspective distortions, such as in selfies. This accuracy is attributed to the TRG's depth-aware landmark prediction architecture, which actively utilizes head translation information during the landmark prediction process. This finding further supports our main idea that head translation estimation should be conducted simultaneously with facial geometry estimation.

The main contributions of this study can be summarized as follows:

- We propose TRG for 6DoF head pose estimation. To the best of our knowledge, this is the first study to introduce an explicit bidirectional interaction structure between head translation and face geometry. Through this innovative structure, TRG simultaneously mitigates ambiguity concerning head depth and face size.
- The proposed strategy for estimating correction parameters for the bounding box demonstrates stable generalization performance on out-of-distribution data in terms of head translation.
- The landmark-to-image alignment strategy demonstrates high accuracy not only in terms of head translation but also regarding head rotation.
- TRG's depth-aware landmark prediction architecture exhibits high landmark prediction accuracy, even in images heavily influenced by perspective transformation, such as selfies.
- Extensive experimental results on the benchmark datasets ARKitFace [25] and BIWI [15] show that TRG outperforms current SotA methods.

2 Related Works

2.1 Landmark-free Approach

The landmark-free approach [1, 3, 4, 14, 21, 31, 47] aims to estimate head pose directly from input image without relying on landmarks. However, most landmark-free approaches [3,4,14,21,31,47] only estimate head rotation and do not consider head translation.

Among them, img2pose [1] not only estimates head rotation but also head translation. It calculates head translation from a proposal and employs a local-to-global transformation strategy to convert the estimated local pose into a global image space. Intrinsics are utilized during the conversion of the local head pose into the global head pose. However, img2pose does not use intrinsics when calculating head translation from a proposal, leading to inaccurate local head poses. This is because utilizing intrinsics is essential when calculating depth from an image, even when dealing with a cropped image. Furthermore, [1] does not utilize face geometry information during inference, which can exacerbate depth ambiguity.

In contrast to landmark-free approaches, our proposed method explicitly utilizes facial geometry information. Specifically, TRG simultaneously mitigates ambiguity regarding face size and head translation through a bidirectional interaction structure. Additionally, TRG does not directly calculate head translation from cropped images but infers bounding box correction parameters instead. It then computes head translation using the inferred correction parameters and intrinsics. The proposed bounding box correction parameter strategy enables stable and accurate inference of head translation.

2.2 Landmark-based Approach

Numerous landmark-based approaches have been proposed [18, 25, 27, 37, 43–45, 56] for estimating a 6DoF head pose or 3D head rotation. [27, 37, 43] proposed methods that simultaneously estimate 2D face landmarks and 3D head rotation by leveraging the synergy between them using learning-based approaches. However, these studies have not explored the synergy between 3D face geometry and head translation.

SynergyNet [44] demonstrated that the parameters for shape and expression [35] can improve 3D sparse landmarks, and these enhanced landmarks can, in turn, improve the 3DMM parameters and head rotation during training. However, during the test time, it utilized a unidirectional information transfer architecture, which does not refine the 3DMM parameters and head rotation from the improved landmarks. Furthermore, SynergyNet is a model based on weak-perspective projections, similar to those in [5, 10, 16, 19, 28, 55]. Such models fundamentally do not compute the actual distance between the camera and the face.

MICA [56], JMLR [18], and PerspNet [25] employ unidirectional information transfer methods that first estimate face geometry and then calculate head pose.

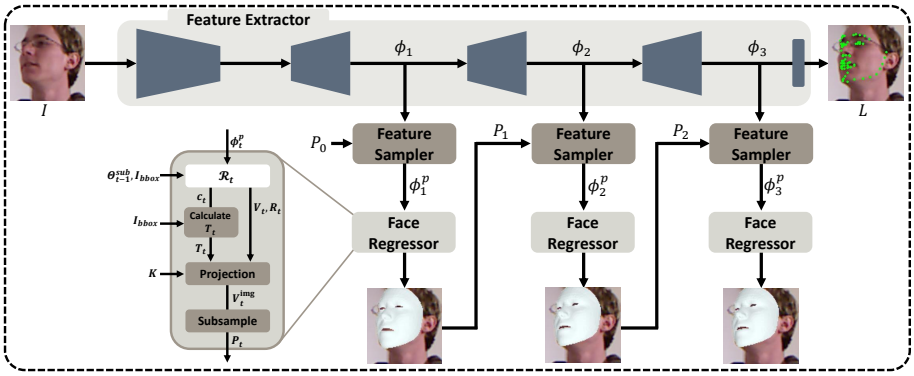

Fig. 2: Overall pipeline of the proposed method.

However, these methods are limited in their ability to reconstruct real-scale face geometry due to depth ambiguity. Furthermore, calculating the 6DoF head pose based on these inaccurate geometry priors makes it difficult to achieve high accuracy.

To address the aforementioned issues, we propose, for the first time, an explicit bidirectional interaction structure between the 6DoF head pose and face geometry. Additionally, unlike other landmark-based approaches, the proposed structure actively utilizes head depth information during the landmark estimation process. This approach demonstrates accurate geometry estimation even for images with strong perspective distortions, such as selfies.

