Multi-task head pose estimation in-the-wild

Roberto Valle, José M. Buenaposada and Luis Baumela

Abstract—We present a deep learning-based multi-task approach for head pose estimation in images. We contribute with a network architecture and training strategy that harness the strong dependencies among face pose, alignment and visibility, to produce a top performing model for all three tasks. Our architecture is an encoder-decoder CNN with residual blocks and lateral skip connections. We show that the combination of head pose estimation and landmark-based face alignment significantly improve the performance of the former task. Further, the location of the pose task at the bottleneck layer, at the end of the encoder, and that of tasks depending on spatial information, such as visibility and alignment, in the final decoder layer, also contribute to increase the final performance. In the experiments conducted the proposed model outperforms the state-of-the-art in the face pose and visibility tasks. By including a final landmark regression step it also produces face alignment results on par with the state-of-the-art.

Index Terms—Head pose estimation, multi-task learning, face alignment, occlusions detection.

1 Introduction

Head pose greatly affects facial appearance. It is one of the parameters that influences to a largest extent the performance of many face analysis tasks. For this reason it is a fundamental step in computer vision algorithms estimating attention [1], identifying social interaction [2], recognizing faces [3] or robustly estimating facial attributes [4], [5]. It is a challenging problem in "in-the-wild" conditions, *i.e.*, in presence of extreme orientations, partial occlusions and varying resolution, illumination, facial hair and makeup. Although it has been often considered as by-product or auxiliary task of facial landmark location [6], recent results prove that it is much more efficient than landmark estimation and it may achieve superior performance in subsequent face analysis tasks, such as recognition [3]. In this paper we present a multi-task approach to head pose estimation in unrestricted images. We exploit the strong dependencies among head pose and landmark-related tasks within a multi-task Convolutional Neural Network (CNN) to produce a top performing model.

The *multi-task learning* (MTL) paradigm encompasses a set of learning techniques that provide effective mechanisms for sharing information among multiple tasks. It enables the use of larger and more diverse data sets, that improve the regularization during training and the generalization of the final model [7]. MTL is intimately related to *transfer learning* (TFL). In TFL a model is trained for one or more auxiliary tasks and subsequently refined for a main target task [8], [9]. Traditionally MTL implies a simultaneous or parallel treatment of all tasks [7], whereas in TFL tasks are learned sequentially. In our approach we combine parallel and sequential learning, so it cannot be clearly cast into one of the above two schemes. We rather generalize the traditional concept of MTL to include both. Following other approaches in the literature [10] we consider different degrees of MTL asymmetry. In this regard TFL is an extremely asymmetric MTL scheme in which auxiliary tasks are only used for pre-training. In our proposal we adopt an asymmetric approach where we seek to optimize head pose using visibility and alignment as auxiliary tasks. However, as we show in our experiments, the co-operation in our model among all three tasks is so high that all of them achieve state-of-the-art results in the most popular benchmarks and improve the performance they would otherwise achieve independently.

A key element in a multi-task CNN is the architecture of the model and the location of each task in the net. A natural approach is to share bottom layers among all tasks, since they model low-level features, whereas top layers, that capture high level features, are specific to each task [11]. In the context of face processing, some approaches have completely separate networks to model each attribute [12], others share all features in a common backbone [13], and others combine feature maps from different parts of the encoder network [5]. In our architecture, an encoder-decoder CNN, we carefully place each task to optimize the final performance. We locate head pose, a holistic task, at the end of the encoder. In this way the network bottleneck acts as embedding representing face pose. Visibility and alignment tasks are located at the decoder end, since they require information about the spatial location of landmarks in the image.

To train our model we leverage on the large face landmarks annotated data sets available. We first train the CNN for the landmarks-based face alignment task. Then we fine-tune it for head pose, face alignment and landmarks visibility. In the most asymmetric incarnation of our model, once trained, we may dispose of the decoder and the associated alignment and visibility tasks to produce a very efficient head pose estimation system. Alternatively, we may keep the full trained model and use the landmarks visibility and alignment outputs of the CNN as input to a novel face landmarks regression module based on an ensemble of regression trees. This model further improves the accuracy of landmarks location by imposing a valid face shape on the set of regressed landmarks.

We evaluate our model for all three tasks using COFW, AFLW and AFLW2000-3D landmark-based data sets. In these experiments our model beats the top competing approaches in the head pose and visibility estimation tasks. It also achieves performance comparable with the state-of-the art for the landmark-based face alignment task. We also evaluate head pose estimation with *Biwi*. Although it was not acquired in-the-wild, this data set is a widely used marker-less benchmark for head pose estimation. Here our results also establish a new state-of-the-art.

In summary, we propose a multi-task approach for head pose estimation. The proposed solution combines a good model architecture, training strategy and a set of complementary tasks that boost final performance. The resulting model achieves top results for all three tasks, head pose, face alignment and visibility. In Fig. 1 we display our predictions for some frames of a video from 300VW [14]. It shows a remarkable tracker-like stability although each frame is processed independently.

2 Related Work

The unique ability of neural networks to transfer and share knowledge among various tasks is one of the reasons for its present success. This is typically done using MTL techniques. In computer vision MTL has been widely used to simultaneously learn related tasks such as semantic segmentation and surface normal prediction [11]. In the facial analysis field, head pose is often used as a pre-processing step to help estimate face landmarks [15], [16], [17]. Other approaches simultaneously estimate head pose with facial landmarks [4], [5], Facial Action Units [18], gender [5] and various other facial attributes [19]. Alternatively, facial attributes estimation have also been combined with landmark detection [13], [19], [20]. In our approach we follow an asymmetric MTL scheme where the primary task is head pose

- *Roberto Valle and Luis Baumela are with the Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Spain.*
E-mail: rvalle@fi.upm.es and lbaumela@fi.upm.es
- *José. M. Buenaposada is with the ETSII, Universidad Rey Juan Carlos, C/ Tulipán s/n, 28933 Móstoles, Spain.*
E-mail: josemiguel.buenaposada@urjc.es

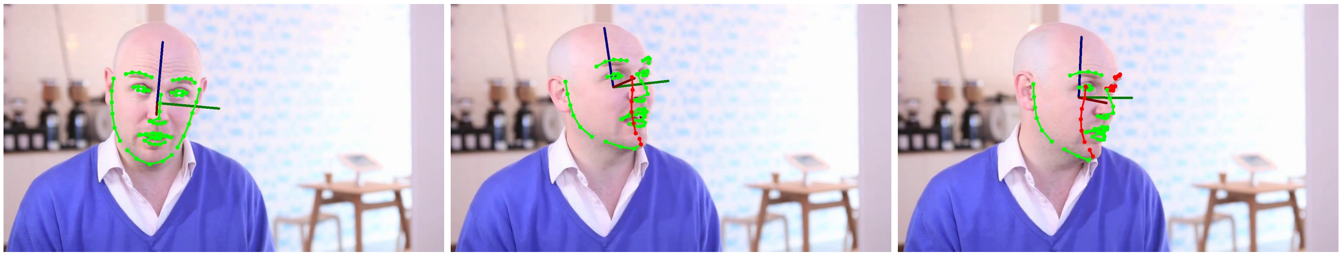

Fig. 1: Simultaneous head pose estimation, facial landmark location and their visibility predictions when processing a video from 300VW [14]. Green and red points show visible and non-visible landmarks respectively. The co-ordinate system qualitatively represents head pose.

estimation and use face landmarks as an auxiliary task that regularize and improve the performance of the primary task.

