Semi-Supervised Unconstrained Head Pose Estimation in the Wild

Huayi Zhou, *Member, IEEE*, Fei Jiang, Jin Yuan,
Yong Rui, *Fellow, IEEE*, Hongtao Lu, *Member, IEEE*, and Kui Jia, *Member, IEEE*

Abstract—Existing research on unconstrained in-the-wild head pose estimation suffers from the flaws of its datasets, which consist of either numerous samples by non-realistic synthesis or constrained collection, or small-scale natural images yet with plausible manual annotations. This makes fully-supervised solutions compromised due to the reliance on generous labels. To alleviate it, we propose the first semi-supervised unconstrained head pose estimation method SemiUHPE, which can leverage abundant easily available unlabeled head images. Technically, we choose semi-supervised rotation regression and adapt it to the error-sensitive and label-scarce problem of unconstrained head pose. Our method is based on the observation that the aspect-ratio invariant cropping of wild heads is superior to previous landmark-based affine alignment given that landmarks of unconstrained human heads are usually unavailable, especially for underexplored non-frontal heads. Instead of using a pre-fixed threshold to filter out pseudo labeled heads, we propose dynamic entropy based filtering to adaptively remove unlabeled outliers as training progresses by updating the threshold in multiple stages. We then revisit the design of weak-strong augmentations and improve it by devising two novel head-oriented strong augmentations, termed pose-irrelevant cut-occlusion and pose-altering rotation consistency respectively. Extensive experiments and ablation studies show that SemiUHPE outperforms its counterparts greatly on public benchmarks under both the front-range and full-range settings. Furthermore, our proposed method is also beneficial for solving other closely related problems, including generic object rotation regression and 3D head reconstruction, demonstrating good versatility and extensibility. Code is in https://github.com/hnuzhy/SemiUHPE.

Index Terms—Head pose estimation, semi-supervised learning, pseudo-label filtering, unsupervised data augmentations

1 Introduction

Human head pose estimation (HPE) from a single RGB image in the wild is a long-standing yet still challenging problem [1], [2]. It has numerous applications such as driver monitoring [3], classroom observation [4], eye-gaze proxy [5], [6], and human intentions detection in social robots [7], [8]. Meanwhile, HPE can also serve as a crucial auxiliary to facilitate other face or head related multi-tasks (*e.g.*, landmark localization [9], [10], face alignment [11]–[14] and face shape reconstruction [15], [16]).

In the era of deep supervised learning, many HPE methods [17]–[20] need a large amount of labeled data to train. However, existing HPE datasets such as 300W-LP [11], BIWI [21] and DAD-3DHeads [16] have their incompatible limitations for real applications. They are either artificially collected or synthesized [11], [21]–[23] so that having huge domain gaps and scarce diversities compared to natural heads. Others are manually annotated by certified experts [16], [24] at a significant time and economic cost with a small scale. Some examples are shown in Fig. 2. In these labeled datasets, 300W-LP [11] covers yaw angles within $(-99°, 99°)$ and BIWI [21] within $(-75°, 75°)$, both focusing mainly on front-range visible faces. In contrast, DAD-3DHeads [16] officially reports 39% front, 52% side, and 9% atypical poses, providing broader coverage but still limited representation of backward and invisible heads. These statistics confirm that unconstrained, full-range head pose estimation remains largely underexplored.

Affected by this, although many supervised learning methods designed and trained based on these datasets have achieved excellent quantitative performance on the test set (such as AFLW2000 [11] and BIWI [21]), these models cannot be directly applied to the real world with complex scenarios. Some methods try to synthesize a large number of multi-view [25], extreme-view [26] or full-view [27] human heads to expand the training set, but these generated images still have hallucination defects and cannot be fully trusted. Different from most existing fully supervised methods, we turn to the semi-supervised learning (SSL) techniques [28]–[31], and propose a semi-supervised unconstrained head pose estimation (SemiUHPE) method, that can leverage extensive easier obtainable yet unlabeled in-the-wild heads [32]–[35] in addition to partially labeled data to improve performance and generalization. The overall framework of our method is illustrated in Fig. 3. It can not only avoid the laborious annotation of 3D head pose on 2D images which itself is ill-posed, but also greatly promote the estimation accuracy of challenging cases in real environments.

Broadly speaking, our work is inspired by the recent proposed semi-supervised rotation regression method FisherMatch [36], [37] focusing on generic objects. FisherMatch inherits two widely used paradigms in semi-supervised classification: the Mean-Teacher [28] framework and pseudo label filtering devised by FixMatch [30]. Its main contribu-

H. Zhou and K. Jia (the corresponding author) are with School of Data Science, The Chinese University of Hong Kong, Shenzhen (e-mail: zhouhuayi@cuhk.edu.cn; kuijia@cuhk.edu.cn). F. Jiang is with Chongqing Academy of Science and Technology (e-mail: fjiang@mail.ecnu.edu.cn). J. Yuna and Y, Rui are with Lenovo Research and Technology (e-mail: yuanjin@seu.edu.cn; yongrui@lenovo.com). H. Lu is with Department of Computer Science and Engineering, Shanghai Jiao Tong University (e-mail: htlu@sjtu.edu.cn).
Manuscript received February 20, 2025; revised September 20, 2025.

Fig. 1. Our unconstrained head pose estimation results on wild challenging heads (*e.g.*, heavy blur, extreme illumination, severe occlusion, atypical pose, and invisible face). Images are all selected from the COCO [32] dataset without head pose labels.

Fig. 2. Examples of front-range datasets 300W-LP [11] (*top*) having synthesized profile faces with many obvious artifacts and BIWI [21] (*middle*) collected in lab environments with only 24 sequences and very limited diversity, and full-range DAD-3DHeads [16] (*bottom*) with laboriously annotated 3D head mesh labels on 2D images.

tion is to use the entropy of prediction with the matrix Fisher distribution [38], [39] as a measure for filtering pseudo labels, which resembles the classification confidence and enables it to handle semi-supervised rotation regression similar to FixMatch. Following this, we focus on the head pose estimation, which is a typical case of the general object rotation regression task. It should be emphasized that this is a quite non-trivial problem. Because comparing to general objects, the HPE task often requires a more sophisticated understanding of 3D spatial relationships and precise numerical output of yaw, pitch, and roll Euler angles. Besides, many less-explored challenging frontal faces and never-touched backwards heads cannot be well-solved by FisherMatch or any other alternative so far. Please refer Fig. 1, Fig. 12 and Fig. 13 for a cursory glance.

To this end, we aim to address unconstrained head pose estimation (UHPE) of omnidirectionality, including heads with common front-range angles and face-invisible back-range angles. We believe that only by solving the problem at any head orientation can we build fundamental HPE algorithms to support the prosperity and progress of downstream applications. Specifically, to tackle the semi-supervised UHPE task, we mainly propose the following three strategies for further improvements:

(1) Aspect-Ratio Invariant Cropping. We observe that previous methods [17]–[20] require aligned faces as inputs, which relies on pre-annotated landmarks. However, this does not apply to our task as back-range heads cannot be aligned and unlabeled data has no landmark labels. Moreover, face alignment during training can cause the affine deformation, thereby hindering the inference on natural heads. Therefore, we recommend using the landmark-free pre-processing of head cropping to maintain aspect-ratio invariant and enhance practical generalizability.

(2) Dynamic Entropy-based Filtering. A key design of FixMatch is the confidence-based pseudo label filtering. FisherMatch develops the prediction entropy-based version for rotation regression. Although a pre-fixed threshold is validated to be effective, dynamic thresholds often make more sense in SSL. Such as adoption of curriculum pseudo labeling [40], label grouping [41] and adaptive threshold [42]. For our task, due to the intermixing of hard and noisy samples in unlabeled wild heads, we consider that gradually updating the threshold as the training converges is a better and more general choice.

(3) Head-Oriented Strong Augmentations. Another key idea of FixMatch is the weak-strong paired augmentations, which feeds the teacher model by weakly augmented unlabeled inputs to guide the student model fed by the same unlabeled yet strongly augmented inputs. Based on it, some SSL methods for detection [43], segmentation [44] and keypoints [45], [46] found that an advanced strong augmentation is quite important. Similarly, we re-examine the properties of unlabeled heads and invent two novel superior augmentations: pose-irrelevant cut-occlusion and pose-altering rotation consistency.