3 Proposed Method

3.1 Overview of the Proposed Method

TRG is designed to iteratively regress head translation $\{T_t \in \mathbb{R}^3\}_{t=1}^3$ and rotation $\{R_t \in \mathbb{R}^6\}_{t=1}^3$ from a single image $I \in \mathbb{R}^{3\times192\times192}$, while also providing the auxiliary output of dense 3D landmarks $\{V_t \in \mathbb{R}^{3\times N^V}\}_{t=1}^3$.

Fig. 2 illustrates the comprehensive structure of TRG, which comprises a feature extractor that generates multi-scale feature maps $\{\phi_t \in \mathbb{R}^{256\times H_t\times W_t}\}_{t=1}^3$ from I, a feature sampler that extracts a landmark-aligned feature vector $\phi_t^p \in \mathbb{R}^{5N_{t-1}^P}$ from the feature map ϕ_t, and a face regressor that regresses head translation T_t, rotation R_t, and dense landmarks V_t from ϕ_t^p. N_{t-1}^P and N^V denote the number of sampling points $P_{t-1} \in \mathbb{R}^{2\times N_{t-1}^P}$ and the number of 3D dense landmarks V_t, respectively. Each of these components−feature extractor, feature sampler, and face regressor−is described in detail in Sections 3.2, 3.3, and 3.4, respectively. Additionally, the loss functions employed in the training are discussed in Section 3.5.

3.2 Feature Extractor

The feature extractor computes multi-scale feature maps $\{\phi_t\}_{t=1}^{3}$ and 2D sparse landmarks $L \in \mathbb{R}^{2\times N^L}$ from a single image I. N^L denote the number of sparse landmarks. The feature extractor comprises ResNet18 [20], three deconvolution layers, a 1×1 convolution layer, and a soft-argmax operation [42]. ResNet18 is initialized with pre-trained weights on ImageNet [11] and is used after removing the final classification layer and the pooling layer. The ϕ_t is computed from the t-th deconvolution layer and fed into the feature sampler. Additionally, the last feature map, ϕ_3 undergoes a transformation into 2D heatmaps through the 1×1 convolution layer. The soft-argmax operation computes L from the resultant heatmaps. These computed landmarks, along with the ground-truth landmarks $L^* \in \mathbb{R}^{2\times N^L}$, are incorporated into the loss function.

3.3 Feature Sampler

The feature sampler computes the landmark-aligned feature vector $\phi_t^p \in \mathbb{R}^{5N_{t-1}^P}$ from the feature map ϕ_t and the corresponding sampling points $P_{t-1} \in \mathbb{R}^{2\times N_{t-1}^P}$. Sampling points P_{t-1} are used to extract point-wise features from the feature map ϕ_t. Here, P_0 is set to 2D grid coordinates. For $t > 0$, P_t is computed using the t-th face regressor. The methodology for deriving these sampling points from the face regressor is described in Section 3.4.

The point-wise feature vector $\phi_t(p_{t-1,n}) \in \mathbb{R}^{256}$ is obtained using bilinear sampling at the location specified by the point $p_{t-1,n} \in \mathbb{R}^2$ on ϕ_t. Here, $p_{t-1,n}$ denotes the n-th column vector of the sampling points P_{t-1}. The N_{t-1}^P point-wise features, denoted as $\{\phi_t(p_{t-1,n})\}_{n=1}^{N_{t-1}^P}$, are then transformed into 5D vectors using a dimension reduction layer $\mathcal{F}(\cdot)$. These vectors are subsequently concatenated to form the landmark-aligned feature vector ϕ_t^p:

$$\phi_t^p = \bigoplus(\{\mathcal{F}(\phi_t(p_{t-1,n}))\}_{n=1}^{N_{t-1}^P}), \tag{1}$$

where $\bigoplus(\cdot)$ denotes concatenation. The dimension reduction layer, $\mathcal{F}(\cdot)$, is structured as a multilayer perceptron (MLP), which comprises three fully connected layers and two Leaky ReLU activations [32, 50]. The obtained landmark-aligned feature vector ϕ_t^p is then fed into the face regressor.

3.4 Face Regressor

The face regressor comprises an MLP $\mathcal{R}_t(\cdot)$ to calculate the head rotation, bounding box correction parameters, and dense landmarks $\Theta_t = \{R_t \in \mathbb{R}^6, c_t \in \mathbb{R}^3, V_t \in \mathbb{R}^{3\times N^V}\}$, a function that computes the head translation $T_t = \{T_t^x, T_t^y, T_t^z\} \in \mathbb{R}^3$ based on the bounding box information $I_{\text{bbox}} = \{\frac{\tau^{x,\text{bbox}}}{f}, \frac{\tau^{y,\text{bbox}}}{f}, \frac{b}{f}\} \in \mathbb{R}^3$ and the correction parameter $c_t = \{s_t, \tilde{\tau}_t^{x,\text{face}}, \tilde{\tau}_t^{y,\text{face}}\}$, and a perspective projection function that calculates the image coordinates of the dense landmarks

$V_t^{img} \in \mathbb{R}^{2\times N^V}$ and the sampling points P_t. V_t and R_t denote the 3D coordinates of the dense landmarks defined in the head space and the head rotation expressed in a 6D representation [52], respectively. T_t^x, T_t^y, and T_t^z represent the head translations along the x-, y-, and z-axes in the camera space, respectively. $\tau^{x,\text{bbox}}$, $\tau^{y,\text{bbox}}$, b, and f denote the x- and y-coordinates of the bounding box center relative to the center of the uncropped image, the size of the bounding box, and the focal length, respectively. s_t, $\tilde{\tau}_t^{x,\text{face}}$, and $\tilde{\tau}_t^{y,\text{face}}$ respectively denote the bounding box scale factor and the normalized offset of the head center relative to the bounding box center in the x- and y- directions.

