
CLERF: Contrastive LEaRning for Full Range Head Pose Estimation

Ting-Ruen Wei
Santa Clara University
Santa Clara, CA

Haowei Liu
Santa Clara University
Santa Clara, CA

Huei-Chung Hu
DOCOMO Innovations
Sunnyvale, CA

Xuyang Wu
Santa Clara University
Santa Clara, CA

Yi Fang
Santa Clara University
Santa Clara, CA

Hsin-Tai Wu
DOCOMO Innovations
Sunnyvale, CA

Abstract

We introduce a novel framework for representation learning in head pose estimation (HPE). Previously such a scheme was difficult due to head pose data sparsity, making triplet sampling infeasible. Recent progress in 3D generative adversarial networks (3D-aware GAN) has opened the door for easily sampling triplets (anchor, positive, negative). We perform contrastive learning on extensively augmented data including geometric transformations and demonstrate that contrastive learning allows networks to learn genuine features that contribute to accurate HPE. On the other hand, we observe that existing HPE works struggle to predict head poses as accurately when test image rotation matrices are slightly out of the training dataset distribution. Experiments show that our methodology performs on par with state-of-the-art models on standard test datasets and outperforms them when images are slightly rotated/ flipped or full range head pose. To the best of our knowledge, we are the first to deliver a true full range HPE model capable of accurately predicting any head pose including upside-down pose. Furthermore, we compared with other existing full-yaw range models and demonstrated superior results.

1 Introduction

In the expansive landscape of machine learning, contrastive learning has marked its territory as a pivotal technique, particularly within the unsupervised learning paradigm [5]. It operates on a simple yet effective principle: teaching models to recognize similar and dissimilar instances, and leveraging large datasets to enhance model understanding and performance. While its application has been widely applied in numerous aspects of computer vision, the venture into Full Range (FR) Head Pose Estimation (HPE) using contrastive learning would be entering uncharted territory. Please note currently available FR models are not in our standard full range because none of them are capable of handling, for example, upside-down head poses. HPE is a complex and crucial task in computer vision [12, 4, 40, 6], which aims to accurately determine the orientation of a person’s head, and the task is important for understanding human behavior and intentions in various applications, from augmented and virtual reality environments to safety systems in vehicles and interactive robotics.

The challenge in applying contrastive learning to FRHPE lies in the sparsity of head poses in the 3D space. It is extremely rare to find an anchor-positive (another head facing the same direction) in the full range 3D space. If we allow anchor-positives to be twenty degrees apart from the anchor at the maximum, the probability of finding an anchor-positive is under $0.0002 (\frac{1}{18^3})$, exemplifying the difficulty of applying contrastive learning.

While many existing works in HPE focus on the limited frontal range [36, 16, 26, 8, 14, 4, 7, 37], with yaw in the range of -90 to 90 degrees, research in FR models [40, 12] are underdeveloped. While it remains a challenge to curate a dataset that covers the FR, which we define as the range of -180 to 180 degrees for all yaw, pitch, and roll (refer to [34]), the FR capability extends coverage from the limited frontal range and is particularly significant in HPE of sports and acrobatic actions. Additionally, we observe that many existing models are sensitive to slight transformations of the test images: minor rotation and/or flip.

To tackle these challenges, we propose a framework to train with contrastive learning a FR model (CLERF) that demonstrates competitive performance on not only original test images but also on slightly augmented versions and other angles that existing models struggle on. Specifically, CLERF generates a synthetic head image with the same yaw and pitch as a real image and geometrically transforms the generated head image to have the exact head orientation as the real image. The process guarantees a positive pair thereby enabling the use of contrastive learning. The advantage of synthetic data lies in its flexibility to represent any head orientation and address a wider range including the angles that are rarely observed in real data. With geometric transformation, we can further expand the coverage to more angles. Through contrastive learning and image augmentation, CLERF aims to learn a good representation, separating neighboring angles from further ones in the FR and making more accurate head pose predictions.

Our main contribution can be summarized as follows:

- We identify the advantage of 3D-aware GAN to generate anchor-positives and facilitate contrastive learning in full range head pose estimation.
- We perform calculations to find parameters of general geometric transformations. Such transformations allow 3D-aware GAN to synthesize positive images to match anchors and expand the head pose coverage to full range.
- We observe that existing models are sensitive to slight transformations of the test images.
- We demonstrate on par performance with state-of-the-art models on original images of the standard test sets and outperformed them on minor variants.
- Our model is capable of handling true FR head pose and outperforms existing full-yaw range models at the full range capability. Code will be released.

2 Related Work

Head pose estimation. Classical approaches, such as template matching and detector arrays, related to head pose estimation can be found in the survey paper [24]. Deformable models [3, 42, 10, 43] were used to create the commonly used synthetic pose dataset such as 300W-LP [42]. Head pose estimation can be divided into two categories, with and without facial landmarks. Traditionally, HPE was made when facial features are visible, with Dlib [19] being one of the pioneers in using face landmarks for prediction. However, such method becomes error-prone when facial landmarks are not detected, especially at large yaw angles. Therefore, with the advent of deep learning, researchers utilized Convolutional Neural Networks (CNN) to predict the three Euler angles directly, as in HopeNet [26] and WHENet [40]. Nonetheless, directly applying regression on the Euler angles leads to discontinuity easily because each distinct angle can be represented by different numbers, e.g., $0^\circ = 360^\circ$. To avoid such discontinuity, 6DRepNet [12], 6DRepNet360 [12], and TriNet [4] predict the 3×2 and 3×3 rotation matrices respectively, while still evaluating the Euler angles for comparison with other HPE models. Other works include Kuhnke et al. [20], who bridged the gap between synthetic and realistic images in head pose estimation using relative pose consistency, and Opal [6], which aligned the different reference systems between the training and testing datasets and proposed a generalized geodesic distance metric as the loss function. SemiUHPE [39] applied weak-strong augmentations in a semi-supervised fashion, leveraging a large amount of unlabeled head poses. Instead of the common CNN approach, TokenHPE [37] utilized a transformer for HPE by predicting orientation tokens.

