Adversarial Learning Semantic Volume for 2D/3D Face Shape Regression in the Wild

Hongwen Zhang, Qi Li, and Zhenan Sun, *Senior Member, IEEE*

***Abstract*— Regression-based methods have revolutionized 2D landmark localization with the exploitation of deep neural networks and massive annotated datasets in the wild. However, it remains challenging for 3D landmark localization due to the lack of annotated datasets and the ambiguous nature of landmarks under the 3D perspective. This paper revisits regression-based methods and proposes an adversarial voxel and coordinate regression framework for 2D and 3D facial landmark localization in real-world scenarios. First, a semantic volumetric representation is introduced to encode the per-voxel likelihood of positions being the 3D landmarks. Then, an end-to-end pipeline is designed to jointly regress the proposed volumetric representation and the coordinate vector. Such a pipeline not only enhances the robustness and accuracy of the predictions but also unifies the 2D and 3D landmark localization so that the 2D and 3D datasets could be utilized simultaneously. Further, an adversarial learning strategy is exploited to distill 3D structure learned from synthetic datasets to real-world datasets under weakly supervised settings, where an auxiliary regression discriminator is proposed to encourage the network to produce plausible predictions for both the synthetic and real-world images. The effectiveness of our method is validated on benchmark datasets 3DFAW and AFLW2000-3D for both 2D and 3D facial landmark localization tasks. The experimental results show that the proposed method achieves significant improvements over the previous state-of-the-art methods.**

***Index Terms*— 2D/3D facial landmark localization, semantic volumetric representation, joint voxel and coordinate regression, auxiliary regression adversarial learning.**

I. Introduction

FACIAL landmark localization is an essential step for the subsequent processing of face images. During the last decades, a significant amount of research has been dedicated to solving this problem. For 2D facial landmark localization, nearly-saturated performance [1], [2] has been achieved on near-frontal face images thanks to the exploitation of deep neural networks and the availability of massive annotated face datasets. However, advances in 3D facial landmark localization remain limited due to the depth ambiguity and the lack of fully annotated face images in the wild.

Over the past few years, regression based methods for facial landmark localization have shown their effectiveness on addressing issues such as occlusions, expressions, and head poses presented in real-world face images. Following the pioneering work of Explicit Shape Regression (ESR) [3], cascaded regression methods [4]–[7] attempt to learn the mapping from shape-index features to landmark coordinates. Although these methods could achieve highly accurate results for nearly frontal face images, their performances are barely satisfactory under the case of bad initializations or face images with large head poses etc. On the other hand, heatmap regression based methods [2], [8] estimate the heatmap for each individual landmark instead. Such a heatmap representation encodes the likelihood of each position being a specific landmark. The heatmap regression strategy avoids the inefficient learning of non-linear mapping from feature space to landmark positions, thus has greatly facilitated solving landmark localization problems including face alignment [2], [8] and human pose estimation [9], [10]. Heatmap regression based methods are also generalized to 3D landmark localization problems as well. In [11], Pavlakos it et al. extend the 2D heatmap to its 3D version and show its effectiveness in the application of 3D human pose estimation. Although these heatmap regression based methods could work well when facial components are visible, they might produce blurred heatmaps when there are invisible landmarks due to occlusions, making it unstable and error-prone to estimate landmark positions from those multi-mode heatmaps. Moreover, for 3D landmark localization, directly employing 3D heatmap for each landmark is cumbersome and memory-demanding especially when the number of target landmarks increases.

Meanwhile, though there are massive real-world datasets annotated with 2D landmarks, it's difficult to obtain ground-truth 3D facial landmarks for face images in the wild. Existing datasets annotated with 3D landmarks are typically comprised of synthetic face images which are created through rendering [12] or profiling algorithms [13], hence the variations on appearance are limited when compared with real-world datasets. Although existing methods could produce reliable results for those synthetic datasets, it is barely satisfactory when applying them to real-world images due to

Manuscript received October 5, 2018; revised February 25, 2019; accepted April 3, 2019. Date of publication April 19, 2019; date of current version July 16, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant U1836217, Grant 61427811, Grant 61573360, Grant 61721004, Grant 61702513, and Grant 61806197, and in part by the National Key Research and Development Program of China under Grant 2017YFC0821602 and Grant 2016YFB1001000. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yun Fu. *(Corresponding author: Zhenan Sun.)*

H. Zhang and Z. Sun are with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: hongwen.zhang@cripac.ia.ac.cn; znsun@nlpr.ia.ac.cn).

Q. Li is with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: qli@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TIP.2019.2911114

1057-7149 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on February 22,2020 at 08:45:42 UTC from IEEE Xplore. Restrictions apply.

the gap between two domains. To alleviate this, the two-step strategy is employed to lift the 2D estimations to 3D shapes. Typical two-step approaches [14]–[16] perform 2D landmark localization at first and then obtain the 3D face shape through depth estimation or 3D face model fitting. Though such a strategy is effective, it is suboptimal and sensitive to the result of 2D landmark localization.

Furthermore, when given 3D face shape estimations on real-world images, human vision is capable to distinguish those faulty estimations from the correct ones. Based on the prior of facial geometric structure, it is easy for the human to tell which part is less implausible and infer the missing part even under extreme occlusions. However, it is non-trivial to incorporate the prior knowledge into deep neural networks. There are several attempts to make use of the prior knowledge of geometric structure via designing handcrafted geometric constraints [17], [18]. In addition, adversarial learning recently is also exploited to improve the performance of 3D human pose estimation [19].

Motivated by the above observations, we propose an end-to-end framework for 2D and 3D facial landmark localization, and exploit adversarial learning to distill the structure of the 3D face shape learned from fully annotated synthetic images to real-world images without depth annotations.

To this end, we first introduce the semantic volumetric representation for the 3D face shape. Compared with the conventional volumetric representation [11], the proposed volumetric representation is more compact while still preserving the semantic information of landmarks. Based on such a volumetric representation, an end-to-end network is proposed for robust and accurate facial landmark localization. Specifically, the backbone of the network consists of two parts, namely a volume estimator and a coordinate regressor, which are used to predict volumetric representations and coordinate vectors of 3D face shapes respectively. With the proposed pipeline, the network could be simultaneously trained with images annotated with both 2D and 3D landmarks. To further leverage information from both 2D and 3D datasets, we propose an auxiliary regression adversarial learning strategy to improve the generalization performance of the network, where the volume estimator is treated as the generator, and an auxiliary regression volume discriminator is employed to encourage the volume estimator to generate plausible volumes. In this way, the proposed framework could be trained in a weakly supervised manner and leverage both synthetic datasets and real-world datasets.

To summarize, the main contributions of this work are listed as follows.

- A semantic volumetric representation is introduced for the 3D face shape. The dimensionality of the proposed representation is fixed regardless of the number of target landmarks. Such a representation provides an effective and efficient solution for generic 3D landmark localization, which could also facilitate related problems such as 3D human pose estimation.
- A joint voxel and coordinate regression pipeline is designed for facial landmark localization. Such a pipeline enables end-to-end training and improves the robustness and accuracy of landmark localization. Moreover, 2D and 3D landmark localization could be unified in the proposed pipeline so that both 2D and 3D annotated datasets could be leveraged simultaneously.
- An auxiliary regression adversarial learning strategy is proposed to better distill the 3D geometric structures learned from synthetic datasets to real-world images. This strategy facilitates the effective and stable training of the network and enhances its performance on 3D landmark localization in challenging scenarios. To the best of our knowledge, we are the first to exploit adversarial learning on the task of 3D facial landmark localization.

An early version of this work appeared in [20]. We have made significant extensions to our previous work in two main aspects. First, we upgrade the compact volumetric representation to its semantic version, obtaining better localization performance. Second, we exploit an adversarial learning strategy under the weakly supervised setting for 3D facial landmark localization in real-world scenarios.

The remainder of this paper is organized as follows. Section II briefly reviews previous works related to ours. Details of the proposed method are presented in Section III. Experimental results are presented in Section IV. Finally, Section V concludes the paper.

II. Related work

A significant amount of work has been introduced for landmark localization in the last decades. In this section, we briefly review previous work related to ours, including methods on 2D and 3D landmark localization, and adversarial learning for dense prediction tasks.

A. 2D Landmark Localization

Treating the 2D landmark localization task as a regression problem has become a common practice in recent years. These methods fall roughly into two categories: coordinate regression based method and heatmap regression based method. For the former category, cascaded regression methods [3]–[5] employ cascaded regressors to learn the mapping from shape-indexed features to increments of landmark coordinate vectors. Deep neural networks such as Convolutional Neural Networks (CNNs) [21], [22] and Recurrent Neural Networks (RNNs) [23], [24] have also been exploited as regressors to predict the facial landmark shapes. To avoid inefficient learning of the pixel-to-coordinate mapping, heatmap regression based methods [2], [8], [10] cast landmark localization as regressing the heatmaps of landmarks instead of coordinate vectors. Methods of this category pursue regressing clear and accurate 2D heatmaps for target landmarks. For example, Stacked Hourglass Networks [10] uses the symmetric topology and intermediate supervision, which has been demonstrated to be effective in both applications of human pose estimation [10] and face alignment [2]. Several state-of-the-art works [1], [2], [25] built upon this architecture achieve nearly saturate performance on 2D landmark localization. Despite their effectiveness, it is usually unstable to estimate the positions from multi-mode heatmaps of those invisible landmarks.

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on February 22,2020 at 08:45:42 UTC from IEEE Xplore. Restrictions apply.

Very recently, instead of adopting the maximum operation, Sun *et al.* [26] propose to infer the landmark coordinate from its heatmap through the integral operation, which allows end-to-end training and shows its effectiveness on human pose estimation.

B. 3D Landmark Localization

Two-step strategy is one of the popular solutions to 3D landmark localization problems which performs 2D landmark estimations at first and then predicts the depth information for these 2D landmarks [14]–[16]. On the other hand, Pavlakos *et al.* [11] introduce the volumetric representation for 3D body joints and show that predicting joints in a discretized 3D space could be more effective for 3D human pose estimation. The volumetric representation proposed in [11] could be viewed as a natural extension of the 2D heatmap, which is highly demanding for memory and computation. Although regressing such a representation in a coarse-to-fine manner could alleviate this problem [11], it still cannot avoid the curse of dimensionality when the number of target landmarks increases. Therefore, this method can not be easily generalized to other 3D object landmark localization problems. In contrast, we propose to encode the positions of all landmarks in a single volume with the dimensionality fixed regardless of the number of landmarks, providing a much more efficient solution for generic 3D landmark localization.