The three above-mentioned strategies are complementary: cropping guarantees a stable and scale-invariant input, dynamic filtering ensures the reliability of semi-supervised training, and augmentations promote robustness against extreme pose and occlusion variations. Their synergy enables our framework to consistently improve accuracy and generalization in unconstrained HPE. In method section, we provide more details of them. We also present how these strategies can be utilized to solve the task generic object rotation regression similar to HPE and help improve another task 3D head reconstruction strongly related to HPE. Despite the simplicity, these seamless adaptations reveal valuable universality and scalability of SemiUHPE. In experiments, we demonstrate the effectiveness and versatility of our proposed strategies by extensive comparison and ablation studies. In addition, we give impressive qualitative estimation results on challenging wild heads, which is promising for downstream applications.

To sum up, we mainly have five contributions: (1) The semi-supervised unrestricted head pose estimation (SemiUHPE) task for wild RGB images is proposed for the

first time. (2) A well-performed framework including novel customized strategies for tackling the SemiUHPE problem is proposed. (3) Benchmarks with many baseline methods for the SemiUHPE are constructed. (4) New SOTA results are achieved under both front-range and full-range HPE settings. (5) Our proposed strategies are verified to be effortlessly applicable to promote two other tasks similar or relevant to unconstrained HPE.

Our content is organized as follows: Firstly, in Section 2, we present related works including various 2D head pose estimation methods, the popular semi-supervised learning algorithms, and some semi-supervised rotation regression researches about general objects. Then, in Section 3, we give a definition of the problem, the description of overall SemiUHPE framework, motivations, explanations of proposed strategies and adaptations for other tasks. After that, in Section 4, we show multiple datasets and settings required for training, corresponding implementation details, quantitative result comparisons under various settings, complete ablation studies, qualitative visualization results and other optional setup attempts. Finally, in Section 5, we summarize findings of this paper and provide an in-depth discussion of some future directions.

2 Related Work

2.1 RGB-based Head Pose Estimation

Human head pose estimation (HPE) using monocular RGB images is a widely researched field [1], [2]. Benefiting from deep CNN, data-driven supervised learning methods tend to dominate this field. Basically, we can divide them into four categories based on landmarks [11], [47]–[50], Euler angles [17]–[19], [24], [51]–[53], vectors [14], [20], [54]–[58] or 3D Morphable Model (3DMM) [10]–[12], [16], [59]–[61]. Euler angles-based methods are essentially hindered by the *gimbal lock*. Vectors-based methods using representations such as unit quaternion [54], rotation vector [14] and rotation matrix [20], [56], [58] can alleviate this drawback and allow full-range predictions. The 3DMM-based methods treat HPE as a sub-task with optimizing multi-tasks when doing 3D face reconstruction, which currently keep SOTA performance on both front-range [10], [61] and full-range [16] HPE. In this paper, we focus on full-range unconstrained HPE and choose rotation matrix as the pose representation.

2.2 Semi-Supervised Learning (SSL)

The SSL aims to improve models by exploiting a small-scale labeled data and a large-scale unlabeled data. It can be categorized into pseudo-label (PL) based [30], [41], [42], [62]–[64] and consistency-based [28], [29], [31], [40], [65], [66]. The PL-based method selects unlabeled images into the training data iteratively by utilizing suitable thresholds to filter out uncertain samples with low-confidence. While, the consistency-based method enforces outputs or intermediate features to be consistent when the input is randomly perturbed. For example, MixMatch [29] combines consistency regularization with entropy minimization to obtain confident predictions. Based on MixMatch, SimPLE [66] exploits similar high confidence pseudo labels. FixMatch [30] integrates both pseudo label filtering and weak-to-strong augmentation consistency. FlexMatch [40], CCSSL [67] and FullMatch [68] extend FixMatch by adopting the curriculum pseudo labeling, contrastive learning and usage of all unlabeled data, respectively. We follow the empirically powerful FixMatch family [40], [42], [67], [68] to propose novel solutions for tackling HPE.

2.3 Semi-Supervised Rotation Regression

This is a less-studied field compared with other popular SSL tasks such as classification and detection. For the 6D pose estimation, Self6D [69], [70] establishes self-supervision for 6D pose by enforcing visual and geometric consistencies on top of unlabeled RGB-D images. NVSM [71] builds a category-level 3D cuboid mesh for estimating pose of rigid object in a synthesis-and-matching way. Recently, based on FixMatch [30], FisherMatch [36] firstly proposes the semi-supervised rotation regression for generic objects. It can automatically learn uncertainties along with predictions by introducing the matrix Fisher distribution [38], [39] to build a probabilistic model of rotation. The entropy of this distribution has been validated to be an efficient indicator of prediction for pseudo label filtering. The improved version [37] has integrated the rotation Laplace distribution [72] which is more robust to the disturbance of outliers and enforces much gradient to the low-error region. Essentially, HPE is subordinate to the rotation regression problem. MFDNet [56] also utilizes the matrix Fisher distribution to model head rotation uncertainty. Kuhnke et al. [73], [74] introduce the relative pose consistency into semi-supervised constrained HPE on dataset BIWI, while we explicitly target the more challenging unconstrained in-the-wild setting with novel cropping, dynamic filtering, and head-oriented augmentations. We follow [36], [56] and adopt this distribution to tackle the SemiUHPE task.

3 Method

Problem Definition. Our SemiUHPE aims to utilize a small set of head images with pose labels $\mathcal{D}^l = \{(\mathbf{x}_i^l, \mathbf{y}_i^l)\}_{i=1}^{N_l}$ and fully explore a large set of unlabeled head images $\mathcal{D}^u = \{\mathbf{x}_i^u\}_{i=1}^{N_u}$. Here, $\mathbf{x}^l$ and $\mathbf{x}^u$ represent labeled and unlabeled RGB images respectively, and $\mathbf{y}_i^l$ is the ground-truth head pose label of $\mathbf{x}^l$ such as a rotation matrix or three Euler angles. The N_l and N_u are the number of labeled and unlabeled head images, respectively. Usually, the labeled set $\mathcal{D}^l$ contains either many accurate annotations yet non-photorealistic images (*e.g.*, 300W-LP [11]), or laboriously hand-annotated labels but limited images (*e.g.*, DAD-3DHeads [16]). The unlabeled set $\mathcal{D}^u$ has much more realistic wild heads with spontaneous expressions and diversified properties, such as COCO [32]. In short, for those challenging heads in the unlabeled set which is defined as $\mathcal{D}^u = \{\mathcal{D}_{id}^u \bigcup \mathcal{D}_{ood}^u\}$, we need to exploit the in-distribution valuable portion $\mathcal{D}_{id}^u$ that is less-explored, and avoid negative impact of the noisy out-of-distribution portion $\mathcal{D}_{ood}^u$.

Clarification of $\mathcal{D}_{id}^u$ and $\mathcal{D}_{ood}^u$. In this work, the terms id and ood are not intended to strictly follow the classical notions of covariate or distributional shift, but rather serve as task-specific operational definitions. Concretely, $\mathcal{D}_{id}^u$ denotes samples that, despite possibly having large head rotations, still provide sufficient discriminative cues (*e.g.*,, side-view

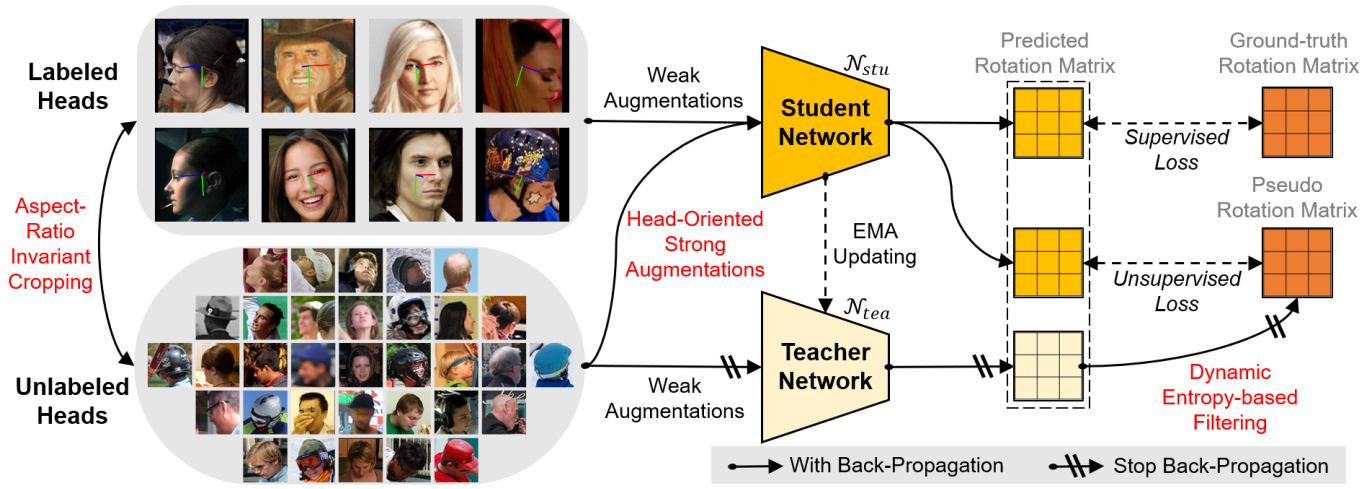


Fig. 3. The framework illustration of our SemiUHPE. We leverage small-scale labeled heads and large-scale unlabeled wild heads to optimize the teacher-student mutual learning Mean-Teacher framework. Three customized strategies are marked with red color. We finally keep the student model which is more efficient and robust for HPE evaluation.

heads with visible contours or head-top structure) such that the network prediction confidence remains relatively high. These heads are valuable for improving robustness although they are rarely covered by existing labeled datasets. By contrast, $\mathcal{D}^u_{ood}$ corresponds to extremely ambiguous cases with multiple sources of visual degradation (*e.g.*, backward heads with heavy occlusion or blur), where even humans may struggle to infer orientation. Such cases are regarded as noisy and are progressively filtered by our dynamic entropy-based strategy.