Pre-training a deep model with a large and general data set such as ImageNet has been a common practice for multiple vision tasks [8]. In the context of face analysis, ImageNet [5], [21], [22], [23], [24], [25], and other large face-related data sets, such those for face recognition [19], [26], [27], [28], have also been extensively used for predicting various facial attributes. More recently, self-supervised tasks have also emerged as powerful unsupervised pre-training mechanisms [29], [30], [31], [32]. For estimating head pose, pre-training with an unsupervised face alignment task yields better results than using a large supervised face recognition data set [29]. This is possibly due to the geometrical cues learned in the alignment process. Following the same reasoning, we hypothesize that face landmark estimation is related to head pose. So, pre-training with the former task may improve the performance of the latter. Moreover, there is a lack of annotated "in-the-wild" head pose data sets. With our approach we leverage on the abundance of in-the-wild landmark-annotated data to train our model. As we show in the experiments, pre-training with a facial landmark estimation task improves head pose accuracy, beating other ImageNet pre-trained competing models [21], [22], [23], [24].

In a MTL strategy the final results depend on the affinity or degree of co-operation among the tasks involved [9], [33]. In extreme situations *negative transfer* may actually hinder the final performance [10], [34]. Many approaches that simultaneously estimate head pose with other facial attributes, *e.g.*, [4], [5], [19], combine various competing tasks in the same network layer. In our experiments we show that head pose does not co-operate with landmark-related tasks when placed in the same layer. To address this issue we propose to use an encoder-decoder CNN and locate head pose, a holistic task, at the encoder end, that represents global face information. We place landmark-related tasks at the decoder end, where spatial information is represented at the finest detail (see Fig.2).

The best head pose estimation algorithms address the problem from a single task perspective. In the simplest case they fine-tune a backbone previously trained on ImageNet [3], [21], [22]. QuatNet and GLDL are respectively the state-of-the-art in AFLW and AFLW2000-3D. They use standard CNN-based models pre-trained in ImageNet. QuatNet combines ordinal and L2 regression losses representing head pose angles with quaternions [23]. GLDL learns a Gaussian distribution per co-ordinate using a Gaussian Labels Distribution Loss (GLDL) [24]. FDN and FSA-Net are the top performers in the Biwi data set. Both approaches stand on specifically taylored network architectures. FDN uses a three-branch network with a feature decoupling module to explicitly learn discriminative features for each pose angle [25]. FSA-Net combines spatially grouped pixel-level features of activation maps from different layers [35]. A recent alternative achieves state-of-the-art results on Biwi training with synthetically generated data [36]. To this end it introduces an adversarial domain adaptation approach for partially shared and continuous label spaces.

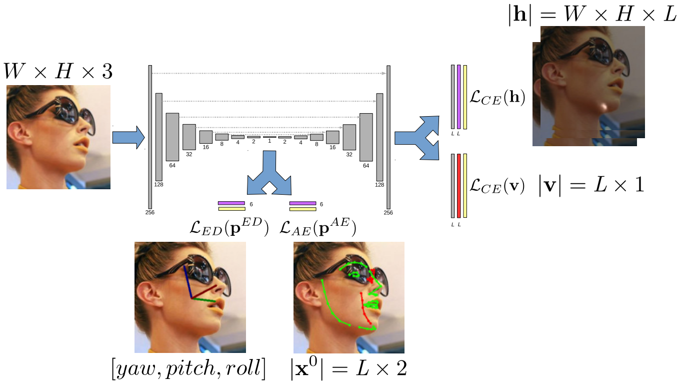
$|\mathbf{h}| = W \times H \times L$
$W \times H \times 3$
256
128
64
32
16
8
4
2
1
2
4
8
16
32
64
128
256
6
6
L L
L L
$\mathcal{L}_{CE}(\mathbf{h})$
$\mathcal{L}_{CE}(\mathbf{v})$
$|\mathbf{v}| = L \times 1$
$\mathcal{L}_{ED}(\mathbf{p}^{ED})$
$\mathcal{L}_{AE}(\mathbf{p}^{AE})$
$[yaw, pitch, roll]$
$|\mathbf{x}^0| = L \times 2$

Fig. 2: Multi-task encoder-decoder for the estimation of head pose, $\mathcal{L}_{ED}(\mathbf{p}^{ED})$, rigid and deformable facial landmarks location, $\mathcal{L}_{AE}(\mathbf{p}^{AE})$ and $\mathcal{L}_{CE}(\mathbf{h})$, and their visibilities, $\mathcal{L}_{CE}(\mathbf{v})$. We locate the head pose and rigid landmarks estimation tasks at the bottleneck layer, and the non-rigid face deformation and visibilities at the decoder end.

We leverage on the ideas discussed above to build a top performing head pose estimation algorithm. Our architecture is an standard encoder-decoder CNN with residual blocks and lateral skip connections. The key element of our proposal is a MTL scheme that combines a set of complementary tasks strategically located in the architecture. With our approach we improve not only the prediction accuracy, but also the computational and data efficiency, compared to training different models with data sets for each task.

3 Multi-task Head Pose Estimation

In this section we present our two-stage framework termed MNN+OR. First, we describe a novel Multi-task CNN (MNN) that estimates head pose, landmark heatmaps and their visibilities (see Fig. 2). Second, we introduce an Occlusion-aware Regressor (OR) that we use to regress the location of facial landmarks (see Fig. 3).

3.1 Multi-task Neural Network (MNN)

The most successful CNN architectures for facial landmark detection use an encoder-decoder network with lateral connections such as U-Net [37] and RCN [38]. Both capture local and global features at different scales. The popular Hourglass architecture [39] has a similar topology with extra convolutional layers in the lateral connections.

In this section, we introduce an architecture termed Multi-task Neural Network (MNN) based on a U-Net encoder-decoder with bottleneck residual blocks [40] instead of its original convolutional layers. The residual block lets us reduce the number of operations and increase depth while preserving the gradient back propagation through. We also include lateral skip connections that link symmetric layers between the encoder and the decoder preserving the spatial information (see the Supplementary Material).

MNN is a symmetric encoder-decoder architecture each with 9 stages. The encoder reduces the spatial extent of the input face image from 256 × 256 to 1 × 1 pixels. In the depth dimension we increase the number of feature maps from 64 in the first layer up to 256 in the bottleneck. We also include BatchNormalization and ReLu after each convolutional layer.

We encourage the encoder to act as feature embedding that learns a holistic face representation, favouring the exchange of information among all tasks. We attach to this layer two losses related to head pose estimation. The decoder learns local features tailored to the estimation of non-rigid landmark locations and their visibilities. Henceforth, we describe these losses and tasks.

