IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON TRENDS AND ADVANCES IN BIO-INSPIRED IMAGE-BASED DEEP LEARNING METHODOLOGIES AND APPLICATIONS

Received February 13, 2020, accepted February 26, 2020, date of publication March 2, 2020, date of current version March 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2977346

An End-to-End Task-Simplified and Anchor-Guided Deep Learning Framework for Image-Based Head Pose Estimation

JING LI[1], JIANG WANG[1], AND FARHAN ULLAH[1,2]
[1]College of Computer Science, Sichuan University, Chengdu 610065, China
[2]Department of Computer Science, COMSATS University Islamabad, Sahiwal Campus, Sahiwal 57000, Pakistan

Corresponding author: Jiang Wang (jiang.wang@stu.scu.edu.cn)

ABSTRACT Image-based Head Pose Estimation (HPE) from an arbitrary view is still challenging due to the complex imaging conditions as well as the intrinsic and extrinsic property of the faces. Different from existing HPE methods combining additional cues or tasks, this paper solves the HPE problem by relieving problem complexity. Our method integrates the deep Task-Simplification oriented Image Regularization (TSIR) module with the Anchor-Guided Pose Estimation (AGPE) module, and formulate the HPE problem into a unified end-to-end learning framework. In this paper, we define anchors as images that strictly obey the ''gravity rule in camera'', which follows the assumption that camera coordinate of the vertical axis should always be consistent with that of the local head coordinate. We formulate image pair as the regularized image produced by TSIR along with its anchor counterpart, both of which are fed into the AGPE module for estimating fine-grained head poses. This paper also proposes an Anchor-Guided Pairwise Loss (AGPL), which describes the interdependent relevance of poses between each pair of images. The proposed method is evaluated and validated with sufficient experiments which show its effectiveness. Comprehensive experiments show that our approach outperforms the state-of-the-art image-based methods on both indoor and outdoor datasets.

INDEX TERMS Head pose estimation, task-simplification oriented image regularization, anchor-guided pose estimation, anchor-guided pairwise loss, deep learning framework.

I. INTRODUCTION

The task of head pose estimation is to infer the orientation of a person's head with respect to the viewpoint of the camera. Head pose information serves as a significantly important component for human-computer interaction, where the humanoid robots in domestic environment are trained to have the ability of estimating head poses such that a more natural interaction with its users [1] can be ensured, and the driving assisted system is able to monitor the driver's field of view as well as his awareness [2]. Head pose estimation also allows systems to monitor social interactions [3], [4], and is able to observe one person's specific target of interest [5].

The inherent difficulties in estimating head pose as well as other related tasks include various imaging conditions, different non-Lambertian lighting sources, the existence of complex background or occlusions and etc [6]. In fact, conventional analysis using human vision in the real world is also confronted with the same problems. For example, human visions in the real-world would fail if they are distant, far away or with extreme bad illumination conditions. Interestingly, humans will not persist in improving their perceptual ability under such harsh conditions. Instead, they solve these problems by walking closer or by lighting up for brighter views. In other words, humans are endowed with the ability to find ways out of the current predicament by relieving problem complexity for better conditions. In computer vision community, many face recognition algorithms employ face alignment operation before the recognition module [7]–[9], which simplifies the problem by enforcing the same semantic regions locating at the same local area share similar facial features. Simple mechanisms of task-simplification on head pose estimation can also be found in literature via image-level operations:

- The illumination conditions are usually normalized via the mean and standard deviation regularization [10].

The associate editor coordinating the review of this manuscript and approving it for publication was Peter Peer.

- Influence of the background image could basically be removed through face localization techniques and image cropping [10]–[12], which eliminates noise from irrelevant issues and mainly focus on the facial area.
- Intrinsic camera parameters like the focal length could be modified by image-level scaling, which is used to normalize the size of the facial area occupying the image.

However, the above simplification schemes are unable to help relieve problem complexity brought about by its underlying huge solution set. Regarding the possible viewing angle towards the camera, which directly reveals image appearance, the issue of what kind of simplification scheme should be adopted remains unclear. To solve the above problem, this paper proposes a novel deep task-simplification method to relieve the complexity of HPE problem. Motivated by real-world situations that objects on earth all follow the law of gravity, otherwise human perception would become harder because of the existence of a much larger pose set. Based on this observation, this paper reduces the overall pose set by regularizing head poses, which strictly obeys the ''gravity rule in camera.''

Suppose $(X_{cam}, Y_{cam}, Z_{cam})$ denotes the rectangular camera coordinate system, where the axis X_{cam} and Y_{cam} are within the imaging plane and Y_{cam} is defined as the direction of ''gravity in camera''. Similarly, we denote the local coordinate system with respect to the head as $(X_{loc}, Y_{loc}, Z_{loc})$. In the local coordinate system, the axis X_{loc} is horizontally parallel to the line connecting eye centers, Y_{loc} is vertically located and is perpendicular X_{loc}, and Z_{loc} is the normal of the X_{loc}-Y_{loc} plane pointing to the front. Our ''gravity rule in camera'' suggests that projection of Y_{cam} coincides with that of Y_{loc}, where the face images after simplification are provided with the characteristics that the visual appearances of them would remain in relatively consistent pattern, like both the line linking two eye centers and the line connecting two mouth corners are practically horizontal. We call the images obeying ''gravity rule in camera'' as anchors. We generate the anchor images by adopting a direct correlation between 3D face and 2D images, based on the assumption that rotating on the 2D image is equivalent to that of rotating along axis Z_{cam}. The work that is most related to ours [13] approximated image-level rotation via transforming the face image to fit the predefined reference shape. Their transformation is solved by calculating the affine matrix between facial landmarks of the input image and a canonical 3D shape. On the one hand, robustness and accuracy of this normalization method rely heavily upon accurate 2D landmark locations, which remains an open issue in the research community, especially with ''in-the-wild'' conditions. Moreover, it is not reasonable that only one frontal template is employed [13], as landmark locations, vary significantly from frontal view to the profile view.

We address the above issues by circumventing the use of any key points or additional canonical shapes. In detail, we formulate our problem by combining a task-simplified regularization module and an anchor-guided estimation module into an end-to-end deep learning-based framework. The main contributions of this paper are summarized as follows:

- We decompose the challenging HPE problem into two inter-related tasks, namely AGPE module based face image regularization and TSIR module based head pose estimation, and integrate them into an end-to-end pipeline via deep learning framework.
- We propose an anchor-guided pairwise loss between the regularized image and its anchor counterpart, which not only improves final prediction but also endows TSIR with fine-grained regularization.
- Our deep task-simplification module estimates image-level rotation transformation without any auxiliary information, e.g. facial landmarks or the additional reference face shape.

The remainder of this paper is organized as follows. We review the most related articles to the filed of head pose estimation in Sec. II. Sec. III describes our proposed method in detail. Sec. IV exhibits experimental results as well as related discussions, and concluding remarks are drawn in the end.