3.1 Framework Overview

As shown in Fig. 3, we adopt the Mean-Teacher [28] in our overall framework. The teacher model $\mathcal{N}_{tea}$ is the exponential moving average (EMA) of the student model $\mathcal{N}_{stu}$. While, the $\mathcal{N}_{stu}$ is supervisedly trained by labeled data, and also unsupervisedly guided by pseudo labels of unlabeled data predicted by the $\mathcal{N}_{tea}$. In this way, two models are enforced by the history consistency. Then, FixMatch [30] extends it by combining weak-strong augmentations and pseudo label filtering for classification. For rotation regression, FisherMatch [36] follows FixMatch and utilizes the *entropy* of predicted *matrix Fisher distribution* as a measure for pseudo label filtering. We briefly review this probabilistic rotation distribution below.

Matrix Fisher Distribution $\mathcal{MF}(\mathbf{R};\mathbf{A})$ This is a favorable probabilistic modeling of deep rotation estimation, which has a bounded gradient [39], [56] and intrinsic advantage than its counterpart Bingham distribution [38]. Specifically, it is a distribution over $\mathcal{SO}(3)$ with probability density function as:

$$p(\mathbf{R}) = \mathcal{MF}(\mathbf{R};\mathbf{A}) = \frac{1}{F(\mathbf{A})}\exp(\mathrm{tr}(\mathbf{A}^T\mathbf{R})) \tag{1}$$

where $\mathbf{A}\in\mathbb{R}^{3\times3}$ is an arbitrary 3×3 matrix and $F(\mathbf{A})$ is the normalizing constant. Then, the mode $\mathbf{R}$ and dispersion $\mathbf{S}$ of the distribution are computed as:

$$\mathbf{R} = \mathbf{U}\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & \det(\mathbf{UV})\end{bmatrix}\mathbf{V}^T \tag{2}$$

where $\mathbf{U}$ and $\mathbf{V}$ are from the singular value decomposition (SVD) of $\mathbf{A} = \mathbf{USV}^T$. Each singular value s_i in $\mathbf{S} = \mathrm{diag}(s_1, s_2, s_3)$ is sorted in descending order, and indicates the concentration strength.

Entropy-based Confidence Measure During training, the network regressor $\mathcal{N}$ takes a single RGB image $\mathbf{x}$ as input and outputs a 3×3 matrix $\mathbf{A}_f = \mathcal{N}(\mathbf{x})$. The predicted $\mathbf{A}_f$ is a matrix Fisher distribution $f \sim \mathcal{MF}(\mathbf{A}_f)$. It contains a predicted rotation and the information of concentration by computing mode $\mathbf{R}_f$ and dispersion $\mathbf{S}_f$ as in Eq. 2, respectively. The *entropy* of this prediction, used as confidence measure of uncertainty, can be computed as:

$$H(f) = \log F_f - \sum\nolimits_{i=1}^{4}\left(z_{f_i}\frac{1}{F_f}\frac{\partial F_f}{\partial z_{f_i}}\right) \tag{3}$$

where F_f is constant wrt. parameter $\mathbf{Z} = \mathrm{diag}(0, z_1, z_2, z_3)$. And $\mathbf{Z}$ is a 4×4 diagonal matrix with $0 \geq z_1 \geq z_2 \geq z_3$. The element z_i is from the corresponding unit quaternion $\mathbf{q}\in\mathcal{S}^3$. Assume $\mathbf{A}_f = \mathbf{U}_f\mathbf{S}_f\mathbf{V}_f^T$, γ is the standard transform from unit quaternion to rotation matrix, $\mathbf{e}_i$ is the i-th column of $\mathbf{I}_4$, and $\mathbf{E}_i = \gamma(\mathbf{e}_i)$, then z_{f_i} is the trace of $\mathbf{E}_i^T\mathbf{S}_i$. More details of the derivation are in [36], [39]. Generally, a lower entropy indicates a more peaked distribution that means less uncertainty and higher confidence. Next, we discuss how to optimize the primary entropy-based filtering strategy for SemiUHPE, as well as design stronger augmentations for unlabeled images $\mathbf{x}^u$.

3.2 Aspect-Ratio Invariant Cropping

For input preprocessing, we call for keeping the aspect-ratio invariant by loosely cropping head-centered images with bounding boxes and padding the out-of-plane area with zero. The reasons are two-fold. Firstly, the naive cropping-resizing may lead to scaling-related flaws [39] such as perceived orientation change. Secondly, existing HPE methods [17]–[20] take it for grant that aligning face with landmarks will bring better results like face recognition [75], [76]. However, this often introduces severe affine deformation that disrupts the natural face (refer Fig. 4). Moreover, face landmarks are not applicable to back-range or wild-collected heads. We will verify the necessity and advantages of aspect-ratio invariant cropping in experiments.

Fig. 4. The illustration of how a naive cropping-resizing leads to perceived orientation and affine deformation.

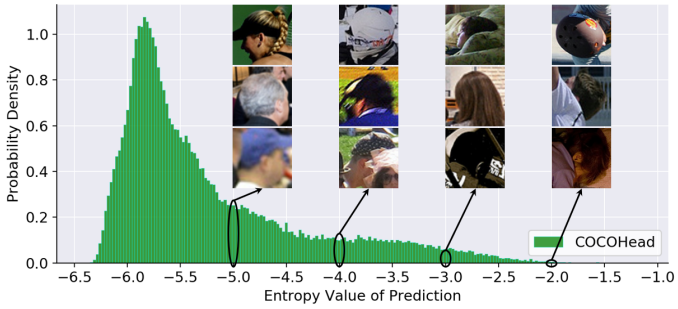


Fig. 5. The illustration of prediction entropies corresponding to their head samples in the unlabeled dataset (*e.g.* COCOHead).

3.3 Dynamic Entropy-based Filtering

In original FisherMatch, it adopts a trivial entropy-based filtering rule on unlabeled data which keeps the prediction as a pseudo label when its entropy is lower than a pre-fixed threshold τ. The obtained unsupervised loss is:

$$\mathcal{L}_{unsup}(\mathbf{x}^u) = \mathbb{1}_{(H(p_{tea})\leq\tau)}\mathcal{L}^{CE}(p_{tea}, p_{stu}) \tag{4}$$

where $\mathbb{1}_{(\cdot)}$ is the indicator function, being 1 if the condition holds and 0 otherwise. $H(p_{tea})$ is the prediction entropy computed as in Eq. 3. $\mathcal{L}^{CE}(\cdot,\cdot)$ is the cross entropy loss to enforce consistency between two continuous matrix Fisher distributions [39]. The p_{tea} and p_{stu} are denoted as $p_{tea} = \mathcal{MF}(\mathbf{A}^u_{tea})$ and $p_{stu} = \mathcal{MF}(\mathbf{A}^u_{stu})$ which are outputted by the teacher model $\mathbf{A}^u_{tea} = \mathcal{N}_{tea}(\mathbf{x}^u)$ and the student model $\mathbf{A}^u_{stu} = \mathcal{N}_{stu}(\mathbf{x}^u)$, respectively.