For 3D landmark localization in the wild, one of the significant challenges is the lack of annotated data. Recently, several attempts have been made to tackle this problem in a weakly supervised manner. For instance, Tung *et al.* [27] combine adversarial priors with the reconstruction loss from re-projection of 3D predictions for weakly supervised 2D-to-3D lifting. Zhou *et al.* [17] introduce a geometric constraint for 3D human pose to regularize the learning of 3D pose estimation under the weakly supervised setting. Further, Yang *et al.* [19] extend it under an adversarial learning framework by introducing a multi-source discriminator to enforce the pose estimator to generate plausible poses on unannotated in-the-wild images. However, little to no work has investigated weakly supervised learning for the task of 3D facial landmark localization. In this work, we treat 2D annotations as weak labels and develop a novel framework to leverage 2D annotated data for 3D facial landmark localization under the weakly supervised setting.

C. Adversarial Learning for Dense Prediction

As an emerging technique, Generative Adversarial Networks (GANs) [28] have achieved impressive results in various tasks such as image generation and editing. A typical GAN comprises two competing modules: a generator and a discriminator. The discriminator learns to distinguish between samples produced by the generator and real samples, while the generator learns to produce samples that are indistinguishable from real ones. Various improvements [29]–[32] have been proposed for more stable and easy training of GANs, and their applications are far beyond image generation. Among them, side information [33], [34] and auxiliary tasks [32] are exploited in adversarial learning to enhance performance of the generator, which also facilitates traditional tasks such as classification [35] and regression [36]. Moreover, the conditional GANs [31], [37] have been adopted as a general-purpose solution for dense prediction problems such as image style transfer, image segmentation, human pose estimation and parsing. These problems typically involve pixel-level mapping from input images to structured label maps, which has inherent relationships with heatmap regression based methods for landmark localization.

For image segmentation, Hung *et al.* [38] design a fully-convolutional discriminator to enforce the outputs of the segmentation network more spatially close to the ground-truth, so that unlabeled images can be leveraged to enhance the segmentation model. Similar ideas have also been applied in human parsing. Liu *et al.* [39] introduce adversarial networks on both feature maps and structured labels for cross-domain human parsing. Luo *et al.* [40] develop the Macro-Micro adversarial network to enforce the local and semantic consistency of the parsing results. For human pose estimation, Chou *et al.* [41] propose to impose the adversarial loss upon heatmaps, which encourages the pose estimator to produce reasonable poses. Chen *et al.* [42] design a multi-task network to generate both pose heatmaps and occlusion heatmaps, where two discriminators are adopted to distinguish between plausible estimations and implausible ones. The key to the success of the previous work is the idea of the adversarial learning strategy that helps the produced label maps more geometrically reasonable. In this work, we exploit adversarial learning to encourage the predicted volumes, which could be viewed as label maps with 3D structures, to be more geometrically reasonable under the weakly supervised setting.

III. Method

As illustrated in Fig. 1, the backbone of the proposed method consists of two parts, i.e. a volume estimator and a coordinate regressor. In addition, an auxiliary regression volume discriminator is proposed for adversarial training of our network. In the following subsections, we firstly introduce the proposed volumetric representation for the 3D face shape, then we present the joint voxel and coordinate regression for unified 2D and 3D landmark localization. Finally, we describe the adversarial training strategy for 3D landmark localization in a weakly supervised manner.

A. Semantic Volumetric Representation

Previous works [11], [43] have shown that encoding the landmark positions into heatmap-like or volumetric representations could provide much more discriminative information than naively concatenating the coordinate vectors of 2D or 3D landmarks. This kind of dense supervision makes it easier for fully convolutional networks to learn pixel to pixel mappings, and has been widely used in tasks such as facial landmark localization [8] and human pose estimation [10], [43].

For 3D landmark localization, the volumetric representation proposed in [11] encodes the position of a specific landmark in a volume with a 3D Gaussian centered around the ground truth position. This idea extends the typically used 2D

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on February 22,2020 at 08:45:42 UTC from IEEE Xplore. Restrictions apply.

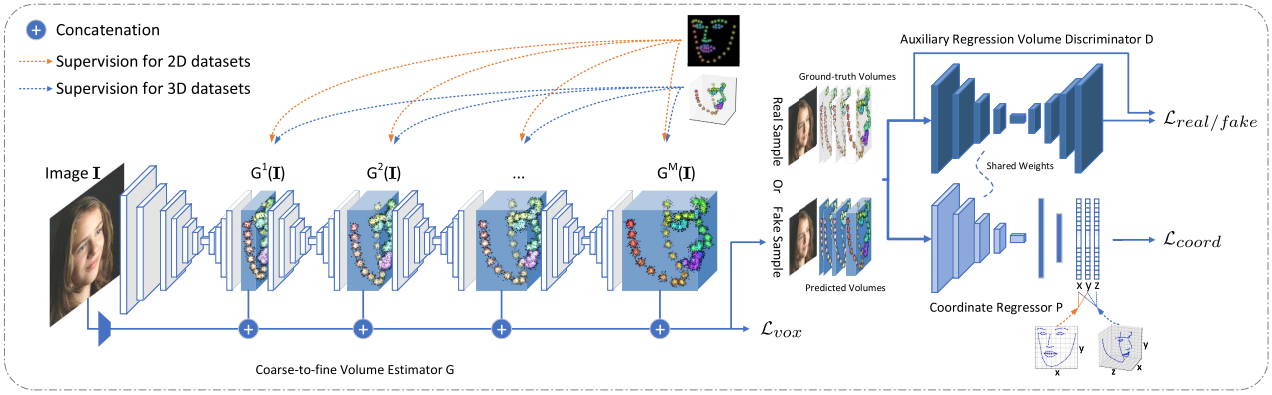

Fig. 1. Illustration of the proposed framework. The backbone network of our method consists of a volume estimator G and a coordinate regressor P. Besides, an auxiliary regression volume discriminator D is employed to encourage the volume estimator to generate plausible volumes, where the encoder of the auto-encoder based discriminator shares weights with the coordinate regressor.

heatmap in a natural manner, which leads to a representation with much larger dimensionality. Although regressing such a representation in a coarse-to-fine manner could alleviate this problem, the curse of dimensionality cannot be avoided when the number of target landmarks increases. Instead of representing each landmark individually in a single volume, we propose the semantic volumetric representation to encode positions of all target landmarks in a more compact manner while preserving their semantic relationship through different colors. Specifically, for the 3D face shape $\boldsymbol{s}$ with N target landmarks $\{\boldsymbol{s}^n\}_{n=1}^N$, N different colors $\{\boldsymbol{c}^n\}_{n=1}^N$ are bound with each landmark, where $\boldsymbol{c}^n = [c_1^n, c_2^n, c_3^n]$ is a triplet denoting the color value of R, G and B channel for the n-th landmark respectively. Coordinates of all target landmarks are converted into a colored (i.e. 3 channels in this case) 3D volume $\mathbf{V}$ with the size of $w \times h \times d$. Let $\mathbf{V}_{l,i,j,k}$ denote the l-th channel color value of the voxel located at (i, j, k). For the n-th landmark located at $\boldsymbol{s}^n = (x^n, y^n, z^n)$, its contribution to $\mathbf{V}_{l,i,j,k}$ can be written as:

$$v_{l,i,j,k}^n = c_l^n \frac{1}{(2\pi)^{\frac{3}{2}}\sigma^3} e^{-\frac{(x^n-i)^2+(y^n-j)^2+(z^n-k)^2}{2\sigma^2}}, \tag{1}$$

where the kernel size σ is set empirically. For the 3D face shape with N target landmarks, the overall contribution to $\mathbf{V}_{l,i,j,k}$ takes as the maximum value in $\{v_{l,i,j,k}^n\}_{n=1}^N$:

$$\mathbf{V}_{l,i,j,k} = \max_n v_{l,i,j,k}^n. \tag{2}$$

Hence, the dimensionality of the representation is fixed regardless of the number of target landmarks. Compared with the compact volumetric representation [20], which could be viewed as a *greyscale* volume discarding semantic meaning of landmark points, the *colored* variant proposed here is more discriminative as shown later in our experiments.

Analogously, for the 2D face shape $\{\boldsymbol{s}^n = (x^n, y^n, 0)\}_{n=1}^N$ without depth annotations (z^n is set as 0 for notation simplicity), we could create the semantic heatmap $\mathbf{H}$ with the size of $w \times h$ in a similar manner. Specifically, the l-th channel color value of the pixel located at (i, j) can be calculated as:

$$\mathbf{H}_{l,i,j} = \max_n c_l^n \frac{1}{2\pi\sigma^2} e^{-\frac{(x^n-i)^2+(y^n-j)^2}{2\sigma^2}}. \tag{3}$$

Note that such a heatmap could also be viewed as a volume with the size of $w \times h \times 1$. In addition, given a 3D volume, the corresponding semantic heatmap could also be obtained through marginalizing the volume along the z dimension:

$$\mathbf{H}_{l,i,j} = \sum_k \mathbf{V}_{l,i,j,k}. \tag{4}$$

In this way, the semantic heatmap could be seen as a byproduct of the proposed semantic volumetric representation. Fig. 2 visualizes the 3D face shape and the corresponding volumetric and heatmap representations. It is worth noting that, in this paper, we also refer to the semantic volumetric representation as semantic volume or volume for simplicity.

B. Joint Voxel and Coordinate Regression

Cascaded regression is wildly employed in 2D landmark localization. Such a strategy could make full use of the regressors and progressively refine the output of the networks. Our method follows this technique and decouples the facial landmark localization problem into the following two subtasks. The first one aims to regress the ideal volumetric representations of landmarks in a coarse-to-fine manner. The second one aims to regress the coordinate vectors of landmarks from the predicted volumetric representations and the input image.