The MLP $\mathcal{R}_t(\cdot)$ estimates the residual for calculating Θ_t from the landmark-aligned feature ϕ_t^p, the previously iterated output $\Theta_{t-1}^{sub} = \{R_{t-1}, c_{t-1}, V_{t-1}^{sub} \in \mathbb{R}^{3\times 305}\}$, and the bounding box information I_{bbox} [29,49,50]. Θ_t is computed by adding the residual estimated by $\mathcal{R}_t(\cdot)$ to Θ_{t-1}. V_{t-1}^{sub} represents the landmarks obtained by subsampling V_{t-1} [36]. The use of V_{t-1}^{sub} for $\mathcal{R}_t(\cdot)$ instead of V_{t-1} reduces the redundancy of the dense landmarks, which improves the performance of the proposed model [6–8, 30, 50].

We model a real human face as being enclosed within a box B of size $0.2m \times 0.2m$, with m denoting meters. The size of this box, when projected into the image space, is represented by b. However, since the assumption about the face size is typically imprecise, $\mathcal{R}_t(\cdot)$ estimates a scale factor s_t to adjust the size of B. Furthermore, $\mathcal{R}_t(\cdot)$ is responsible for determining the normalized offsets of the head center $\tilde{\tau}_t^{x,\text{face}}, \tilde{\tau}_t^{y,\text{face}}$. These offsets represent the values obtained by normalizing the image space translation from the bounding box center to the head center with b. The calculation of T_t from c_t and I_{bbox} is expressed as:

$$\begin{aligned} T_t^x &= \frac{0.2s_t}{b}\tau^{x,\text{bbox}} + 0.2s_t\tilde{\tau}_t^{x,\text{face}}, \\ T_t^y &= \frac{0.2s_t}{b}\tau^{y,\text{bbox}} + 0.2s_t\tilde{\tau}_t^{y,\text{face}}, \quad T_t^z = \frac{0.2s_t}{b}f. \end{aligned} \tag{2}$$

The derivation of Eq. 2 can be found in the supplementary material. The image coordinates of the dense landmarks, V_t^{img}, are computed by projecting V_t, as follows:

$$V_t^{img} = \Pi(V_t, R_t, T_t, K), \tag{3}$$

where $\Pi(\cdot)$ and $K \in \mathbb{R}^{3\times 3}$ denote the perspective projection and the intrinsic camera parameters, respectively. The sampling points P_t are obtained by subsampling V_t^{img}.

3.5 Loss Functions

We detail the loss functions employed to train TRG, ensuring accurate predictions of face geometry and head pose. The training process utilizes several loss functions for dense landmarks: head space coordinate loss $\mathcal{L}_{\text{head}}$, camera space coordinate loss $\mathcal{L}_{\text{cam}}$, and image space coordinate loss $\mathcal{L}_{\text{img}}$. For a precise estimation of head rotation, a head rotation loss $\mathcal{L}_{\text{rot}}$ is also adopted. As iteration

progresses, the loss functions are doubled as follows:

$$\begin{aligned}
\mathcal{L}_{\text{head}} &= \sum_{t=1}^{3} 2^{t-3}\left(\frac{1}{N^V}\sum_{n=1}^{N^V}\|V_{t,n} - V_n^*\|_1\right), \\
\mathcal{L}_{\text{cam}} &= \sum_{t=1}^{3} 2^{t-3}\left(\frac{1}{N^V}\sum_{n=1}^{N^V}\|V_{t,n}^{cam} - V_n^{*,cam}\|_1\right), \\
\mathcal{L}_{\text{img}} &= \sum_{t=1}^{3} 2^{t-3}\left(\frac{1}{N^V}\sum_{n=1}^{N^V}\|V_{t,n}^{img} - V_n^{*,img}\|_1\right), \\
\mathcal{L}_{\text{rot}} &= \sum_{t=1}^{3} 2^{t-3}\left(\|R_t^{mat} - R^{*,mat}\|_F\right),
\end{aligned} \tag{4}$$

where $*$ and $V_{t,n}$ represent the ground truth and the n-th column vector of V_t, respectively. $V_t^{cam} = R_t^{mat}V_t + T_t \in \mathbb{R}^{3\times N^V}$ and V_t^{img} represent the camera space coordinates and the image space coordinates of the t-th dense landmarks, respectively. $R_t^{mat} \in \mathbb{R}^{3\times 3}$ represents the 3D head rotation in matrix form, and $\|\cdot\|_F$ denotes the Frobenius norm.