II. RELATED WORK

Existing methods on HPE vary from model-based approaches [14], [15] to appearance-based ones [16], [17]. This section is limited to appearance-based approaches as they are the most relevant methods to our work. We classify the appearance-based methods into two categories and give detailed review to each of them: 1) Methods seeking for additional assistance and 2) Methods focusing on improving the prediction power of HPE algorithms themselves.

A. METHODS REQUIRING ADDITIONAL ASSISTANCE

Using additional assistance on solving HPE problem can roughly be classified as either a pre-estimation based method that utilizes visual cues as an auxiliary, or simultaneous estimation based methods that explore the inherent dependencies from inter-related tasks.

1) ADDITIONAL CUE BASED METHODS

In recent years, deep learning based methods [18] have dominated landmark localization methods [11], [19], [20], due to their robustness and representative ability to extract specific task related features. As a by-product of landmark localization, head pose problems are usually solved as a 2D-3D fitting problem with a pre-defined average 3D face model available. However, the precision is likely to be affected by different error sources [10], including the accuracy of the 2D landmark estimation algorithm, 3D error caused by the fixed average model, and solution accuracy of the optimization method. Xia *et al.* [13] proposed a pose estimation method via auxiliary assistance of facial landmarks. In their work, they combine landmark-based heatmaps with grayscale images in channel level for head pose prediction. Keypoint

based methods can indeed help boost the performance of head pose estimation, but only if each of the landmarks is accurately localized. However, due to the unstable performance of face alignment algorithms under challenging conditions such as large view angles or extreme illumination conditions, the reliance on accurate landmark localization hinders the way to general applications in realistic scenarios.

2) MULTI-TASK BASED METHODS

Head pose estimation relates a lot to other facial analysis problems. Kumar *et al.* [21] presented an iterative method for both keypoint estimation and pose prediction, based on the observations that landmark locations change accordingly with the rotating head. In order to better prevent the overfitting problem and improve the generalization ability of the training network, there have emerged many works that try to solve more than one facial related tasks jointly in one multi-task network. Zhu and Ramanan [22] based on a mixture of trees with a shared pool of parts and simultaneously solved face detection, pose estimation, and landmark localization, which regards each keypoint as a part and use global mixtures to capture topological changes induced by head poses. Hyperface [23] fused the learned features from intermediate layers of deep CNN, which is then fed into a multi-task learning network. In addition to the pose and landmark estimation, Hyperface [23] also predicted the face bounding box location and facial gender information. Beyond Hyperface [23], Ranjan *et al.* [24] conducted even more tasks like smile detection, age estimation and face recognition via the powerful convolutional neural network which regularizes the shared parameters of CNN and builds a synergy among different domains and tasks. Zhu *et al.* [11] predicted both face orientation and keypoints via reconstructing the 3D face model. They learn parameters of the morphable models [25], which regard the reconstruction problem as a linear combination of several pre-defined bases of the face model.

B. METHODS FOCUSING ON ALGORITHM PERFORMANCE

The HPE methods have attracted increasing attention of researchers, especially when deep learning-based methods have become more and more prevalent in computer vision-related tasks. Behera *et al.* [26] learned multi-level features by exploring different image regions, which combine multiple local regions with the whole image for discrete pose classification. Limitation of pure classification based method is that they only predict the approximate range of head pose intervals, and they lack the ability for fine-grained estimation, which would inevitably obstruct the way to wider applications. FacePoseNet [27] directly regressed a 6DoF vector for both camera localization and orientation; the estimated pose further serves as auxiliary information for face alignment and recognition. Yet, no validation results on the precision of the estimated pose were reported. Ruiz *et al.* [10] proposed a multi-loss method directly from image intensities by estimating weights of evenly distributed pose intervals, and the final estimation is obtained by weighted sum along pose spans. Yang *et al.* [12] adopted a hierarchical coarse-to-fine classification strategy, followed by a soft stage wise regression scheme, where the final fine-grained predictions are obtained through extracting intermediate features with further aggregation and regression.

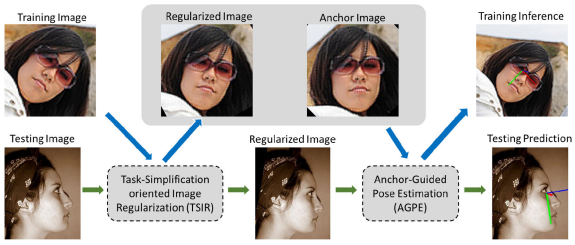

FIGURE 1. Overview of our HPE framework. The input image is first fed into the TSIR module to generate the regularized image. By pairing the regularized image with its anchor counterpart, the final prediction is obtained through AGPE module. The blue arrows indicate the training flow, while the green ones indicate the testing flow.

We argue that HPE is a complicated and comprehensive problem. Other mechanisms on solving HPE other than improving algorithm prediction power or introducing prior knowledge should also be considered. In this paper, we proposed a pure image-based task-simplification strategy for fine-grained pose estimation by reducing the complexity of HPE.

III. PROPOSED METHOD

Fig. 1 shows the overview of our proposed framework. The input of our algorithm contains only one single RGB image, and no other auxiliary information like depth or facial landmarks are used. We formulate the complex HPE problem by integrating a TSIR module and an AGPE module into the end-to-end learning framework. The TSIR module aims at reducing the scale of the overall solution set and conducting the in-plane rotation upon input image to get a regularized image following the "gravity rule in camera." And the following AGPE module would output the fine-grained $3D$ rotation angles of the input image concerning Yaw, Pitch and Roll.

In order to make the problem well defined, the training database is denoted as $\{I_{train}, I_{ach}, l_y, l_p, l_r\}$, where I_{train} is the training image and I_{ach} denotes its corresponding anchor image. The remaining symbols are groundtruth pose labels that l_y for Yaw angle, l_p for Pitch angle and l_r for Roll angle respectively. In the training phase, the input training image I_{train} first goes into *TSIRN* to produce the regularized image I_{reg}, and then the paired image of (I_{reg}, I_{ach}) is fed into *AGPEN* that

$$\hat{l_y}, \hat{l_p}, \hat{l_r} = AGPEN(TSIRN(I_{train}), I_{ach}) \tag{1}$$

where $\hat{l_y}, \hat{l_p}, \hat{l_r}$ denotes the estimated labels for Yaw, Pitch and Roll, and the symbols *TSIRN* and *AGPEN* are short for TSIR Net and AGPE Net respectively. Note that each face image in the pair share the same network architecture and parameters. For a given image I_{test} in the test phase, its estimation could