For our task, the unlabeled set $\mathcal{D}^u$ has many challenging heads. It is quite difficult to distinguish whether a sample in $\mathcal{D}^u$ belongs to $\mathcal{D}^u_{id}$ or $\mathcal{D}^u_{ood}$ for the teacher model through a fixed threshold. For example, as shown in Fig. 5, we calculated the predicted entropies of the teacher model $\mathcal{N}_{tea}$ for all samples in $\mathcal{D}^u$. The $\mathcal{N}_{tea}$ has quite certain predictions for most samples (lower entropies). While, samples with high uncertainty (higher entropies) are divided into two types: hard heads still in $\mathcal{D}^u_{id}$ or noisy heads in $\mathcal{D}^u_{ood}$. The former includes cases with severe occlusion or atypical pose which are infrequent in $\mathcal{D}^l$ yet possible to be correctly predicted. The latter contains extreme noisy cases with unrecognizable pose due to insufficient context or incorrect category. Moreover, the predictive ability of $\mathcal{N}_{tea}$ improves with the deepening of training, which means the difficulty and uncertainty of the same sample for $\mathcal{N}_{tea}$ is also changing.

Therefore, we propose the dynamic entropy-based filtering to improve the pseudo-label quality and enhance the model's robustness in real-world. Specifically, with the assumption that $\mathcal{D}^u = \{\mathcal{D}^u_{id} \bigcup \mathcal{D}^u_{ood}\}$, we choose to retain a portion of unlabeled data in $\mathcal{D}^u$ for unsupervised training. For each mini-batch inputs, we still need a concrete threshold τ_k to filter predictions, where τ_k is progressively updated throughout K stages. The τ_t is calculated as:

$$\tau_k = \text{percentile}\langle H(\mathcal{MF}(\mathcal{N}^k_{tea}(\mathbf{x}^u_i)))|^{N_u}_{i=1}, \delta\rangle \tag{5}$$

where δ is the percentage of remained unlabeled data, and closely related to the unknown $\mathcal{D}^u_{ood}$ in $\mathcal{D}^u$. $\text{percentile}\langle\cdot,\cdot\rangle$ returns the value of δ^{th} percentile. $\mathcal{N}^k_{tea}$ is the teacher model in k-th stage ($k \in \{1, 2, ..., K\}$). Then, we revise Eq. 4 as:

$$\mathcal{L}'_{unsup}(\mathbf{x}^u) = \mathbb{1}_{(H(p^k_{tea})\leq\tau_k)}\mathcal{L}^{CE}(p^k_{tea}, p_{stu}) \tag{6}$$

where $p^k_{tea} = \mathcal{MF}(\mathcal{N}^k_{tea}(\mathbf{x}^u))$. Usually, with a given percentage δ, the dynamic entropy threshold τ_k will decrease as the stage k increases. And the optimal value of δ is inversely proportional to the images number of $\mathcal{D}^u_{ood}$ in $\mathcal{D}^u$. It should be emphasized that in our context, the distinction between $\mathcal{D}^u_{id}$ and $\mathcal{D}^u_{ood}$ reflects the difficulty and reliability of pose inference rather than classical covariate shift, which helps the dynamic threshold adaptively retain plausible but hard samples while filtering extremely noisy ones. Representative $\mathcal{D}^u_{id}$ and $\mathcal{D}^u_{ood}$ examples under our definition are visualized in Figs. 5, 11, 12, 14 to illustrate the filtering effect in practice.

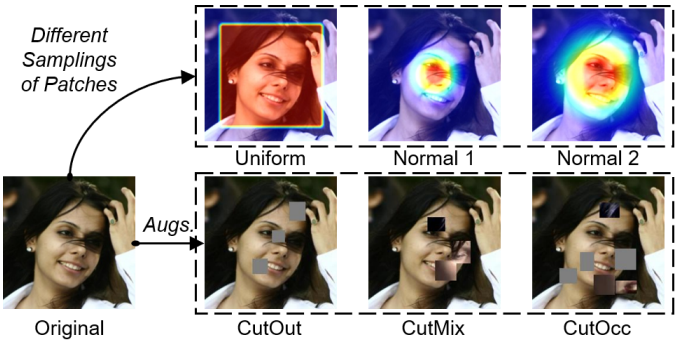


Fig. 6. The illustration of patches sampling with different distributions, and the novel pose-irrelevant cut-occlusion (CutOcc) augmentation.

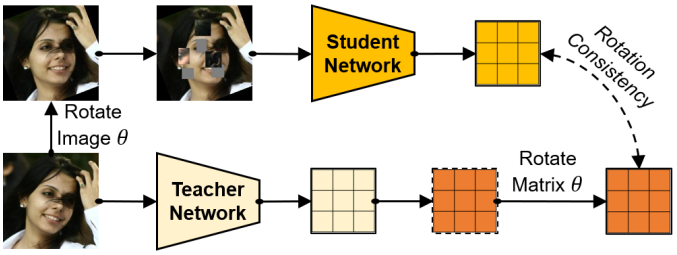


Fig. 7. The illustration of proposed pose-altering rotation consistency (RotCons) augmentation. CutMix-related patches are fetched from other samples in the same mini-batch.

3.4 Head-Oriented Strong Augmentations

FisherMatch follows the original weak-strong augmentations in FixMatch, and defines both the weak augmentation T_{weak} and strong augmentation T_{strong} as random cropping-resizing only with different scale factors. Empirically, many SSL methods [31], [45], [46] have found that the core of weak-strong augmentations paradigm in FixMatch is a more advanced strong augmentation than T_{strong}. We thus propose two novel strong augmentations for unlabeled heads.

3.4.1 *Pose-Irrelevant Cut-Occlusion*

Firstly, we explore pose-irrelevant augmentations such as popular CutOut [77], Mixup [78] and CutMix [79]. CutOut

simulates random occlusion. Mixup combines global features in different samples. CutMix balances both occlusion and crossed features. Considering that self- or emerged-occlusion is common in wild heads, we extend the CutOut and CutMix by sampling target patches with head-centered distributions. Heuristically, we provide three proposals of sampling distribution $\mathcal{S}$: Uniform distribution with a small distance from the boundary ($\mathcal{S}_{\text{Uniform}}$), Normal distribution with a smaller variance ($\mathcal{S}_{\text{Normal1}}$) and a larger variance ($\mathcal{S}_{\text{Normal2}}$). Among them, we verified in experiments that the $\mathcal{S}_{\text{Normal2}}$ is superior for its stronger head-centered concentration of noise addition. Then, we propose to conduct CutOut and CutMix in sequence to obtain an advanced combination named Cut-Occlusion (CutOcc). The motivation is to leverage the synergistic effect between two existing components. As shown in Fig. 6, CutOcc is also visually understandable.

3.4.2 Pose-Altering Rotation Consistency

Although HPE is sensitive to rotation, we can still perform in-plane rotation augmentation in $\mathcal{SO}(3)$ similar to SSL keypoints detection [45], [46], [80]. For a batch of unlabeled heads $\{\mathbf{x}_i^u\}_{i=1}^{B_u}$, we present unsupervised rotation consistency augmentation as shown in Fig. 7. On one hand, we rotate each $\mathbf{x}_i^u$ with a random angle θ from $(-30°, 30°)$, with selectively conducting the subsequent pose-irrelevant operation CutOcc. The strongly augmented $\overline{\mathbf{x}^u}$ are then fed into the student model $\mathcal{N}_{stu}$, which outputs parameters in the form of matrix $\overline{p_{stu}^i} \in \mathbb{R}^{3\times3}$. On the other hand, we directly feed weakly augmented $\widetilde{\mathbf{x}^u}$ into the teacher model $\mathcal{N}_{tea}$, and obtain the predicted matrix $\widetilde{p_{tea}^i}$. Then, we need to rotate $\widetilde{p_{tea}^i}$ around the Z-axis with corresponding degree for consistency training. We summarize these steps as follows.

$$
\begin{aligned}
&\overline{p_{stu}^i} = \mathcal{MF}(\mathcal{N}_{stu}(\overline{\mathbf{x}_i^u})), \quad \overline{\mathbf{x}_i^u} = T_{\text{CutOcc}}(T_{\text{Rot}_\theta}(\mathbf{x}_i^u)) \\
&\widetilde{p_{tea}^i} = \mathcal{MF}(\mathcal{N}_{tea}(\widetilde{\mathbf{x}_i^u})), \quad \widetilde{\mathbf{x}_i^u} = T_{\text{weak}}(\mathbf{x}_i^u) \\
&\widehat{p_{tea}^i} = \mathbf{M}_\theta \widetilde{p_{tea}^i}, \quad \mathbf{M}_\theta = \begin{bmatrix} \cos(\theta) & \sin(\theta) & 0 \\ -\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}
\end{aligned} \tag{7}
$$

where T_{CutOcc} and T_{Rot_θ} means our proposed strong augmentations CutOcc and in-plane rotation, respectively. $\mathbf{M}_\theta$ represents the in-plane rotation matrix corresponding to θ. And $\widehat{p_{tea}^i}$ is the aligned prediction. We finally enforce consistency between distributions $\widehat{p_{tea}^i}$ and $\overline{p_{stu}^i}$.

