2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Latent Embedding Clustering for Occlusion Robust Head Pose Estimation

José Celestino, Manuel Marques and Jacinto C. Nascimento
Institute for Systems and Robotics, Instituto Superior Técnico, Lisboa, Portugal

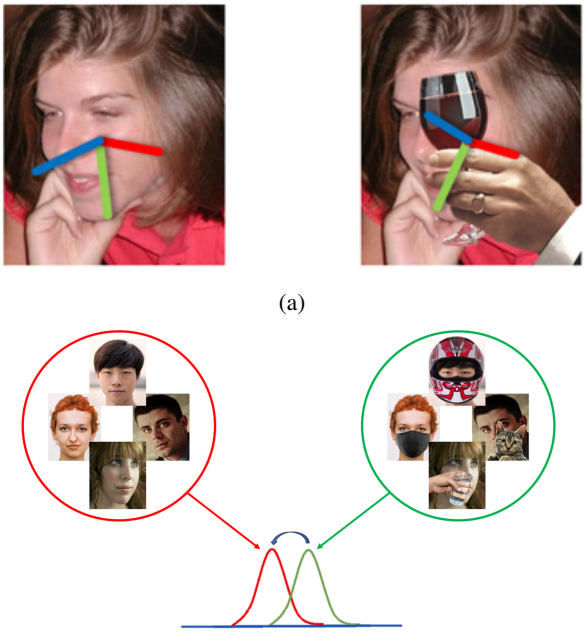
***Abstract*— Head pose estimation has become a crucial area of research in computer vision given its usefulness in a wide range of applications, including robotics, surveillance, or driver attention monitoring. One of the most difficult challenges in this field is managing head occlusions that frequently take place in real-world scenarios. In this paper, we propose a novel and efficient framework that is robust in real world head occlusion scenarios. In particular, we propose an unsupervised latent embedding clustering with regression and classification components for each pose angle. The model optimizes latent feature representations for occluded and non-occluded images through a clustering term while improving fine-grained angle predictions. Experimental evaluation on in-the-wild head pose benchmark datasets reveal competitive performance in comparison to state-of-the-art methodologies with the advantage of having a significant data reduction. We observe a substantial improvement in occluded head pose estimation. Also, an ablation study is conducted to ascertain the impact of the clustering term within our proposed framework.**

I. INTRODUCTION

Head pose estimation (HPE) can be roughly defined as the prediction of the relative orientation (and position) of the human's head with respect to the camera. HPE became a relevant topic in computer vision, being crucial in providing crucial information for an ever-growing range of applications, including human-computer/robot interaction [4], surveillance systems [23], driver attention monitoring [10], [28], [3], virtual/augmented reality [15], health care [2] and marketing [25]. Many of the applications above suffer from a substantial in-the-wild setback that has scarcely been investigated in detail: *the presence of occlusions*. Caused by external objects, facial accessories, or even body parts, they are often inevitable and introduce significant difficulties in capturing reliable facial features leading to inaccurate head pose estimates, (*e.g.* Fig. 1a). Consequently, existing methods often struggle to handle occlusions effectively, which results in unreliable performance in an unconstrained real-world environment.

Deep learning has become a crucial and an unavoidable tool to address this task [14], [38], [34], [20]. One of the most recent trends in HPE is the use of *latent space regression* (LSR), *e.g.* [6] that has shown to be effective in improving the robustness and generalization to occluded and non-occluded images. Basically, this methodology applies regression for latent embeddings so that the model learns a similar latent representation for occluded and non-occluded samples of the same pose.

However, we can point out two main limitations of this approach: **(1)** it has a high computational cost since it is a fully supervised learning approach requiring the non-occluded ground truth latent space sample for each occluded replica, and **(2)** for data augmentation purposes it always requires occluded/non-occluded image pairs to extract the correspondent non-occluded ground truth labels. We term this as *constrained data augmentation*.



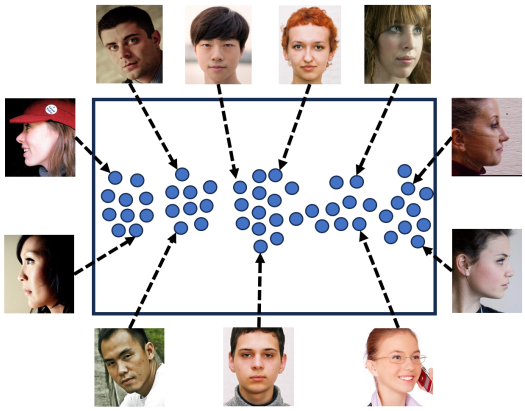
(a)

(b)

Fig. 1: (a) Occlusions strongly affect state-of-the-art methods for head pose estimation. (b) We propose to improve the occluded HPE estimation by introducing a novel framework containing a multi-loss Euler framework with an unsupervised clustering of the latent space. This is achieved through a minimization of the difference between probability distributions for occluded and non-occluded images.