1) Coarse-to-fine Volume Estimator: The volume estimator G learns the mapping from pixels of the input image $\mathbf{I}$ to the volumetric representations $\mathcal{V}$: $G(\mathbf{I}) \rightarrow \mathcal{V}$. Inspired by previous works [10], [11] on 2D and 3D human pose estimation, we also adopt the Stacked Hourglass Networks [10] with intermediate supervision and skip connection. Specifically, the volume estimator consists of M stacked Hourglass modules [10] of which supervisions are ground-truth volumes denoted as $\mathcal{V} = \{\mathbf{V}^m\}_{m=1}^M$. Then, the volume estimator is trained using the voxel-wise mean squared error loss:

$$\mathcal{L}_{vox}(\mathcal{V}|\mathbf{I}) = \sum_m^M \sum_{l,i,j,k} \left\| G^m(\mathbf{I})_{l,i,j,k} - \mathbf{V}_{l,i,j,k}^m \right\|^2, \tag{5}$$

where $G^m(\cdot)$ denotes the volume outputted by the m-th Hourglass module.

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on February 22,2020 at 08:45:42 UTC from IEEE Xplore. Restrictions apply.

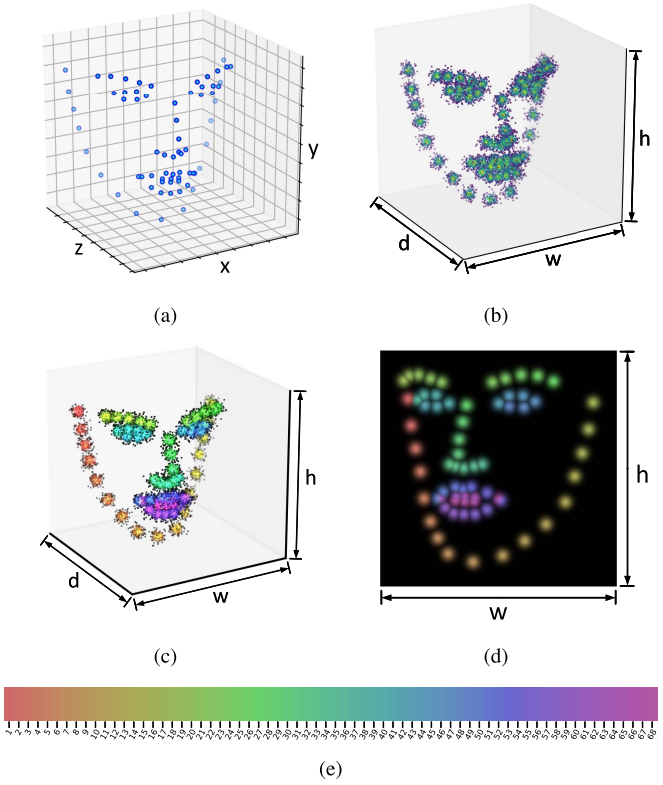

Fig. 2. Visualization of different representations. (a) 3D plot of a face shape in the coordinate system; (b) The compact volumetric representation; (c) The semantic volumetric representation; (d) The semantic heatmap obtained by marginalizing the corresponding semantic volumetric representation; (e) The correspondence between landmark indexes and colors. For both compact and semantic volumetric representation, the voxel values are indicated by the density of the point cloud. For the semantic volumetric representation and heatmap, voxels/pixels around different landmarks are colored with the corresponding colors. For simplicity, volumetric representation is also referred to as volume in this paper. Best viewed in electronic form.

As pointed out in [11], making accurate prediction along the z dimension is much more challenging than another two dimensions. Hence, regressing volumes with the increasing resolution along the z dimension in a coarse-to-fine manner could be more effective and robust. In practice, the resolution d of $\mathbf{V}^m$ takes the number from preset values and progressively increases along with m.

2) Coordinate Regressor: Typical coordinate regression based methods predict the coordinate vectors or their increments directly from image features, while typical heatmap regression based methods [8], [10], [44] retrieve the coordinates of landmarks from the peak points of the corresponding heatmaps. In our case, however, the conventional "taking-maximum" operation is no longer applicable since the positions of all landmarks are encoded into a single volume. By combining the two forms of regression based methods, we propose to infer coordinates of landmarks from both volumetric representations and the corresponding input image for robust landmark localization.

To this end, we adopt a coordinate regressor P to learn the mapping from volumetric representations $\mathcal{V} = \{\mathbf{V}^m\}_{m=1}^M$ and the input image $\mathbf{I}$ to the corresponding coordinate vector $\boldsymbol{s}$: $P(\mathcal{V}, \mathbf{I}) \rightarrow \boldsymbol{s}$. Inspired by the work on 3D object recognition [45] and hand pose estimation [46], we employ 3D convolution kernels instead of 2D kernels in our coordinate regressor. The 3D convolution is typically used to extract features from both spatial and temporal dimensions for video analysis problems [47], hence it could be more naturally adopted to extract 3D information from the volumetric representation. The proposed coordinate regressor consists of five 3D convolutional layers, with batch normalization and Leaky ReLU activation added in between and a fully connected layer at the end. For training, we employ $\mathcal{L}_2$ regression loss to measure the error of predicted coordinate vectors:

$$\mathcal{L}_{coord}(\boldsymbol{s}|\mathcal{V}, \mathbf{I}) = \|\boldsymbol{s} - P(\mathcal{V}, \mathbf{I})\|_2^2, \tag{6}$$

where $\boldsymbol{s}$ denotes the concatenated vector of the ground-truth landmark coordinates.

3) Unified 2D/3D Facial Landmark Localization: Since the semantic heatmap and 2D coordinate vector are byproducts under the proposed framework, we can naturally adopt a mixed training technique utilizing both 2D and 3D datasets simultaneously. Let $\mathcal{I}_{2D}$ and $\mathcal{I}_{3D}$ denote the 2D and 3D datasets respectively. Each training sample $\{\mathbf{I}, \boldsymbol{s}, \mathcal{V}\}$ consists of the input image $\mathbf{I}$, the ground-truth face shape $\boldsymbol{s}$, and the ground-truth volumes $\mathcal{V}$ created according to $\boldsymbol{s}$, where $\boldsymbol{s} = \{(x^n, y^n)\}_{n=1}^N$ and $\mathcal{V} = \{\mathbf{H}^m\}_{m=1}^M$ for 2D datasets, $\boldsymbol{s} = \{(x^n, y^n, z^n)\}_{n=1}^N$ and $\mathcal{V} = \{\mathbf{V}^m\}_{m=1}^M$ for 3D datasets. Then, the loss function for the volume estimator G is unified as:

$$\mathcal{L}_{vox}(\mathcal{V}|\mathbf{I}) = \begin{cases} \sum_m^M \sum_{l,i,j} \left\| \sum_k G^m(\mathbf{I})_{l,i,j,k} - \mathbf{H}^m_{l,i,j} \right\|^2, & if\ \mathbf{I} \in \mathcal{I}_{2D} \\ \sum_m^M \sum_{l,i,j,k} \left\| G^m(\mathbf{I})_{l,i,j,k} - \mathbf{V}^m_{l,i,j,k} \right\|^2, & if\ \mathbf{I} \in \mathcal{I}_{3D}. \end{cases} \tag{7}$$

The loss function for the coordinate regressor P is also unified as:

$$\mathcal{L}_{coord}(\boldsymbol{s}|\mathcal{V}, \mathbf{I}) = \begin{cases} \left\| [\boldsymbol{s}]^{2D} - [P(\mathcal{V}, \mathbf{I})]^{2D} \right\|_2^2, & if\ \mathbf{I} \in \mathcal{I}_{2D} \\ \|\boldsymbol{s} - P(\mathcal{V}, \mathbf{I})\|_2^2, & if\ \mathbf{I} \in \mathcal{I}_{3D}, \end{cases} \tag{8}$$

where $[\cdot]^{2D}$ denotes the operator extracting 2D coordinates from 3D coordinate vector. With the proposed unified framework, the network can be jointly trained with both synthetic images fully annotated with 3D landmarks and real-world images annotated with only 2D landmarks.

4) Two-stage Training: Instead of training the whole network from scratch, we adopt a two-stage training scheme which is more stable and effective in our experiments. The two subnetworks mentioned above are pre-trained separately for each sub-task beforehand and finally fine-tuned as an integrated one. Specifically, at the pre-training stage, the volume estimator is trained with input images and ground-truth volumes. Meanwhile, the coordinate regressor is trained with ground-truth volumes (concatenated with the downsampled input images) and the corresponding coordinate vectors.

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on February 22,2020 at 08:45:42 UTC from IEEE Xplore. Restrictions apply.

At the fine-tuning stage, the coordinate regressor is attached to the volume estimator, and the whole network is fine-tuned with joint supervision of both ground-truth volumes and coordinate vectors. Formally, the whole network is trained in an end-to-end manner using the weighted sum of two regression loss terms at the final stage:

$$\mathcal{L} = \mathcal{L}_{vox}(\mathcal{V}|\mathbf{I}) + \lambda_{coord}\mathcal{L}_{coord}(\boldsymbol{s}|\hat{\mathcal{V}}, \mathbf{I}), \tag{9}$$

where $\hat{\mathcal{V}} = \{G^m(\mathbf{I})\}_{m=1}^{M}$ denotes the predicted volumes, and λ_{coord} is used to balance the two terms.

C. Auxiliary Regression Adversarial Learning

Existing 2D annotated face datasets consist of real-world images covering a diverse range of poses, expressions, occlusions and illuminations. For the task of 3D landmark localization, those 2D annotations could be regarded as weak labels due to the lack of depth information. For better leveraging 2D and 3D annotated datasets, we further adopt the adversarial training strategy to distill the geometric structures from 3D annotated datasets to real-world images. To this end, the volume estimator is treated as a generator which aims at producing plausible volumes for both synthetic and real-world images. Meanwhile, a volume discriminator is adopted to distinguish ground-truth volumes from those generated by the volume estimator. During training, the volume estimator is learned to estimate volumes that are indistinguishable from the ground-truth ones. In this way, the volume estimator is regularized so that the estimations on real-world images are imposed to have a similar distribution with the ground-truth volumes.