Holistic tasks. The location of the loss functions associated to our tasks is essential given that the feature maps in different layers of the CNN represent the image information at different levels of abstraction and aggregation.

Since the head pose is a global attribute, we compute it from the bottleneck layer at the encoder end. Our objective here is to estimate the six parameters of the rigid transformation, $\mathbf{p} \in \mathbb{R}^6$, representing the relative pose between the head and the camera. To this end, we include two fully connected layers, $\mathbf{p}^{ED}$ and $\mathbf{p}^{AE}$, with 6 outputs each at the end of the encoder (see Fig. 2). We optimize these layers with two loss functions,

$$\mathcal{L}_{ED}(\mathbf{p}^{ED}) = \sum_{i=1}^{N} ||\tilde{\mathbf{p}}_i - \mathbf{p}_i^{ED}||_2, \tag{1}$$

$$\mathcal{L}_{AE}(\mathbf{p}^{AE}) = \sum_{i=1}^{N}\left(\sum_{l=1}^{L}\left(\frac{\tilde{\mathbf{w}}_i^l}{||\tilde{\mathbf{w}}_i||_1} \cdot ||\tilde{\mathbf{x}}_i^l - \pi(\mathbf{p}_i^{AE}, \mathbf{X}^l)||_2\right)\right), \tag{2}$$

where N denotes the number of images, L the number of landmarks, $\mathbf{p}_i^{ED}$ and $\mathbf{p}_i^{AE}$ the predicted pose parameters for the i-th training image using each loss, $\tilde{\mathbf{p}}_i$ the ground truth head pose parameters for the i-th training image, $\tilde{\mathbf{x}}_i^l \in \mathbb{R}^{L\times 2}$ the l-th landmark ground truth co-ordinates for the i-th training image, $\mathbf{X}^l \in \mathbb{R}^{L\times 3}$ the 3D co-ordinates of the l-th landmark, and π the camera projection.

Each loss plays an important role in our model. On the one hand, $\mathcal{L}_{ED}(\mathbf{p}^{ED})$ directly minimizes the euclidean error of pose parameters and provides an accurate and unambiguous pose estimation, $\mathbf{p}^{ED}$. On the other hand, $\mathcal{L}_{AE}(\mathbf{p}^{AE})$ measures the alignment error produced by the rigid projection of the mean 3D face model, $\mathbf{x}_i = \pi(\mathbf{p}_i^{AE}, \mathbf{X})$. The latter provides a better landmark initialization for the OR stage. However, the pose estimated, $\mathbf{p}^{AE}$, has projection ambiguity and estimation error caused by $\mathbf{X}$ not being the actual 3D landmark location, but that of the mean face. The combination of both losses provides unambiguous and accurate pose regression, as well as accurate rigid landmark localization.

Position-dependent tasks. Facial landmarks detection and their visibility estimation require both global and abstract features with a fine spatial resolution. Therefore, we use the feature maps at the end of the MNN decoder to estimate these attributes (see Fig. 2). For the landmark location task we introduce a convolutional layer producing $[256\times256\times L]$ feature maps and a softmax activation layer to generate heatmaps, such that $\sum_p^{256\times256} \mathbf{h}(p) = 1$. For the visibility task, we add a pooling layer with kernel size 256 × 256 to generate the vector of L visibilities associated to our landmarks, $\mathbf{v}$. To train this model we use the cross-entropy loss,

$$\mathcal{L}_{CE}(\mathbf{h}) = \sum_{i=1}^{N}\left(\sum_{l=1}^{L}\left(\frac{\tilde{\mathbf{w}}_i^l}{||\tilde{\mathbf{w}}_i||_1}\sum_{p=1}^{256\times256}\left(-\tilde{\mathbf{h}}_i^l(p)\cdot\log(\mathbf{h}_i^l(p))\right)\right)\right), \tag{3}$$

$$\mathcal{L}_{CE}(\mathbf{v}) = \sum_{i=1}^{N}\left(\sum_{l=1}^{L}\left(\frac{\tilde{\mathbf{w}}_i^l}{||\tilde{\mathbf{w}}_i||_1}\sum_{p=1}^{2}\left(-\tilde{\mathbf{v}}_i^l(p)\cdot\log(\mathbf{v}_i^l(p))\right)\right)\right), \tag{4}$$

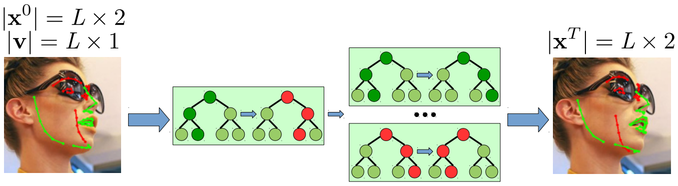
$|\mathbf{x}^0| = L \times 2$
$|\mathbf{v}| = L \times 1$
$|\mathbf{x}^T| = L \times 2$

Fig. 3: The OR is initialized with the 3D face model projected landmarks, $\mathbf{x}^0$, and their visibilities, $\mathbf{v}$. It incrementally updates the landmark location discarding the predictions of those regression trees whose features are extracted around occluded landmarks, shown in red.

where N is the number of images, L the number of landmarks, $\tilde{\mathbf{h}}_i^l$, $\mathbf{h}_i^l$ the l-th ground truth and predicted heatmaps for the i-th training image, and $\tilde{\mathbf{v}}_i^l$, $\mathbf{v}_i^l$ the l-th ground truth and predicted visibilities for the i-th training image.

To handle unlabelled landmarks we include $\tilde{\mathbf{w}}^l$, a landmark mask indicator variable ($\tilde{\mathbf{w}}_i^l = 1$ when the l-th landmark is annotated, and $\tilde{\mathbf{w}}_i^l = 0$ otherwise). This loss also enables data augmentation with large rotations, translations and scales, labelling landmarks falling outside of the bounding box as missing ($\tilde{\mathbf{w}}_i^l = 0$).

Multi-task loss. The loss function $\mathcal{L}(\mathbf{p}^{ED}, \mathbf{p}^{AE}, \mathbf{h}, \mathbf{v})$ computes a global error obtained from the pose parameters $\mathbf{p}^{ED}$, $\mathbf{p}^{AE}$, the landmark heatmaps, $\mathbf{h}$, and the visibilities, $\mathbf{v}$, by combining them using a weighted sum of the losses,

$$\begin{aligned}\mathcal{L}(\mathbf{p}^{ED}, \mathbf{p}^{AE}, \mathbf{h}, \mathbf{v}) = \alpha_{\mathbf{p}_1}\mathcal{L}_{ED}(\mathbf{p}^{ED}) + \alpha_{\mathbf{p}_2}\mathcal{L}_{AE}(\mathbf{p}^{AE}) +\\ \alpha_{\mathbf{h}}\mathcal{L}_{CE}(\mathbf{h}) + \alpha_{\mathbf{v}}\mathcal{L}_{CE}(\mathbf{v}).\end{aligned} \tag{5}$$

We empirically tune the weights $\alpha_{\mathbf{p}_1}$, $\alpha_{\mathbf{p}_2}$, $\alpha_{\mathbf{h}}$ and $\alpha_{\mathbf{v}}$ to balance the importance of all tasks. To this end, we train each task individually and determine the relative loss magnitudes when the learning process converges and ponder them accordingly.