Head pose dataset creation. The 300W across Large Poses (300W-LP) dataset [42] was based on 300W [27], which standardized multiple alignment datasets with 68 landmarks, including AFW [41], LFPW [2], HELEN [38], IBUG [27], and XM2VTS [23]. The Euler angles labeled are extracted from the rotation matrix estimated from its 3D Dense Face Alignment, which applies the morphable model

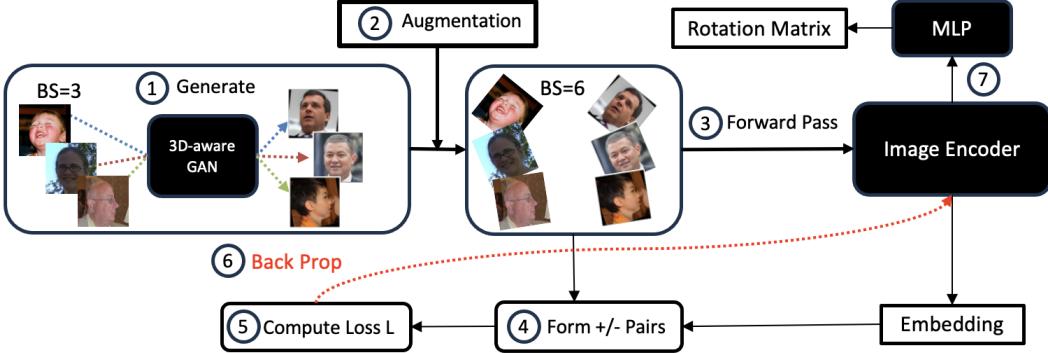


Figure 1: Proposed method for contrastive learning in full range head pose representation, with a batch size of 3 instances for illustration purposes. Steps 1~6 comprise a training iteration for the representation model which is frozen in the downstream MLP training in step 7.

3DMM [3] to obtain the standard and rotated 3D faces tailored for an image. While 300W-LP offered pose labels limited to frontal views, the CMU Panoptic Dataset [18] provided extensive 3D facial landmarks for multiple individuals captured from cameras spanning an entire hemisphere, which can be converted to a head pose. WHENet [40] pioneered techniques to transform the CMU Panoptic Dataset into a comprehensive head pose estimation dataset covering a larger range of angles.

Contrastive Learning. To the best of our knowledge, no prior work has utilized contrastive learning in head pose estimation, so we discuss a few contrastive learning applications in related computer vision areas. For gaze estimation, GazeCLR [17] revised the NT-Xent loss [5] on multi-view images to learn gaze representation and improve the cross-domain gaze estimation performance. CRGA [32] utilized contrastive regression in gaze estimation for a domain adaptation task. As for hand pose estimation, PeCLR [29] modified the SimCLR [5] framework for equivariance contrastive learning on 3D hands and improved the performance of the original models. In face recognition, PCL [21] learned face representations by disentangling the pose from the original face in computing the contrastive loss. For human activity representation, P-HLVC [28] leveraged human pose dynamics to learn human activities that are resistant to domain shifts. In human pose estimation, Honari et al. [13] separated time-variant from time-invariant features and only applied contrastive learning on time-invariant features. ICON [35] enforced the consistency between individual keypoints belonging to the same category across images and also the consistency between pair relations across images in a multi-person scenario.

3 Methodology

Our proposed methodology involves three key components: anchor-positive generation to empower contrastive learning, geometric transformation to cover the full range space, and contrastive learning to strengthen the fine-grained head pose predictions. An overview is shown in Figure 1.

3.1 Anchor-Positive Synthetic Image Generation

Given a training sample, we generate with 3D-aware GAN a synthetic image that has the same yaw and pitch values and apply proper rotation to match the roll value (Figure 1 step 1). With the mathematics listed in Hu et al. [15], we can solve for a triad of $(yaw, pitch, \phi)$ from the rotation matrix. The rotation associated with the roll angle ϕ based on the rotation matrix $R \in SO(3)$ (refer to [33]) representing rotation with yaw and pitch is the following:

$$\begin{aligned} R_{rotate}(\phi) &= R_{extrinsic}(\phi) \times R \\ &= \begin{bmatrix} \cos(\phi) & \sin(\phi) & 0 \\ -\sin(\phi) & \cos(\phi) & 0 \\ 0 & 0 & 1 \end{bmatrix} \times R \end{aligned} \quad (1)$$

where R_{rotate} is the rotation matrix of the rotated image. We provide detailed formulation below.