It is worth noting that Kuhnke et al. [73], [74] also employed rotation-based consistency within semi-supervised HPE, but their methods were designed for constrained benchmarks (e.g., BIWI), whereas our proposed T_{CutOcc} and T_{Rot_θ} are specifically tailored for unconstrained in-the-wild heads with severe occlusion and atypical orientations.

3.5 Adaptation of SemiUHPE

As discussed above, HPE is essentially a 3D rotation regression task. Therefore, SemiUHPE can be applied to any type of object without modification, and is roughly equivalent to a *semi-supervised generic object rotation regression* (SemiObjRot) framework. In experiments, we will quantitatively validate that SemiObjRot is superior to those specially designed methods [36], [37] for regressing the rotation of common objects. Next, we explain how to simply adapt SemiUHPE so that it can be used to address the *semi-supervised 3D head reconstruction* (Semi3DHead) task.

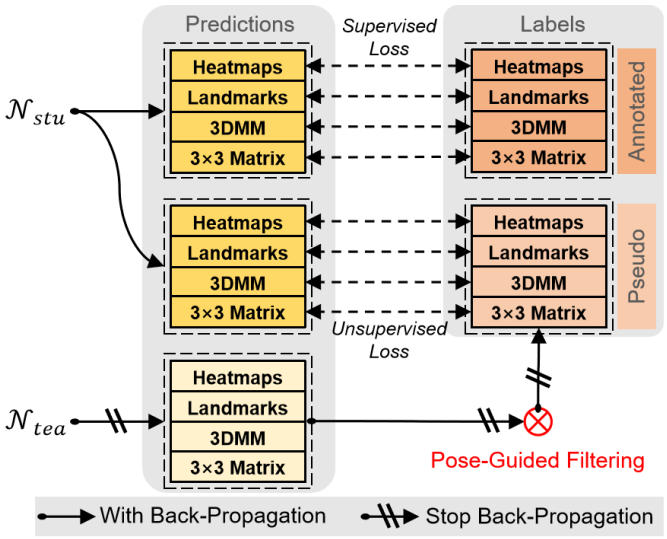


Fig. 8. Overall architecture of our Semi3DHead framework. It extends DAD-3DNet [16] by incorporating multiple input branches: (i) heatmaps and landmarks inherited from DAD-3DNet, (ii) a 3D Morphable Model (3DMM) branch using the FLAME model [81], and (iii) an additional 3×3 rotation matrix branch introduced in this work to directly regress head pose and filter out out-of-distribution samples.

Specifically, current mainstream 3D head reconstruction methods [12], [16], [82] rely on the 3D Morphable Model (3DMM) (*e.g.*, BFM [83] and FLAME [81]) to transform it into a concise regression problem, which mainly predicts head shape and pose parameters. Then, motivated by the success of HPE using only monocular images [18], [20], [58], we decide to add one additional HPE sub-branch on the basic reconstruction network. This branch can also be optimized smoothly for predicting 3D head pose (*e.g.*, represented by the rotation matrix in $\mathcal{SO}(3)$), which is similar but independent of the head rotation in outputted 3DMM parameters. Without loss of generality, as shown in Fig. 8, we choose DAD-3DNet [16] as the basic main network and construct a modified semi-supervised version for 3D head reconstruction. It leverages small-scale labeled 3D heads and large-scale unlabeled wild heads to optimize the teacher-student mutual learning Mean-Teacher framework. Although DAD-3DNet has multiple output branches (such as 3DMM, heatmaps and landmarks), we can still follow the training idea of SemiUHPE. And we also experimentally confirmed that the newly added branch does not hinder the overall model performance.

To filter unreliable pseudo-labeled heads, we introduce a new pose regression branch and design a pose-guided filtering strategy. As illustrated in Fig. 9 (left), one straightforward way is to measure the discrepancy between the predicted rotation matrix from the head-pose branch and the rotation matrix embedded in the 3DMM branch. This discrepancy can be quantified by the *geodesic distance* on the rotation group $\mathcal{SO}(3)$, which corresponds to the minimal angular difference between two rotations (approximated by the Frobenius norm $\|\mathbf{I}-\mathbf{R}_1\mathbf{R}_2^T\|_F$). A smaller geodesic distance indicates higher consistency and reliability. However, relying solely on matrix differences may not always provide a statistically robust measure of uncertainty. To address this,

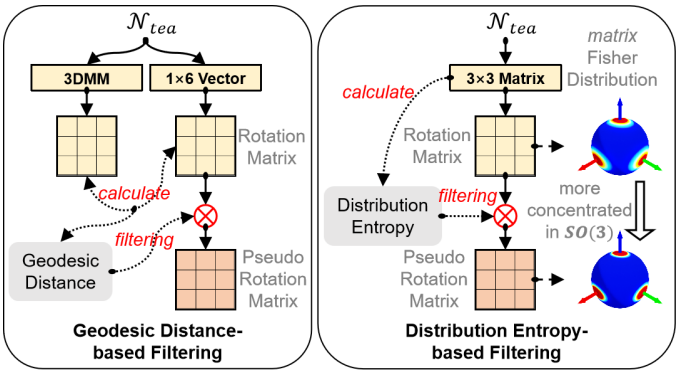


Fig. 9. The illustration of two pose-guided pseudo-label filtering strategies based on geodesic distance (left) or distribution entropy (right).

we further model the head pose prediction probabilistically by fitting it with a matrix Fisher distribution, which naturally captures the distribution of rotations on $\mathcal{SO}(3)$. In this setting, the *distribution entropy* serves as an indicator of prediction confidence: low entropy means the distribution is sharp and reliable, while high entropy corresponds to ambiguous or noisy samples. As shown in Fig. 9 (right), we use this entropy-based criterion to dynamically filter out high-uncertainty samples. There are two advantages to applying the matrix Fisher distribution. On one hand, its output is an arbitrary 3×3 matrix with homeomorphic parameters to $\mathbb{R}^9$ [39], which is easier to estimate than the 3D rotations in $\mathcal{SO}(3)$ topology [56]. On the other hand, we can measure the confidence of predicted distribution by calculating its entropy, and devise a self-reliant more robust filtering mechanism. Therefore, while different rotation representations can be mathematically transformed into each other, we adopt the 3×3 rotation matrix as the default representation. We will verify that the adapted SemiUHPE can benefit from both filtering strategies, with the latter being superior.

3.6 Training Protocol

Whether it is the original SemiUHPE or its adapted version, our training has two consecutive phases.

Phase1: We train the student model $\mathcal{N}_{stu}$ to learn a rotation regressor on labeled set $\mathcal{D}^l$ with a supervised loss:

$$\mathcal{L}_{sup}(\mathbf{x}^l,\mathbf{y}^l) = -\log(\mathcal{MF}(\mathbf{y}^l;\mathcal{N}_{stu}(\mathbf{x}^l)) \tag{8}$$

where $-\log(\cdot)$ means the negative log likelihood (NLL) of the mode predicted by $\mathcal{N}_{stu}$ in the distribution of ground-truth label $\mathbf{y}^l$. We save the best performed student model for cloning an identical teacher model for the next phase.

Phase2: After supervised training, we obtain a pair of teacher and student networks with the same initialization. Now, we begin the semi-supervised phase on both labeled set $\mathcal{D}^l$ and unlabeled set $\mathcal{D}^u$. The total loss is:

$$\mathcal{L} = \mathcal{L}_{sup}(\mathbf{x}^l,\mathbf{y}^l) + \lambda\mathcal{L}'_{unsup}(\mathbf{x}^u) \tag{9}$$

where $\mathcal{L}_{sup}$ and $\mathcal{L}'_{unsup}$ are calculated as in Eq. 8 and Eq. 6, respectively. The λ is a weight of unsupervised loss, which is set to 1 in all experiments. Besides, we usually allocate different iterations for two phases based on the complexity of labeled $\mathcal{D}^l$ and unlabeled $\mathcal{D}^u$ datasets.

4 Experiments

4.1 Datasets and Settings

We introduce two kinds of datasets for the HPE task.

Labeled Datasets: we adopt the popular benchmark 300W-LP [11] which has 122,450 samples with half flipping as the train-set and AFLW2000 [11] as the val-set for comparing with the mainstream front-range HPE methods. We also utilize a recent 3D head reconstruction dataset DAD-3DHeads [16] with three subsets (37,840 images in train-set, 4,312 images in val-set, and 2,746 images in test-set) for implementing the full-range unconstrained HPE.