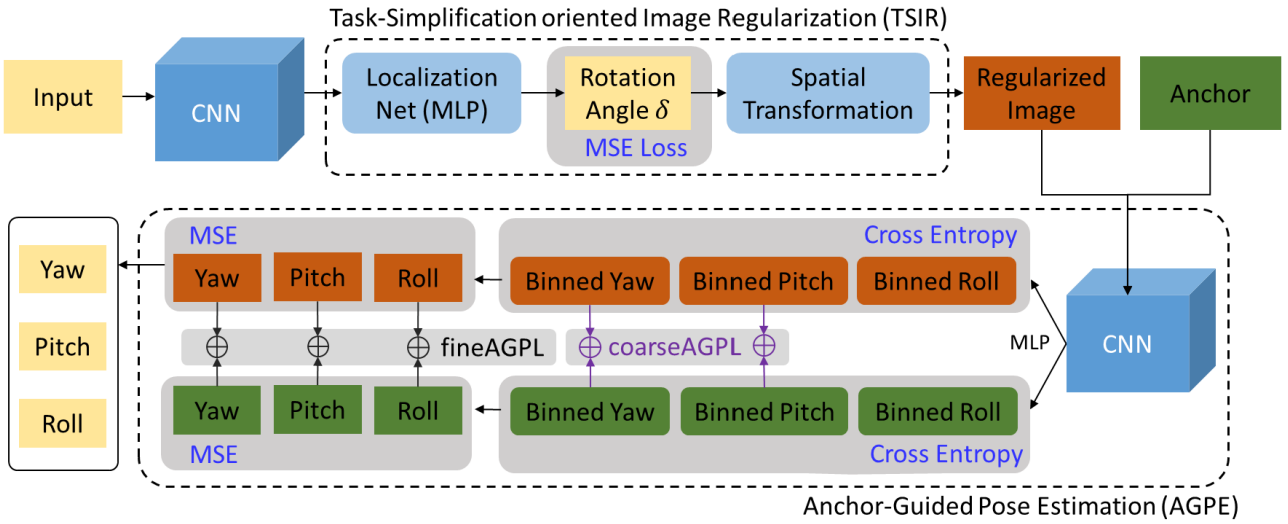

FIGURE 2. Architecture of the feed-forward part of the proposed framework. Our feature extraction module employ CNN-based architecture for both tasks. We use MLP for rotation angle regression as well as facial pose classification for Yaw, Pitch and Roll angle. Both MSE loss and the Cross-Entropy loss are conducted between the estimation and the ground-truth. The anchor-guided pairwise loss is calculated between the anchor and its regularized counterpart with respect to Yaw, Pitch and Roll. The symbol $\oplus$ denotes pairwise operation on the loss function, please refer to Sec. III-D for more details.

be formulated as

$$\hat{l_y}, \hat{l_p}, \hat{l_r} = AGPEN(TSIRN(I_{test})) \tag{2}$$

In the remaining part of this section, details of the proposed method would be elaborated at length. We would first show our proposed end-to-end deep architecture and then explain sequentially how the rotation consistency is guaranteed between $2D$ image coordinate system and $3D$ space coordinate system, how the anchor images are obtained following the "gravity rule in camera" and how anchors work for the final pose estimation.

A. END-TO-END DEEP ARCHITECTURE DESIGN FOR HPE

Fig. 2 shows the network architecture employed in our proposed method. Two networks are adopted for the task of head pose estimation. TSIR is the task-simplification module, which predicts the optimal rotation angle through which the regularized face image could be obtained and serves as the input of AGPE. We use the Spatial Transformer Network (STN) [28] as the main component of the TSIR module, which is a learnable network that allows spatial manipulations and image-level rotation. The $2D$ rotation angle $\hat{\delta}$ is estimated via CNN architectures like ResNet [29] for feature extraction which is followed by Multi-Layer Perception (MLP) for regression. Not that the other choices like ShuffleNetV2 [30] or self-designed networks for feature extraction would work just as well.

To obtain the regularized image, the affine transformation matrix

$$A_{2D} = \begin{bmatrix} \cos\hat{\delta} & \sin\hat{\delta} \\ -\sin\hat{\delta} & \cos\hat{\delta} \end{bmatrix} \tag{3}$$

is applied to the input image for spatial transformation via bilinear interpolation technique. Note that there is an implicit assumption lies in the TSIR module, saying that rotating $\hat{\delta}$ on the $2D$ image is equivalent to that of rotating in $3D$ camera coordinate system along Z_{cam}. We would provide theoretical analysis on when this assumption would hold in Sec. III-B.

The AGPE module first uses CNN (i.e. ResNet) to extract independent features given the regularized and anchor image pairs. After that, three MLPs sharing the same architecture are followed, which corresponds to the three poses, respectively. Finally, the combined losses of both pose classification and regression are employed. Suppose $[-\Omega, \Omega]$ denotes range span of facial poses for each Euler angle, which is divided evenly with interval k, the head pose estimation problem could thus be transformed to one discrete classification problem, with a total of $\lceil 2\Omega/k \rceil$ bins for each pose interval. The classification probability could be obtained via the stable softmax layer, followed by the MLP mentioned above. And the weighted sum of the probability over the pre-defined separated bins is used to represent the final estimation. Regarding loss functions, we argue that existing methods usually formulate loss functions independent of images. In this paper, we observe that there exists a strong relevance on Euler angles between the input image pair. Details of the proposed pairwise pose loss are described in Sec. III-D.

B. ORDERING-SENSITIVE 2D-3D ROTATION CONSISTENCY

Fig. 3 shows the schematic diagram of the rotation consistency between $2D$ image space and $3D$ camera space, indicated by x-y and X-Y-Z respectively. Note that the image plane x-y is parallel to that of the X-Y plane in camera coordinate, which is perpendicular to that of the Z axis. The orange path shows $2D$ rotation pipeline, where the $2D$

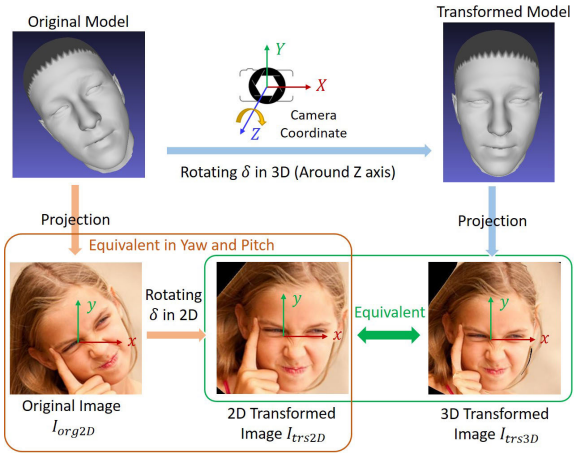

FIGURE 3. Schematic diagram of rotation consistency between 2*D* and 3*D*. The 3*D* camera coordinate of *X*, *Y*, and *Z* axis corresponds to Pitch, Yaw and Roll respectively. The image plane on 2*D* *x*-*y* coordinate is parallel to *X*-*Y* plane on the camera coordinate. Image obtained by rotating δ around *Z* axis that followed by projection operation is equivalent to that of rotating δ directly on 2*D* image.