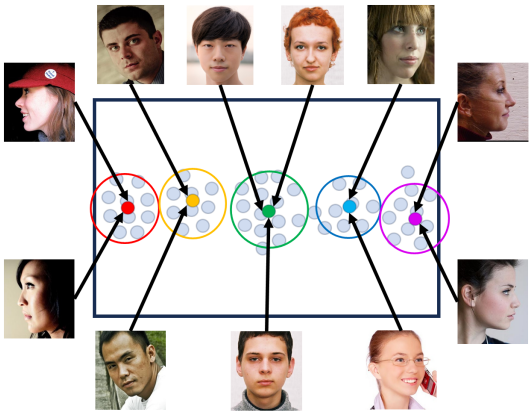
In this paper, we propose a novel methodology to improve occlusion robustness in HPE that does not suffer from the two limitations above. To accomplish this, we draw inspiration from recent work on embedding clustering for deep learning in computer vision [32], [12] and combine unsupervised latent embedding clustering with fine-grained Euler angle regression. In this way, we are able to improve feature representation for pose estimation while fine-tuning the latent embedding space via minimization of a clustering loss (Fig. 1b), without requiring the high cost of having

979-8-3503-9494-8/24/$31.00



LSR: N ground truth latent points for N training images.

(a)



Ours: K ground truth latent points for N training images, where K is the number of clusters and K<<N.

(b)

Fig. 2: Comparison between [6] (a) and our proposed framework (b). While in [6], N ground truth latent points are needed for training with N images, our method only requires $K << N$ ground truth points to train with the same N images. In our proposal, K are cluster centers points in the latent space.

labeled embedding data for each training image.

We can underscore the following advantages of our proposal. First, we only require $K << N$ ground truth latent embedding points for training N images, instead of having one ground truth point per image (see Figure 2). Furthermore, our method allows for expansion of the occluded training dataset as it will be detailed in the experimental evaluation (see Sec.IV-B).

In summary, our contributions are as follows:

- Novel methodology with combined optimization of fine-grained Euler angle estimations and unsupervised latent embedding clustering for refinement of the feature embedding space and improvement of estimation robustness against the occlusions,
- Low computational cost at estimating HPE (see limitation **(1)** above),
- State-of-the-art results for occluded images in synthetic and natural benchmark datasets,
- Competitive performance regarding state-of-the art, (*e.g.* [6]), with a substantial reduction of the ground truth embeddings,
- Opposing to the *constrained data augmentation* (see limitation **(2)** above), our proposal has the ability to augment the occluded training dataset with images that are not occluded replicas of non-occluded images initially used to establish the ground truth labels, and
- Ablation study regarding the impact of the novel application of unsupervised embedding clustering in the HPE context.

The paper is organized as follows: Section II revises related literature. Section III describes the proposed methodology to achieve occlusion robustness in HPE. In Section IV, we compare the performance of the proposed framework with SoTA methodologies in three benchmark datasets: (i) BIWI [19], (ii) AFLW2000 [39], with both occluded and non-occluded images; and (iii) Pandora [3], with real-life occlusions. We also perform ablation studies regarding the novel use of unsupervised embedding clustering. In Section V, we present our concluding remarks and potential venues for future improvement.

II. RELATED WORK

In this section, we delve into existing literature of relevant research closely related to our current study.

A. Head Pose Estimation

Regarding the literature in HPE, the state of the art can be framed in two different classes of strategies: (i) based on facial landmarks detection and model fitting; or (ii) through a deep learning model based on image features. The existing research is very extensive and diverse for the two classes above.

The model-based class of approaches tries to fit a head mesh to facial landmarks/keypoints collected from the image. Landmarks represent model fixed points that define the contours of specific face regions, (*e.g.* mouth, eyes, nose) in a head mesh. Keypoints, on the other hand, are tracked head feature points that do not have predetermined locations in the image domain. The work in [9] follows a fusion method that combines landmark and keypoint detection to improve performance when compared to using only one of them alone. Other recent works related to this class of approaches propose fitting landmarks and keypoints to morphable head models, instead of rigid ones. The idea is to improve the facial model fitting to different individuals, since generalization issues may occur in rigid models. The

authors of [37] propose a method that requires only four non-coplanar keypoints and includes a 3D face morphing method. The work of [5] also aims to estimate the head pose from a small set of head keypoints. Other examples are [11], where pose estimation is approached as a 3D Morphable Model (3DMM) parameter regression problem, and [20], which fits a 3D FLAME head mesh and simultaneously learns the head pose, shape and expression.

In the wake of advancements in deep learning, new learning-based approaches have emerged for addressing the HPE problem without need of landmarks or keypoints. The authors of [22] describe the advantages that these methods have over landmark-to-pose methodologies: they do not depend on head meshes, choice of landmark detection or 2D to 3D alignment methods. Several recent works have followed this class of approach, *e.g.* [22], [38], [33], [34], [1], [16], [14], [17], [8]. The authors of [34] solve HPE as a soft-stage-wise regression problem, inspired by [35]. In [1], the authors propose simultaneous face detection and head pose estimation with 6 degrees of freedom. This allows for more robust face detection with low computational cost regarding HPE. The works of [14] and [17] try to address common problems of HPE methods. In [14], the authors try to counter the ambiguity problem of rotation labels by using a rotation matrix representation and a geodesic loss. The method in [17] tackles the perspective distortion in face images caused by the misalignment of the face with the camera coordinate system. The authors from [22] propose a deep learning strategy that employs a backbone neural network augmented with three fully-connected layers, each one used to predict a different Euler angle using a multi-loss approach that combines a classification loss with a weighted regression loss for fine-grained predictions. This approach inspired other works, such as [38] which extends the idea to a full 360° range of yaws by synthetically extending the training dataset and introducing a new wrapped loss. Instead of regressing Euler angles, the work of [16] proposes a quaternion-based multi-regression loss method. This approach incorporates ordinal regression, enabling us to tackle the non-stationary nature of HPE, where variations in pose differ within each angle interval.