If connectivity between dense landmarks is defined in the dataset, we utilize this information to apply an edge length loss. We empirically found that applying the edge length loss $\mathcal{L}_{\text{ed}}$ [18, 33] to V_3, estimated by the final face regressor, improves the model's performance in estimating face geometry. The edge length loss $\mathcal{L}_{\text{ed}}$ can be written as:

$$\mathcal{L}_{\text{ed}} = \sum_{M}\sum_{\{n,m\}\subset M} |\|V_{3,n} - V_{3,m}\|_2 - \|V_n^* - V_m^*\|_2|, \tag{5}$$

where M denotes a triangle. Additionally, to improve the quality of the feature map, we apply the sparse 2D landmark loss $\mathcal{L}_L$ to the landmarks L obtained from ϕ_3 as follows:

$$\mathcal{L}_L = \frac{1}{N^L}\sum_{n=1}^{N^L}\|L_n - L_n^*\|_1. \tag{6}$$

The final loss function to train TRG can be written as:

$$\mathcal{L} = \lambda_{\text{head}}\mathcal{L}_{\text{head}} + \lambda_{\text{cam}}\mathcal{L}_{\text{cam}} + \lambda_{\text{img}}\mathcal{L}_{\text{img}} + \lambda_{\text{rot}}\mathcal{L}_{\text{rot}} + \lambda_{\text{ed}}\mathcal{L}_{\text{ed}} + \lambda_L\mathcal{L}_L, \tag{7}$$

where λs represent the weights of the loss functions. λ_{head}, λ_{cam}, λ_{img}, λ_{rot}, λ_{ed}, and λ_L are set to 20, 2, 0.01, 10, 2, and 1.25, respectively.

4 Experimental Results

4.1 Implementation Details

The spatial dimensions H_t, W_t of the feature map ϕ_t were set to $\frac{192}{2^{5-t}}$. The number of sampling points N_t^P was set to $18 \times 18 = 324$ when $t = 0$, and to 305

when $t > 0$. N^V and N^L were set to 1220 and 68, respectively. For the ARKitFace training dataset [25], we selected a random sample and used its corresponding ground-truth dense 3D landmarks and head rotation as the initial landmarks V_0 and head rotation R_0 for the TRG. The initial correction parameter c_0 was set to $\{s_0 = 1, \tilde{\tau}_0^{x,\text{face}} = 0, \tilde{\tau}_0^{y,\text{face}} = 0\}$. For the TRG training, both the ARKitFace training data [25] and 300W-LP [55] were utilized. Unless otherwise stated, the performances of models trained using both datasets are presented. When a fair comparison with the state-of-the-art methods is required, results from models trained solely on the ARKitFace training dataset are also provided. Please refer to the supplementary material for more implementation details.

4.2 Datasets

ARKitFace [25] is a dataset that provides the 6DoF head poses, the dense 3D landmarks, and intrinsic camera parameters. It is collected from selfie scenarios, with data gathered at a camera-to-face distance ranging from 0.3 to 0.9 meters, resulting in images significantly influenced by strong perspective transformations. Following previous work [25], we used 717,840 frames from 400 subjects for training, and 184,884 frames from 100 subjects for testing.
300W-LP [55] is an extended synthetic dataset derived from the 300W [39], which itself is composed of several standardized datasets, including AFW [54], HELEN [51], IBUG [40], and LFPW [2]. Through face profiling, the 300W-LP dataset provides 122,450 synthesized images from approximately 4,000 original pictures.
BIWI [15] provides 6DoF head poses, a 3D neutral face mesh for each subject, and intrinsic camera parameters. Since BIWI does not provide ground-truth face meshes for each frame, our evaluation focuses solely on the head poses. BIWI serves exclusively as test data to assess the effectiveness of our method. We evaluated the performance of our proposed model by following the protocol used in previous studies [25, 46].

4.3 Evaluation Metrics

For head rotation accuracy assessment, we follow the approach used in previous studies [1,23,25,46], measuring rotation errors separately for roll, pitch, and yaw. Additionally, to provide a comprehensive understanding of the head rotation estimation performance, we also present the mean absolute error (MAE_r) and geodesic error (GE) [9]. For evaluating the accuracy of head translation, we calculate the errors for translation along the x-, y-, and z-axes, represented as t_x, t_y, and t_z errors, respectively. Similar to head rotation, we present the mean absolute error performance for head translation, denoted as MAE_t. Following previous research [25], we utilize the average 3D distance (ADD) metric [22] to present a holistic evaluation of the method's performance in estimating both rotation and translation:

$$\text{ADD} = \frac{1}{N^V} \sum_{n=1}^{N^V} \|(R_3^{mat} V_n^* + T_3) - (R^{*,mat} V_n^* + T^*)\|_2. \tag{8}$$

Table 1: Ablation study of TRG on ARKitFace and BIWI. We explored the effects of the bidirectional interaction structure and utilizing the correction parameter. We also investigated the importance of utilizing face geometry in the 6DoF head pose estimation process and the effectiveness of the landmark-to-image alignment method. "MS" means multi-scale features.

Method	ARKitFace				BIWI		
	Mean ↓	MAE_r ↓	MAE_t ↓	ADD ↓	MAE_r ↓	MAE_t ↓	ADD ↓
1-iter (w/o MS)	1.69	1.00	3.70	8.93	3.28	13.74	32.28
2-iter (w/o MS)	1.66	0.89	**3.61**	8.72	2.95	13.77	31.28
3-iter (w/o MS)	**1.57**	**0.88**	3.63	8.71	**2.59**	13.67	31.52
T_t-prediction	1.66	0.92	4.64	11.66	8.81	1.7K	5.1K
Landmark-free baseline	-	1.03	3.86	9.34	3.87	18.42	42.22
Grid sampled baseline	1.59	0.95	3.74	8.99	2.98	14.58	35.04
TRG (Ours)	1.58	0.91	3.62	**8.68**	2.75	**12.97**	**29.46**

To assess the 3D landmark prediction accuracy of our proposed method, we evaluate the median and average distances between the estimated and ground-truth dense landmarks [25]. The effectiveness of our method is evaluated based on the estimated values of V_3, R_3, and T_3 from the final face regressor at $t = 3$. The unit for median, mean, translation error, and ADD is in millimeters, and the unit for rotation error is in degrees.