transformed image I_{trs2D} is obtained by rotating δ on the originally captured image I_{org2D} within image plane. The 3*D* rotation pipeline is shown in blue arrows, where the original 3*D* face model first rotates δ along Z axis. By projecting the rotated model onto image plane, we get I_{trs3D} which is equivalent to I_{trs2D} via the orange flow in terms of their Eular angles. Therefore, the relationship between orientation prediction on I_{trs2D} and the original image I_{org2D} can be formulated as following

$$\begin{cases} l_y(I_{org2D}) = l_y(I_{trs2D}) \\ l_p(I_{org2D}) = l_p(I_{trs2D}) \\ l_r(I_{org2D}) = l_r(I_{trs2D}) - \delta \end{cases} \tag{4}$$

Thus, the orientation estimation of I_{org2D} could be indirectly obtained via estimating Euler angle of I_{trs2D}.

To make the above equivalence tenable, it is necessary to change the order of the rotation axes for the Euler angle. Existing datasets for head pose estimation usually utilize the rotation axes for Euler angle as Z-Y-X. The Z-Y-X ordering means that the final orientation of the head is obtained by first rotating around Z axis, followed by Y and X axis sequentially. In other words, after rotating along Z axis, the Euler angles for Y and X axis, namely Yaw and Pitch, would change accordingly. However, once the rotation axes for Euler angle change to X-Y-Z ordering, the rotation along Z axis becomes the last operation, and it is thus independent of Yaw and Pitch. Assuming Euler angles under Z-Y-X ordering is given by (ϕ, γ, θ), indicating Pitch, Yaw and Roll respectively. The rotation matrix along X, Y and Z are expressed as

$$R_x(\phi) = \begin{bmatrix} 1 & 0 & 0 \\ \cos\phi & -\sin\phi & 0 \\ 0 & \sin\phi & \cos\phi \end{bmatrix}$$

$$R_y(\gamma) = \begin{bmatrix} \cos\gamma & 0 & \sin\gamma \\ 0 & 1 & 0 \\ -\sin\gamma & 0 & \cos\gamma \end{bmatrix}$$

$$R_z(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5}$$

The final rotation matrix is calculated as

$$R = R_z(\theta) * R_y(\gamma) * R_x(\phi) \tag{6}$$

Suppose that the rotation angles under X-Y-Z ordering are denoted as (α, β, η), based on Eqn. 5, we could obtain

$$\begin{aligned} \sin\beta &= R_{13} \\ \cos\beta &= \sqrt{{R_{11}}^2 + {R_{12}}^2} \\ \sin\alpha &= -R_{23}/\cos\beta \\ \cos\alpha &= R_{33}/\cos\beta \\ \sin\eta &= -R_{12}/\cos\beta \\ \cos\eta &= R_{11}/\cos\beta \end{aligned} \tag{7}$$

New rotation representations then could be calculated via the mapping function H that

$$\begin{aligned} \alpha &= H(\sin\alpha, \cos\alpha) \\ \beta &= H(\sin\beta, \cos\beta) \\ \eta &= H(\sin\eta, \cos\eta) \end{aligned} \tag{8}$$

and H is formulated as

$$H(\sin\omega, \cos\omega) = \begin{cases} -\arccos\omega & \sin\omega < 0, \cos\omega < 0 \\ \arcsin\omega & \sin\omega < 0, \cos\omega > 0 \\ \arcsin\omega & \sin\omega > 0, \cos\omega > 0 \\ \arccos\omega & \sin\omega > 0, \cos\omega < 0 \end{cases} \tag{9}$$

for any $\omega \in (\alpha, \beta, \eta)$. Note that we use this two-step operation based on sine and cosine functions rather than the tangent functions to make sure that the large rotation angles would be recovered correctly and integrally.

C. "GRAVITY RULE IN CAMERA" BASED ANCHOR GENERATION

The anchor image is generated by rotating δ_0 along Z axis. To achieve "gravity rule in the camera," we enforce projection of the rotating axis Y_{cam} be coincident with axis y in image coordinate. Suppose $M = R_x(\alpha) * R_y(\beta) * R_z(\eta)$ is the rotation matrix following X-Y-Z ordering, the anchor's rotation angle then could be calculated through the projected axis $Y_{cam}^{proj} = [M_{12}, M_{22}, 0]$ and $y_{2D} = [0, 1, 0]$

$$\delta_0 = sign * < Y_{cam}^{proj}, y_{2D} > \tag{10}$$

where $sign = 1$ means clockwise rotation, while $sign = -1$ means anticlockwise rotation, which could be computed as:

$$sign = \begin{cases} 1 & if \cos < V, Z > \ is\ greater\ than\ 0 \\ -1 & if \cos < V, Z > \ is\ less\ than\ 0 \end{cases} \tag{11}$$

where $V = Y_{cam}^{proj} \times y_{2D}$ and $Z = [0, 0, 1]$. When vector V is pointing into X-Y plane under the Right-Hand Coordinate System, the rotating operation should be conducted clock wisely since the cosine value between V and Z is less than 0 and vice versa.

D. ANCHOR-GUIDED PAIRWISE LOSS

Eqn. 4 has made it clear that there is a close relationship between the Euler angle representatives of the input image and the rotated one. It is reasonable for us to build the relationship among the original image I_{inp}, the regularized image I_{reg} and the anchor image I_{ach} according to Eqn. 4 that

$$\begin{cases} \hat{l}_y(I_{inp}) = \hat{l}_y(I_{reg}) = \hat{l}_y(I_{ach}) \\ \hat{l}_p(I_{inp}) = \hat{l}_p(I_{reg}) = \hat{l}_p(I_{ach}) \\ \hat{l}_r(I_{inp}) = \hat{l}_r(I_{reg}) - \hat{\delta} = \hat{l}_r(I_{ach}) - \delta_0 \end{cases} \tag{12}$$

Eqn. 12 provides a straight forward view of formulating interrelationship between I_{reg} and I_{ach}, which inspires us to design the Anchor-Guided Pairwise Loss (AGPL). Considering different representations of the pose using either discrete pose classification or continuous pose regression, we design two forms of losses under the guidance of the anchor image; namely, coarseAGPL originated from binned prediction and fineAGPL from fine-grained regression. Note that the final representation of AGPL is obtained by combining the two formulations. Suppose $\hat{b_s}$ is the binned prediction and $\hat{l_s}$ represents the regressed predicting pose value. The superscript $s \in \{y, p\}$ denotes head pose of either Yaw or Pitch, which follows the condition that the underlying angles between I_{reg} and its anchor I_{ach} should be the same. To make full use of this constraint, the cross-entropy loss between the binned predictions should satisfy that

$$L_c^{s\in\{y,p\}} = -\frac{1}{N}\sum_{\mathbb{B}}\sum_{i=1}^{\lceil 2\Omega/k \rceil} \hat{b_s}(I_{ach}, i) * log(\hat{b_s}(I_{reg}, i)) \tag{13}$$

and Mean Square Error (MSE) for fine-grained estimations requires

$$L_f^{s\in\{y,p\}} = \frac{1}{N}\sum_{\mathbb{B}} |\hat{l_s}(I_{ach}) - \hat{l_s}(I_{reg})|^2 \tag{14}$$

Here $\mathbb{B}$ denotes the face image dataset and N is the total number of images in $\mathbb{B}$.