B. Occlusions in HPE

Even though occlusions are one of the roadblocks encountered in HPE, not many works have tackled this problem in depth. The method in [29] uses an iterative Lucas-Kanade optical flow tracker to track the displacement of a face feature concerning the center of the head, but still requires the mouth to be free from occlusions. The work in [31] estimates landmark visibility probabilities to perform occlusion prediction and includes prior occlusion pattern loss to improve performance. The authors of [30] combine the estimation of facial landmarks, head pose, and facial deformation under occlusions, but only evaluate yaw estimation. The work in [6] includes the generation of synthetically occluded datasets and proposes combining a multi-loss for fine-grained predictions with latent embedding regression. This allows to improve the latent feature representation of synthetically occluded images to that of the original non-occluded images, and therefore, enhances the pose prediction with occlusions.

C. Unsupervised Deep Clustering

Unsupervised clustering analysis has become quite relevant learning strategy in several fields, *e.g.* data science, machine learning, and computer vision, to quote a few. Within the scope of our application, it focuses on grouping image data without ground truth labels and based on some feature similarity metric. On the one hand, simpler methods like k-means [18] are effective for clustering, however, struggle to learn representations and cluster data when the input feature space dimension is high. Thus, transforming the data to a lower dimensional space is a viable solution. On the other hand deep neural networks have become a fundamental tool to perform feature dimension reduction. Thus, performing data clustering in low dimensional subspaces seems a natural approach to be adopted. One way to achieve this is to use *deep clustering* [21]. Deep clustering, however, remains a relatively recent field of study, but some works are available in the field. For instance, the unsupervised Deep Embedding Clustering (DEC) method [32] uses neural networks to achieve accurate clustering. It learns a mapping from the input data to a relevant lower dimensional feature space using an auto-encoder with an image reconstruction loss and iteratively optimizes a clustering objective. The authors of [12] argue that relying solely on the clustering loss may not guarantee the preservation of local structure, potentially resulting in the corruption of the feature space. They propose incorporating both the clustering and the reconstruction losses of the auto-encoder. This approach aims to optimize cluster assignment while concurrently updating features to preserve local structure for clustering. The work of [24] employs augmented data and integrates the clustering loss with an instance-wise contrastive loss. This combination aims to maximize the similarity of positive pairs while penalizing negative ones. Additionally, an anchor loss is introduced to maximize agreement between raw samples and augmented cluster assignments.

III. Methodology

The proposed methodology aims to address the difficulties mentioned in Sec. I. Specifically, we aim to use few ground truth latent embeddings, say K, to obtain the representative of N images, having $K << N$. To accomplish this, we use unsupervised latent embedding clustering motivated by the ideas of [32], [12], [24]. We integrate the clustering objective with additional functions to ensure the preservation of accurate feature representation and prevent corruption of the latent feature space. In our case, this involves the use of multi-loss functions for each estimation angle. Also, as it will be seen in the experimental evaluation, we address the limitation (2), that is, to relax about the ground truth constraint requirements when augmenting the training data. We refer to our work as *Latent Embedding Clustering for Head Pose Estimation* (LEC-HPE).

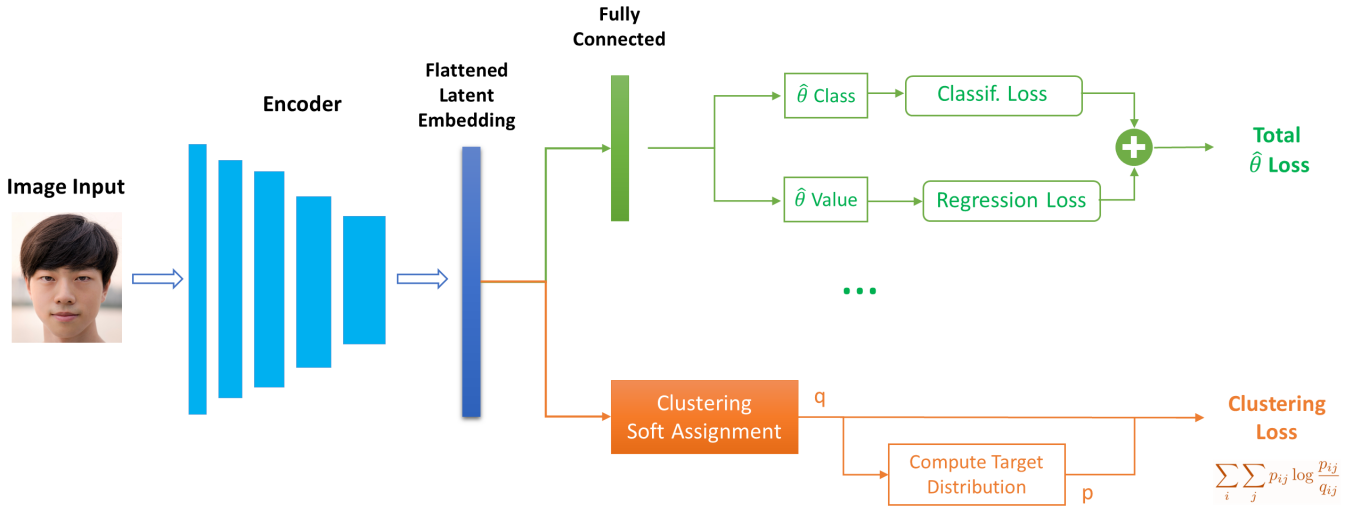


Fig. 3: Network structure for LEC-HPE. The architecture includes a branch for clustering of feature space embeddings and one multi-loss branch for each predicted Euler angle ($\hat{\theta} \in \{yaw, pitch, roll\}$), to ensure continuous feature learning and avoid distortion of the latent space.