4.4 Ablation Experiments

Effectiveness of bidirectional interaction structure. In this experiment, we delve into the significance of explicit bidirectional interaction between the 6DoF head pose and face geometry. To investigate this, we observe the model's performance variations based on the number of interactions between these two types of information. For our experiments, we designed 1-iteration, 2-iteration, and 3-iteration baselines and then compared their performance. The 1-iteration baseline model simultaneously regresses the face geometry and head pose using $\mathcal{R}_1(\cdot)$ but without an iterative inference process. The 2- and 3-iteration baseline models enhance this process by incorporating the iterative inference approach. They project the predicted dense landmarks onto the image feature, with all other aspects remaining consistent with the 1-iteration baseline. Similar to the 1-iteration baseline, they utilize only ϕ_1 and do not employ multi-scale features. The key distinction between the 3-iteration models and TRG lies in the utilization of multi-scale features.

The evaluation on the ARKitFace test data, as presented in Table 1, indicates that the performance in estimating the face geometry and head pose improves with the increasing number of iterations. This improvement is attributed to the reduction in ambiguity between the face geometry and 6DoF head pose as the number of bidirectional interactions increases. The BIWI evaluation results further corroborate the effectiveness of the bidirectional interaction method.

Use of correction parameter. In this experiment, we investigate the rationale behind estimating the correction parameter c_t instead of directly estimating head

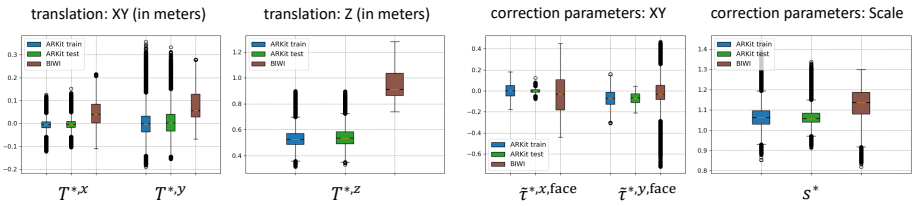

Fig. 3: The distribution of ground-truth translation and correction parameters in ARKitFace and BIWI. The colors blue, green, and brown represent the distributions of the ARKitFace training data, ARKitFace test data, and BIWI dataset, respectively. The symbol $*$ denotes ground-truth.

translation T_t. To elucidate this, we compare the performance of two models: the T_t-prediction baseline, which directly estimates head translation T_t and TRG. According to Table 1, while the T_t-prediction baseline demonstrates accurate estimation of head translation on the ARKitFace test data, its performance significantly declines on the BIWI dataset. We attribute this discrepancy to the differing translation distributions between the ARKitFace and BIWI datasets.

The first and second columns of Fig. 3 illustrate the ground-truth head translation distributions for ARKitFace and BIWI. While the translation distribution in the ARKitFace training data closely matches its test data, it significantly differs from that of BIWI. This discrepancy is particularly noticeable in the z-axis translations, indicating substantial divergence between the ARKitFace training data and BIWI. To achieve generalization from the ARKitFace training data to BIWI, a model must effectively extrapolate the z-axis translation. However, as evidenced by Table 1, this extrapolation poses a significant challenge for the direct translation estimation model.

The third and fourth columns of Fig. 3 visualize the distribution of the ground-truth correction parameters for both ARKitFace and BIWI datasets. A key observation here is that the variation in the correction parameter distribution is significantly smaller compared to the translation distribution. Based on these observations, we can conclude that shifting the estimation target from T_t to c_t effectively reduces distribution discrepancies. This strategic redefinition enhances the model's generalizability, particularly for data that fall outside the training distribution, as evidenced in Table 1.

The importance of utilizing facial geometry and the effectiveness of landmark-to-image alignment technique. For the purpose of our experiment, we designed a landmark-free baseline that does not estimate facial geometry $\{V_t\}_{t=1}^3$. Given the absence of facial geometry information, the landmark-free baseline is unable to utilize landmark-to-image alignment techniques. Consequently, it extracts grid sampled features from $\{\phi_t\}_{t=1}^3$ and inputs them into a face regressor. However, due to significant structural differences from TRG, we mitigate these differences by also designing a grid sampled baseline for incremental comparison. This grid sampled baseline is similar to the TRG, except