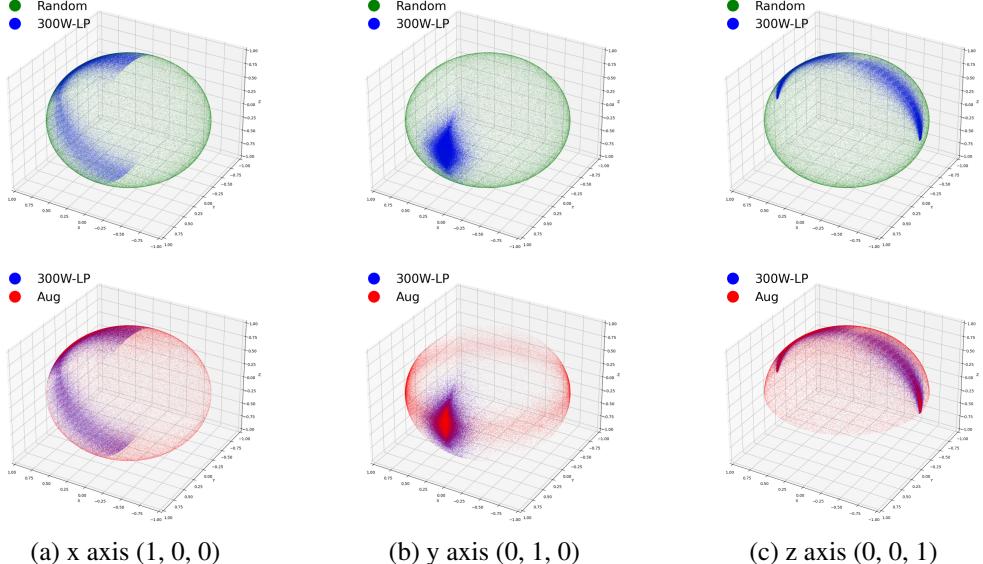


Figure 2: Visualization of 300W-LP dataset [43] after randomized rotation and flipping augmentations. A point on the sphere $\in \mathbb{R}^3$ is formed by projecting a rotation matrix to the 3D sphere, which multiplies unit vectors in (a) x, (b) y, and (c) z coordinate axes. The top row shows random rotation matrices (in green) along with those in the 300W-LP dataset (in blue), and the bottom row shows the augmented 300W-LP dataset (in red). The geometric transformations expand the original dataset for wider coverage.

a novel two-stage self-adaptive image alignment for robust training, a tri-grid neural volume that effectively manages the representation challenges of both the face and the back of the head, and the integration of prior knowledge from 2D image segmentation to improve 3D model training. Panohead allows the generation of a head pose facing a specific yaw and pitch, which is a crucial component in facilitating contrastive learning. As a starting dataset, we generated 25,984 images with 812 unique-looking people and 32 images each. Each set of 32 images is uniformly distributed within the yaw and pitch space of [-3.14 radians, 3 radians] and [-1.5 radians, 0.1 radians]. The rotation matrix R follows the reference system of 300W-LP as below:

3.2 Geometric Transformation for Full Range HPE

The second key lies in the geometric transformation on head pose images and their corresponding rotation matrices to enable FRHPE (Figure 1 step 2). Besides rotation defined in Equation 1, we follow the mathematics in Hu et al. [15] and apply flipping across L_θ on the XY-plane with a given rotation $R \in SO(3)$ to obtain the following rotation matrix:

$$\begin{aligned} R_{flip}(\theta) &= Flip_\theta \times R \times Flip_X \\ &= \begin{bmatrix} \cos(2\theta) & \sin(2\theta) & 0 \\ \sin(2\theta) & -\cos(2\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \times R \times \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned} \quad (2)$$

where θ is the counter-clockwise angle between the positive x-axis and the line L that we flip across. Figure 2 shows the resulting distribution of the 300W-LP dataset after our rotation and flipping (in red), compared to that of random rotation matrices (in green) and the original distribution (in blue). As the 300W-LP dataset lacks head poses facing away from the camera (large yaw angles), the resulting distribution does not fully match the uniform distribution (a semi-sphere versus a sphere in Figure 2(c)). Nonetheless, our rotation and flipping augmentation fills the gap and transforms head poses to cover a much wider range than the original images.

Theorem 3.1. *Let H be a image geometric transformation function and A, B present two valid rotation matrices, then $d(A, B) = d(H(A), H(B))$.*

To maintain the guaranteed existence of a positive pair, we apply the same rotation and flipping augmentation to both images in a positive pair. Since this is a crucial condition for contrastive learning, we prove Theorem 3.1 in Appendix A that geometric transformations preserve the geodesic distance d for any two rotation matrices $A, B \in SO(3)$, with d defined as the following [12]:

$$d(A, B) = \cos^{-1}\left(\frac{\text{tr}(A \times B^T) - 1}{2}\right) \quad (3)$$

where tr and T represent the trace and transpose, respectively. In other words, for any chosen anchor, positive, negative triplet (a, p, n) , applying the same rotation and flipping augmentations results in another valid triplet (a', p', n') . We present our strategy for sampling triplets for contrastive learning.

Corollary 3.1.1. *Strategy for sampling (anchor, positive, negative) triplets under the mathematical foundation in [15]:*

1. Take an image I_A as anchor and generate a synthetic image I_P with rotation matrix $R' = (R_{\text{pitch}} \times R_{\text{yaw}})$.
2. For anchor images provided with rotation matrices R as labels, we consider solving the corresponding triad of (yaw, pitch, roll) for rotation matrix $R = R_{\text{roll}} \times (R_{\text{pitch}} \times R_{\text{yaw}})$. Then we rotate the image I_P with R_{roll} to create a positive image with rotation matrix $R_{\text{roll}} \times R' = R$. Refer to Figure 6 in Appendix B for examples.
3. For any existing anchor, positive, negative triplet (a, p, n) , applying random rotation or flipping augmentation H returns a new valid triplet $(H(a), H(p), H(n))$.

Armed with the result above, we can easily sample lots of valid triplets and their augmented versions for contrastive learning.