A. Latent Embedding Clustering for Head Pose Estimation

The overall architecture for the LEC-HPE model is illustrated in Fig. 3. This structure includes a backbone encoder and a total of four separate branches subject to optimization. Three branches are used for the prediction of each Euler angle $\hat{\theta} \in \{yaw, pitch, roll\}$, and include fully-connected layers that output logits to be processed within a multi-loss framework for classification and fine-grained estimation. The remaining branch is responsible for clustering and fine-tuning the latent space. In the following subsections, we will describe the proposal in more detail.

We aim to align the latent space encoding for occluded images closely with that of non-occluded images. In order to avoid requiring unique latent embedding labels for each training image, which can be computationally expensive, we perform clustering of the latent space. Formally, let us consider a dataset X with N image samples x_i *i.e.*, $\{x_i \in X\}_{i=1}^{N}$. The images in X are the input to the encoder which transforms the data with a non-linear mapping $f_\Omega : X \rightarrow L$, where Ω are the learning parameters of the network and L the latent feature space. In the L space, the embedded points of semantically similar data samples are closer together. The proposed clustering algorithm acts on the latent feature space L, by learning a set of K clusters centers, *i.e.*, $\{c_j \in L\}_{j=1}^{K}$. Thus, we approximate each embedded point $l_i \in L$ to a cluster centroid c_j which can be learned and refined in an unsupervised manner. The number of clusters (K) will be much smaller than the size of the training dataset X, avoiding a high computation cost. That is, our strategy is not dependent on occluded/non-occluded image pairs and allows for the expansion of the training dataset without the need for latent labels. Thus, relaxing on the constrain data augmentation as observed, for instance, in [6]), where each each occluded training image requires the correspondent non-occluded embedded ground truth.

B. Two-stage Training

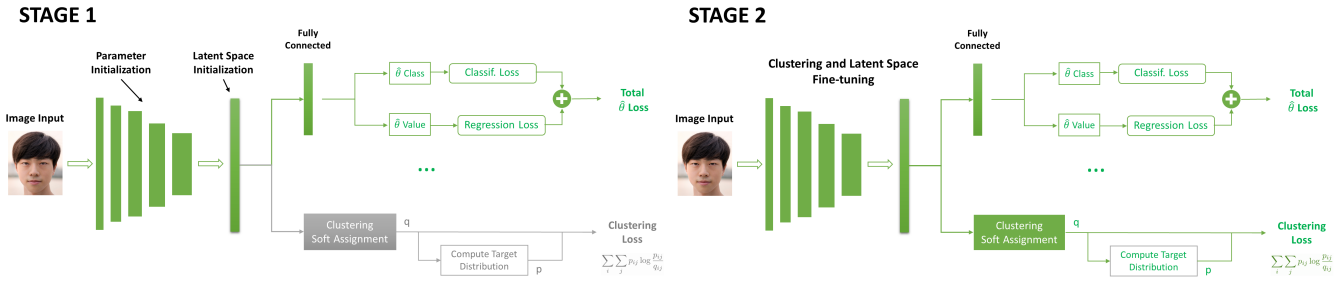
Our training strategy is two-fold, as illustrated in Fig. 4. **Stage 1**: in the first stage, we perform parameter and feature space initialization by optimizing the model for bin prediction using a classification loss and fine-grained Euler angle estimation using a regression loss; **Stage 2**: in the second stage, we add a clustering term. While in [32] the authors drop the loss function used for parameter/feature space initialization and replace it with the clustering loss, we maintain the multi-loss functions involved in the first stage. Similar to previous studies [24], [12], we also confirm that relying solely on the clustering loss in the second stage results in the degradation of the latent feature space (see results in Sec. IV-D). The purpose of this stage is to perform clustering and fine-tune the latent space while ensuring that it preserves its relevance for the head pose estimation task.

C. Fine-grained Losses for Feature Learning and Preservation

Within the proposed architecture illustrated in Fig. 3, there are three branches related to the Euler angle pose estimation. The top (green) branch is dedicated to a specific Euler angle $\hat{\theta} \in \{yaw, pitch, roll\}$ and incorporates a fully-connected layer that extends the backbone network. This layer outputs a vector of raw predictions (logits) for classification purposes. A Softmax activation function converts the logits to a vector of normalized probabilities with one value for each possible class:

$$S(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{n} e^{y_j}} \tag{1}$$

where y_i is the logit for class $i \in \{1, ..., n\}$. Each class corresponds to a classification bin that covers a pose range in degrees, *e.g.* [0º, 3º]. From here, the output of the Softmax



(a) Stage 1: Parameter and latent space initialization.

(b) Stage 2: Clustering and latent space finetuning.

Fig. 4: Two-stage training: Green highlights indicate the model components involved in each stage, while grey highlights represent parts not included in the respective stages.

activation function is processed in two separate branches: one for classification and the other for regression.