1) Auxiliary Regression Volume Discriminator: The discriminator D is designed to tell whether the volumes are geometric plausible or not. Since a plausible face shape for a particular image may still be inaccurate for another face image, the input of the discriminator contains the concatenation of the volumes and their corresponding face images. For training the discriminator, those inputs comprising the ground-truth volumes are treated as real samples, and those comprising volumes predicted by the estimator G on both 2D and 3D datasets are treated as fake samples. Moreover, the discriminator adopts an autoencoder-like architecture and computes reconstruction errors $\mathcal{L}_{real}$ and $\mathcal{L}_{fake}$ for real samples and fake samples, respectively. Specifically, for real samples, the discriminator is optimized to reconstruct volumes similar to the given ground-truths, i.e., minimizing $\mathcal{L}_{real}$. On the other hand, for fake samples, the discriminator is optimized to reconstruct volumes different to the given estimations, i.e., maximizing $\mathcal{L}_{fake}$. In this way, the discriminator could be viewed as an energy function which assigns low energy to real samples and high energy to fake ones. Formally, both $\mathcal{L}_{real}$ and $\mathcal{L}_{fake}$ adopt the voxel-wise mean squared error loss, which could be written as:

$$\begin{aligned}\mathcal{L}_{real}(\mathcal{V}, \mathbf{I}) &= \sum_{m=1}^{M} \left\| \mathbf{V}^m - D^m(\mathcal{V}, \mathbf{I}) \right\|_2^2 \\ \mathcal{L}_{fake}(\hat{\mathcal{V}}, \mathbf{I}) &= \sum_{m=1}^{M} \left\| \hat{\mathbf{V}}^m - D^m(\hat{\mathcal{V}}, \mathbf{I}) \right\|_2^2,\end{aligned} \tag{10}$$

Algorithm 1 The Training Process of the Proposed Method

Require:

Training images $\mathbf{I}$, the corresponding ground-truth face shape $\boldsymbol{s}$ and volumes $\mathcal{V}$;

▷ **Pre-training stage**

*/*Using training samples:* $\{\mathbf{I}, \boldsymbol{s}, \mathcal{V}\} \in \mathcal{I}_{3D}$**/*

The volume estimator G is optimized using Eq. (5) with the images $\mathbf{I}$ and the ground-truth volumes $\mathcal{V}$;

The coordinate regressor P is optimized using Eq. (6) with images $\mathbf{I}$, the ground-truth volumes $\mathcal{V}$ and face shape $\boldsymbol{s}$;

▷ **Adversarial learning stage**

*/*Using training samples:* $\{\mathbf{I}, \boldsymbol{s}, \mathcal{V}\} \in \mathcal{I}_{2D} \cup \mathcal{I}_{3D}$**/*

repeat

  Forward D and P by $D(\mathcal{V}, \mathbf{I})$ and $P(\mathcal{V}, \mathbf{I})$;

  Forward G by $\hat{\mathcal{V}} = G(\mathbf{I})$;

  Forward D and P by $D(\hat{\mathcal{V}}, \mathbf{I})$ and $P(\hat{\mathcal{V}}, \mathbf{I})$;

  Optimize D and P by minimizing Eq. (11);

  Optimize G by minimizing Eq. (14);

until convergence is reached

where $D^m(\cdot, \cdot)$ denotes the m-th volume reconstructed by the discriminator.

As shown in previous work on GAN technique, utilizing side information in the GAN framework could improve the training procedure [32]. Motivated by this, we assign the discriminator an auxiliary task, that is, regressing landmark coordinates from the input images and volumes. In practice, the encoder of the discriminator shares weights with the coordinate regressor P, and the discriminator is optimized along with the coordinate regression loss. Therefore, the overall loss function for training the discriminator is written as follows:

$$\begin{aligned}\mathcal{L}_D = &\sum_{\mathbf{I} \in \mathcal{I}_{3D}} \mathcal{L}_{real}(\mathcal{V}, \mathbf{I}) + \lambda_{aux}\mathcal{L}_{coord}(\boldsymbol{s}|\mathcal{V}, \mathbf{I}) \\ &+ \sum_{\mathbf{I} \in \mathcal{I}_{2D} \cup \mathcal{I}_{3D}} \left(-\mu_t \mathcal{L}_{fake}(\hat{\mathcal{V}}, \mathbf{I}) + \lambda_{aux}\mathcal{L}_{coord}(\boldsymbol{s}|\hat{\mathcal{V}}, \mathbf{I}) \right),\end{aligned} \tag{11}$$

where the weight λ_{aux} is used to balance the auxiliary regression loss and reconstruction loss. Inspired by previous work [30], [41], we introduce a variable $\mu_t \in [0, 1]$ to control the learning procedure of the discriminator and generator. Formally, the variable μ_t is initialized as 0 and updated at each training step t using the following equation:

$$\mu_{t+1} = \mu_t + \lambda_\mu(\gamma\, \mathcal{L}_{real}(\mathcal{V}, \mathbf{I}) - \mathcal{L}_{fake}(\hat{\mathcal{V}}, \mathbf{I})), \tag{12}$$

where λ_μ and γ are two hyper-parameters. During training, μ_t is used to control the emphasis on $\mathcal{L}_{fake}$, and adjusted proportionally according to the gap between $\gamma\, \mathcal{L}_{real}$ and $\mathcal{L}_{fake}$. In this way, the discriminator is optimized more adaptively so that the training of GAN could be more stable.

2) Adversarial Training of Volume Estimator: Under the proposed framework, the volume estimator G is treated as a generator and aims to produce geometrically plausible volumes for given images. The discriminator punishes the volume estimator when its predictions are far from being satisfactory.

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on February 22,2020 at 08:45:42 UTC from IEEE Xplore. Restrictions apply.

Specifically, besides the volume regression loss, the volume estimator G is optimized with the training signals back-propagated from the discriminator and the coordinate regressor, including the coordinate regression loss $\mathcal{L}_{coord}(\boldsymbol{s}|\hat{\mathcal{V}}, \mathbf{I})$ and the adversarial loss $\mathcal{L}_{adv}(\hat{\mathcal{V}}, \mathbf{I})$, where

$$\mathcal{L}_{adv}(\hat{\mathcal{V}}, \mathbf{I}) = \sum_{m=1}^{M} \left\| \hat{\mathbf{V}}^m - D(\hat{\mathbf{V}}^m, \mathbf{I}) \right\|_2^2. \quad (13)$$

The total loss for the volume estimator G is written as:

$$\mathcal{L}_G = \sum_{\mathbf{I} \in \mathcal{I}_{2D} \cup \mathcal{I}_{3D}} \mathcal{L}_{vox}(\mathcal{V}|\mathbf{I}) + \lambda_{coord}\mathcal{L}_{coord}(\boldsymbol{s}|\hat{\mathcal{V}}, \mathbf{I}) + \lambda_{adv}\mathcal{L}_{adv}(\hat{\mathcal{V}}, \mathbf{I}). \quad (14)$$

We also pretrain the volume estimator and coordinate regressor before adopting the adversarial training strategy. At the pre-training stage, the volume estimator and coordinate regressor are optimized by the loss function Eq. (5) and Eq. (6) respectively, using the synthetic dataset fully annotated with 3D landmarks. After that, the synthetic dataset with 3D annotations and the real-world dataset with 2D annotations are mixed for training. During this stage, the generator (i.e. volume estimator) and discriminator are optimized alternately under the proposed auxiliary regression adversarial learning framework. The whole training process is summarized in Algorithm 1.

IV. Experiments

In this section, we evaluate the proposed method on both 2D and 3D facial landmark localization tasks. We first describe the implementation details of the proposed method. Then, we introduce the datasets as well as the evaluation metrics used in our experiments. After that, we present the experimental results of the proposed method. In the end, we conduct ablation study to investigate the effectiveness of components proposed in our method.

A. Experimental Settings

The proposed network takes a 256×256 face image as input and outputs predictions of the volumetric representation and the coordinate vector. Following the setting of [11], four Hourglass modules are stacked together as the volume estimator (i.e. $M = 4$) and output volumes with the size of $64 \times 64 \times d$, where the resolution d in z-dimension is chosen from the set $\{1, 2, 4, 64\}$ successively. The Gaussian kernel size in Eq. (1) is empirically set as $\sigma = 1$. The weights used in Eqs. (9), (11) and (14) are set as $\lambda_{coord} = 1$, $\lambda_{aux} = 0.1$ and $\lambda_{adv} = 0.001$ respectively. Among them, parameters λ_{coord} and λ_{aux} are simply selected in order to make values of the corresponding terms have similar scales with respect to the overall loss function. The parameter λ_{adv} is selected to be relatively small so that the original regression terms could still play dominant roles in $\mathcal{L}_G$. The hyper-parameters used in Eq. (12) are set as $\lambda_\mu = 0.001$ and $\gamma = 0.7$ respectively in our experiments, which is followed the setting of BEGAN [30]. During training, data augmentation techniques, such as rotation (± 40Â°), scaling (0.7 to 1.3), color jittering ($\pm 30\%$ channel-wise) and flipping, are applied randomly to input images. In the experiments, the network is pre-trained for 20 epochs on a 3D dataset, and then the model is fine-tuned for 10 epochs on the same 3D dataset or trained for 20 epochs on the mixture of 2D and 3D datasets. We adopt the ADAM [48] optimization algorithm with an initial learning rate of 2.5×10^{-4} to train the model, and reduce the learning rate to 2.5×10^{-5} after 20 epochs. Our approach is implemented using PyTorch.

B. Time Complexity

In the proposed framework, only the volume estimator and coordinate regressor are involved during testing. Our model (with 4 Hourglass modules) can run in realtime, with the speed of 23fps tested on an NVIDIA TITAN Xp GPU. The most time-consuming part of our model is the volume estimator since it takes about 10ms for one Hourglass module to process an image. The introduced execution time of the coordinate regressor is minor since it accounts for less than 2ms during testing. In practical usage, we can reduce the number of Hourglass modules from 4 to 2 to double the speed while obtaining comparable results (see Table V). It is also worth noting that the speed of our approach using either the semantic or compact volume is nearly the same since only several layers are different in the network architecture.

C. Datasets

We evaluate the proposed method on multiple face datasets including synthetic datasets with 3D annotations and real-world datasets with 2D annotations. Details of these datasets used in our experiments are listed below.

3DFAW [12] is provided by organizers of the 3D Face Alignment in the Wild (3DFAW) Challenge [12]. It contains more than 23000 face images from BU-4DFE [49], BP4D-Spontaneous [50] and MultiPIE [51]. For each face image in the dataset, 66 3D facial landmarks as well as the face bounding box are annotated. The 3D points are annotated consistently using a model-based structure-from-motion technique [52]. The 3DFAW dataset is divided into three subsets: training set, validation set and test set, containing 13969, 4725 and 4912 face images, respectively. Our method is trained on the training set and tested on both the validation and test set. It should be noted that the ground-truth 3D landmarks of the test set are not publicly available. Hence the evaluation results on the test set are provided by 3DFAW Challenge organizers via the CodaLab platform[1].