3.3 Contrastive Learning

After data augmentation, we pass the batch of images to the image encoder E to obtain their embedding vectors (Figure 1 step 3) and form triplets. Though the image generation step guarantees an anchor-positive, we include neighboring head orientations that have a geodesic similarity $\cos(d(R, R'))$ higher than the threshold T_{GD} as additional anchor-positives, where R and R' are the rotation matrices of the anchor and another image in the mini-batch (Figure 1 step 4). Furthermore, we filter and retain the hard and semi-hard negatives that violate a margin v according to the Euclidean distance between the embedding vectors. Subsequently for back-propagation (Figure 1 step 6), we compute the Circle Loss [30] L (Figure 1 step 5) that aims to increase the similarity between positive pairs and decrease that between negative pairs, through a re-weighting process to focus on the less-optimized samples, with L defined as the following:

$$\log\left(1 + \sum_{i=1}^{N_p} \exp(\gamma * \max(0, 1 + m - s_i)(s_i - m)) \sum_{j=1}^{N_n} \exp(-\gamma * \max(0, m + s_j)(s_j - 1 + m))\right) \quad (4)$$

where γ is a scaling factor, m is a relaxation margin, s is the cosine similarity of the embedding vectors, N_p and N_n are the number of positive and negative pairs respectively.

After the representation model is trained, we freeze it and train a downstream multi-layer perceptron (MLP) to project the representation to the fine-grained head pose angles, represented by a rotation matrix (Figure 1 step 7). To ensure a unitary matrix, we transform the six-dimensional MLP output to the nine-dimensional rotation matrix through the Gram-Schmidt process as implemented in 6DRepNet [12]. With a trained representation model and its downstream MLP, we evaluate test datasets and compare against baseline models.

Table 1: Model performance against baseline models across six datasets. CLERF achieved on par performance with the baseline models on the original AFLW2000 and BIWI, and outperformed all baseline models when test images are slightly rotated and flipped (SA AFLW2000 and SA BIWI). In the full range, CLERF outperformed other full range models on heavily-rotated images by a large margin. FR indicates whether the model is full range, and the optimal results are highlighted in bold.

Model	FR	Yaw	Pitch	Roll	Mean	Yaw	Pitch	Roll	Mean
		AFLW2000				BIWI			
FSA-Net [36]	×	4.50	6.08	4.64	5.07	4.27	4.96	2.76	4.00
HopeNet [26]	×	6.47	6.56	5.44	6.16	5.17	6.98	3.39	5.18
TokenHPE [37]	×	5.44	4.36	4.08	4.66	4.51	3.95	2.71	3.72
6DRepNet [12]	×	3.27	4.58	2.98	3.61	3.23	5.32	2.78	3.78
WHENet [40]	✓	5.11	6.24	4.92	5.42	3.99	4.39	3.06	3.81
6DRepNet360 [12]	✓	3.58	5.28	3.46	4.11	3.28	6.06	3.08	4.14
CLERF	✓	4.22	6.18	4.67	5.02	3.57	4.49	3.13	3.73
		SA AFLW2000				SA BIWI			
FSA-Net [36]	×	18.59	16.02	17.04	17.22	7.09	9.42	6.20	7.57
HopeNet [26]	×	6.44	9.31	6.08	7.28	10.80	10.07	9.32	10.07
TokenHPE [37]	×	6.29	7.29	6.56	6.70	6.84	7.12	5.19	6.39
6DRepNet [12]	×	8.41	8.80	7.68	8.30	6.45	7.09	6.79	6.78
WHENet [40]	✓	13.11	12.88	15.06	13.68	9.51	10.99	9.52	10.00
6DRepNet360 [12]	✓	5.69	6.76	5.34	5.93	10.21	7.88	6.51	8.20
CLERF	✓	4.56	6.10	4.86	5.36	6.57	4.52	4.21	5.10
		FA AFLW2000				FA BIWI			
WHENet [40]	✓	22.04	23.03	39.82	28.30	30.77	22.43	41.65	31.95
6DRepNet360 [12]	✓	14.00	16.93	21.96	17.63	25.97	17.90	34.04	25.97
CLERF	✓	4.84	5.79	4.31	4.98	7.68	7.89	6.93	7.50

4 Experiments

4.1 Datasets

Training. 300W-LP [43] contains 122,450 images from multiple databases with faces mainly in the frontal range. We utilize PanoHead [1] as the 3D-aware GAN to generate anchor-positives for the representation model training only.

Testing. Following previous works, we test the models on AFLW2000 [43] and BIWI [9]. These datasets mainly contain frontal faces with yaw in the range of [-90, 90] degrees. Therefore, to facilitate a comprehensive analysis of our full range capability, we additionally evaluated on four variants of the previous test sets: slightly-augmented (SA) AFLW2000, SA BIWI, fully-augmented (FA) AFLW2000, and FA BIWI. The SA version is obtained by rotating each image clockwise by 10 degrees and flipping it along the line that is 85 degrees counter-clockwise from the positive x-axis (an example is shown in the second row of Figure 4). As for the FA version, we randomly rotate between -180 to 180 degrees and flip along the line that is between 0 to 90 degrees counter-clockwise from the positive x-axis (an example is shown in the third row of Figure 4).