Regarding the classification branch, we apply a cross-entropy loss which takes the normalized probabilities and measures the distance from the ground truth values:

$$L_{class} = -\sum_{i}^{C} t_i \log\left(S(y_i)\right) \tag{2}$$

where t_i is the truth value (0 or 1), C is the number of classes and $S(y_i)$ is the Softmax probability for the i^{th} class.

For the regression branch, we use the Softmax probabilities to compute the predicted angle in degrees from the expected value, as defined in [38]:

$$\hat{\theta}_{Euler} = r \sum_{i=1}^{C} S(y_i) \left(i - \frac{1+C}{2} \right) \tag{3}$$

where $\hat{\theta}_{Euler}$ is the predicted Euler angle in degrees and r is the angle width of each bin, (*e.g.* if the bin is [0º, 3º], then $r = 3$). Afterwards, we apply a mean squared error loss (MSE) as the regression objective between the predicted and ground truth Euler angle:

$$L_{reg} = \frac{1}{N} \sum_{i=1}^{N} \left\| \theta_{i_{gt}} - \hat{\theta}_{i_{Euler}} \right\|_2^2 \tag{4}$$

where N is the number of predictions, $\theta_{i_{gt}}$ is the ground truth and $\hat{\theta}_{i_{Euler}}$ the predicted Euler angle.

The classification and regression losses are then combined, using a regularization coefficient, to yield the final total loss for each Euler angle ($L_{yaw}, L_{pitch}, L_{roll}$):

$$L_{\hat{\theta}} = L_{class\_\hat{\theta}}(\hat{\theta}_{pred}, \theta_{gt}) + \alpha\, L_{reg\_\hat{\theta}}(\hat{\theta}_{pred}, \theta_{gt}) \tag{5}$$

where $L_{class\_\hat{\theta}}$ and $L_{reg\_\hat{\theta}}$ are classification and regression losses for $\hat{\theta} \in \{yaw, pitch, roll\}$, $\hat{\theta}_{pred}$ and θ_{gt} are predicted and ground truth values, respectively and $\alpha > 0$ is the regularization coefficient that manages the trade-off between the two terms. The regression loss is used to achieve fine-grained estimations, while the classification loss helps the model to predict the vicinity of the pose. These Euler objective functions allow the model to learn the best non-linear mapping between the data space and the latent space for the task of head pose estimation during the first stage of training, while aiding the model to preserve an adequate mapping during the second stage (see results in Sec. IV-D).

D. Unsupervised Latent Clustering

The remaining branch of the architecture defined in Fig. 3 concerns the clustering of the latent feature space. Once **Stage 1** (Fig. 4a) of the training procedure is completed, the encoder has learned an initial non-linear mapping $f_\Omega : X \to L$, which transforms the input image data space X to a relevant representation for head pose estimation in the latent space L. Therefore, we can move to **Stage 2** (Fig. 4b) and perform clustering during training to fine-tune the feature space. As in [32], we perform k-means [18] in the latent feature space mapped from the entire training input data $Z = f_\Omega(X)$, to obtain the initial cluster centers $\{c_i\}_{i=1}^{K}$, where K is the number of clusters.

With initialized cluster centroids, we measure the pairwise similarity q_{ij} between a latent embedded point l_i and a cluster center c_j according to [32] and also used in the t-Distributed Stochastic Neighbor Embedding (t-SNE) [27] technique:

$$q_{ij} = \frac{(1 + \|l_i - c_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|l_k - c_l\|^2)^{-1}} \tag{6}$$

As described in [32], this quantified similarity can be interpreted as the probability of assigning the i^{th} sample to the j^{th} cluster center and therefore considered a soft/probabilistic assignment. From this soft assignment q_{ij} we can compute the target distribution p_{ij}:

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_{j'} (q_{ij'}^2 / \sum_i q_{ij'})} \tag{7}$$

With q_{ij} and p_{ij}, we can optimize cluster centers and fine-tune the encoder parameters by minimizing a Kullback-Leibler (KL) divergence objective between the target distribution P and soft assignment Q:

$$L_{clustering} = \mathrm{KL}(P \parallel Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{8}$$

E. Overall Loss

The overall loss is the sum of all described objective functions with regularization coefficient β to adjust the impact of the clustering term in the optimization of the model:

$$L_{total} = L_{yaw} + L_{pitch} + L_{roll} + \beta L_{clustering} \quad (9)$$

where $\beta = 0$ during the first training stage, and $\beta > 0$ during the second training stage.

IV. Experimental Evaluation

This section describes carefully the extensive experimental evaluation involving the use of several benchmark HPE datasets.

A. Datasets

We utilize four distinct benchmark in-the-wild head pose datasets for training and testing purposes: 300W-LP [36], BIWI [19], AFLW2000 [39] and Pandora [3].

- The **300W-LP dataset** is a synthetic dataset used for training and comprises an extensive collection of 61225 face samples. It spans a wide range of individuals, illumination conditions, and poses, which makes it ideal for training head pose estimation models [38], [6].
- The **BIWI dataset** is used for testing and offers over 15000 images that capture head poses from 20 individuals along wide ranges, mainly for yaw and pitch. Widely recognized as a benchmark dataset for head pose estimation challenges [22], [16], [38], it provides depth and RGB images (640x480 pixels) and ground truth pose annotations for each image.
- The **AFLW2000 dataset** is used for testing and consists of 2000 in-the-wild images with diverse head poses and strong variations in lighting and background conditions. It contains ground truth annotations of the Euler angles of the head poses.
- For all the aforementioned datasets, and similarly to [6], we generate synthetic occlusions to obtain the occluded versions of the 300W-LP, BIWI, and AFLW2000 datasets.
- The **Pandora dataset** [3] is used for testing and includes real-life occlusions. It simulates driving poses from the perspective of a camera situated inside a dashboard. It contains over 250000 RGB (1920x1080 pixels) and depth images (512x424) with corresponding annotations, capturing head poses in a wide range of the Euler angles. For our specific tests, we use 9619 occluded head images, featuring actors wearing various common occlusion garments such as sunglasses, scarves, caps, and masks.