Unlabeled Datasets: we utilize COCO [32] with wild human heads as the unlabeled set for boosting both the front-range and full-range HPE. COCO does not have the head box label. Thus, we utilize its variation COCO-HumanParts [84] with labeled head boxes, and preprocess it as in BPJDet [85], [86] to generate COCOHead, which has about 74K samples after removing heads smaller than 30 pixels. These left heads cover diverse scenarios and cases. Besides, we also adopt this processing paradigm to extract valid head images from other human-related datasets including WiderFace [33], CrowdHuman [34] and OpenImageV6 [31], which can be used for further ablation tests and model improvement in following subsections.

Then, we designed three increasingly difficult SSL settings along with a group of compared methods (*e.g.*, both fully supervised and semi-supervised) as described below:

Setting1: 300W-Self In this setting, we use partially annotated 300W-LP as the labeled set and the left part as the unlabeled set. The labeled ratio is selected from (2%, 5%, 10%, 20%). The Phase1 and Phase2 has 20K and 40K iterations, respectively. The number of threshold updating stages K in Phase2 is 4. Batch size for the labeled set B_l and unlabeled set B_u is 32 and 128, respectively. The remained percentage δ of unlabeled data is 0.95. The test-set is AFLW2000. For a fair comparison, we keep samples with Euler angles within $\pm90^\circ$ following the previous front-range HPE methods [18]–[20]. The evaluation metric is Mean Absolute Error (MAE) of Euler angles. We thus convert predictions into Euler angles for comparing.

Setting2: 300W-COCOHead Still for the front-range HPE, we combine all labeled 300W-LP with additional unlabeled faces in COCOHead for further performance boosting. We set Phase1 and Phase2 with 180K and 60K iterations, respectively. Parameters K, B_l, B_u and δ are set to 6, 16, 128 and 0.75, respectively. The test-set is also AFLW2000. The others are the same as **Setting1**.

Setting3: DAD-COCOHead This is for the full-range HPE task. We use the train-set of DAD-3DHeads as the labeled set, and COCOHead as the unlabeled set. We set Phase1 and Phase2 with 100K and 100K iterations, respectively. And parameters K, B_l, B_u and δ are set to 5, 16, 128 and 0.75, respectively. We report testing results on both val-set and test-set of DAD-3DHeads. Following [16], measures of the ground-truth matrix $\mathbf{R}_1$ and predicted $\mathbf{R}_2$ are (1) Frobenius norm of the matrix $\mathbf{I}-\mathbf{R}_1\mathbf{R}_2^T$, and (2) the angle in axis-angle representation of $\mathbf{R}_1\mathbf{R}_2^T$ (*a.k.a* the geodesic distance between two rotation matrices).

Compared Methods: In order to make a more comprehensive comparison, apart from our proposed SemiUHPE,

we also implemented the fully supervised version (Sup.) using only labeled data but adopting an advanced matrix Fisher representation [39], and the semi-supervised baseline FisherMatch [36] (Base.) following our settings. For the network backbone, it is selected from three purely ConvNets-based candidates including ResNet50 [87], RepVGG [88] and EfficientNetV2-S [89]. These backbones have similar network parameters and are widely used in other HPE methods [18]–[20]. Besides, we also included two additional baselines to strengthen the comparisons: (1) FisherMatch+ [37] (Base.+), where we replace the matrix Fisher distribution with a rotation Laplace distribution; (2) Sample Adaptive Augmentation [90] (SAA), an SSL method from image classification that introduces stronger adaptive augmentations. Both Base.+ and SAA were re-implemented on top of the best-performing settings of FisherMatch for SemiUHPE. For all experiments, we implemented them using PyTorch 1.13.1 on one single RTX 3090 (24 GB) or A800 (80 GB) GPU. The Adam optimizer is used. The learning rate in Phase1 is 1e-4, and reduced to 1e-5 in Phase2. During training and testing, all head images are padded and resized into shape 224×224 as inputs.

Adapted SemiUHPE: Similarly, we introduce the semi-supervised experimental data and settings for generic object rotation regression (SemiObjRot) and 3D head reconstruction (Semi3DHead). For SemiObjRot, we follow the setting in FisherMatch [36] and conduct SSL comparing tests on the dataset Pascal3D+ [91]. This dataset contains real images from Pascal VOC and ImageNet of 12 rigid object classes. We evaluate 6 vehicle categories (aeroplane, bicycle, boat, bus, car, motorbike) which have relatively evenly distributed poses in azimuth angles, and set the number of labeled images as 7, 20 and 50 for each category respectively. For Semi3DHead, we conduct SSL comparing tests on the dataset DAD-3DHeads [16], where part of the train-set is used as labeled data (*e.g.*, 5%, 10% or 20%), the rest is used as unlabeled data, and the val-set is used to report performance. The comparison methods are roughly the same as SemiUHPE.

4.2 Implementation Details

Following above-defined three SSL settings, we provide many more details of implementing our proposed SemiUHPE and compared methods.

Confidence Threshold in FisherMatch: In order to fairly compare SemiUHPE with the baseline method FisherMatch [36], we must select an appropriate pre-defined confidence threshold for it. In **Setting1**, we have determined the optimal threshold $\tau = -5.4$ by running several controlled experiments as shown in Fig. 15a. Then, in **Setting2**, we selected the final computed dynamic threshold $\tau_6 = -4.3$ of SemiUHPE (with either ResNet50 [87], RepVGG [88] or EfficientNetV2-S [89] as the backbone) as the pre-fixed threshold for FisherMatch. Similarly, in **Setting3**, we selected the final computed dynamic threshold $\tau_5 = -4.8$, $\tau_5 = -4.7$ and $\tau_5 = -4.8$ of SemiUHPE with the backbone ResNet50, RepVGG and EfficientNetV2-S as the pre-fixed threshold, respectively. As shown in Table 2, Table 3 and Table 4, the performance of implemented FisherMatch is comparable or better than fully supervised methods, which explains the correctness and effectiveness of this strong baseline. Nonetheless, our SemiUHPE can always exceed FisherMatch by a clear margin.

TABLE 1
Hyper-parameters in paired weak-strong augmentations.

Is Used?		Aug.	Parameter	Probability
Weak	Strong			
✓	✓	Flip	Horizontally	0.5
✓	✓	Blur	—	0.05
✓	✗	Scale	$s \in [0.8, 1.25]$	1.0
✗	✓	Scale	$s \in [0.6, 1.5]$	1.0
✗	✓	T_{CutOcc}	CutOut: 3 Holes CutMix: 3 Holes	1.0
✗	✓	T_{Rot_θ}	$\theta \in (-30°, 30°)$	1.0

Rotation Consistency for dataset 300W-LP: For the pose-altering rotation consistency part T_{Rot_θ} in our proposed head-oriented strong augmentation, we observed that it does not have a positive effect on semi-supervised HPE task related to the 300W-LP [11] dataset. The main factor causing this obstacle is that the 300W-LP dataset does not use an accurate 3×3 rotation matrix to represent the head pose label, but rather chooses the Euler angles with inherent flaws, including the *gimbal lock* and *discontinuity*. This further prevents us from aligning the two matrices after random rotation. Therefore, we did not apply T_{Rot_θ} in **Setting1** or **Setting2**. This defect does not exist in **Setting3** that uses the labeled set DAD-3DHeads [16] with the rotation matrix as its head pose label.

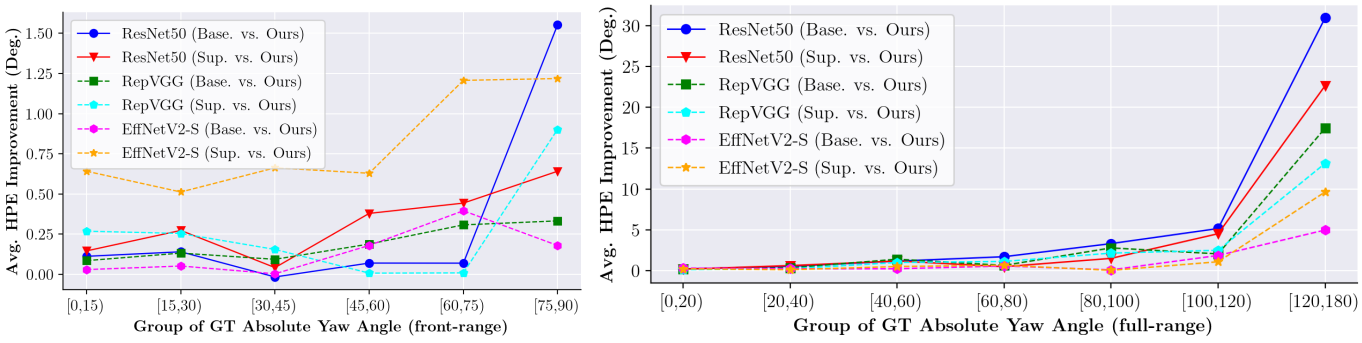
Parameters of Augmentations: Hyper-parameters related to augmentations are shown in Table 1. Specifically, Flip and Blur mean left-right flipping and simple image filter, respectively. Scale means the random resized crop. Both T_{CutOcc} and T_{Rot_θ} are proposed by us. And the number of mask holes in CutOut [77] or CutMix [79] means generated random patches. Generally, it is laborious and impractical to undergo ablation tests for searching optimal values of these parameters. Therefore, these hyper-parameters are empirically initialized.