As illustrated in Eqn. 12, applying the paired constraint enforced upon Yaw and Pitch would fail when it comes to Roll angle. However, we observe that the equality would hold when we calculate $\hat{l_s}$, $s \in \{r\}$ backward to the original image. We formulate the paired loss on Roll as

$$L_f^{s\in\{r\}} = \frac{1}{N}\sum_{\mathbb{B}} |\hat{l}_{ach2org} - \hat{l}_{reg2org}|^2 \tag{15}$$

where $\hat{l}_{ach2org} = \hat{l_s}(I_{ach}) - \delta_0$ is Roll angle of the original image estimation from anchor image, and $\hat{l}_{reg2org} = \hat{l_s}(I_{reg}) - \hat{\delta}$ describes that from the regularized image. Eqn. 15 promotes the prediction power on Roll estimation especially at the initial epochs in the training phase. Meanwhile, from the task-simplification perspective of view, the large rotation error produced by TSIR may hinder the way to the task of simplification and induce performance decline on AGPE. In other words, $L_f^{s\in\{r\}}$ has contributed to increasing the prediction power for both modules of TSIR and AGPE.

Despite our proposed AGPL, we also apply cross entropy loss $L_{CE}^{s\in\{y,p,r\}}$ on binned probability of the Euler angle and MSE loss $L_{MSE}^{s\in\{y,p,r\}}$ on the regressed values to supervise the training network, as suggested in [10]. As for TSIR, we add a $L2$-norm-based regression loss $L_{rot}^{s\in\{y,p,r\}}$ between the estimated rotation angle and its ground-truth to enforce TSIR on learning proper mapping representation from the input face image to its regularized counterpart. The total loss function is formulated as

$$\begin{aligned} L_{total} = \alpha_f \sum_{s\in\{y,p,r\}} L_f^s + \alpha_c \sum_{s\in\{y,p\}} L_c^s \\ + \alpha_{CE} \sum_{s\in\{y,p,r\}} L_{CE}^s + \alpha_{MSE} \sum_{s\in\{y,p,r\}} L_{MSE}^s \end{aligned} \tag{16}$$

It is notable that anchor image acts on different perspectives. On the one hand, it serves as a guidance for how the original image should be regularized based on the inferred optimal transformation operation. On the other hand, it also helps AGPE to find a more distinguishable pattern that could be easier to recognize.

IV. EXPERIMENTAL RESULTS

A. DATASETS

Three popular datasets for head pose estimation including the 300W-LP [11], AFLW2000 [11] and BIWI [31] are employed in our experiment.

The 300W-LP (300 faces in-the-Wild across Large Pose) is a synthetic dataset generated from 300W [32], which is a combination of several datasets for landmark based face alignment, including LFPW (Labeled Face Parts in the Wild) [33], AFW (Annotated Faces in the Wild) [22], HELEN (the HELEN facial feature dataset) [34], and XM2VTS (the eXtended Multi Modal Verification for Teleservices and Security applications) [35]. The synthesized image is obtained using face profiling techniques, which project each personalized 3D face model to different views through gradually rotating the model until it reached the profile view. Additional flipping operation on face images after projection is carried out to further augmenting the dataset. In total about 122, 450 face image samples labeled in 300W-LP, with the groundtruth pose be inferred via fitting the 3D face model and the 6DoF of pose parameters.

AFLW2000 provides face images and corresponding groundtruth head poses for the first 2000 images in AFLW (Annotated Facial Landmarks in the Wild) [36]. AFLW2000 is a very challenging dataset which provides a large-scale collection of annotated face images gathered from the web, exhibiting a large variety of changes in pose, expression, ethnicity, age, gender as well as general imaging and

environmental conditions. The groundtruth pose in AFLW2000 is labeled and organized in the same way as 300W-LP. Following Hopenet, we exclude the 31 images whose angles exceed the interval $[-99°, 99°]$.

BIWI (the BIWI kinect head pose dataset) contains 24 sequences of 20 different people from RGBD cameras. The dataset includes a total of 15,000 frames including RGB images, depth images and the annotations. The head pose range covers about ± 75 degrees for Yaw and ± 60 degrees for Pitch. Groundtruth label is presented in the form of the 3D location of the head and 3D rotation matrix. Different from the other two datasets which are collected from in the wild, all images in BIWI are captured under indoor environment.

B. IMPLEMENTATION DETAILS

We implement the proposed method within Pytorch framework on a computer with one CPU and one GPU. A total of 30 epochs are used for training the proposed network. We use ResNet50 as our backbone network for both training and testing, and the Adam [38] optimizer with the initial learning rate be set to $1e-3$ and $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$. The learning rate is reduced for every 6 epochs by a factor of 0.1. The batch size in both training and testing phase is set to 32. The interval length k in Eqn.13 is set to be 3. We use two different datasets for training, one is the synthetic 300W-LP dataset and the other is BIWI, each of which corresponds to different models, namely In-The-Wild model and Laboratory model. We augment the training data by random cropping and filtering to strengthen the power of our method on handling images with scaling and blurring. We also randomly rotate the training images to cover a large variation of in-plane rotations. For testing phase, Mean Absolute Error (MAE) is used to measure the prediction error that

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \|\hat{l}_i - l_i\|_1 \tag{17}$$

The symbols l_i and $\hat{l}_i$ are groundtruth and the final estimation of either Yaw, Pitch or Roll angle with respect to the ith image, and N is the total number of images in the testing set. As the available groundtruth poses in 300W-LP and AFLW2000 dataset are presented in Z-Y-X ordering based Euler angle while it is provided with 3D rotation matrix in BIWI dataset, we follow equations in Sec. III-B to transform them into consistent representation of X-Y-Z ordering based Euler angle for both the training and testing phase. We test In-The-Wild model upon both AFLW2000 and BIWI to show generalization power of our method. For the Laboratory model, we use three different 70-30 splits of video frames in BIWI dataset. We adopt MTCNN [39] for face detection in the preprocess step for all experiments, and we manually label bounding box of the images where MTCNN fails.