300W-LP [13] contains 61225 synthetic face images across large poses ranging from -90Â° to 90Â°. Those images are synthesized from 300W [53] using the 3D morphable model based profiling algorithm proposed in [13]. For each face, 68 3D landmarks are retrieved from the parameters of the 3D morphable model, using the released code of [13]. In our experiments, the depth values are normalized to have zero mean. The 300W-LP dataset is only used for training, and our method is evaluated on the AFLW2000-3D dataset mentioned below.

[1] https://competitions.codalab.org/competitions/10261

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on February 22,2020 at 08:45:42 UTC from IEEE Xplore. Restrictions apply.

AFLW [54] is a large-scale real-world dataset for facial landmark localization, which contains 25,993 faces covering large variations in appearance and environmental conditions. The original dataset provides up to 21 annotated points visible on each face. To obtain 2D landmarks under 3D perspective, we use the algorithm proposed in [2] to augment the annotations to 68 landmarks. In our experiments, 20000 faces exclusive of those in AFLW2000-3D [13] are used as 2D training data.

AFLW2000-3D [13] contains 2000 face samples selected from the AFLW [54] dataset, introduced by Zhu et al. [13] along with the 300W-LP dataset. The 68 3D landmarks annotated in AFLW2000-3D are consistent with those of 300W-LP. We use the AFLW2000-3D dataset only for testing in our experiments, following the common protocol in the literature [13], [55].

D. Evaluation Metrics

For fair comparison, same evaluation metrics are adopted as in previous works [12], [13].

For evaluation on 3DFAW datasets, the Ground Truth Error (GTE) and Cross View Ground Truth Consistency Error (CVGTCE) are used to measure the performance as recommenced in the 3DFAW Challenge [12]. GTE is defined as the average point-to-point Euclidean error normalized by the distance between the outer corners of the eyes, which could be computed as:

$$GTE(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \frac{1}{N}\sum_{n=1}^{N}\frac{\left\|\boldsymbol{s}^n - \hat{\boldsymbol{s}}^n\right\|_2}{r_i} \tag{15}$$

where $\boldsymbol{s}$ and $\hat{\boldsymbol{s}}$ are the prediction and ground truth respectively, and r_i denotes the normalized distance of the i-th image.

CVGTCE is proposed in the 3DFAW Challenge and aims at evaluation of the cross-view consistency of the predicted landmarks, which is defined as follows:

$$CVGTCE(\boldsymbol{s}, \hat{\boldsymbol{s}}, \mathbf{p}) = \frac{1}{N}\sum_{n=1}^{N}\frac{\left\|(\alpha\mathbf{R}\boldsymbol{s}^n + \mathbf{t}) - \hat{\boldsymbol{s}}^n\right\|_2}{r_i} \tag{16}$$

where the parameter $\mathbf{p} = \{\alpha, \mathbf{R}, \mathbf{t}\}$ denotes the rigid transformation, i.e. scale, rotation and translation, which are obtained by minimizing the follow objective function:

$$\{\alpha, \mathbf{R}, \mathbf{t}\} = \arg\min_{\alpha,\mathbf{R},\mathbf{t}}\sum_{n=1}^{N}\left\|\hat{\boldsymbol{s}}^n - \left(\alpha\mathbf{R}\boldsymbol{s}^n + \mathbf{t}\right)\right\|_2 \tag{17}$$

For evaluation on AFLW2000-3D, the metric is chosen as the Normalized Mean Error (NME), which is defined as the average point-to-point Euclidean error normalized by the square root of the bounding box size. The formulation of NME could be written as the same as Eq. (15), where the normalized distance r_i is adapted as the bounding box size. Note that the bounding box size is calculated from 2D landmarks for both tasks of 2D and 3D facial landmark localization, which is consistent with previous work [2], [56].

TABLE I
COMPARISON OF GROUND TRUTH ERROR (GTE) ON THE VALIDATION SET OF 3DFAW

Method	GTE (%)
SDM+3DMM [16]	6.34
Gou et al. [16]	5.90
Bulat et al. [15]	4.94
JVCR [20]	4.36
Ours	**4.07**

E. 3D Facial Landmark Localization

In this subsection, we compare our approach with existing methods on synthetic dataset 3DFAW and real-world dataset AFLW2000-3D for the task of 3D facial landmark localization.

1) Evaluation on 3DFAW: Since both the training and testing samples of 3DFAW are comprised of synthetic images, we train the network with 3D dataset without using the adversarial learning strategy. The evaluation on 3DFAW consists of two parts. The first part is evaluated on the validation set, and the second part is the performance evaluation on the test set where results are provided by the challenge organizers.

Table I shows the comparison results of GTE on the validation set. It can be observed that the proposed method outperforms others, most of which except our previous work [20] are based on the two-step strategy. For comparison with top ranked methods on the 3DFAW Challenge, we further evaluate our method on the test set. Comparisons of both CVGTCE and GTE on the 3DFAW test set are reported in Table II. Note that the ground truth 3D landmarks of the test set are not available to the participants, and the numbers for all methods are taken from the CodaLab leaderboard and literature [12], [58]. Our method achieves the best result in comparison with other methods, including the previously top-ranked method [15] and Tulyakov *et al.* [58] which is built upon a 3D variant of cascaded regression method. The previous top two methods, Bulat and Tzimiropoulos [15] and Tulyakov *et al.* [58], belong to heatmap regression based methods and coordinate regression based methods respectively. Our method outperforms them considerably since the proposed joint voxel and coordinate regression pipeline combines merits of the robustness of heatmap regression based methods and the accuracy of coordinate regression based methods. In addition, the end-to-end regression of 3D landmark shapes also contributes to the superior performance over those methods using the two-step strategy, which typically perform the 2D landmark localization followed by depth prediction. Our method proposed in this paper also outperforms our previous work [20], which uses compact volumes as regression targets of the volume estimator. It can be seen in Table II that there is around 7% improvement over the previously best method [15] and 3% improvement over our previous work JVCR [20]. Fig. 5 shows example results of the proposed method on the 3DFAW validation set.

2) Evaluation on AFLW2000-3D: We further evaluate our method on AFLW2000-3D to demonstrate the effectiveness of our method on face images with large poses and appearance variations. In this case, our model is trained on 300W-LP

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on February 22,2020 at 08:45:42 UTC from IEEE Xplore. Restrictions apply.

TABLE II

COMPARISONS OF CROSS VIEW GROUND TRUTH CONSISTENCY ERROR (CVGTCE) AND GROUND TRUTH ERROR (GTE) ON THE TEST SET OF 3DFAW

Method	CVGTCE (%)	GTE (%)
Zavan et al. [57]	5.90	10.80
Gou et al. [16]	4.94	6.20
Zhao et al. [14]	3.97	5.88
Bulat et al. [15]	3.47	4.56
Tulyakov et al. [58]	3.80	5.10
JVCR [20]	3.46	4.35
Ours	**3.29**	**4.24**

TABLE III

COMPARISON OF NORMALIZED MEAN ERROR (NME) ON AFLW2000-3D

Method	NME(%)
3DDFA [13]	7.51
DeFA [59]	6.23
3D-FAN [2]	5.24
PRN [56]	4.70
JVCR [20]	4.39
Ours	**4.11**

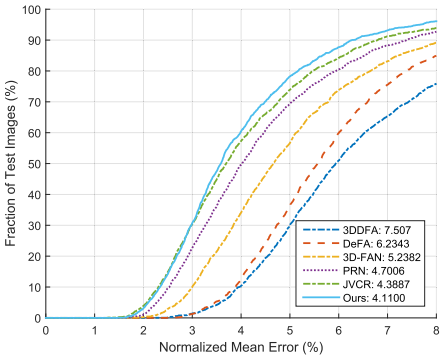

Fig. 3. Comparison of Cumulative Errors Distribution (CED) curves on AFLW2000-3D. 68 landmarks with 3D coordinate are considered in the evaluation. Curve of other methods are borrowed from literature [56].

and AFLW datasets using the proposed adversarial learning strategy after the pre-training stage on 300W-LP. Comparisons of NME and CED curves against other state-of-the-art methods are shown in Table III and Fig. 3 respectively. Our method outperforms all previous methods including two-step strategy based method 3D-FAN [2] and 3D face model based methods such as 3DSTN [55] and PRN [56]. Likewise, our method also obtains significant improvement over the most recent state-of-the-art [56] and our previous work [20]. Besides the proposed end-to-end pipeline, our success could also be attributed to the adversarial learning of the model since it helps to distill structural constraints from 3D annotated datasets to real-world images. This would be further demonstrated in our ablation experiments later. Example results of our method on AFLW2000-3D are depicted in Fig. 6. It can be seen that our method is robust to occlusions and large appearance variations occurred in real-world face images.

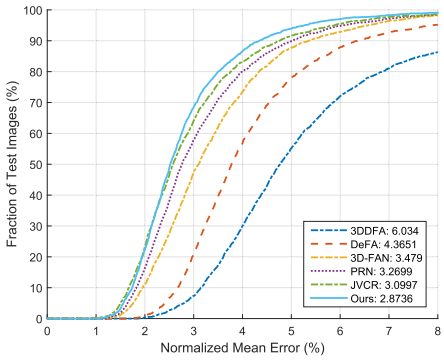

Fig. 4. Comparison of Cumulative Errors Distribution (CED) curves on AFLW2000-3D. 68 landmarks with 2D coordinate are considered in the evaluation. Curve of other methods are borrowed from literature [56].