SemiObjRot and Semi3DHead: For SemiObjRot, we inherit most of the training parameter settings in FisherMatch, except that the newly added parameter δ is set to 0.95 and strong augmentations are replaced with our proposed ones. Besides, we evaluate the experiments by the mean error, the median error (in degrees) and the accuracy within 30° between the prediction and the ground truth. For Semi3DHead, we follow DAD-3DNet and **Setting1** to conduct SSL training experiments. The evaluation metrics are updated into Reprojection NME, Z_n accuracy, Chamfer Distance and Pose Error for the 3D head learning task. More details can be found in [36] and [16].

4.3 Quantitative Comparison

Below, we report results under three settings as well as improvement analysis of our SemiUHPE. If there is no special reminder, all baselines (Sup., Base., Base.+ and SAA) are implemented with our proposed aspect-ratio invariant cropping. The last is the results of SemiObjRot and Semi3DHead.

Front-range HPE Results under Setting1. As shown in Table 2, our SemiUHPE can surpass both the supervised methods [39] and baseline FisherMatch [36] with either backbone. When using ResNet50, our SemiUHPE is always



(a) (b)

Fig. 10. The average HPE improvement on (a) AFLW2000 under **Setting2** and (b) DAD-3DHeads val-set under **Setting3**.

TABLE 2
Euler angles errors on AFLW2000. Models are trained on 300W-LP with different ratios of label. The best result is in red color. SL and SSL are fully supervised learning and semi-supervised learning, respectively.

Type	Method	Backbone	2%	5%	10%	20%	All
SL	Sup.	ResNet50	4.347	3.987	3.831	3.619	3.578
	Sup.	RepVGG	4.252	3.817	3.724	3.579	3.498
	Sup.	EffNetV2-S	4.009	3.678	3.517	3.444	3.379
SSL	Base.	ResNet50	4.190	3.798	3.609	3.492	—
	Base.	RepVGG	4.023	3.730	3.489	3.445	—
	Base.	EffNetV2-S	3.991	3.596	3.448	3.372	—
	Base.+	ResNet50	4.102	3.702	3.514	3.475	—
	SAA	ResNet50	4.038	3.654	3.508	3.469	—
	Ours	ResNet50	3.956	3.629	3.487	3.463	—
	Ours	RepVGG	3.953	3.607	3.510	3.424	—
	Ours	EffNetV2-S	3.835	3.526	3.377	3.348	—

the best out of other three SSL methods (Base., Base.+ and SSA). Meanwhile, the smaller the ratio of labeled samples used (from 20% to 2%), the more significant the reduction in MAE errors. Moreover, with using 20% labeled 300W-LP, the performance of semi-supervised baseline and our SemiUHPE can always exceed the supervised method using all labeled 300W-LP. These results are also comparable to some SOTA supervised methods in Table 3. We attribute this to the stronger robustness brought by partially integrated unsupervised training of unlabeled data. It also indicates that pure supervised learning may cause final models overfitting on the train-set, while SSL can improve the adaptability and generalization to a certain extent.

Front-range HPE Results under Setting2. As shown in Table 3, with the supporting of SSL, our method based on ResNet50 can achieve a low MAE result **3.37**, which is comparable to other supervised SOTA front-range HPE methods such as DAD-3DNet [16] with MAE **3.66**, SynergyNet [10] with MAE **3.35** and DSFNet-f [61] with MAE **3.25**. Please note that all of them are 3DMM-based which require dense face landmark labels and complex 3D face reconstruction pipelines. While, our SemiUHPE only requires reasonable excavation of unlabeled wild heads, which is more practical and scalable in real applications. Comparing to other three SSL baselines (Base., Base.+ and SSA), our method still outperforms them under this full-range setting. When using a stronger backbone such as RepVGG as in 6DRepNet [20], our method can obtain a lower result with MAE **3.35**, which further explains its generality and advantage. In particular, after adopting the superior backbone EfficientNetV2-S, we can further reduce the MAE errors into **3.31**, which achieved

TABLE 3
Euler angles errors on AFLW2000. Models are all trained on 300W-LP. Extra means additional annotations (*e.g.*, Landmarks (LMs) or 3DMM). The marker † indicates using additional labeled training data. The best and second-best result is in red and blue color, respectively.

Type	Method	Reference	Extra	Pitch	Yaw	Roll	MAE
SL	3DDFA [11]	CVPR'16	3DMM	5.98	4.33	4.30	4.87
	HopeNet [17]	CVPRW'18	No	6.56	6.47	5.44	6.16
	QuatNet [54]	TMM'18	No	5.62	3.97	3.92	4.50
	FSA-Net [18]	CVPR'19	No	6.08	4.50	4.64	5.07
	WHENet-V [19]	BMVC'20	No	5.75	4.44	4.31	4.83
	FDN [51]	AAAI'20	No	5.61	3.78	3.88	4.42
	3DDFA-V2 [12]	ECCV'20	3DMM	5.26	4.06	3.48	4.27
	MNN [13]	TPAMI'20	LMs	4.69	3.34	3.48	3.83
	Rankpose [57]	BMVC'20	No	4.75	2.99	3.25	3.66
	TriNet [55]	WACV'21	No	5.77	4.20	4.04	4.67
	MFDNet [56]	TMM'21	No	5.16	4.30	3.69	4.38
	Img2Pose [14]	CVPR'21	LMs	5.03	3.43	3.28	3.91
	SADRNet [59]†	TIP'21	3DMM	5.00	2.93	3.54	3.82
	SynergyNet [10]	3DV'21	3DMM	4.09	3.42	2.55	3.35
	6DRepNet [20]	ICIP'22	No	4.91	3.63	3.37	3.97
	DAD-3DNet [16]†	CVPR'22	3DMM	4.76	3.08	3.15	3.66
	TokenHPE [58]	CVPR'23	No	5.54	4.36	4.08	4.66
	DSFNet-f [61]†	CVPR'23	3DMM	4.28	2.65	2.82	3.25
	CIT-v1 [92]	IJCV'23	LMs	4.38	2.68	3.45	3.50
	2DHeadPose [24]†	NN'23	No	4.47	2.85	2.82	3.38
	HeadDiff [93]	TIP'24	No	4.55	3.15	3.03	3.57
	OPAL (6D) [94]	PR'24	No	4.59	2.85	3.04	3.49
	Sup. (ResNet50)	—	No	4.58	3.20	2.95	3.58
	Sup. (RepVGG)	—	No	4.50	3.18	2.81	3.50
	Sup. (EffNetV2-S)	—	No	4.40	2.99	2.75	3.38
SSL	Base. (ResNet50)	CVPR'22	No	4.61	2.99	3.00	3.53
	Base. (RepVGG)	CVPR'22	No	4.46	2.84	2.77	3.36
	Base. (EffNetV2-S)	CVPR'22	No	4.39	2.87	2.79	3.35
	Base.+ (ResNet50)	TPAMI'25	No	4.58	2.96	2.83	3.46
	SAA (ResNet50)	CVPR'23	No	4.56	2.96	2.79	3.44
	Ours (ResNet50)	—	No	4.52	2.89	2.71	3.37
	Ours (RepVGG)	—	No	4.43	2.86	2.75	3.35
	Ours (EffNetV2-S)	—	No	4.39	2.79	2.77	3.31

the second best performance so far.

Full-range HPE Results under Setting3. As shown in Table 4, results of Img2Pose [14] and DAD-3DNet [16] are obtained from the paper [16], except for results of DAD-3DNet on the val-set which are evaluated by us using its official model. The supervised method is implemented on DAD-3DHeads train-set. Generally, our SemiUHPE significantly surpasses the compared DAD-3DNet and retrained baselines without using any 3D information of face or head, which again proves its superiority and versatility. By expanding the unsupervised dataset (*e.g.*, only COCOHead) to a larger one, which is the combination of COCOHead, CrowdHuman and OpenImageV6 with totally about 403K

TABLE 4
HPE results on the *val-set* and *test-set* of challenging DAD-3DHeads dataset. The marker * indicates additional *ublabeled* data is used.