4.2 Experimental Setup

Representation Model. We use the improved version of Swin Transformer Base model [22] with weights pre-trained on ImageNet in PyTorch and an output embedding of size 1024 as the image encoder E . We train it for 30 epochs on 20,000 images from 300W-LP dataset and 20,000 PanoHead-generated images with the Adam optimizer and a learning rate of 10^{-5} on a single NVIDIA Tesla V100 32GB GPU. We utilize the PyTorch Metric Learning library [25] for triplet sampling and loss computation. For hyperparameter values, we apply $T_{GD} = 0.8$ and $v = 0.1$ for triplet filtering and $m = 0.4$ and $\gamma = 80$ for loss L .

Table 2: Supervised learning against contrastive learning. We validate the contrastive learning strategy by comparing the result to that of supervised learning and observe a significant improvement. The optimal results are highlighted in bold.

Model	AFLW				BIWI			
	Yaw	Pitch	Roll	Mean	Yaw	Pitch	Roll	Mean
CLERF-Supervised	4.65	6.48	4.81	5.31	5.56	6.21	2.58	4.78
CLERF	4.22	6.18	4.67	5.02	3.57	4.49	3.13	3.73

MLP. Our downstream MLP consists of four fully connected layers of 256 units and a skip connection from the input layer to the last fully connected layer. We train the MLP for 40 epochs with mean geodesic distance d as the loss function and an exclusive subset of the training dataset as the validation set for early stopping.

Data Augmentation. With a probability of 0.5, we randomly rotate the image clockwise by a degree between 0 and 90, and with another probability of 0.3, we randomly flip the image along an axis that is between 0 and 90 degrees counter-clockwise from the positive x axis. For pixel-wise augmentation, we deploy the popular techniques including translation, resizing, down-sampling for low resolution, hue change, sharpness, grayness, contrast limited adaptive histogram equalization, and brightness. Furthermore, we conduct the following pixel-wise transformations: RGB shift, channel shuffle, gamma correction, color jitter, Gaussian noise, and Gaussian blur. For pixel removal, we utilize center crop and coarse dropout. These methods alter the pixel values, reinforcing a more robust training outcome.

Baseline Models and Evaluation Metric. We consider many existing HPE models as our baseline models. FSA-Net [36] utilized feature aggregation and soft stagewise regression. HopeNet [26] applied multiple losses on a convolutional neural network. TokenHPE [37] leveraged a transformer to learn the relationship within the facial part. 6DRepNet [12] relied on the geodesic loss function and its variant, 6DRepNet360 [12], expanded to full range. WHENet [40], the other full range model, wrapped the loss function to stabilize the learning of large yaw angles. Following these works, we evaluate the models with mean absolute error (MAE) on each of yaw, pitch, and roll and compute the average of the three MAEs as the Mean. Since no models, including ours, are specialized in one of yaw, pitch, or roll predictions, we mainly compare the Mean.

5 Empirical Results

5.1 Main Evaluation

We present the evaluation of all models on six datasets in Table 1. On AFLW2000, CLERF performed nearly on par with 6DRepNet, down by 1.4 degrees in Mean. On BIWI, CLERF performed on par with TokenHPE, falling short by 0.01 degree in Mean. We recognize that having FR coverage, CLERF optimizes for the entire range, leaving relatively less focus on the frontal range that non-full range models directed all of their resources to learn. We act on this to construct a more comprehensive analysis by slightly rotating and flipping the same test images (SA AFLW2000 and SA BIWI), keeping them within the same frontal range that non-full range models trained on. Results show that CLERF outperformed all baseline models, a sign that existing models might have been fixated on the exact angles of the original test images. With 0.57 and 1.29 degrees better than the runner-up for SA AFLW2000 and SA BIWI, respectively, CLERF demonstrates robustness and superiority. Additionally, we examine the full range capability by rotating and flipping the test images on a larger scale and only compare CLERF to full range baseline models. CLERF remains the highest performing, leading by more than 10 degrees in Mean. An interestingly observation is that CLERF had a minor improvement of 0.03 degrees in Mean on FA AFLW2000 compared to the original version, an evidence that CLERF optimizes for the entire range.

5.2 Ablation Studies

Contrastive Learning. We validate the contrastive learning motivation by comparing against the traditional supervised learning approach. We concatenate the representation model with the

Table 3: Ablation study on flip and rotation as data augmentation methods. Each of flip and rotation significantly improves the test performance on both the original and SA test images. The integration of flip and rotation achieved the best outcome. The optimal results are highlighted in bold. (Imp. % indicates the percentage of improvement in Mean.)

Augmentation	AFLW2000		SA AFLW2000		BIWI		SA BIWI		
	Method	Mean	Imp. %	Mean	Imp. %	Mean	Imp. %	Mean	Imp. %
CLERF		6.88	-	9.93	-	5.48	-	8.99	-
+ Flip only		5.98	13%	7.78	22%	4.96	9%	7.08	21%
+ Rotate only		5.66	18%	6.19	38%	4.64	15%	5.77	36%
+ Rotate & Flip		5.02	27%	5.36	46%	3.73	32%	5.10	43%

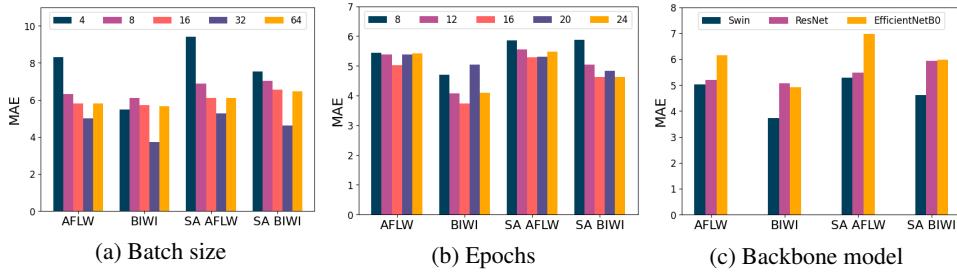


Figure 3: Change of model performance across different choices in (a) batch size, (b) epochs, and (c) backbone model. A batch size of 32, 16 epochs, and Swin Transformer make the best combination. (AFLW2000 is abbreviated as AFLW.)

downstream MLP and train on geodesic loss with the same dataset. The comparison result is shown in Table 2, and we observe a significant improvement with our contrastive learning approach. This demonstrates that contrasting between neighboring angles and further angles boosts the learning of head orientations. By first grouping up the neighboring head poses and separating those that are further away, we enable the downstream MLP to more accurately project the representation to fine-grained head poses.