F. 2D Facial Landmark Localization

We also compare our method with other state-of-the-arts on AFLW2000-3D for 2D facial landmark localization task. In this case, only 2D coordinates are involved in the evaluation and Normalized Mean Error (NME) is chosen as the evaluation metric. Note that the model is trained in the same way as the model evaluated on AFLW2000-3D in the 3D case. The CED curves for various methods are shown in Fig. 4. The proposed method achieves superior performance compared with others. To further evaluate the proposed method across poses, we report NME on three subsets which are divided according to their head yaw angles. The comparison with existing methods for 21 and 68 landmarks are shown in Table IV. The results of RCPR [60], ESR [3], and SDM [4] are obtained from [13] and these methods have been retrained on 300W-LP for adaptation to large poses. Our method is barely surpassed by 3D-FAN [2] when evaluating 21 landmarks, where only visible landmarks are involved in the evaluation and performances of existing methods are nearly saturated. The inferior performance on the evaluation of 21 landmarks is also in part due to the inconsistency of the annotation schemes, since the ground-truth 21 landmarks with visible labels provided in AFLW are not perfectly aligned with the corresponding subset of 68 landmarks. When considering all 68 landmarks, the testing case is consistent with the training and our method outperforms others considerably. It can be seen from Table IV that the proposed method achieves superior performance especially for middle and large poses.

G. Ablation Study

Several components proposed in our method jointly contribute to the success of our method. To evaluate the efficacy of each component, we conduct ablation experiments using different configurations of our method. We first give comprehensive investigations on the introduced semantic volumetric representation, then validate the two-stage training scheme, and finally give deeper analyses of the proposed auxiliary regression adversarial learning strategy.

1) Semantic volumetric representation: For ablation experiments in this part, we train the network with the 300W-LP dataset without using the adversarial learning strategy and

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on February 22,2020 at 08:45:42 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
COMPARISON OF NME (%) ON AFLW2000-3D. ONLY 2D COORDINATES ARE INVOLVED IN THE EVALUATION

	AFLW2000-3D (21 pts, visible only)					AFLW2000-3D (68 pts)				
Method	[0°, 30°]	[30°, 60°]	[60°, 90°]	mean	std	[0°, 30°]	[30°, 60°]	[60°, 90°]	mean	std
RCPR [60]	5.43	6.58	11.53	7.85	3.24	4.26	5.96	13.18	7.80	4.74
ESR [3]	5.66	7.12	11.94	8.24	3.29	4.60	6.70	12.67	7.99	4.19
SDM [4]	4.75	5.55	9.34	6.55	2.45	3.67	4.94	9.76	6.12	3.21
3DDFA [13]	5.00	5.06	6.74	5.60	0.99	3.78	4.54	7.93	5.42	2.21
3DDFA+SDM [13]	4.75	4.83	6.38	5.32	0.92	3.43	4.24	7.17	4.94	1.97
Yu et al. [61]	5.94	6.48	7.96	6.79	1.05	3.62	6.06	9.56	6.41	2.99
3DSTN [55]	3.55	3.92	5.21	4.23	0.87	3.15	4.33	5.98	4.49	1.42
RDR [62]	3.63	4.29	5.31	4.41	0.85	-	-	-	-	-
HyperFace [63]	3.93	4.14	4.71	4.26	0.41	-	-	-	-	-
3D-FAN [2]	**3.44**	**3.76**	**4.27**	**3.82**	**0.42**	2.77	3.48	4.60	3.62	0.92
DenseFA [59]	-	-	-	-	-	-	-	-	4.50	-
PRN [56]	3.80	4.10	4.95	4.28	0.59	2.75	3.51	4.61	3.62	0.94
JVCR [20]	3.74	4.02	4.82	4.19	0.56	2.94	3.46	4.53	3.64	0.81
Ours	3.46	3.84	4.44	3.91	0.49	**2.69**	**3.08**	**4.15**	**3.31**	**0.76**

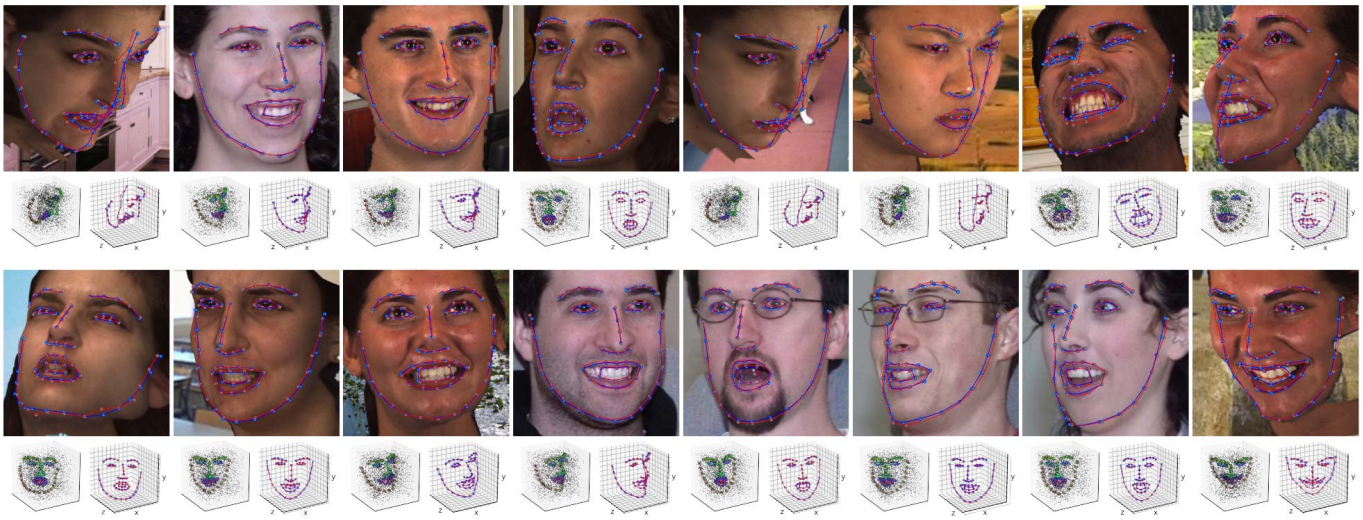

Fig. 5. Example results of the proposed method on 3DFAW. The upper row shows the face images as well as the 2D facial landmark localization results. The lower row shows the predicted volumetric representations (left) and 3D facial landmarks (left). Red and blue indicate the ground-truth and estimated landmark shapes respectively.

report results on the AFLW2000-3D dataset. To demonstrate the advantage of the proposed semantic volumetric representation over the compact volumetric representation, we replace the supervision of the volume estimator with the compact volumetric representation.

a) Number of hourglass modules: We vary the number of stacked Hourglass modules from 1 to 8 for comprehensive analyses. As shown in Table V, approaches using semantic volumes as the supervision outperform those using the compact ones, which is consistent across different numbers of Hourglass modules adopted in the volume estimator. Moreover, the one stacked network supervised by semantic volumes obtains the result comparable to the two or three stacked network supervised by compact volumes, though nearly a half to a third of parameters are retained. This suggests that the semantic volume is preferred especially in practical applications where the computational power is limited. It is also worth noting that the improvement brought by using more Hourglass modules is marginal and becomes saturated especially when the stacked number is greater than 4.

b) Number of target landmarks: We also investigate the impact of the number of target landmarks. The following three subsets selected from the original 68 landmark annotations are used as new target landmark shapes for training and evaluation. The first subset contains 5 points including eye centers, nose tip and mouth corners. The second subset contains 21 points semantically similar to the annotation scheme of AFLW. The third subset contains the remaining 51 points after removing the 17 points of the face's boundary from 68 points. The landmark definitions are consistent during the training and evaluation phases. For fair comparison, the bounding box size calculated from 68 points is used as the normalization distance for the evaluation of all subsets. Results of our method with respect to the different number of target landmarks are reported in Table VI. It can be seen that the proposed semantic volumetric representation is superior to the compact one regardless of the number of target landmarks. What's more, the semantic volumes bring significantly more improvement over the compact volumes when decreasing the number of target landmarks. This is because the compact volumes are

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on February 22,2020 at 08:45:42 UTC from IEEE Xplore. Restrictions apply.

Fig. 6. Example results of the proposed method on AFLW2000-3D. Images are arranged in the same layout as Fig. 5.

TABLE V

3D FACIAL LANDMARK LOCALIZATION PERFORMANCE (NME %) OF APPROACHES SUPERVISED WITH COMPACT VOLUMES VERSUS SEMANTIC VOLUMES ON AFLW2000-3D.

# Stacks	Resolution d Assignment	Type of Volume	
		Compact	Semantic
1	[64]	4.52	4.41
2	[1, 64]	4.40	4.34
3	[1, 2, 64]	4.38	4.31
3	[1, 4, 64]	4.37	4.32
4	[1, 2, 4, 64]	4.29	4.25
4	[1, 4, 16, 64]	4.31	4.30
6	[1, 2, 4, 8, 16, 64]	4.30	4.26
8	[1, 2, 4, 8, 16, 32, 64, 64]	4.31	4.28

TABLE VI

3D FACIAL LANDMARK LOCALIZATION PERFORMANCE (NME %) OF ABLATION APPROACHES REGARDING THE DIFFERENT NUMBER OF TARGET LANDMARKS ON AFLW2000-3D

# Target Landmarks	Type of Volume	
	Compact	Semantic
5	3.54	3.33
21	4.09	3.95
51	3.61	3.50
68	4.29	4.25

more ambiguous when the target landmark shapes become sparser, which makes it more difficult for the coordinate regressor to infer the relationships between landmarks. The semantic volume overcomes this issue since such a representation encodes the landmark relationships and contains more complete information.

c) Appearance of the semantic volume: The color assignment strategy and kernel size σ contribute to the variant appearance of the semantic volume. Intuitively, assigning contrasting colors to adjacent landmarks and adopting a smaller kernel size would make the volume clearer from the perspective of human being. To evaluate whether these alternative options could help in boosting the performance, we randomize the landmark indexes to assign colors and vary the kernel size σ from 0.5 to 5 as well. Table VII reports performance of approaches adopting different color assignment strategies and kernel sizes. As can be seen, adopting either sorted or random color assignment strategy could result in similar performance, which means that the appearance of volumes has no direct impact on the final results. This could be due to the fact that the volumes are intermediate representations in our network and the coordinate regressor is learned in an end-to-end manner. On the other side, adopting larger kernel sizes of the volume is feasible in our model. When $\sigma \geqslant 1$, approaches with different kernel sizes achieve similar performance. However, the performance gets inferior when the kernel size is too small ($\sigma < 1$). One reasonable explanation is that the volumes become too sparse when the kernel size is too small, making it harder for the coordinate regressor to infer landmark shapes from them.