C. EVALUATION ON THE IN-THE-WILD MODEL

Our In-The-Wild Model is trained on in-the-wild face images of 300W-LP and is first tested on AFLW2000. Face detection and cropping followed by rotation ordering consistency operation is conducted upon both datasets. Tab. 1 shows comparison results on AFLW2000 upon different state-of-the-art methods. FacePoseNet [27] directly regressed a vector for camera orientation. Hopenet [10] combined multiple losses for fine-grained head pose estimation. FSA [12] proposed a multi-stage based structure aggregation method. The remaining comparison methods use additional landmark information beyond images. FAN [19] and GT-based method treat HPE as solving a 2D-3D fitting problem via landmarks. 3DDFA [11] convert head pose estimation problem into 3D face reconstruction problem by simultaneously estimate shape and camera related parameters. Among all the landmark-free methods, ours performs the best, with an average prediction error of 3.83. Our method also out-performs most of the landmark-assisted methods except for Xia *et al.* [13]. However, Xia *et al.* [13] relies heavily on accuracy of facial landmarks, whose performance would sharply drop down when it comes to real applications.

TABLE 1. MAE in degree among different methods on the AFLW2000 dataset using In-The-Wild model.

Method	Yaw	Pitch	Roll	Avg
Landmark-free				
FacePoseNet [27]†	17.9	19.2	12.2	16.4
Hopenet [10]*	6.47	6.60	5.44	6.16
FSA [12]*	4.50	6.08	4.64	5.07
Ours*	**2.78**	**5.06**	**3.65**	**3.83**
Landmark-assisted				
FAN [19]	6.36	12.2	8.71	9.11
GT-based	5.92	11.7	8.27	8.65
3DDFA [11]	5.40	8.53	8.25	7.39
Xia et al. [13]*	**0.63**	**2.05**	**1.70**	**1.46**

* was trained on 300W-LP dataset.
† was trained on VGG face dataset [37].

In order to provide an in-depth analysis, Fig. 4 shows MAE distribution with respect to different pose intervals. We divide the whole pose span into 6 intervals of $[-99, -60)$, $[-60, -30)$, $[-30, 0)$, $[0, 30)$, $[30, 60)$ and $[60, 99)$, and we calculate MAE over each of these intervals based on the testing set of AFLW2000. Fig. 4 suggests that the overall error follows a quadratic distribution where the near frontal images are of lowest error while profile views exhibit highest error. However, our method is competitive especially for large pose intervals where a huge boost of error decreasing emerges compared to Hopenet and FSA. The result suggests that our task-simplification based method works is capable at solving images with large poses. Fig. 5 further exhibits some qualitative results for profile view images with illumination changes as well as occlusions, which further demonstrates robustness of our method.

To show generalization of the proposed method, Tab 2 exhibits comparison results on BIWI dataset using our In-The-Wild model. Our performance on BIWI also achieved state-of-the-art result compared with both landmark-free and landmark-assisted methods, where KEPLER [21]

TABLE 2. MAE in degree among different methods on the BIWI dataset using in-the-wild model.

Method	Yaw	Pitch	Roll	Avg
Landmark-free				
FacePoseNet [27]†	26.2	23.9	24.4	24.8
Hopenet [10]*	4.18	6.61	3.27	4.90
FSA [12]*	4.27	4.96	2.76	4.00
Ours[1]	**4.12**	**4.65**	**3.11**	**3.96**
Landmark-assisted				
3DDFA [11]	36.2	12.3	8.78	19.1
KEPLER [21]‡	8.80	17.3	16.2	13.9
FAN [19]	8.53	7.48	7.63	7.89

* was trained on 300W-LP dataset.
† was trained on VGG face dataset.
‡ was trained on AFLW dataset.

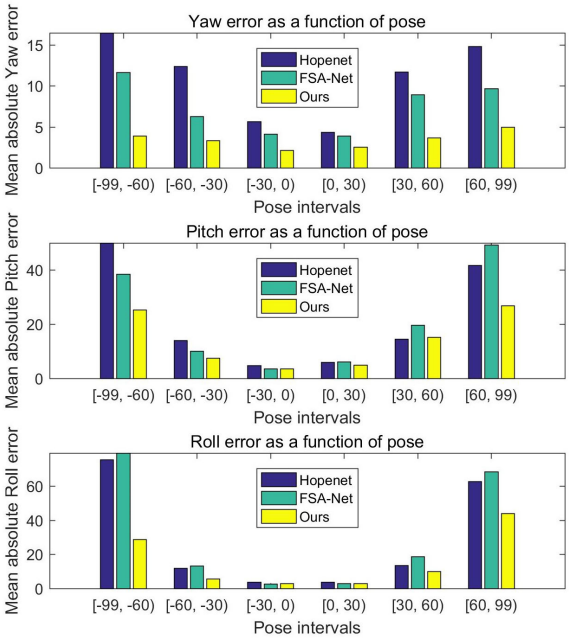

FIGURE 4. Mean absolute error distribution with respect to different pose intervals in AFLW2000 dataset.

simultaneously regresses keypoints and poses using heatmap-CNN architecture. For better exhibiting the process on the In-The-Wild model, Fig. 6 shows the average training loss curve as well as testing loss curves on both In-The-Wild and the Laboratory testing datasets as a function of the epoch number.

Original work on Hopenet and FSA both use Euler angle of Z-Y-X. To analysis how rotation ordering affect the performance of HPE, we transform ordering of Euler angle from Z-Y-X to X-Y-Z for Hopenet and FSA. Other settings for both Hopenet and FSA remain unchanged according to their original paper. Tab. 3 shows that X-Y-Z ordering based representation can help decrease the average prediction error to certain extent. The possible rationale behind is that with X-Y-Z ordering, 3D rotation around Z axis is consistent with that of rotating on the 2D image, where the head rotation along X and Y axis is fixed. In such case, it is more likely that X-Y-Z ordering based representation is capable to

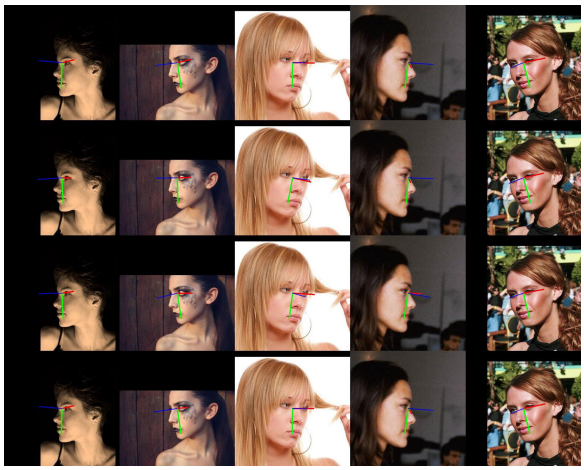

FIGURE 5. Pose estimation on the AFLW2000 dataset. From top to bottom, they are the groundtruth, results of Hopenet, results of FSA-Net and our results. The blue line indicates the direction that the subject is facing. The green line represents the downward direction and the red line pointing to the side.