Batch Size. Batch size is a crucial hyperparameter in contrastive learning when anchor-positives and anchor-negatives are identified within a mini-batch. With sparsity in head orientations, a large batch size tends to increase the number of anchor-positives and the ratio of anchor-positives to anchor-negatives. Figure 3(a) shows the model performance across different batch sizes. We observe a consistent pattern across all four datasets: a larger batch size results in better performance up until a batch size of 32, which corresponds to 32 images from 300W-LP and 32 PanoHead images generated on-the-fly.

Epochs. As the representation model aims to learn a good representation, we study how the number of training epochs affects the learned representation, with results shown in Figure 3(b). Model performance peaked at 16 epochs and declined afterwards. This is likely due to the limited varieties appearing in the synthetic images generated through PanoHead that accumulated over the number of epochs.

Image Encoder Backbone. While we select the Swin Transformer as our image encoder for many of its strengths, we compare to two other backbone models to analyze our proposed approach across different image encoders, as shown in Figure 3(c). The Swin Transformer consistently outperformed others, with ResNet50 [11] generally performing better than the EfficientNetB0 model [31], which aligns with their descending number of model parameters.

Rotation and Flip Augmentation. An important component in our approach that enables FRHPE is the rotation and flip as data augmentation techniques. To quantify their significance, we present the respective improvements in Table 3. The combination of rotation and flip performed the best, with rotation or flip alone making their respective contributions. These geometric transformations resulting in better performance not only in AFLW2000 and BIWI, but also in their SA versions.

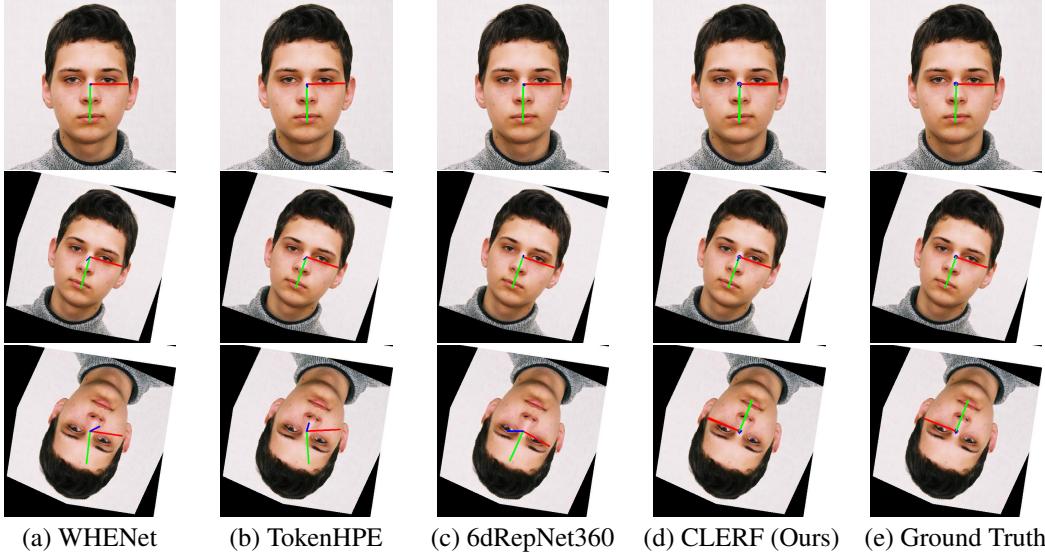


Figure 4: Head pose test predictions of (d) CLERF and (a)~(c) three baseline models versus the (e) ground truth on the original image (first row) and its SA (second row) and FA (third row) versions. Head pose is represented by the three lines colored in red, blue, and green. We observe barely noticeable differences between the model predictions in the original and SA test image, but CLERF much more accurately predicted the FA head pose.

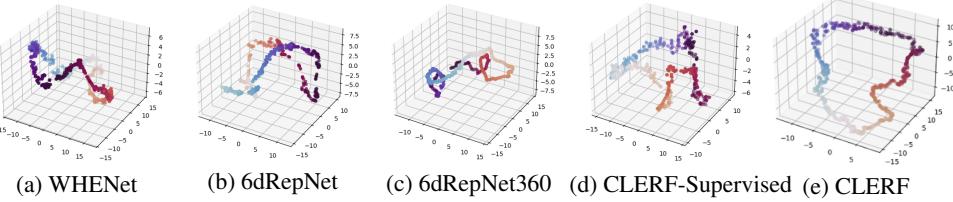


Figure 5: 3D TSNE plot of the embedding vectors of a video across (a)~(e) different models. The video shows a person turning around for a total of 360 degrees. The similarity of colors indicates their temporal proximity.