3.2 Occlusion-aware Regressor (OR)

To achieve top results in the facial landmarks detection task we use an Ensemble of Regression Trees (ERT) that regularizes the MNN result by enforcing it to be a valid face shape [41]. To this end, we introduce an Occlusion-aware Regressor (OR). It is different from other landmark regressors in the literature [42], [43], [44], [45] in that our approach leverages on the robust landmark location and visibility estimation available at the MNN decoder to regress the landmark co-ordinates with top accuracy.

OR initialization. We use the head pose estimated by the MNN (see Section 3.1) to project AFLW mean 3D face model onto the image using $\mathbf{x}_i^0 = \pi(\mathbf{p}_i^{AE}, \mathbf{X})$, a $L \times 2$ matrix (see Fig. 3). This provides the OR with an initial estimation of the scale, and position of the target face shape. With this initialization we ensure that $\mathbf{x}_i^0$ is a valid face shape. This guarantees that the predictions in the next step of the algorithm, using an ERT, will also be valid face shapes [41]. Here, we also initialize the visibilities according to the head pose (*i.e.*, self-occlusions due to extreme head pose orientations) and the MNN prediction (*i.e.*, occlusions), instead of regressing the visibility in the ERT cascade like [42], [44], [45].

Non-rigid face shape deformation. Since the OR is initialized with the rigid face shape in the correct pose (see Fig. 3), to align the face it only needs to estimate the remaining non-rigid deformation of the face. To handle occlusions we incorporate the visibility labels for each i-th training image, $\{\mathbf{v}_i\}_{i=1}^N$, estimated by the the MNN. The initial shape is progressively refined in the cascade in S stages by extracting shape indexed features on the heatmaps $\{\phi(\mathbf{h}_i, \mathbf{v}_i, \mathbf{x}_i^{s-1})\}_{i=1}^N$ following a coarse-to-fine procedure like [45], where $\mathbf{x}_i^{s-1}$ represents the shape of the i-th sample on the previous stage. The novelty of

OR is that, the 2D displacements estimated by trees whose associated landmark is occluded are not added to the final estimation (see Fig. 3).

4 Experiments

To evaluate our approach we perform experiments using four in-the-wild landmark-related data bases and one head pose data set acquired in laboratory conditions. COFW [42] focuses on occlusions. It provides 1345 faces annotated with the positions and the binary occlusion labels for 29 landmarks. On average 28% of the landmarks are occluded. AFLW [46] provides a collection of 25993 faces, with 21 facial landmarks annotated depending on their visibility. For our experiments we discard some images with reported annotation errors [47]. We divide AFLW test subset into intervals of [0°, 30°], [30°, 60°] and [60°, 90°] according to head absolute yaw angle. AFLW2000-3D [48] consists of 2000 faces from AFLW semi-automatically re-annotated with 68 3D facial landmarks. We divide it into intervals of [0°, 30°], [30°, 60°] and [60°, 90°]. Each interval consists of 1306, 462 and 232 faces respectively. It has been typically used for testing head pose and facial landmark location algorithms using 300W-LP as train set [48]. This last data set provides 61225 synthesized face images from 300W [49], also re-annotated with 68 3D landmarks using the same algorithm. The semi-automatic pipeline used to label 300W-LP and AFLW2000-3D has been criticised for not producing accurate annotations for extreme poses and occluded faces [50]. For this reason we only use 300W-LP/AFLW2000-3D for comparing with the state-of-the-art that follows this protocol.

Although it was not acquired in-the-wild, we also evaluate our model with Biwi-Kinect [51]. It contains 15677 images from 24 sequences of 20 subjects acquired in a controlled environment with a Kinect sensor. Since Biwi does not contain landmark annotations, we follow the protocol presented in [22] using 300W-LP as train set.

4.1 Evaluation Metrics

We use the Mean Absolute Error (MAE) metric to quantify the head pose estimation error,

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left(|\tilde{\mathbf{p}}_i - \mathbf{p}_i|\right), \tag{6}$$

where N is the number of face images and $\tilde{\mathbf{p}}_i$, $\mathbf{p}_i$ represent the ground truth and predicted pose parameters respectively.

We also use the Normalized Mean Error (NME) as a metric to measure the shape estimation error

$$\text{NME} = \frac{100}{N}\sum_{i=1}^{N}\left(\sum_{l=1}^{L}\left(\frac{\tilde{\mathbf{w}}_i^l}{||\tilde{\mathbf{w}}_i||_1}\cdot\frac{||\mathbf{x}_i^l - \tilde{\mathbf{x}}_i^l||_2}{d_i}\right)\right), \tag{7}$$

where $\tilde{\mathbf{x}}_i$ and $\mathbf{x}_i$ are respectively the ground truth and estimated shape for the i-th training image and d_i is a normalization value. We use different values of d_i: the distance between eye pupils (*pupils*) and the bounding box height (*height*).

Finally, we report recall percentage at 80% precision to compare landmarks visibility prediction with other published methods.

4.2 Implementation Details

We train our models using Adam with an initial learning rate $\alpha = 10^{-3}$, which is halved whenever the loss plateaus for 15 epochs. We shuffle each training set and split it into 90% train and 10% validation. We also augment our training data by applying to each sample the following random operations: in plane rotation between ±45°, scaling by ±15%, translation by ±5% of the bounding box size, mirroring face image horizontally and colour change multiplying each HSV channel by a random value between $[0.5, 1.5]$. Additionally, we include synthetic rectangular occlusions to enforce the encoder-decoder to learn visibility.

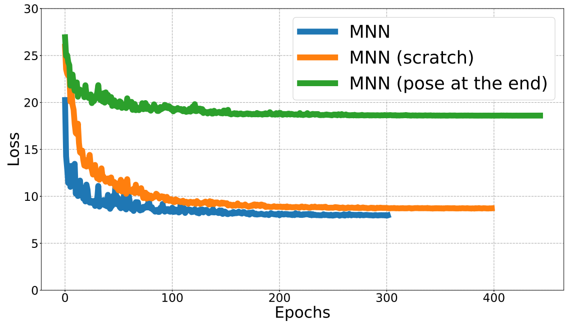

Fig. 4: Blue, orange and green colored learning curves compare the overall validation loss, $\mathcal{L}$, obtained with MNN by fine-tuning from landmarks, training from scratch, and locating the rigid pose losses at the end of the decoder respectively.

When provided we crop faces using the data set bounding box annotations. In 300W-LP/AFLW2000-3D and Biwi we use respectively the rectangle enclosing the annotated landmarks and the thresholded depth image. These detections are enlarged by 30% and resized to 256×256 pixels. In the landmark-annotated data sets we use POSIT [52] with a set of 2D (image) and 3D (face model) landmark correspondences to compute the head pose. We use as model the mean 3D face shape provided with AFLW [46].