B. Implementation Details

In the network structure illustrated in Fig. 3, we use ResNet-50 [13] as the backbone encoder. This 50-layer convolutional neural network has proven to be effective and suitable for head pose estimation within deep learning frameworks [22], [6]. Specifically, we use a ResNet-50 pre-trained on ImageNet [7]. We follow the training procedure described in Section III-B and illustrated in Fig. 4. In **Stage 1**, we use the original non-occluded 300W-LP dataset to initialize the encoder parameters and the latent feature space. In **Stage 2**, we use the synthetically occluded version of the same dataset. Both training stages are trained for 25 epochs, using Adam optimization with a scheduled learning rate initialized at 10^{-5}, $\epsilon = 10^{-8}$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The regularization coefficient α for the regression components (see (5)) is set to 1. We trained our model with initialization for $K = 10$ clusters, which led to optimal results (see Sec. IV-C). All training and testing input images are pre-processed using a face detector to crop the face region.

Opposing to [6], we no longer need to use the latent ground truth of the images with which we train the **Stage 1**. Thus, we no longer depend on image pairs (*i.e.*, non-occluded and respective occluded replicas) and the correspondent ground truth *constraint* requirements when augmenting the training data (see limitation **(2)** in Sec.I).

The images are also resized to the input dimension of the ResNet-50 backbone (*i.e.* 224x224x3) and normalized using the mean and standard deviation of ImageNet for all color channels. We train our model with 80-20 training/validation splits and a training batch size of 128. The fully-connected layers predict 66 different classification bins in the range $\pm 99^{o}$. We augment the training dataset by randomly flipping, blurring, down-sampling, and up-sampling images.

C. Head Pose Estimation Results on the BIWI, AFLW2000 and Pandora Datasets

We assess the estimation error of our methodology and compare it to the state-of-the-art methods in both the original and synthetic occluded variants of the BIWI and AFLW2000 datasets. Additionally, we evaluate the performance on the Pandora dataset, which includes real-life occlusions. Regarding the AFLW2000 datasets, 31 images were excluded as they fell outside the range accommodated by our bin classification. The head pose estimation mean squared error (MAE) results are displayed in Tables I, II and III, for BIWI, AFLW2000, and Pandora, respectively.

Notice that in Tables I, II and III we include the method LSR in [6], as this method uses a ground truth for each image, that is, having $K = N$. For instance, in the case of 300W-LP training dataset, this involves $K = 61225$ embedding ground truth samples. Our approach only requires $K = 10$, representing cluster centroids. This means that the results in [6] are derived by accessing the entire latent information, whereas our method only has access to a much smaller number of cluster centroids in that latent space. Thus, the results of the method in [6] constitute our upper bound performance since we do not anticipate our method

BIWI[19]	Non-occluded Images				Occluded Images				Combined
	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE	MAE
FSA-Net [34]	5.420	5.568	4.515	5.168	10.987	9.848	7.846	9.560	7.364
6DRepNet [14]	4.238	4.580	3.337	4.051	7.883	14.983	9.665	10.844	7.252
DAD-3D [20]	**3.668**	3.855	3.537	**3.687**	**5.532**	7.924	7.478	6.978	5.328
Lightweight [17]	5.135	3.561	4.106	4.267	10.784	9.654	10.309	10.249	7.292
Hopenet [22]	4.375	**3.559**	**3.348**	3.761	6.725	8.616	7.338	7.560	5.661
LSR [6]	*4.291*	*3.086*	*3.179*	*3.519*	*5.429*	*4.823*	*3.467*	*4.573*	*4.046*
LEC-HPE (ours)	4.602	3.897	3.759	4.086	6.153	**6.576**	**4.605**	**5.778**	**4.932**

TABLE I: Head pose estimation MAE (º) results with BIWI. LSR method values are in grey italics since they constitute upper-bound target performance. Bold values are best results apart from target performance.

AFLW2000 [39]	Non-occluded Images				Occluded Images				Combined
	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE	MAE
FSA-Net [34]	5.109	6.462	3.356	5.642	13.664	10.880	10.067	11.537	8.590
6DRepNet [14]	3.230	4.658	3.091	3.660	8.904	9.799	7.408	8.704	6.182
DAD-3D [20]	**3.134**	**4.630**	**3.090**	**3.618**	7.084	14.953	13.680	11.906	7.762
Lightweight [17]	4.267	5.015	3.722	4.335	10.771	9.278	8.335	9.461	6.898
Hopenet [22]	4.965	5.250	3.956	4.724	12.438	10.277	8.586	10.434	7.579
LSR [6]	*3.813*	*5.420*	*4.003*	*4.412*	*4.741*	*6.254*	*4.765*	*5.253*	*4.833*
LEC-HPE (ours)	3.843	5.177	4.038	4.353	**5.211**	**6.481**	**5.064**	**5.589**	**4.971**

TABLE II: Head pose estimation MAE (º) results with AFLW2000. LSR method values are in grey italics since they constitute upper-bound target performance. Bold values are best results apart from target performance.