5.3 Case Studies and Visualization

Qualitative Analysis. After the empirical results, we visually illustrate our model performance with three test cases, each representing one from the original, SA, and FA versions of AFLW2000 test set. We compare to the ground truth and baseline models, as shown in Figure 4. The first row presents an original image with very similar head pose predictions across all models, which is consistent with the similar Mean values listed in Table 1. Second row presents the SA version of the same image and we observe more noticeable deviation for WHENet and TokenHPE on the blue line. As for the FA version of the original image shown in the third row, CLERF fully adapted to the heavy rotation while the baseline models struggled, despite that 6dRepNet360 and WHENet aimed at being full range models. This illustrates that although CLERF is on average one degree behind the baseline models on the original AFLW2000 test set, CLERF adapts and produces accurate results in the full range, making the one degree negligible, not to mention the other test set, BIWI, where CLERF performed on par with baseline models on the original images and outperformed them in the SA and FA versions.

TSNE Visualization. For comprehensive analysis, we take a further step to visualize with 3D t-distributed stochastic neighbor embedding (TSNE) the embedding vectors of a sequence of images showing a person turning in a full circle. Figure 5 compared to the supervised and other baseline models, using a cyclic color map that visualizes neighboring frames with similar colors. An optimal pattern consists of a clear circle with the colors following the cyclic order. CLERF (Figure 5(e))

exhibited clear separation for angles that are further away and kept nearby angles close as shown by the continuous change of colors. On the other hand, lacking only the contrastive learning component, the supervised model (Figure 5(d)) showed the continuity in color changes but did not separate opposite angles as effectively. We made a similar observation with the other baseline models (in Figure 5(a)~(c)).

6 Conclusion

In head pose estimation, the sparsity of head poses has hindered the use of contrastive learning. We proposed a novel idea to tackle this issue and utilize contrastive learning to separate the representation of opposite angles, thereby improving head pose estimation. By generating anchor-positives through a 3D-aware generative adversarial network and the proper geometric transformations, we achieved on par performance with current state-of-the-art models. Though many previous works optimized their performance on the frontal range, we observed a significant decline in performance when the test images are slightly rotated or flipped. In this realm, our model outperformed all baseline models. Furthermore, we validated our full range capability on heavily-transformed images and observed superior performance over other full range models.

References

- [1] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20950–20959, 2023.
- [2] Peter N. Belhumeur, David W. Jacobs, David J. Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 545–552. IEEE Computer Society, 2011.
- [3] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003.
- [4] Zhiwen Cao, Zongcheng Chu, Dongfang Liu, and Yingjie Chen. A vector-based representation to enhance head pose estimation, 2020.
- [5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [6] Alejandro Cobo, Roberto Valle, José M Buenaposada, and Luis Baumela. On the representation and methodology for wide and short range head pose estimation. *Pattern Recognition*, 149:110263, 2024.
- [7] Donggen Dai, Wangkit Wong, and Zhuojun Chen. Rankpose: Learning generalised feature with rank supervision for head pose estimation. *arXiv preprint arXiv:2005.10984*, 2020.
- [8] Naina Dhingra. Lwposr: Lightweight efficient fine grained head pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, pages 1495–1505, 2022.
- [9] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *International journal of computer vision*, 101:437–458, 2013.
- [10] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

- [12] Thorsten Hempel, Ahmed A Abdelrahman, and Ayoub Al-Hamadi. Toward robust and unconstrained full range of rotation head pose estimation. *IEEE Transactions on Image Processing*, 33:2377–2387, 2024.
- [13] Sina Honari, Victor Constantin, Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Temporal representation learning on monocular videos for 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6415–6427, 2022.
- [14] Heng-Wei Hsu, Tung-Yu Wu, Sheng Wan, Wing Hung Wong, and Chen-Yi Lee. Quatnet: Quaternion-based head pose estimation with multiregression loss. *IEEE Transactions on Multimedia*, 21(4):1035–1046, 2018.
- [15] Huei-Chung Hu, Xuyang Wu, Yuan Wang, Yi Fang, and Hsin-Tai Wu. Mathematical foundation and corrections for full range head pose estimation. *arXiv preprint arXiv:2403.18104*, 2024.
- [16] Bin Huang, Renwen Chen, Wang Xu, and Qinbang Zhou. Improving head pose estimation using two-stage ensembles with top-k regression. *Image and Vision Computing*, 93:103827, 2020.
- [17] Swati Jindal and Roberto Manduchi. Contrastive representation learning for gaze estimation. In *Annual Conference on Neural Information Processing Systems*, pages 37–49. PMLR, 2023.
- [18] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [19] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.
- [20] Felix Kuhnke and Jörn Ostermann. Domain adaptation for head pose estimation using relative pose consistency. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2023.
- [21] Yuanyuan Liu, Wenbin Wang, Yibing Zhan, Shaoze Feng, Kejun Liu, and Zhe Chen. Pose-disentangled contrastive learning for self-supervised facial representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9717–9728, 2023.
- [22] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.
- [23] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maître. Xm2vtsdb: The extended m2vts database. 1999.
- [24] Erik Murphy-Chutorian and Mohan Trivedi. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31:607–26, 05 2009.
- [25] Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. Pytorch metric learning. *ArXiv*, abs/2008.09164, 2020.
- [26] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.
- [27] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pages 397–403. IEEE Computer Society, 2013.
- [28] David Schneider, Saquib Sarfraz, Alina Roitberg, and Rainer Stiefelhagen. Pose-based contrastive learning for domain agnostic activity representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3433–3443, 2022.