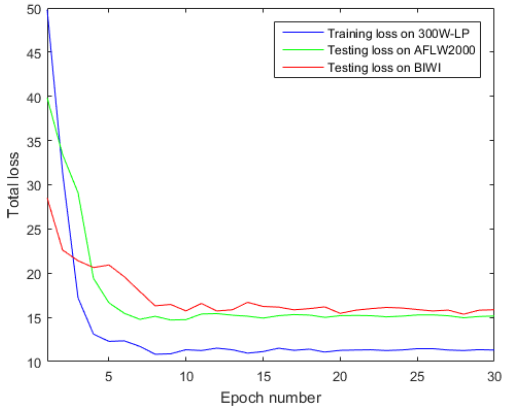

FIGURE 6. Evolution of the average training and testing losses as a function of the number of epochs.

TABLE 3. Ordering related comparison on AFLW2000 dataset using In-The-Wild model.

Method	Yaw	Pitch	Roll	Avg
ZYX ordering				
Hopenet [10]	6.47	6.60	5.44	6.16
FSA [12]	4.50	6.08	4.64	5.07
XYZ ordering				
Hopenet [10]	6.68	6.08	4.18	5.65
FSA [12]	4.75	5.37	3.65	4.59
Ours	**2.78**	**5.06**	**3.65**	**3.83**

build approximately one-to-one relationship between image appearance and head poses, which provide an easier pattern to recognize. Among all methods using X-Y-Z ordering, we find that our method performs the best, which shows its superiority compared to the others.

The backbone network of Resnet is complex and time-consuming which is impractical to be applied directly on robots or mobile devices. We take efficiency into consideration by using ShuffleNetV2 to replace the backbone Resnet and remain the other settings unchanged for both training and

TABLE 4. MAE in degree among different methods on the BIWI dataset using Laboratory model via cross validation.

Method	Yaw	Pitch	Roll	Avg
Landmark-free				
DeepHeadPose [40]	5.67	5.18	-	-
Liu et al. [41]	6.10	6.00	5.70	5.93
Drouard et al. [42]	4.24	5.43	4.13	4.60
FSA [12]	2.89	4.29	3.60	3.60
Hopenet [10]	3.29	3.39	3.00	3.23
Ours	**1.02**	**1.18**	**1.99**	**1.40**
Landmark-assisted				
Xia et al. [13]	**2.83**	**5.52**	**2.86**	**3.74**

testing. Our testing procedure takes an average of 0.02 seconds per image on TSIR module and 0.036 seconds per image on the AGPE. Note that we set the batch size equal to 1 when measuring time consumption. Furthermore, our light-weighted In-The-Wild model obtains MAE of 4.06 on AFLW2000 dataset, which is more or less comparable to that of 3.83 on Resnet, demonstrating that our method doesn't draw much benefit from network architecture. Considering that many face alignment methods nowadays are also with good real-time capacity as well as accuracy, we also show efficiency of our method compared with landmark-based simplification. In order to avoid error accumulation, we use groundtruth landmarks provided by the datasets for comparison. We estimate rotation angle between each of the input face shape and a reference shape using Singular Value Decomposition (SVD) based algorithm. We train the model by combining the landmark-based simplification and AGPE and choose ShuffleNetV2 as the backbone network. Although TSIR takes more time than landmark-based simplification, TSIR is sufficient to be applied in most of the applications. In addition, the landmark-based simplification combining AGPE obtains MAE of 3.92 on the AFLW2000 dataset, which is comparable to that of ours. We argue that performance of landmark-based method depends on the underlying face alignment algorithm, while our method doesn't rely on other assistance.

D. EVALUATION ON THE LABORATORY MODEL

To validate our method on constraint environment, we also train our model on the indoor captured dataset BIWI, with a pre-processing step of face detection followed by face cropping through the scaled face bounding box. Among the comparative methods, DeepHeadPose [40] use multi-modal RGBD data as input and formulated the problem of head pose estimation as one of classification of gazing direction via CNN. Liu *et al.* [41] trained to learn head features on synthetic head images using rendering techniques and solve head pose estimation as a regression problem. Drouard *et al.* [42] used a mixture of linear regressions with partially-latent output which learned to map high-dimensional feature vectors into the joint space of head pose angles and bounding box shifts. As shown in Tab 4, our method outperforms state-of-the-art by reducing the error from 3.23 to 1.40. Our method also surpasses the landmark-based method of [13], which further show advantage of the landmark-free methods.

TABLE 5. Ablation study in terms of MAE on AFLW2000 dataset.

Method	Yaw	Pitch	Roll	Avg
Baseline model	3.99	6.72	5.58	5.43
Baseline model + AGPL	2.98	6.17	5.06	4.74
Baseline model + TSIR	3.55	5.49	4.17	4.40
Baseline model + TSIR + AGPL	**2.78**	**5.06**	**3.65**	**3.83**

E. COMPARISON WITH BASELINE MODEL

In this section we conduct some ablation studies to verify effectiveness of our proposed task-simplification scheme and the novel anchor-guided pairwise loss. All ablation studies are trained using 300W-LP dataset and are tested upon AFLW2000. We first provide our baseline model which ignore the task-simplification module. With the training face images under X-Y-Z ordering based pose representation, we train the baseline model via a multi-loss formulation followed by [10]. To demonstrate effectiveness of our task-simplification module, we add TSIR module in the aforementioned baseline model. We also show the effect on HPE using AGPL by employing it on the baseline model. Different from that in Sec. III-D, here the input image pair is consisted with the original input face image and its anchor counterpart. We compare the above three experiments with our method, which combines the baseline model, task-simplification module of TSIR and the pairwise loss function of AGPL. Tab. 5 shows that both TSIR and AGPL have contributed to improve the performance. The task-simplification scheme is more capable of estimating Roll angles, while AGPL works better on Yaw angle prediction. The combination of TSIR and AGPL show further ability on three poses' prediction which all achieves the best result in terms of MAE.

V. CONCLUSION

This paper proposes a novel algorithm for HPE by combining the task-simplification mechanism and anchor-guided estimation method into one unified learning framework. Our method infers head poses from an image alone, without additional visual cues like facial landmarks or depth maps. Our TSIR module approximates the "gravity rule in camera" by estimating rotation transformation on original face images for task simplification. The regularized image, which is paired with its anchor counterpart, are fed into our AGPE module for final pose estimation. We propose the novel anchor-guided pairwise loss function called AGPL, which not only promotes head pose estimation with higher precision but also reduces rotation error on the TSIR module. As our main concern lies in pose regularization, this method may fail when under extreme lighting or with relatively low image quality. Furthermore, considering mobile applications that require less computational cost, we would expand our method on light-weighted networks with accuracy ensures.