Pandora [3]	Occluded Images			
	Yaw	Pitch	Roll	MAE
FSA-Net [34]	11.736	8.607	**6.691**	9.011
6DRepNet [14]	10.896	6.897	7.500	8.431
DAD-3D [20]	9.348	7.437	8.474	8.420
Lightweight [17]	10.133	8.690	7.064	8.629
Hopenet [22]	10.442	7.239	6.925	8.202
LSR [6]	*9.096*	*5.657*	*6.215*	*6.989*
LEC-HPE (ours)	**9.071**	**5.885**	6.918	**7.291**

TABLE III: Head pose estimation MAE (º) results with Pandora. LSR method values are in grey italics since they constitute upper-bound target performance. Bold values are best results apart from target performance.

to outperform these results. Our goal is to closely approach their performance while significantly reducing the required ground truth data.

To obtain the optimal number of K centroids, we applied the elbow method [26] with the initialized latent feature space and tested HPE in models with $K \in \{5, 10, 15, 20, 30, 40\}$. Both the elbow method and the tests indicated $K = 10$ as the model order that provides the best performance.

Overall, our method delivers competitive performance compared to [6] and surpasses other SoTA methodologies in occluded scenarios by a significant margin.

When compared to the upper-bound method (LSR), the occluded average MAE was approximately only 0.3º worse in AFLW2000 and Pandora. Furthermore, the non-occluded average error was identical to LSR and even surpassed it in AFLW2000. This led to a combined average MAE which is also competitive with LSR results across all datasets. The combined average MAE for AFLW2000 is only 2.9% worse than that of LSR. Regarding other state-of-the-art methods, the LEC-HPE estimation results in occluded images improve results by a substantial margin of 36%, 17%, and 11%, when compared to the best-performing method (apart from upper-bound LSR) in AFLW2000, BIWI, and Pandora, respectively. Regarding the combined MAE, our model surpasses the best-performing method (apart from LSR) by 19.6% in AFLW2000, 11.1% in Pandora, and 7% in BIWI.

D. Study on the Impact of Clustering

BIWI [19]	Non-occluded	Occluded	MAE
$\beta = 0$	4.653	6.434	5.544
$\beta = 10$	4.509	6.096	5.303
$\beta = 100$	**4.086**	**5.778**	**4.932**
$\beta = 1000$	5.035	6.676	5.856

TABLE IV: LEC-HPE MAE (º) with different β values in the BIWI dataset.

We carried out an ablation study on the impact of the clustering term by varying the β regularization coefficient in the overall training loss implemented in **Stage 2** (see (9)). We trained LEC-HPE with four different values for

AFLW2000 [39]	Non-occluded	Occluded	MAE
$\beta = 0$	5.069	6.243	5.656
$\beta = 10$	4.987	6.095	5.541
$\beta = 100$	**4.353**	**5.589**	**4.971**
$\beta = 1000$	5.610	6.593	6.102

TABLE V: LEC-HPE MAE (º) with different β values in the AFLW2000 dataset.

Pandora [3]	Occluded
$\beta = 0$	7.938
$\beta = 10$	7.787
$\beta = 100$	**7.291**
$\beta = 1000$	8.169

TABLE VI: LEC-HPE MAE (º) with different β values in the Pandora dataset.

this coefficient, $\beta = \{0, 10, 100, 1000\}$. The MAE estimation results are listed in Tables IV, V and VI.

When $\beta = 0$, the second training stage does not include the clustering loss and is identical to the first training stage. This means that we only use the top branch in Fig. 3.

The results across all datasets reveal that including the clustering term ($\beta = 10$ and $\beta = 100$) in the second stage reduces the estimation error for occluded and non-occluded images. In particular, when $\beta = 100$ is employed, optimal outcomes are observed, resulting in an error reduction ranging from 8% to 12% compared to when the clustering term is not utilized ($\beta = 0$). Estimation errors escalate with $\beta = 1000$, indicating that an excessively strong influence of the clustering term on the overall loss leads to the deterioration of the properly initialized latent feature space. This confirms the need to maintain the fine-grained Euler angle losses in the second training stage to avoid the model deviating from the best performance for the fine-grained head pose estimation task.

V. Conclusion

In this paper, we present an efficient methodology to address the challenge of occlusion in head pose estimation, a significant hurdle in this sub-field of computer vision. From the experimental evaluation, the proposed framework improves occlusion robustness by combining unsupervised latent embedding clustering in the latent feature space with a fine-grained Euler angle multi-loss scheme.

The main idea behind LEC-HPE is to improve feature representation for pose estimation while refining the latent embedding space through clustering, eliminating the need for labeled embedding data for each training image. This approach offers a more efficient alternative compared to some of the most recent occlusion-focused state-of-the-art work, without requiting a constrained expansion of the training dataset.

We demonstrate experimentally that we can achieve similar results without the need to have ground truth for each latent embedding label. Our results also surpass the state of the art for standard head pose estimation by a significant margin in occluded images. We perform an ablation study to quantitatively evaluate the impact of the clustering term and verify that using this term improves the estimation results. We also confirm the need to include the fine-grained Euler angles scheme to avoid latent space corruption.

Although not fully addressed in this paper, further work will include a process for automatic selection of the optimal parsimonious number of cluster centroids in the low dimensional latent space. Also, we plan to test this methodology with smaller and more efficient backbone encoders for low-power applications. We also intend to evaluate the use of clustering losses for the classification component in the multi-loss scheme.