At runtime our implementation of MNN+OR processes test images on average at a rate of 12.8 FPS using a NVidia GeForce GTX 1080Ti (11GB) GPU and a dual Intel Xeon Silver 4114 CPU at 2.20GHz (2×10 cores/40 threads, 128 GB), where the MNN takes 66 ms and the OR 12 ms per face using C++, Tensorflow and OpenCV libraries. We may also dispose of the MNN decoder and the OR regressor to build a very efficient head pose estimation module. The resulting model infers head pose using the GPU at a rate of 62.5 FPS.

4.3 Ablation study

In this section, we analyze the contribution of each component in our framework in the final performance.

4.3.1 *Task location*

In the first experiment we evaluate the importance of locating the head pose losses at the MNN bottleneck. To this end we adopt a MTL strategy pre-training the model with the landmark location task. The green and blue curves in Fig. 4 show respectively the loss achieved when locating both rigid pose losses, $\mathcal{L}_{ED}$ and $\mathcal{L}_{AE}$, at the end of the decoder ($\mathcal{L} = 18.6$) and at the end of the encoder ($\mathcal{L} = 7.9$). In the second case we achieve a reduction of 57.5% in the final loss. We infer that this gain is caused by two reasons. First, the superiority of the holistic features extracted from the embedding in the encoder-decoder bottleneck. Second, because head pose and landmark-related tasks do not co-operate when located in the same layer. Hard parameter sharing among these tasks decreases the final performance. From now on, we attach the rigid head estimation losses, $\mathcal{L}_{ED}$ and $\mathcal{L}_{AE}$, at the end of the encoder.

4.3.2 *Training strategy*

For these experiments, we incorporate two 2D landmark-based in-the-wild data sets. 300W [49] provides 68 manually annotated facial landmarks. We followed the most established approach and divide the annotations into 3148 training and 689 testing images (public competition). Thereafter, we also perform experiments on the 300W private benchmark, using previous 3837 images for training and 600 newly updated images as testing set. WFLW [53] consists of 7500 extremely challenging training and 2500 testing images divided into six subgroups, pose, expression, illumination, make-up, occlusion and blur, with 98 fully manual annotated landmarks. Since these data sets do not provide any head pose label, we compute it using POSIT [52] with AFLW [46] mean 3D face shape.

Here we evaluate our model under different training strategies. In the simplest case we follow a single task approach and minimize $\mathcal{L}_{ED}(\mathbf{p}^{ED})$ in Eq. (1) (Pose row in Table 1). We also consider several symmetric and asymmetric MTL schemes. In the symmetric case we train our model from scratch with all three tasks, minimizing $\mathcal{L}(\mathbf{p}^{ED}, \mathbf{p}^{AE}, \mathbf{h}, \mathbf{v})$ in (5) (Sym row in Table 1, orange stroke in Fig. 4). We also look at an asymmetric MTL scheme in which we pre-train with the image alignment task, optimizing $\mathcal{L}_{CE}(\mathbf{h})$ in Eq. (3). Once this training converges, we include the head pose and visibility tasks and optimize $\mathcal{L}(\mathbf{p}^{ED}, \mathbf{p}^{AE}, \mathbf{h}, \mathbf{v})$ (Pre+Sym row in Table 1, blue stroke in Fig. 4). Finally, in the most asymmetric MTL situation, we pre-train with the image alignment task, optimizing $\mathcal{L}_{CE}(\mathbf{h})$. Upon convergence, we then only optimize the head pose task, $\mathcal{L}_{ED}(\mathbf{p}^{ED})$ (Pre+Pose row in Table 1).

The orange and blue curves in Fig. 4 respectively display the difference between using the symmetric MTL training scheme ($\mathcal{L} = 8.7$) against the asymmetric MTL that pre-trains with the landmarks task followed by a symmetric MTL with all three tasks ($\mathcal{L} = 7.9$). In our problem pre-training regularizes the learning process and achieves a 9% reduction in the final loss, $\mathcal{L}$.

Further, in Table 1 we show head pose estimation results for different landmark-based data sets and training strategies. On average we achieve the largest improvement in mean MAE when changing from single task learning (first row) to MTL (three bottom rows). In the worst case, when moving from single task to the symmetric MTL case, we achieve a 7.5% reduction in mean MAE. The asymmetric approaches, that involve a pre-training step with the landmark face alignment task, achieve the best results, with a reduction of 11.9% in the average mean MAE with respect to the single task approach. There is no difference whether after pre-training we refine the model only for the pose task or for all three tasks. Hence, the second model will be the selected configuration and training strategy in our experiments.

Method		300W pub	300W priv	COFW	AFLW	WFLW	Avg
Single task	Pose	1.91	2.22	2.67	3.43	2.46	2.54
Multi-task	Sym	1.76	1.97	2.57	3.35	2.10	2.35
	Pre+Sym	1.59	1.96	2.36	3.22	2.08	2.24
	Pre+Pose	1.56	1.96	2.34	3.23	2.11	2.24

TABLE 1: Head pose mean MAEs for different training strategies. First row (Pose) single task encoder. Second row (Sym) symmetric MTL for all three tasks. Third row (Pre+Sym) MTL learning scheme pre-training with face landmarks. Fourth row (Pre+Pose) asymmetric MTL scheme pre-training with landmarks fine-tuned with pose.

We also evaluate the importance of the MTL scheme for visibility estimation using the COFW data set. Fist, we train MNN only for the visibility task. In this case, we achieve a recall of 21.75% at 80% precision for occlusion detection. This is a poor result, far worse than most published results (see Table 3), possibly caused by the small size of the training data set. However, using our selected MTL strategy, we get a recall of 72.12%, a large improvement in recall at the typical 80% precision point. So, with a small training data set, such as COFW, the combination of multiple related tasks within a MTL scheme boosts the final performance.

4.3.3 Occlusion-aware regressor

Here we evaluate the contribution of the OR stage to the performance of the landmark location task. We report the NME of, 1) MNN alone, locating each landmark by the maximum response on its heatmap; 2) the full MNN+OR framework. We show in Tables 4, 5 and 6 the performances in COFW, AFLW and AFLW2000-3D data sets. These results prove the importance of the OR stage to regularize the MNN landmark predictions. Model MNN+OR reduces the NME of model MNN in COFW by 10.8%, in AFLW by 3% and in AFLW2000-3D by 6.9%. The improvement grows proportionally with the presence of self-occluded parts (*i.e.*, AFLW2000-3D) and non-visible landmarks (*i.e.*, COFW) in the data sets.

4.4 Comparison with the state-of-the-art

In this section we evaluate our model in the most challenging benchmarks for all three tasks.

4.4.1 Head pose

In Table 2 we compare our head pose estimation proposal with the best published results in the literature. We train two MNN models. For AFLW there is no standard protocol to determine the training and testing partitions. We use the benchmark proposed in [21]. For testing in AFLW2000-3D and Biwi we train our model with 300W-LP, like [22], [23], [24], [25], [35]. However, we use the pose estimated from the correct 300W-LP landmarks from [50].