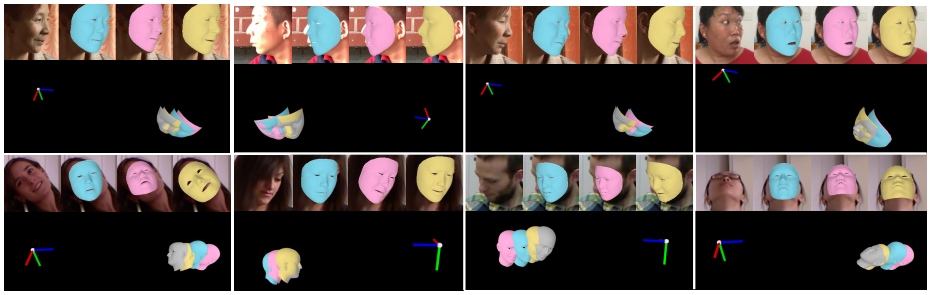

Fig. 4: Qualitative comparison on the ARKitFace and BIWI datasets. The first and second rows show visualized results for ARKitFace and BIWI, respectively. The colors cyan, pink, gold, and gray represent JMLR, PerspNet, TRG, and ground truth, respectively. The red, green, and blue axes respectively represent the X, Y, and Z axes of the camera coordinate system.

it does not employ the landmark-to-image alignment method, indicating that the primary distinction from the landmark-free baseline lies in whether facial geometry is estimated.

According to our findings, as presented in the Table 1, the landmark-free baseline underperforms compared to the grid sampled baseline. This supports our hypothesis that landmark information should be integrated during the 6DoF head pose estimation process. Furthermore, our results demonstrate that TRG outperforms the grid sampled baseline, affirming the superiority of our landmark-to-image alignment strategy.

Table 2: Comparison with previous methods for 6DoF head pose estimation on ARKitFace test dataset. Models trained with multiple datasets are marked with the symbol ⋆, and retrained models are indicated by the symbol †.

Method	MAE_r	GE	MAE_t	ADD
img2pose [1, 25]	5.55	-	7.02	20.54
Direct 6DoF Regress [25]	1.87	-	9.06	21.39
Refined Pix2Pose [25, 34]	2.35	-	14.00	36.44
JMLR [18] † ⋆	1.16	2.39	4.86	11.87
PerspNet [25]	0.99	1.81	4.18	10.01
TRG (Ours)	0.92	**1.80**	3.64	8.74
TRG (Ours) ⋆	**0.91**	1.84	**3.62**	**8.68**

Table 3: Comparison with previous methods for dense 3D landmark estimation on ARKitFace test dataset.

Method	Median	Mean
PRNet [17]	1.97	2.05
3DDFA-v2 [19]	2.35	2.31
Deng *et al.* [13]	2.46	2.55
JMLR [18] † ⋆	1.86	1.94
PerspNet [25]	1.72	1.76
TRG (Ours)	**1.55**	1.61
TRG (Ours) ⋆	**1.55**	**1.58**

4.5 Comparison with State-of-the-Art Methods

In this experiment, we conducted a benchmark of our proposed method against existing approaches for 6DoF head pose estimation. The evaluation results on the ARKitFace and BIWI datasets are presented in Tables 2, 3 and 4. Model retrained for this comparison is marked with the symbol †. Multiple datasets

were used for the model, which could be trained on multiple datasets. However, PerspNet was trained exclusively using the ARKitFace train dataset due to the difficulty of using two datasets [25,55] with differing 3D face mesh topologies. To ensure a fair comparison, we also present the results of TRG trained solely on the ARKitFace train dataset. Models trained on multiple datasets are denoted with the symbol $\star$.

Evaluation on ARKitFace [25]. Img2pose directly infers the 6DoF head pose from images without utilizing face geometry information. However, the absence of face geometry information can lead to increased face size ambiguity, potentially worsening the performance of head pose inference, as can be seen in Table 2.

JMLR and PerspNet do not incorporate head pose information during the face geometry inference process. The predicted face geometry, derived without considering head pose information, is relatively inaccurate (Table 3). Consequently, methods that predict the 6DoF head pose based on this relatively imprecise geometry yield inaccurate results (Table 2). In contrast, TRG actively integrates face geometry information into the head pose estimation process. According to Table 2, TRG achieves state-of-the-art in head pose estimation, attributed to its explicit bidirectional interaction structure. Furthermore, owing to its depth-aware landmark prediction architecture, TRG maintains stable face landmark prediction accuracy even in selfie scenarios, as shown in Table 3. Fig. 4 visually illustrates the performance of TRG and existing models [18,25] for head pose estimation and face landmark prediction. When the geometries predicted by each model are aligned with the image, they appear to be well-aligned. However, a stark contrast in model performance becomes evident when comparing the ground-truth geometry with the predicted geometries in the 3D camera space. JMLR and PerspNet struggle to accurately predict the actual size of a human face, resulting in high translation errors.

Evaluation on BIWI [15]. According to Table 4, TRG significantly outperforms existing optimization-based methods [18,25,56] in head translation estimation. This superior performance is attributed to TRG's design, which effectively leverages the synergy between face geometry and head translation. Furthermore, TRG's landmark-to-image alignment method enables it to achieve high head rotation estimation accuracy, surpassing even methods that solely estimate 3D head rotation. Fig. 4 qualitatively demonstrates TRG's exceptional head pose estimation performance. To visualize how closely the predicted head pose matches the ground-truth pose, we utilized the ground-truth neutral mesh and the predicted head pose.