VI. Acknowledgements

This work was supported by LARSyS funding (DOI: 10.54499/LA/P/0083/2020, 10.54499/UIDP/50009/2020, and 10.54499/UIDB/50009/2020], through Fundação para a Ciência e a Tecnologia and by the SmartRetail project [PRR - C645440011-00000062], through IAPMEI - Agência para a Competitividade e Inovação.

References

[1] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7613–7623, 2021.

[2] K. Amara, M. A. Guerroudji, O. Kerdjidj, N. Zenati, and N. Ramzan. Holotumour: 6dof phantom head pose estimation based deep learning and brain tumour segmentation for ar visualisation and interaction. *IEEE Sensors Journal*, 2023.

[3] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara. Poseidon: Face-from-depth for driver pose estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5494–5503. IEEE, 2017.

[4] A. Candeias, T. Rhodes, M. Marques, J. P. ao Costeira, and M. Veloso. Vision augmented robot feeding. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[5] G. Cantarini, F. F. Tomenotti, N. Noceti, and F. Odone. Hhp-net: A light heteroscedastic neural network for head pose estimation with uncertainty. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3341–3350, 2021.

[6] J. Celestino, M. Marques, J. C. Nascimento, and J. P. Costeira. 2d image head pose estimation via latent space regression under occlusion settings. *Pattern Recognition*, 137:109288, 2023.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[8] N. Dhingra. Lwposr: Lightweight efficient fine grained head pose estimation. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1204–1214, 2022.

[9] J. M. Diaz Barros, B. Mirbach, F. Garcia, K. Varanasi, and D. Stricker. *Real-Time Head Pose Estimation by Tracking and Detection of Keypoints and Facial Landmarks*, pages 326–349. 07 2019.

[10] A. Fernández, R. Usamentiaga, J. L. Carús, and R. Casado. Driver distraction using visual-based sensors and algorithms. *Sensors*, 16(11), 2016.

[11] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li. Towards fast, accurate and stable 3d dense face alignment. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 152–168, Cham, 2020. Springer International Publishing.

[12] X. Guo, L. Gao, X. Liu, and J. Yin. Improved deep embedded clustering with local structure preservation. In *Ijcai*, volume 17, pages 1753–1759, 2017.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2496–2500, 2022.

[15] W. Hoff and T. Vincent. Analysis of head pose accuracy in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 6(4):319–334, 2000.

[16] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee. Quatnet: Quaternion-based head pose estimation with multiregression loss. *IEEE Transactions on Multimedia*, 21(4):1035–1046, 2019.

[17] X. Li, D. Zhang, M. Li, and D.-J. Lee. Accurate head pose estimation using image rectification and a lightweight convolutional neural network. *IEEE Transactions on Multimedia*, pages 1–1, 2022.

[18] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[19] K. S. Mader. Biwi Kinect Head Pose Database, 2018.

[20] T. Martyniuk, O. Kupyn, Y. Kurlyak, I. Krashenyi, J. Matas, and V. Sharmanska. Dad-3dheads: A large-scale dense, accurate and diverse dataset for 3d head alignment from a single image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[21] Y. Ren, J. Pu, Z. Yang, J. Xu, G. Li, X. Pu, P. S. Yu, and L. He. Deep clustering: A comprehensive survey. *arXiv preprint arXiv:2210.04142*, 2022.

[22] N. Ruiz, E. Chong, and J. M. Rehg. Fine-grained head pose estimation without keypoints. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2155–215509, 2018.

[23] K. Sankaranarayanan, M.-C. Chang, and N. Krahnstoever. Tracking gaze direction from far-field surveillance cameras. In *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 519–526, 2011.

[24] G. Sheng, Q. Wang, C. Pei, and Q. Gao. Contrastive deep embedded clustering. *Neurocomputing*, 514:13–20, 2022.

[25] K. Smith, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez. Tracking the visual focus of attention for a varying number of wandering people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1212–1229, 2008.

[26] R. L. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[27] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[28] Y. Wang, G. Yuan, and X. Fu. Driver's head pose and gaze zone estimation based on multi-zone templates registration and multi-frame point cloud fusion. *Sensors*, 22:3154, 04 2022.

[29] M. Wenzel and W. Schiffmann. Head pose estimation of partially occluded faces. pages 353– 360, 06 2005.

[30] Y. Wu, C. Gou, and Q. Ji. Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5719–5728, 2017.

[31] Y. Wu and Q. Ji. Robust facial landmark detection under significant head poses and occlusion. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3658–3666, 2015.

[32] J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[33] Y. Xu, C. Jung, and Y. Chang. Head pose estimation using deep neural networks and 3d point cloud. *Pattern Recognition*, 121:108210, 07 2021.

[34] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1087–1096, 2019.

[35] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, and Y.-Y. Chuang. Ssr-net: A compact soft stagewise regression network for age estimation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1078–1084. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[36] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4010–4019, 2017.

[37] H. Yuan, M. Li, J. Hou, and J. Xiao. Single image-based head pose estimation with spherical parametrization and 3d morphing. *Pattern Recognition*, 103:107316, 02 2020.

[38] Y. Zhou and J. Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. In *British Machine Vision Conference (BMVC 2020)*, 2020.

[39] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Li. Face alignment across large poses: A 3d solution. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, 2016.