We outperform the state-of-the-art in AFLW (3.22 MAE), which represents an 11% mean MAE reduction over QuatNet [23], the best reported result in the literature. Moreover, in our MTL strategy we only use AFLW annotations, whereas QuatNet and many competing approaches use additional training data [21], [22], [23], [24], [29]. In the experiment evaluated on AFLW2000-3D and Biwi our approach establishes two new top results, again, with no extra training data. While in Biwi we reduce in 6.9% FDN's [25] MAE, in AFLW2000-3D our result only improves by 2.3% GLDL's [24]. This is caused by the inaccurate AFLW2000-3D annotations in extreme head poses [50]. While our approach was trained with poses estimated from the corrected landmarks, our competitors were trained on the original 300W-LP annotations, poisoned with the same errors. We re-annotated AFLW2000-3D with the poses estimated from the correct landmarks. We denote this data set AFLW2000-3D-POSIT. When we evaluate our mdel with it, the mean MAE goes down to 1.71.

The results in Table 2 must be considered with caution. It is obvious that landmark detection and head pose estimation tasks are clearly more connected if the pose is calculated from the landmarks. Moreover, the MAEs of AFLW and AFLW2000-3D-POSIT have a negative (optimist) bias. This is because in them the head pose in the train and test sets is computed with the same semi-automatic estimation procedure. The same argument applies to all our competitors in AFLW2000-3D. However, our results would be positively (pessimistically) biased in this data set, since some of its annotations are not correct. Hence, our unbiased MAE for AFLW2000-3D would be between the 3.83 and 1.71 bounds. In contrast, the experiment with Biwi involves different train/test data sets and annotation procedures. Hence, it provides the most accurate MAE estimations. Although, in a data set taken in laboratory conditions.

4.4.2 Facial landmarks visibility

In Table 3 we compare the landmarks visibility estimation that we obtain with MNN, against the best published results in the literature. To evaluate this task we use COFW [42], as far as we known, the only data set with annotated occlusions.

With our approach, we get a 72.12% recall at 80% precision, a new state-of-the-art for this data set. The notable improvement with respect to the closest competitor, 3DDE [45], is caused by two key differences. First, the MTL strategy boost the performance of the landmark visibility task. Second, in 3DDE the visibility is estimated by the landmark ERT regressor, like in [42], [44], whereas in the

Method	AFLW				AFLW2000-3D				AFLW2000-3D-POSIT				Biwi			
	yaw	pitch	roll	mean	yaw	pitch	roll	mean	yaw	pitch	roll	mean	yaw	pitch	roll	mean
FAb-Net [29]	10.70	7.13	5.14	7.65	-	-	-	-	-	-	-	-	-	-	-	-
Kepler [4]	6.45	5.85	8.75	7.01	-	-	-	-	-	-	-	-	8.08	17.2	16.1	13.8
Hyperface [5]	7.61	6.13	3.92	5.88	-	-	-	-	-	-	-	-	-	-	-	-
HopeNet [22]	6.26	5.89	3.82	5.32	6.47	6.55	5.43	6.15	-	-	-	-	4.81	6.60	3.26	4.89
GLDL [24]	6.00	5.31	3.75	5.02	**3.02**	5.06	3.68	3.92	-	-	-	-	4.12	5.61	3.14	4.29
HF-ResNet [5]	6.24	5.33	3.29	4.95	-	-	-	-	-	-	-	-	-	-	-	-
CCR [54]	5.22	5.85	2.51	4.52	-	-	-	-	-	-	-	-	-	-	-	-
Amador *et al.* [21]	5.59	4.79	2.83	4.40	-	-	-	-	-	-	-	-	-	-	-	-
FSA-Caps-Fusion [35]	-	-	-	-	4.50	6.08	4.64	5.07	-	-	-	-	4.27	4.96	2.76	4.00
QuatNet [23]	**3.93**	4.31	2.59	3.61	3.97	5.61	3.92	4.50	-	-	-	-	4.01	5.49	2.93	4.14
FDN [25]	-	-	-	-	3.78	5.61	3.88	4.42	-	-	-	-	4.52	4.70	2.56	3.93
MNN	4.16	**3.07**	**2.43**	**3.22**	3.34	**4.69**	**3.48**	**3.83**	**2.15**	**1.40**	**1.58**	**1.71**	**3.98**	**4.61**	**2.39**	**3.66**

TABLE 2: Head pose MAEs for AFLW, AFLW2000-3D and Biwi. AFLW200-3D-POSIT is the outcome of re-annotating AFLW2000-3D with the corrected landmarks annotations from [50].

Method	Full occlusion
RCPR [42]	40
Wu *et al.* [43]	44.43
Wu *et al.* [55]	49.11
ECT [56]	63.4
3DDE [45]	63.89
MNN	**72.12**

TABLE 3: Recall of landmarks visibility estimation methods at 80% precision using COFW.

Method	[0°, 30°] height	[30°, 60°] height	[60°, 90°] height	Full height
CCR [54]	-	-	-	5.72
Hyperface [5]	3.93	4.14	4.71	4.26
Kepler [4]	-	-	-	2.98
AIO [19]	2.84	2.94	3.09	2.96
HF-ResNet [5]	2.71	2.88	3.19	2.93
Binary-CNN [60]	2.77	2.60	2.64	2.85
PCD-CNN [17]	2.33	2.60	2.64	2.49
3DDE [45]	2.10	2.00	2.04	2.06
CHR2C [57]	2.07	**1.86**	**1.81**	1.98
MNN	2.12	1.90	1.89	2.03
MNN+OR	**2.05**	**1.86**	1.85	**1.97**

TABLE 5: Face alignment NME using AFLW.

proposed approach the visibility is estimated in the MNN model. This result proves again the relevance of our MTL approch.

4.4.3 Facial landmark location

We compare the MNN+OR framework with the state-of-the-art in face landmark regression. To this end we use results reported for 2D and 3D face alignment data sets. We use COFW and AFLW to provide a reference comparison with data sets involving 2D landmarks and AFLW2000-3D for 3D landmarks.

We analyze in Table 4 the MNN+OR landmark location performance in COFW, the common benchmark to evaluate occlusions. Here, we achieve a performance comparable to the best reported result for this data set, CHR2C [57], based on two stacked U-Net-like models.

Method	Full pupils
RCPR [42]	8.50
TCDCN [20]	8.05
Wu *et al.* [43]	6.40
Wu *et al.* [55]	5.93
ECT [56]	5.98
PCD-CNN [17]	5.77
SHN [39]	5.6
Wing [58]	5.44
ODN [59]	5.30
3DDE [45]	5.11
CHR2C [57]	5.09
MNN	5.65
MNN+OR	**5.04**

TABLE 4: Face alignment NME using COFW.

In Table 5 we compare MNN+OR with previous literature using AFLW images. This is a challenging database due to the large number of faces with extreme poses and occluded landmarks, which are not annotated. In this case, again, we achieve a performance comparable to the best reported result in the literature.