TABLE VII

3D FACIAL LANDMARK LOCALIZATION PERFORMANCE (NME %) OF APPROACHES ADOPTING DIFFERENT COLOR ASSIGNMENT STRATEGY AND KERNEL SIZE σ ON AFLW2000-3D

Size of σ	Color Assignment Strategy	
	Sorted	Random
0.5	4.40	4.44
1	4.25	4.26
2	4.28	4.27
5	4.24	4.25

2) Two-stage training: The two-stage training scheme is shown to be more stable and effective in our experiments. As an alternative, the one-stage training scheme refers to training the whole network from scratch. Table IX reports the results of approaches using different training schemes. It can be seen that the approach adopting the two-stage training scheme

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on February 22,2020 at 08:45:42 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII

2D FACIAL LANDMARK LOCALIZATION PERFORMANCE (NME %) OF APPROACHES USING DIFFERENT TRAINING STRATEGIES ON AFLW2000-3D

Method	AFLW2000-3D (68 pts)				
	[0°, 30°]	[30°, 60°]	[60°, 90°]	mean	std
Baseline	2.89	3.43	4.35	3.56	0.74
Baseline-mix	2.78	3.35	4.29	3.47	0.76
Adversarial	2.69	3.08	4.15	3.31	0.76

TABLE IX

3D FACIAL LANDMARK LOCALIZATION PERFORMANCE (NME %) OF APPROACHES ADOPTING ONE-STAGE VERSUS TWO-STAGE TRAINING SCHEME ON AFLW2000-3D

Training Scheme	NME(%)
One-stage	5.82
Two-stage	4.25

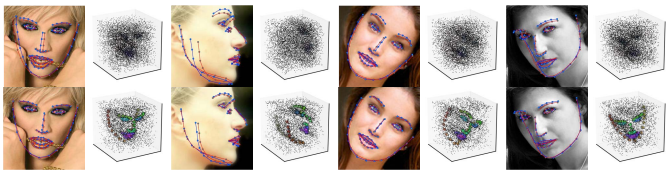

Fig. 7. Qualitative comparison of ablation approaches adopting one-stage (the top row) and two-stage (the bottom row) training schemes on the AFLW dataset.

achieves higher performance. This could be further confirmed by visualization of the volumes estimated by approaches using different schemes. As observed in Fig. 7, the network could produce much clearer volumes when adopting the two-stage training scheme.

3) Auxiliary regression adversarial learning: Finally, we conduct experiments to investigate the efficacy of the auxiliary regression adversarial learning strategy proposed in this paper. Experimental settings using different learning strategies are denoted as follows:

- *Baseline* refers to the approach adopting the backbone network where only the 3D dataset (300W-LP) is used for training.
- *Baseline-mix* refers to the approach adopting the unified training strategy where 3D and 2D datasets (300W-LP and AFLW) are used for training.
- *Adversarial* refers to the approach adopting the proposed auxiliary regression adversarial learning strategy where 3D and 2D datasets (300W-LP and AFLW) are used for training.

a) Adversarial learning: For 3D facial landmark localization, the performance of approaches using different strategies and different numbers of stacked Hourglass are shown in Fig. 8. It can be observed that the unified training strategy is helpful to improve the generalization of the network on real-world images where the 2D dataset is involved in the training procedure. Moreover, the adversarial training strategy further improves the results since the 3D structural constraints are better distilled from the 3D dataset to 2D dataset. Qualitative

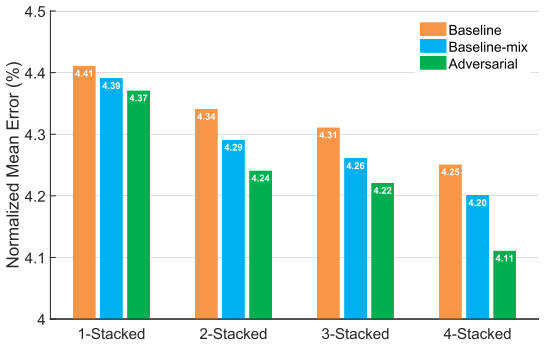

Fig. 8. 3D facial landmark localization performance (NME %) of approaches using different training strategies on AFLW2000-3D.

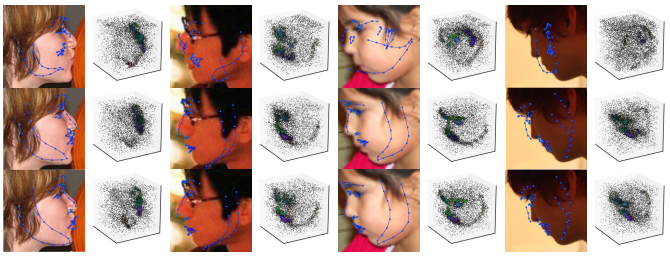

Fig. 9. Qualitative comparison of ablation approaches on the AFLW dataset. The three rows are results of (from top to bottom) *Baseline*, *Baseline-mix* and *Adversarial* respectively.

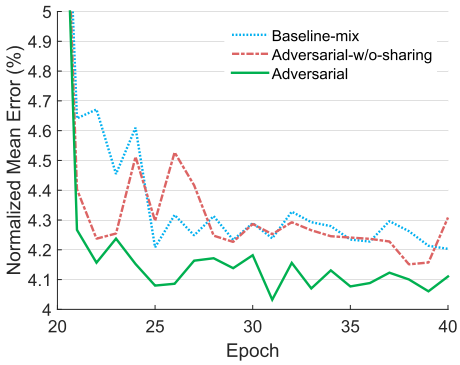

Fig. 10. NME curves of ablation approaches across training epochs on AFLW2000-3D.

comparison of ablation approaches is also shown in Fig. 9. The unified training of 2D and 3D datasets improves the 3D predictions on images taken under challenging scenarios, and the results are further refined through adversarial learning. Meanwhile, the proposed strategies also facilitate 2D facial landmark localization task as well. To further validate this, Table VIII reports performance of ablation approaches across different head poses for 2D facial landmark localization. It can be seen that the adversarial training strategy could bring significant improvements over baseline methods for all ranges of the head pose.

b) Auxiliary regression task: The auxiliary regression task of discriminator contributes to stable training of the proposed network. To validate this, the discriminator and coordinate regressor are isolated so that there is no weight sharing between them. Such an experiment setting is denoted as *Adversarial-w/o-sharing*. For different ablation approaches,

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on February 22,2020 at 08:45:42 UTC from IEEE Xplore. Restrictions apply.

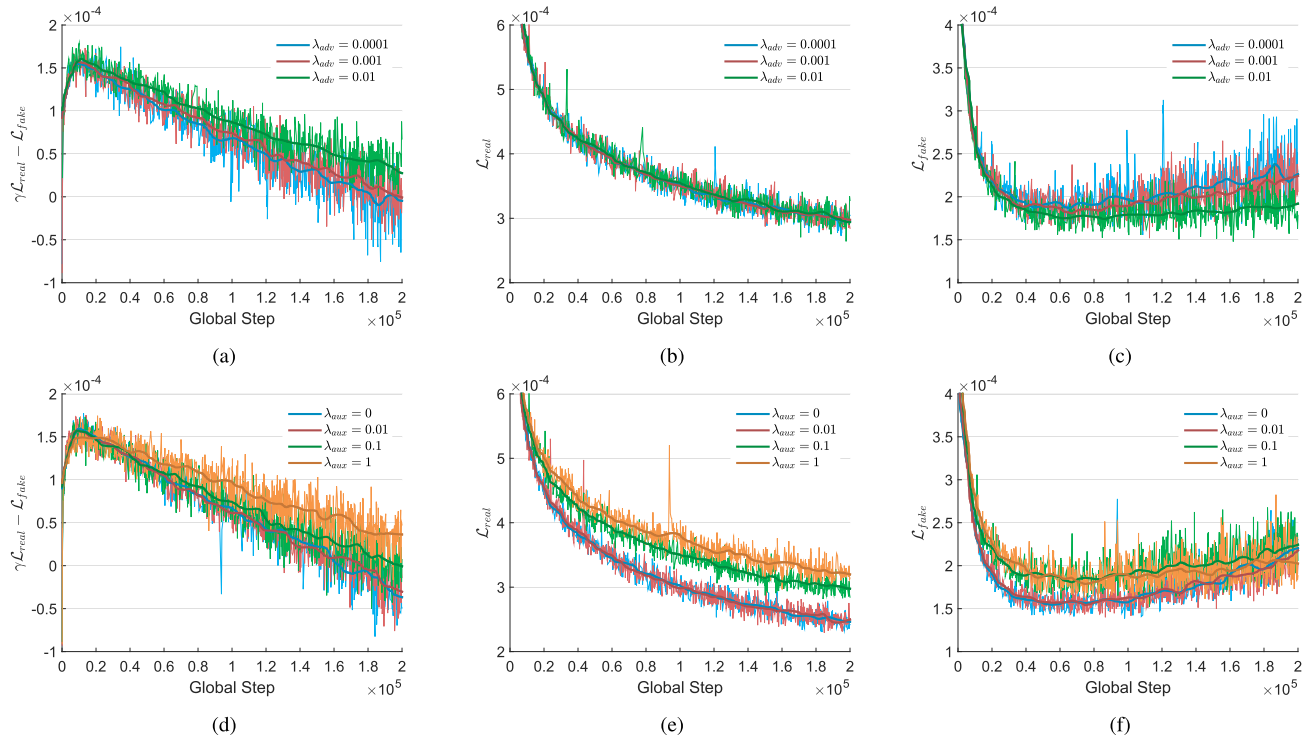

Fig. 11. Curves of $\gamma\,\mathcal{L}_{real} - \mathcal{L}_{fake}$, $\mathcal{L}_{real}$ and $\mathcal{L}_{fake}$ vs. global steps during the adversarial learning procedure. (a)(b)(c) The parameter λ_{adv} varies from 0.0001 to 0.01 with the parameter λ_{aux} fixed as 0.1. (d)(e)(f) The parameter λ_{aux} varies from 0 to 1 with the parameter λ_{adv} fixed as 0.001. Note that $\lambda_{aux} = 0$ indicates the case where the discriminator and coordinate regressor are isolated so that there is no weight sharing between them.

NME across different training epochs on AFLW2000-3D are reported in Fig. 10. It can be seen that the proposed auxiliary regression adversarial learning strategy could facilitate stable training as the model converges faster and achieves a lower error.