- [29] Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11230–11239, October 2021.
- [30] Yifan Sun, Changmao Cheng, Yuhang Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407, 2020.
- [31] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.
- [32] Yaoming Wang, Yangzhou Jiang, Jin Li, Bingbing Ni, Wenrui Dai, Chenglin Li, Hongkai Xiong, and Teng Li. Contrastive regression for domain adaptation on gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19376–19385, 2022.
- [33] Wikipedia contributors. Charts on so(3) — Wikipedia, the free encyclopedia. [https://en.wikipedia.org/w/index.php?title=Charts_on_SO\(3\)&oldid=1099134153](https://en.wikipedia.org/w/index.php?title=Charts_on_SO(3)&oldid=1099134153), 2022. [Online; accessed 21-May-2024].
- [34] Wikipedia contributors. Euler angles — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Euler_angles&oldid=1223812775, 2024. [Online; accessed 20-May-2024].
- [35] Xiaoxia Xu, Yingguo Gao, Xingjia Pan, Ke Yan, Xiaoyu Chen, and Qi Zou. Inter-image contrastive consistency for multi-person pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3063–3071, 2023.
- [36] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1087–1096, 2019.
- [37] Cheng Zhang, Hai Liu, Yongjian Deng, Bochen Xie, and Youfu Li. Tokenhpe: Learning orientation tokens for efficient head pose estimation via transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8897–8906, 2023.
- [38] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pages 386–391. IEEE Computer Society, 2013.
- [39] Huayi Zhou, Fei Jiang, and Hongtao Lu. Semi-supervised unconstrained head pose estimation in the wild. *arXiv preprint arXiv:2404.02544*, 2024.
- [40] Yijun Zhou and James Gregson. Whenet: Real-time fine-grained estimation for wide range head pose, 2020.
- [41] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 2879–2886. IEEE Computer Society, 2012.
- [42] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 146–155. IEEE Computer Society, 2016.
- [43] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

A Proof of Theorem 3.1

Proof. Suppose A and B are rotation matrices $\in SO(3)$ and $\theta \leq 90^\circ$. According to Equation 2, flipping across L_θ on A and B are $A_{flip}(\theta) = Flip_\theta \times A \times Flip_X$ and $B_{flip}(\theta) = Flip_\theta \times B \times Flip_X$. From Equation 3, $tr(A \times B^T)$ is the unique factor that can alter the geodesic distance. To show that the geodesic distance between $A_{flip}(\theta)$ and $B_{flip}(\theta)$ equals to the geodesic distance between A and B , it suffices to show the equality $tr(A \times B^T) = tr(A_{flip}(\theta) \times B_{flip}(\theta)^T)$ holds. Let's consider $A_{flip}(\theta) \times B_{flip}(\theta)^T$ first.

$$\begin{aligned}
& A_{flip}(\theta) \times B_{flip}(\theta)^T \\
&= (Flip_\theta \times A \times Flip_X) \times (Flip_\theta \times B \times Flip_X)^T \\
&= Flip_\theta \times A \times Flip_X \times Flip_X^T \times B^T \times (Flip_\theta)^T \\
&= Flip_\theta \times A \times (Flip_X \times Flip_X^T) \times B^T \times (Flip_\theta)^T \\
&= Flip_\theta \times A \times B^T \times (Flip_\theta)^T
\end{aligned} \tag{5}$$

Next, the commutativity of the trace of a matrix guarantees the following:

$$\begin{aligned}
tr(A_{flip}(\theta) \times B_{flip}(\theta)^T) &= tr(Flip_\theta \times A \times B^T \times (Flip_\theta)^T) \\
&= tr(A \times B^T \times (Flip_\theta \times Flip_\theta^T)) \\
&= tr(A \times B^T).
\end{aligned} \tag{6}$$

Therefore, $d(A, B) = d(A_{flip}(\theta), B_{flip}(\theta))$. Hence, flipping preserves the geodesic distance. Similarly, we can apply Equation 1 to prove the equality $tr(A, B) = tr(A_{rotated}(\phi), B_{rotated}(\phi))$ holds for the rotations associated with rotating an image by an angle ϕ based on A and B , so rotation preserves the geodesic distance as well. \square

B Examples for Anchor-Positive Synthetic Image Generation

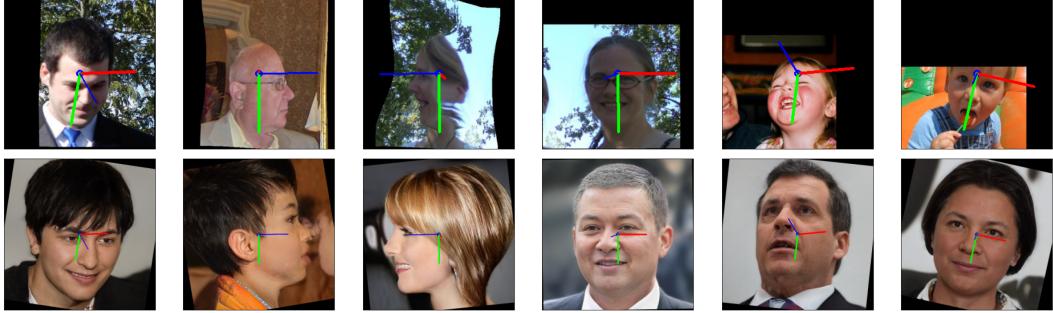


Figure 6: The first row shows the rotation matrices of 300W-LP dataset images. The second row shows that we can create synthetic images of the same head poses as the first row. In a contrastive learning scenario, the first row are the anchors and the second row are the anchor-positives while anchor-negatives can be selected easily due to the sparsity of head poses.