4.6 Limitations

In the process of deriving depth from images using the proposed method, the requirement for camera intrinsics emerges as a necessary component. This necessity indicates that, in the absence of camera intrinsics, while it is still possible to estimate relative depth among faces in an image, achieving precise depth measurement poses a challenge. To address this challenge and ensure accurate depth determination between the face and the camera, incorporating algorithms that

Table 4: Comparison with previous methods for 6DoF head pose estimation on BIWI dataset. The models were evaluated using BIWI solely for testing purposes, without utilizing it as training data. We used the camera intrinsics provided by BIWI for the evaluation of the head pose estimation performance of MICA [56].

Method	Yaw	Pitch	Roll	MAE_r	GE	t_x	t_y	t_z	MAE_t	ADD
Dlib [26]	11.86	13.00	19.56	14.81	-	-	-	-	-	-
3DDFA [55]	5.50	41.90	13.22	19.07	-	-	-	-	-	-
EVA-GCN [45]	4.01	4.78	2.98	3.92	-	-	-	-	-	-
HopeNet [38]	4.81	6.61	3.27	4.89	9.53	-	-	-	-	-
QuatNet [23]	4.01	5.49	2.94	4.15	-	-	-	-	-	-
Liu *et al.* [31]	4.12	5.61	3.15	4.29	-	-	-	-	-	-
FSA-Net [46]	4.27	4.96	2.76	4.00	7.64	-	-	-	-	-
HPE [24]	4.57	5.18	3.12	4.29	-	-	-	-	-	-
WHENet-V [53]	3.60	4.10	2.73	3.48	-	-	-	-	-	-
RetinaFace [12] ⋆	4.07	6.42	2.97	4.49	-	-	-	-	-	-
FDN [48]	4.52	4.70	2.56	3.93	-	-	-	-	-	-
MNN [43]	3.98	4.61	2.39	3.66	-	-	-	-	-	-
TriNet [3]	3.05	4.76	4.11	3.97	-	-	-	-	-	-
6DRepNet [21]	3.24	4.48	2.68	3.47	-	-	-	-	-	-
Cao *et al.* [4]	4.21	3.52	3.10	3.61	-	-	-	-	-	-
TokenHPE [47]	3.95	4.51	2.71	3.72	-	-	-	-	-	-
Cobo *et al.* [9]	4.58	4.65	2.71	3.98	7.30	-	-	-	-	-
img2pose [1] ⋆	4.57	3.55	3.24	3.79	7.10	-	-	-	-	-
Direct 6DoF Regress [25]	16.49	14.03	5.81	12.11	-	62.36	85.01	366.52	171.30	562.38
Refined Pix2Pose [25, 34]	5.75	5.06	11.23	7.35	-	16.82	21.30	255.36	97.83	356.32
MICA [56] ⋆	5.40	7.17	3.80	5.46	-	9.32	13.66	60.13	27.70	68.03
JMLR [18] † ⋆	6.31	6.17	3.72	5.40	8.61	8.66	7.27	32.63	16.19	39.71
PerspNet [25]	3.10	**3.37**	2.38	2.95	5.61	**4.15**	**6.43**	46.69	19.09	100.09
TRG (Ours)	3.28	3.52	1.87	2.89	5.68	8.41	7.38	27.13	14.31	32.10
TRG (Ours) ⋆	**3.04**	3.44	**1.78**	**2.75**	**5.35**	7.83	6.99	**24.07**	**12.97**	**29.46**

estimate intrinsics becomes essential. This aspect of requiring camera intrinsics for depth calculations highlights an area for further exploration and adaptation in our method, especially when intrinsic parameters are not readily available.

5 Conclusion

This study proposed a novel approach by introducing the TRG to predict a 6DoF head pose from a single image. Through extensive experimentation, we demonstrated the effectiveness of the explicit bidirectional interaction between the 6DoF head pose and the dense 3D face landmarks, a core feature of the TRG architecture. We further established that our method of estimating the correction parameters significantly enhances the generalizability of the model in cross-dataset evaluations. Evaluation on the ARKitFace and BIWI datasets showed TRG's superior performance in head pose estimation compared to existing state-of-the-art methods. Our extensive experiments have also highlighted the strength of TRG's depth-aware landmark prediction structure, particularly in images heavily influenced by perspective transformation, facilitating accurate estimation of face geometry. Based on these findings, our future work will focus on accurately reconstructing detailed facial geometries from close-up facial photos, such as selfies, further pushing the boundaries of facial analysis technology.

Acknowledgement

This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00219700, Development of FACS-compatible Facial Expression Style Transfer Technology for Digital Human, 90%) and National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1F1A1066170, Physically valid 3D human motion reconstruction from multi-view videos, 10%).