To gain deeper insights into the role the auxiliary regression task plays on the adversarial learning procedure, we conduct experiments varying the parameters λ_{adv} and λ_{aux}, which are two essential parameters for training the generator and discriminator respectively. Curves of $\mathcal{L}_{real}$ and $\mathcal{L}_{fake}$ as well as $\gamma\,\mathcal{L}_{real} - \mathcal{L}_{fake}$ vs. global steps are shown in Fig. 11, where we fix one parameter of them and vary another parameter for comparison. As observed from Fig. 11a and 11d, increasing either λ_{adv} or λ_{aux} has similar effect on the training procedure. Both these two cases enlarge the gap between $\gamma\,\mathcal{L}_{real}$ and $\mathcal{L}_{fake}$, which means that they all contribute to hindering the discriminator from differentiating fake samples between the real ones. This could help to stabilize the adversarial learning procedure since GAN is typically unstable when the discriminator gets too strong too quickly. However, the inherent mechanisms of how these two parameters work upon the model are different. When increasing λ_{adv}, the generator (i.e. volume estimator) is encouraged to produce fake samples with lower reconstruction error. This could be concluded from Fig. 11b and 11c, where curves of $\mathcal{L}_{real}$ are similar, while curves of $\mathcal{L}_{fake}$ become lower with the increasing of λ_{adv}. On the other hand, when increasing λ_{aux}, the discriminator assigns higher reconstruction error for both real and fake samples as shown in Fig. 11e and 11f. This could be explained by the extra regularization imposed by the auxiliary regression task, which makes the discriminator's job harder. Such regularization facilitates the stable training and results in a better model with higher performance.

V. Conclusion

In this paper, we propose the adversarial voxel and coordinate regression framework for 2D and 3D facial landmark localization. First, the semantic volumetric representation is introduced to encode positions of all landmarks in a single volume while still preserving their semantic information. The dimensionality of such a *color indexed* representation could be reduced greatly compared with the conventional *channel indexed* volumetric representation. By combining the merits of both heatmap regression and coordinate regression based methods, the proposed joint voxel and coordinate regression provides a promising solution for robust and accurate facial landmark localization. Meanwhile, 2D and 3D landmark localization problems could be unified in the proposed framework so that 2D and 3D datasets could be leveraged simultaneously. To further utilize different types of existing datasets, we exploit adversarial learning to distill the 3D structure of face shapes learned from fully annotated datasets to real-world images without depth annotations. The proposed auxiliary regression adversarial learning strategy effectively enhances the performance of landmark localization in challenging scenarios. Experimental results on both 2D and 3D landmark localization demonstrate the effectiveness of the proposed method. In future work, we will consider exploiting the proposed pipeline in the context of 3D human pose estimation.

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on February 22,2020 at 08:45:42 UTC from IEEE Xplore. Restrictions apply.

ACKNOWLEDGMENT

The authors would like to thank the associate editor and anonymous reviewers for their constructive comments and suggestions, and Zhihang Li, Jie Cao, and Dr. Wanli Ouyang for helpful discussions.

REFERENCES

[1] J. Yang, Q. Liu, and K. Zhang, "Stacked hourglass network for robust facial landmark localisation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 79–87.

[2] 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.*, Oct. 2017, pp. 1021–1030.

[3] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *Int. J. Comput. Vis.*, vol. 107, no. 2, pp. 177–190, 2014.

[4] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.

[5] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 FPS via regressing local binary features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1685–1692.

[6] S. Zhu, C. Li, C. Change Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4998–5006.

[7] S. Zhu, C. Li, C.-C. Loy, and X. Tang, "Unconstrained face alignment via cascaded compositional learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3409–3417.

[8] A. Bulat and G. Tzimiropoulos, "Convolutional aggregation of local evidence for large pose face alignment," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 86.1–86.12.

[9] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 717–732.

[10] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 483–499.

[11] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7025–7034.

[12] L. A. Jeni, S. Tulyakov, L. Yin, N. Sebe, and J. F. Cohn, "The first 3D face alignment in the wild (3DFAW) challenge," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 511–520.

[13] 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.*, Jun. 2016, pp. 146–155.

[14] R. Zhao, Y. Wang, C. F. Benitez-Quiroz, Y. Liu, and A. M. Martinez, "Fast and precise face alignment and 3D shape reconstruction from a single 2D image," in *Proc. Eur. Conf. Comput. Vis. Workshops*. Cham, Switzerland: Springer, 2016, pp. 590–603.

[15] A. Bulat and G. Tzimiropoulos, "Two-stage convolutional part heatmap regression for the 1st 3D face alignment in the wild (3DFAW) challenge," in *Proc. Eur. Conf. Comput. Vis. Workshops*. Cham, Switzerland: Springer, 2016, pp. 616–624.

[16] C. Gou, Y. Wu, F.-Y. Wang, and Q. Ji, "Shape augmented regression for 3D face alignment," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 604–615.

[17] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3D human pose estimation in the wild: A weakly-supervised approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 398–407.

[18] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2621–2630.

[19] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3D human pose estimation in the wild by adversarial learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5255–5264.

[20] H. Zhang, Q. Li, and Z. Sun, "Joint Voxel and coordinate regression for accurate 3D facial landmark localization," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2018, pp. 2202–2208.

[21] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3476–3483.

[22] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 918–930, May 2016.

[23] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 57–72.

[24] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4177–4187.

[25] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1831–1840.

[26] X. Sun, B. Xiao, S. Liang, and Y. Wei, "Integral human pose regression," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2018, pp. 529–545.

[27] H.-Y. F. Tung, A. W. Harley, W. Seto, and K. Fragkiadaki, "Adversarial inverse graphics networks: Learning 2D-to-3D lifting and image-to-image translation from unpaired supervision," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4364–4372.

[28] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[29] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," in *Proc. Int. Conf. Learn. Represent.*, 2017. [Online]. Available: https://openreview.net/forum?id=ryh9pmcee.

[30] D. Berthelot, T. Schumm, and L. Metz. (2017). "BEGAN: Boundary equilibrium generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1703.10717

[31] M. Mirza and S. Osindero. (2014). "Conditional generative adversarial nets." [Online]. Available: https://arxiv.org/abs/1411.1784

[32] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2642–2651.

[33] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[34] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.

[35] Z. Ding, Y. Guo, L. Zhang, and Y. Fu, "One-shot face recognition via generative learning," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2018, pp. 1–7.

[36] M. Rezagholiradeh and M. A. Haidar, "Reg-Gan: Semi-supervised learning based on generative adversarial networks for regression," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2018, pp. 2806–2810.

[37] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.

[38] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2018, p. 65.

[39] S. Liu *et al.*, "Cross-domain human parsing via adversarial feature and label adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7146–7153.

[40] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Macro-micro adversarial network for human parsing," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 418–434.

[41] C.-J. Chou, J.-T. Chien, and H.-T. Chen, "Self adversarial training for human pose estimation," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Nov. 2018, pp. 17–30.

[42] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial PoseNet: A structure-aware convolutional network for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1221–1230.

[43] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1799–1807.

[44] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4724–4732.

[45] Z. Wu *et al.*, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1912–1920.

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on February 22,2020 at 08:45:42 UTC from IEEE Xplore. Restrictions apply.

[46] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "3D convolutional neural networks for efficient and robust hand pose estimation from single depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jul. 2017, pp. 5679–5688.

[47] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015.

[49] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *Proc. IEEE Int. Conf. Automatic Face Gesture Recognit. (FG)*, 2008, pp. 1–6.

[50] X. Zhang *et al.*, "BP4D-spontaneous: A high-resolution spontaneous 3D dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 692–706, Oct. 2014.

[51] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[52] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3D face alignment from 2D video for real-time use," *Image Vis. Comput.*, vol. 58, pp. 13–24, Feb. 2017.

[53] 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*, 2013, pp. 397–403.

[54] M. Köstinger, 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*, Nov. 2011, pp. 2144–2151.

[55] C. Bhagavatula, C. Zhu, K. Luu, and M. Savvides, "Faster than real-time facial alignment: A 3D spatial transformer network approach in unconstrained poses," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4000–4009.

[56] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3D face reconstruction and dense alignment with position map regression network," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 557–574.

[57] F. H. de Bittencourt Zavan, A. C. Nascimento, L. P. e Silva, O. R. P. Bellon, and L. Silva, "3D face alignment in the wild: A landmark-free, nose-based approach," in *Proc. Eur. Conf. Comput. Vis. Workshops*. Cham, Switzerland: Springer, 2016, pp. 581–589.

[58] S. Tulyakov, L. A. Jeni, J. F. Cohn, and N. Sebe, "Viewpoint-consistent 3D face alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2250–2264, Sep. 2018.

[59] Y. Liu, A. Jourabloo, W. Ren, and X. Liu, "Dense face alignment," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 1619–1628.

[60] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1513–1520.

[61] R. Yu, S. Saito, H. Li, D. Ceylan, and H. Li, "Learning dense facial correspondences in unconstrained images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4733–4742.

[62] S. Xiao *et al.*, "Recurrent 3D-2D dual learning for large-pose facial landmark detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1633–1642.

[63] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: 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.

Hongwen Zhang received the B.E. degree in automation from the South China University of Technology, Guangzhou, China, in 2015. He is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include biometrics, pattern recognition, computer vision, and machine learning.

Qi Li received the B.E. degree in automation from the China University of Petroleum, Qingdao, China, in 2011 and the Ph.D. degree in pattern recognition and intelligent systems from CASIA in 2016. He is currently an Assistant Professor with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, China. His research interests focus on face preprocessing, computer vision, and machine learning.

Zhenan Sun received the B.S. degree in industrial automation from the Dalian University of Technology, Dalian, China, in 1999, the M.S. degree in system engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2002, and the Ph.D. degree in pattern recognition and intelligent systems from the Chinese Academy of Sciences, Beijing, China, in 2006. Since 2006, he has been a Faculty Member with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, where he is currently a Professor. He has authored or coauthored more than 200 technical papers. His current research interests include biometrics, pattern recognition, and computer vision. He is a fellow of the IAPR and serves as the Chair for IAPR Technical Committee on Biometrics. He is an Associate Editor of the IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE.

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on February 22,2020 at 08:45:42 UTC from IEEE Xplore. Restrictions apply.