Finally, in Table 6, we also evaluate our model using a 3D data set. To this end we train our model with 300W-LP and test in AFLW2000-3D. In this case, we achieve 2.58 NME in the *Full* set. This result sets the new state-of-the-art for this data set, with a 16.2% reduction in NME with respect to the best published result in the literature, MHM [61] (3.08 NME), based on a two-stage cascade of heatmap regressors. Even without the final OR regressor, the MNN model alone already improves in 10% the previous best result. Our two-stage hybrid strategy is specially effective in 3D face alignment, where the OR stage is initialized using the extremely accurate head pose estimated by the MNN (see Fig. 3).

Method	[0°, 30°] height	[30°, 60°] height	[60°, 90°] height	Full height
RCPR [42]	4.26	5.96	13.18	7.80
3DSTN [62]	3.15	4.33	5.98	4.49
3DDFA [48]	2.84	3.57	4.96	3.79
PRN [63]	2.75	3.51	4.61	3.62
Binary-CNN [60]	2.47	3.01	4.31	3.26
MHM [61]	**2.36**	2.80	4.08	3.08
MNN	2.71	2.53	3.48	2.77
MNN+OR	2.54	**2.24**	**3.34**	**2.58**

TABLE 6: Face alignment NME using AFLW2000-3D.

5 Conclusions

In this paper we have presented a supervised multi-task approach to head pose, facial landmark location and visibility estimation. It is based on a heatmap encoder-decoder CNN, MNN, followed by an ensemble of regression trees to estimate the landmark co-ordinates. Rather than using head pose as a by-product or auxiliary task for landmark estimation, in our approach landmark-related tasks are used to boost head pose estimation. However, they are only required at training time. During testing we may dispose of the decoder and landmark regression modules to produce an extremely efficient head pose regressor with the best reported accuracy in the literature. In our head pose estimation experiments with landmark-based data sets we improve the best reported result in AFLW, QuatNet [23], and GLDL [24], the top performer in AFLW2000-3D. We also establish a new state-of-the-art in Biwi, a data set acquired in laboratory conditions and accurately annotated from depth images.

The MNN model and the MTL training strategy are fundamental to achieve top performance with our framework. In the ablation analysis we show that we get the largest improvement when switching from single task to a multi-task approach. We can further improve the performance if we adopt an asymmetric MTL scheme and pre-train the MNN with the face landmark estimation task. This confirms previous results showing that pre-training a model with a hard problem significantly improves the performance of other related tasks [64]. Also, our ablation shows that hard parameter sharing between head pose and face landmark estimation is detrimental of the final performance. This also confirms that multi-task and transfer (pre-training) relationships are different [33]. To address this issue and to provide each task with the appearance information aggregated at the best level of abstraction, we hook up the head pose loss to the encoder end, whereas the losses of spatially related tasks, such as landmark location and visibility, are attached to the decoder end.

Our model also reaches top performance in the two landmark related tasks. In visibility estimation it achieves 72.12% recall at 80% precision in COFW. A 13% improvement over the previous reported state-of-the-art, 3DDE [45]. We also compute the location of face landmarks using a novel occlusion-aware regressor (OR), that estimates face deformation from the heatmaps of visible landmarks. The full MNN+OR achieves results comparable to the state-of-the-art, 3DDE and CHR2C [45], [57], when evaluated in AFLW and COFW. In AFLW2000-3D, where self-occlusions play a key role, it sets a reduction of 16% over the previous state-of-the-art, MHM [61].

A fundamental problem to build a head pose estimation algorithm is the lack of training data. We propose a MTL strategy that takes advantage of the data bases available with manually labelled face landmarks. Pre-training with large object or face recognition data sets are alternative popular means to address this issue. We have proved that in the context of head pose estimation our proposal beats this strategy. This is due to the better co-operation between head pose and face landmark tasks. To further increase the robustness and accuracy of head pose estimation, our approach may be combined with self-supervised training [29], [30], [31], [32] and the use of synthetically generated data sets [36].

It is difficult to establish an state-of-the-art for head pose estimation in-the-wild, due to the lack of accurately annotated data sets. Present in-the-wild head pose evaluation methodologies are based on landmark data bases, such as AFLW and 300W-LP/AFLW2000-3D. The semi-automatic pipeline used to process them introduces errors in the train and test set annotations that bias the evaluation. Using re-annotated 300W-LP/AFLW2000-3D head poses we were able to upper and lower bound the performance of our approach in this situation. Biwi's evaluation methodology, based on train/test data sets acquired with different technologies and annotation algorithms, provides a more realistic performance estimation in laboratory conditions.

The proposed approach not only provides a satisfactory prediction accuracy but also a good computational efficiency. Instead of evaluating three different models, one for each task, we use a single encoder-decoder CNN, with an extremely efficient ERT, to simultaneously solve all three tasks at a rate of 12.8 FPS. However, if we are only interested in estimating head pose as a preliminary face processing step [3], our encoder-only model achieves 62.5 FPS.

Acknowledgments

The authors acknowledge funding from the Spanish Ministry of Economy and Competitiveness, project TIN2016-75982-C2-2-R. José M. Buenaposada was also partially funded by the *Comunidad de Madrid* project RoboCity2030-DIH-CM (S2018/NMT-4331). They also thank the anonymous reviewers for their comments and Felix Kuhnke for his help in interpreting Biwi annotations.

References

[1] L. M. Bergasa, J. M. Buenaposada, J. Nuevo, P. Jiménez, and L. Baumela, "Analysing driver's attention level using computer vision," in *11th International IEEE Conference on Intelligent Transportation Systems, ITSC*, 2008, pp. 1149–1154.

[2] S. O. Ba and J. Odobez, "Multiperson visual focus of attention from head pose and meeting contextual cues," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 101–116, 2011.

[3] F. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. G. Medioni, "Faceposenet: Making a case for landmark-free face alignment," in *Proc. International Conference on Computer Vision Workshops*, 2017, pp. 1599–1608.

[4] A. Kumar, A. Alavi, and R. Chellappa, "KEPLER: simultaneous estimation of keypoints and 3D pose of unconstrained faces in a unified framework by learning efficient H-CNN regressors," *Image and Vision Computing*, vol. 79, pp. 49–62, 2018.

[5] 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. on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2019.

[6] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *International Journal of Computer Vision*, vol. 127, no. 2, pp. 115–142, 2019.

[7] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1998.

[8] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 512–519.

[9] A. R. Zamir, A. Sax, W. B. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3712–3722.

[10] G. Lee, E. Yang, and S. Hwang, "Asymmetric multi-task learning based on task relatedness and loss," in *Proc. International Conference on Machine Learning*, 2016, pp. 230–238.

[11] I. Kokkinos, "Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5454–5463.

[12] F. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "Deep, landmark-free FAME: Face alignment, modeling, and expression estimation," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 930–956, 2019.

[13] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, no. 11, pp. 2597–2609, 2018.

[14] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Proc. International Conference on Computer Vision Workshops*, 2015, pp. 1003–1011.

[15] M. Dantone, J. Gall, G. Fanelli, and L. V. Gool, "Real-time facial feature detection using conditional regression forests," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2578–2585.