References

1. Albiero, V., Chen, X., Yin, X., Pang, G., Hassner, T.: img2pose: Face alignment and detection via 6dof, face pose estimation. In: CVPR (2021)
2. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. IEEE TPAMI **35**(12), 2930–2940 (2013)
3. Cao, Z., Chu, Z., Liu, D., Chen, Y.: A vector-based representation to enhance head pose estimation. In: WACV (2021)
4. Cao, Z., Liu, D., Wang, Q., Chen, Y.: Towards unbiased label distribution learning for facial pose estimation using anisotropic spherical gaussian. In: ECCV (2022)
5. Chai, Z., Zhang, T., He, T., Tan, X., Baltrusaitis, T., Wu, H., Li, R., Zhao, S., Yuan, C., Bian, J.: Hiface: High-fidelity 3d face reconstruction by learning static and dynamic details. In: ICCV (2023)
6. Cho, J., Youwang, K., Oh, T.H.: Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In: ECCV (2022)
7. Chun, S., Park, S., Chang, J.Y.: Learnable human mesh triangulation for 3d human pose and shape estimation. In: WACV (2023)
8. Chun, S., Park, S., Chang, J.Y.: Representation learning of vertex heatmaps for 3d human mesh reconstruction from multi-view images. In: ICIP (2023)
9. Cobo, A., Valle, R., Buenaposada, J.M., Baumela, L.: On the representation and methodology for wide and short range head pose estimation. PR **149**, 110263 (2024)
10. Danecek, R., Black, M.J., Bolkart, T.: EMOCA: Emotion driven monocular face capture and animation. In: CVPR (2022)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
12. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: CVPR (2020)
13. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: CVPRW (2019)
14. Dhingra, N.: Lwposr: Lightweight efficient fine grained head pose estimation. In: WACV (2022)
15. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Van Gool, L.: Random forests for real time 3d face analysis. IJCV **101**, 437–458 (2013)
16. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM TOG **40**(4), 1–13 (2021)
17. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: ECCV (2018)

18. Guo, J., Yu, J., Lattas, A., Deng, J.: Perspective reconstruction of human faces by joint mesh and landmark regression. In: ECCVW (2022)
19. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: ECCV (2020)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
21. Hempel, T., Abdelrahman, A.A., Al-Hamadi, A.: 6d rotation representation for unconstrained head pose estimation. In: ICIP (2022)
22. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: ACCV (2013)
23. Hsu, H.W., Wu, T.Y., Wan, S., Wong, W.H., Lee, C.Y.: Quatnet: Quaternion-based head pose estimation with multiregression loss. IEEE TMM **21**(4), 1035–1046 (2018)
24. Huang, B., Chen, R., Xu, W., Zhou, Q.: Improving head pose estimation using two-stage ensembles with top-k regression. IVC **93**, 103827 (2020)
25. Kao, Y., Pan, B., Xu, M., Lyu, J., Zhu, X., Chang, Y., Li, X., Lei, Z.: Toward 3d face reconstruction in perspective projection: Estimating 6dof face pose from monocular image. IEEE TIP **32**, 3080–3091 (2023)
26. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: CVPR (2014)
27. Kumar, A., Alavi, A., Chellappa, R.: Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In: FG (2017)
28. Li, H., Wang, B., Cheng, Y., Kankanhalli, M., Tan, R.T.: Dsfnet: Dual space fusion network for occlusion-robust 3d dense face alignment. In: CVPR (2023)
29. Li, Z., Liu, J., Zhang, Z., Xu, S., Yan, Y.: Cliff: Carrying location information in full frames into human pose and shape estimation. In: ECCV (2022)
30. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: CVPR (2021)
31. Liu, Z., Chen, Z., Bai, J., Li, S., Lian, S.: Facial pose estimation by deep learning from label distributions. In: ICCVW (2019)
32. Maas, A.L., Hannun, A.Y., Ng, A.Y., et al.: Rectifier nonlinearities improve neural network acoustic models. In: ICML (2013)
33. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In: ECCV (2020)
34. Park, K., Patten, T., Vincze, M.: Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In: ICCV (2019)
35. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: AVSS (2009)
36. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3d faces using convolutional mesh autoencoders. In: ECCV (2018)
37. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE TPAMI **41**(1), 121–135 (2019)
38. Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. In: CVPRW (2018)
39. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: ICCVW (2013)
40. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: ICCVW (2013)

41. Shao, M., Sun, Z., Ozay, M., Okatani, T.: Improving head pose estimation with a combined loss and bounding box margin adjustment. In: FG (2019)
42. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: ECCV (2018)
43. Valle, R., Buenaposada, J.M., Baumela, L.: Multi-task head pose estimation in-the-wild. IEEE TPAMI **43**(8), 2874–2881 (2020)
44. Wu, C.Y., Xu, Q., Neumann, U.: Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In: 3DV (2021)
45. Xin, M., Mo, S., Lin, Y.: Eva-gcn: Head pose estimation based on graph convolutional networks. In: CVPR (2021)
46. Yang, T.Y., Chen, Y.T., Lin, Y.Y., Chuang, Y.Y.: Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In: CVPR (2019)
47. Zhang, C., Liu, H., Deng, Y., Xie, B., Li, Y.: Tokenhpe: Learning orientation tokens for efficient head pose estimation via transformers. In: CVPR (2023)
48. Zhang, H., Wang, M., Liu, Y., Yuan, Y.: Fdn: Feature decoupling network for head pose estimation. In: AAAI (2020)
49. Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., Liu, Y.: Pymaf-x: Towards well-aligned full-body model regression from monocular images. IEEE TPAMI (2023)
50. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: ICCV (2021)
51. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: ICCVW (2013)
52. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR (2019)
53. Zhou, Y., Gregson, J.: Whenet: Real-time fine-grained estimation for wide range head pose. arXiv preprint arXiv:2005.10353 (2020)
54. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: CVPR (2012)
55. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: CVPR (2016)
56. Zielonka, W., Bolkart, T., Thies, J.: Towards metrical reconstruction of human faces. In: ECCV (2022)