REFERENCES

[1] R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel, "Natural human-robot interaction using speech, head pose and gestures," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, vol. 3, Oct. 2004, pp. 2422–2427.

[2] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 2, pp. 300–311, Jun. 2010.

[3] C.-W. Chen, R. C. Ugarte, C. Wu, and H. Aghajan, "Discovering social interactions in real work environments," in *Proc. Face Gesture*, Santa Barbara, CA, USA, Mar. 2011, pp. 933–938.

[4] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 1226–1233.

[5] J. Leroy, F. Rocca, M. Mancas, and B. Gosselin, "Second screen interaction: An approach to infer tv watcher's interest using 3D head pose estimation," in *Proc. 22nd Int. World Wide Web Conf.*, Janeiro, Brazil, May 2013, pp. 465–468.

[6] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.

[7] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1701–1708.

[8] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 815–823.

[9] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Placid, NY, USA, Mar. 2016, pp. 1–10.

[10] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 2018, pp. 2074–2083.

[11] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 146–155.

[12] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang, "FSA-net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1087–1096.

[13] J. Xia, L. Cao, G. Zhang, and J. Liao, "Head pose estimation in the wild assisted by facial landmarks based on convolutional neural networks," *IEEE Access*, vol. 7, pp. 48470–48483, 2019.

[14] P. Martins and J. Batista, "Accurate single view model-based head pose estimation," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Amsterdam, The Netherlands, Sep. 2008, pp. 1–6.

[15] M. Krinidis, N. Nikolaidis, and I. Pitas, "3-D head pose estimation in monocular video sequences using deformable surfaces and radial basis functions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 2, pp. 261–272, Feb. 2009.

[16] I. Chamveha, Y. Sugano, D. Sugimura, T. Siriteerakul, T. Okabe, Y. Sato, and A. Sugimoto, "Appearance-based head pose estimation with scene-specific adaptation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV)*, Workshops, Barcelona, Nov. 2011, pp. 1713–1720.

[17] K. Diaz-Chito, J. Martinez Del Rincon, A. Hernandez-Sabate, and D. Gil, "Continuous head pose estimation using manifold subspace embedding and multivariate regression," *IEEE Access*, vol. 6, pp. 18325–18334, 2018, doi: 10.1109/ACCESS.2018.2817252.

[18] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[19] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial Landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1021–1030.

[20] X. Guo, S. Li, J. Zhang, J. Ma, L. Ma, W. Liu, and H. Ling, "PFLD: A practical facial landmark detector," *CoRR*, vol. abs/1902.10859, pp. 1–11, Feb. 2019.

[21] A. Kumar, A. Alavi, and R. Chellappa, "KEPLER: Keypoint and pose estimation of unconstrained faces by learning efficient H-CNN regressors," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May 2017, pp. 258–265.

[22] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2879–2886.

[23] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyper face: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[24] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An All-In-One convolutional neural network for face analysis," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May 2017, pp. 17–24.

[25] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graph. Interact. Techn.*, Los Angeles, CA, USA, 1999, pp. 187–194.

[26] A. Behera, A. G. Gidney, Z. Wharton, D. Robinson, and K. Quinn, "A CNN model for head pose recognition using wholes and regions," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Lille, France, May 2019, pp. 1–2.

[27] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "FacePoseNet: Making a case for landmark-free face alignment," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Venice, Italy, Oct. 2017, pp. 1599–1608.

[28] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., Montreal, QC, Canada, Dec. 2015, pp. 2017–2025.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[30] N. Ma, X. Zhang, H. Zheng, and J. Sun, "Shufflenet V2: Practical guidelines for efficient CNN architecture design," in *Proc. 15th Eur. Conf.*, vol. 11218, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Munich, Germany. Cham, Switzerland: Springer, Sep. 2018, pp. 122–138, doi: 10.1007/978-3-030-01264-9_8.

[31] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 437–458, Aug. 2012.

[32] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-Wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Sydney, NSW, Australia, Dec. 2013, pp. 397–403.

[33] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Colorado Springs, CO, USA, Jun. 2011, pp. 545–552.

[34] V. Le, J. Brandt, Z. Lin, L. D. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proc. 12th Eur. Conf. Comput. Vis.*, Florence, Italy, (Lecture Notes in Computer Science), vol. 7574, A. W. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Cham, Switzerland: Springer, 2012, pp. 679–692.

[35] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maître, "Xm2vtsdb: The extended m2vts database," in *Proc. 2nd Int. Conf. Audio-Video Biometric Person Authentication (AVBPA)*, vol. 964, Mar. 1999, pp. 965–966.

[36] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 2144–2151.

[37] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Swansea, U.K., X. Xie, M. W. Jones, G. K. L. Tam, Eds., Sep. 2015, p. 41.1–41.12.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, Y. Bengio and Y. LeCun, Eds., May 2015, pp. 1–15.

[39] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," *CoRR*, vol. abs/1604.02878, pp. 1–5, Apr. 2016.

[40] S. S. Mukherjee and N. M. Robertson, "Deep head pose: Gaze-direction estimation in multimodal video," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2094–2107, Nov. 2015.

[41] X. Liu, W. Liang, Y. Wang, S. Li, and M. Pei, "3D head pose estimation with convolutional neural network trained on synthetic images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 1289–1293.

[42] V. Drouard, R. Horaud, A. Deleforge, S. Ba, and G. Evangelidis, "Robust head-pose estimation based on partially-latent mixture of linear regressions," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1428–1440, Mar. 2017.

JING LI received the B.E. degree in computer science and technology from Sichuan University, Chengdu, China, in 2009, and the Ph.D. degree from the Key Laboratory of Machine Perception, Peking University, Beijing, China, in 2014. Since 2014, she works at the College of Computer Science, Sichuan University. Her research interests include face image analysis, 3D modeling, and evolutionary game theory.

JIANG WANG was born in Jiangsu, China, in 1988. He received the M.E. degree in information and communication engineering from the University of Electronic Science and Technology of China, in 2014. He is currently pursuing the Ph.D. degree in computer science from the School of Computer Science, Sichuan University, Chengdu, China. His research interests include face recognition and face image analysis.

FARHAN ULLAH received the B.S. degree in computer science from the University of Peshawar, Pakistan, in 2008, and the M.S. degree in computer science from CECOS University, Peshawar, Pakistan, in 2012. He is currently pursuing the Ph.D. degree in computer science from the School of Computer Science, Sichuan University, Chengdu, China. He received the Research Productivity Award from the COMSATS Institute of Information Technology (CIIT), Sahiwal, Pakistan, in 2016. His research work is published in various renowned journals of Springer, Elsevier, Wiley, MDPI, and Hindawi. His research interests include software similarity, information security, and data science.