[16] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson, "Face alignment assisted by head pose estimation," in *Proc. British Machine Vision Conference*, 2015, pp. 130.1–130.13.

[17] A. Kumar and R. Chellappa, "Disentangling 3D pose in a dendritic CNN for unconstrained 2D face alignment," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 430–439.

[18] Y. Zhou, J. Pi, and B. E. Shi, "Pose-independent facial action unit intensity regression based on multi-task deep transfer learning," in *Proc. International Conference on Automatic Face and Gesture Recognition*, 2017, pp. 872–877.

[19] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *Proc. International Conference on Automatic Face and Gesture Recognition*, 2017, pp. 17–24.

[20] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 5, pp. 918–930, 2016.

[21] E. Amador, R. Valle, J. M. Buenaposada, and L. Baumela, "Benchmarking head pose estimation in-the-wild," in *Proc. Iberoamerican Congress on Pattern Recognition*, 2017, pp. 45–52.

[22] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2074–2083.

[23] H. Hsu, T. Wu, S. Wan, W. H. Wong, and C. Lee, "Quatnet: Quaternion-based head pose estimation with multi-regression loss," *IEEE Trans. on Multimedia*, 2018.

[24] Z. Liu, Z. Chen, J. Bai, S. Li, and S. Lian, "Facial pose estimation by deep learning from label distributions," in *Proc. International Conference on Computer Vision Workshops*, 2019.

[25] H. Zhang, M. Wang, Y. Liu, and Y. Yuan, "FDN: Feature decoupling network for head pose estimation," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 12 789–12 796.

[26] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. International Conference on Computer Vision*, 2015, pp. 3730–3738.

[27] Y. Zhong, J. Sullivan, and H. Li, "Face attribute prediction using off-the-shelf CNN features," in *International Conference on Biometrics (ICB)*, 2016, pp. 1–7.

[28] J. Cao, Y. Li, and Z. Zhang, "Partially shared multi-task convolutional neural network with local constraint for face attribute learning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4290–4299.

[29] O. Wiles, A. S. Koepke, and A. Zisserman, "Self-supervised learning of a facial attribute embedding from video," in *Proc. British Machine Vision Conference*, 2018, p. 302.

[30] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee, "Unsupervised discovery of object landmarks as structural representations," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2694–2703.

[31] J. Thewlis, S. Albanie, H. Bilen, and A. Vedaldi, "Unsupervised learning of landmarks by descriptor vector exchange," in *Proc. International Conference on Computer Vision*, 2019, pp. 6360–6370.

[32] S. Jeon, D. Min, S. Kim, and K. Sohn, "Joint learning of semantic alignment and object landmark detection," in *Proc. International Conference on Computer Vision*, 2019, pp. 7293–7302.

[33] T. Standley, A. R. Zamir, D. Chen, L. Guibas, J. Malik, and S. Savarese, "Which tasks should be learned together in multi-task learning?" *CoRR*, vol. abs/1905.07553, 2019.

[34] Z. Wang, Z. Dai, B. Póczos, and J. G. Carbonell, "Characterizing and avoiding negative transfer," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 293–11 302.

[35] T. Yang, Y. Chen, Y. Lin, and Y. Chuang, "FSA-Net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1087–1096.

[36] F. Kuhnke and J. Ostermann, "Deep head pose estimation using synthetic images and partial adversarial domain adaption for continuous label spaces," in *Proc. International Conference on Computer Vision*, 2019, pp. 10 163–10 172.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI*, 2015, pp. 234–241.

[38] S. Honari, J. Yosinski, P. Vincent, and C. J. Pal, "Recombinator networks: Learning coarse-to-fine feature aggregation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5743–5752.

[39] J. Yang, Q. Liu, and K. Zhang, "Stacked hourglass network for robust facial landmark localisation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 2025–2033.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[41] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[42] X. P. Burgos-Artizzu, P. Perona, and P. Dollar, "Robust face landmark estimation under occlusion," in *Proc. International Conference on Computer Vision*, 2013, pp. 1513–1520.

[43] Y. Wu, C. Gou, and Q. Ji, "Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5719–5728.

[44] R. Valle, J. M. Buenaposada, A. Valdés, and L. Baumela, "A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment," in *Proc. European Conference on Computer Vision*, 2018, pp. 609–624.

[45] ——, "Face alignment using a 3D deeply-initialized ensemble of regression trees," *Computer Vision and Image Understanding*, vol. 189, p. 102846, 2019.

[46] M. Koestinger, 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 *IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011, pp. 2144–2151.

[47] S. Jin and E. Learned-Miller, "Automatic detection of ground-truth labeling error for AFLW," http://people.cs.umass.edu/~souyoungjin/AFLW_gt_error_detection.htm, September 2015.

[48] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3D total solution," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 78–92, 2017.

[49] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016.

[50] 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. International Conference on Computer Vision*, 2017, pp. 1021–1030.

[51] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 437–458, 2013.

[52] D. DeMenthon and L. S. Davis, "Model-based object pose in 25 lines of code," *International Journal of Computer Vision*, vol. 15, no. 1-2, pp. 123–141, 1995.

[53] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2129–2138.

[54] W. Zhang, H. Zhang, Q. Li, F. Liu, Z. Sun, X. Li, and X. Wanu, "Cross-cascading regression for simultaneous head pose estimation and facial landmark detection," in *Biometric Recognition*, 2018, pp. 148–156.

[55] Y. Wu and Q. Ji, "Robust facial landmark detection under significant head poses and occlusion," in *Proc. International Conference on Computer Vision*, 2015, pp. 234–241.

[56] H. Zhang, Q. Li, Z. Sun, and Y. Liu, "Combining data-driven and model-driven methods for robust facial landmark detection," *IEEE Trans. Information Forensics and Security*, vol. 13, pp. 2409–2422, 2018.

[57] R. Valle, J. M. Buenaposada, and L. Baumela, "Cascade of encoder-decoder CNNs with learned coordinates regressor for robust facial landmarks detection," *Pattern Recognition Letters*, vol. 136, pp. 326–332, 2020.

[58] Z. Feng, J. Kittler, M. Awais, P. Huber, and X. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2235–2245.

[59] M. Zhu, D. Shi, M. Zheng, and M. Sadiq, "Robust facial landmark detection via occlusion-adaptive deep networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3486–3496.

[60] A. Bulat and G. Tzimiropoulos, "Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources," in *Proc. International Conference on Computer Vision*, 2017, pp. 3726–3734.

[61] J. Deng, Y. Zhou, S. Cheng, and S. Zafeiriou, "Cascade multi-view hourglass model for robust 3D face alignment," in *Proc. International Conference on Automatic Face and Gesture Recognition*, 2018, pp. 399–403.

[62] 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. International Conference on Computer Vision*, 2017, pp. 4000–4009.

[63] 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. European Conference on Computer Vision*, 2018, pp. 557–574.

[64] A. T. Tran, C. V. Nguyen, and T. Hassner, "Transferability and hardness of supervised classification tasks," in *Proc. International Conference on Computer Vision*, 2019, pp. 1395–1405.