Type	Method	$\|\mathbf{I}-\mathbf{R}_1\mathbf{R}_2^T\|_F\downarrow$		Angle error (degree)↓	
		val-set	test-set	val-set	test-set
SL	Img2Pose [14]	—	0.226	—	9.122
	DAD-3DNet [16]	0.130	0.138	5.456	5.360
	Sup. (ResNet50)	0.133	0.138	5.543	5.234
	Sup. (RepVGG)	0.128	0.134	5.321	5.020
	Sup. (EffNetV2-S)	0.121	0.125	4.993	4.728
SSL	Base. (ResNet50)	0.138	0.145	5.794	5.312
	Base. (RepVGG)	0.130	0.137	5.425	5.182
	Base. (EffNetV2-S)	0.121	0.127	5.013	4.799
	Ours (ResNet50)	0.116	0.127	4.800	4.810
	Ours (RepVGG)	0.116	0.126	4.778	4.760
	Ours (EffNetV2-S)	0.112	0.124	4.636	4.636
	Ours (EffNetV2-S)*	0.109	0.124	4.518	4.408

heads, our method (with marker *) finally achieved lower pose estimation errors on both val-set and test-set. More than that, we give detailed comparison under different conditions in Table 5. Our method is still clearly ahead in many challenging cases such as atypical poses, compound expressions, heavy occlusions and non-strandard light. These are common scenes of in-the-wild head images, which are less-touched by the mainstream HPE studies [18], [20], [93] and benchmarks [11], [21]. Without bells and whistles, we achieved uncontroversial SOTA results on the full-range HPE dataset DAD-3DHeads. And all visualization results shown in this paper are estimated based on the optimal model obtained under this setting.

Performance Improvement Analysis of SemiUHPE. To systematically understand the source of improvement after boosting FisherMatch with our proposed strategies, we further calculate and plot the head pose error gains on datasets AFLW2000 for *front-range* HPE under **Setting2** (Fig. 10a) and DAD-3DHeads for *full-range* HPE under **Setting3** (Fig. 10b). We split all the testing images into different groups according to their GT yaw angles and calculate the improved error degree within each group. We can find that the improvement by our SemiUHPE gets more obvious as the head pose becomes more challenging and has a larger yaw angle (e.g., $> 90°$). This capability benefits from unsupervised joint training using a large number of human heads collected in any orientation.

Results of SemiObjRot. The experiment results on Pascal3D+ dataset of generic object rotation regression are shown in Table 6. The results illustrate that, with using the effective teacher-student mutual learning framework as well as our proposed SSL strategies (including the dynamic entropy-based pseudo label filtering and two advanced strong augmentations), our SemiObjRot significantly outperforms various baselines and three state-of-the-art methods (including FisherMatch, FisherMatch+, and the adapted SAA) under all different numbers of labeled images. This is strong evidence of the undisputed superiority and valuable versatility of our approach.

Results of Semi3DHead. The results of our modified Semi3DHead on the 3D head reconstruction task compared with other baselines are summarized in Table 7. Although 3D head reconstruction seems to be a completely different task, it has an intrinsic connection and synergy with HPE [10]. Therefore, after adding the HPE prediction branch, our

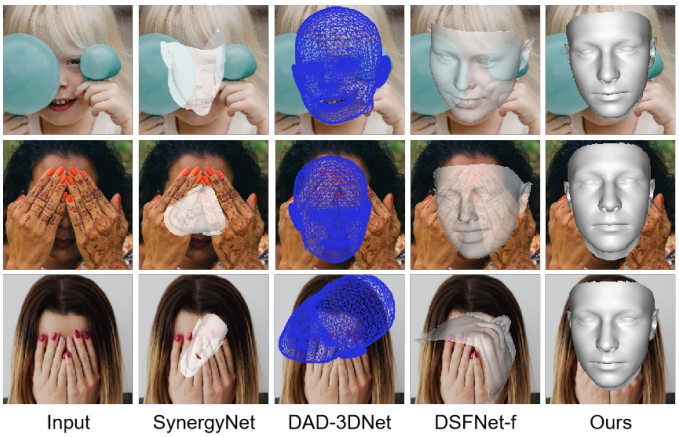


Fig. 11. Qualitative comparison results of front-range head pose estimation (HPE) methods. Our method is more robust to severe occlusion. Images are taken from the paper DSFNet-f [61].

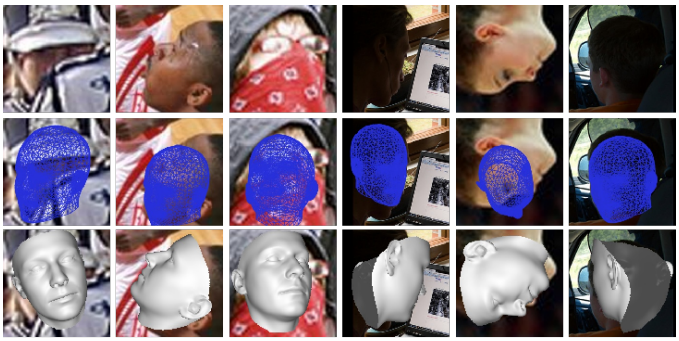


Fig. 12. Qualitative comparison results of full-range head pose estimation (HPE) methods on wild head images. Our method is more robust to severe occlusion, atypical pose and invisible face than DAD-3DNet [16].

Semi3DHead always outperforms the other two baselines (Mean-Teacher [28] and FixMatch [30]) regardless of the pseudo-label filtering scheme used. In addition, by observing the Mean-Teacher after applying the strong augmentation T_{CutOcc} (even some metrics are better than FixMatch based on pseudo-label filtering), we can see the universality of T_{CutOcc}. The two pseudo-label filtering methods we designed are better than the FixMatch based on static thresholds, indicating that reasonable dynamic filtering strategies are universally applicable. At the same time, the entropy-based one using the matrix Fisher distribution is significantly better, which is consistent with the previous analysis. In the future, we can try to extend Semi3DHead to more 3D head reconstruction datasets to fully leverage its ability to mine unlabeled head images.

4.4 Qualitative Comparison

To further explain the superiority of our SemiUHPE, we present the qualitative comparison with SOTA counterparts including SynergyNet [10], DAD-3DNet [16] and DSFNet-f [61]. As shown in Fig. 11 and Fig. 12, our method can obtain impressive HPE results on front yet severely occluded faces and in-the-wild challenging heads. Moreover, we present comparison on never-before-experienced yet challenging samples from the test-set of DAD-3DHeads. As shown in Fig. 13, although these images have high-definition, DAD-3DNet may make significant mistakes, leading to disordered results of head reconstruction. While, our method usually

TABLE 5
Detailed HPE errors on the *test-set* of DAD-3DHeads with four subsets: challenging atypical poses (Pose), compound expressions (Expr.), heavy occlusions (Occl.) and non-standard light (Light). All of our results are returned after the prediction results are submitted to the official and compared with the undisclosed ground-truth labels. The best and second-best result is in red and blue color, respectively.

Type	Method	$\|\mathbf{I}-\mathbf{R}_1\mathbf{R}_2^T\|_F\downarrow$ / Angle error (degree)↓				
		Overall	Pose	Expr.	Occl.	Light
SL	3DDFA-V2 [12]	0.527 / —	0.790 / —	0.455 / —	0.542 / —	— / —
	RingNet [82]	0.438 / —	1.076 / —	0.294 / —	0.551 / —	— / —
	DAD-3DNet [16]	0.138 / 5.360	0.343 / —	0.112 / —	0.203 / —	— / —
	Sup. (ResNet50)	0.138 / 5.234	0.327 / 10.782	0.111 / 4.351	0.181 / 7.134	0.129 / 5.094
	Sup. (RepVGG)	0.134 / 5.020	0.325 / 9.961	0.108 / 4.233	0.179 / 6.880	0.131 / 4.904
	Sup. (EffNetV2-S)	0.125 / 4.728	0.274 / 9.055	0.105 / 4.105	0.170 / 6.449	0.123 / 4.969
SSL	Base. (ResNet50)	0.145 / 5.312	0.400 / 11.109	0.111 / 4.356	0.194 / 7.583	0.138 / 5.523
	Base. (RepVGG)	0.137 / 5.182	0.349 / 10.932	0.110 / 4.305	0.176 / 7.175	0.131 / 5.080
	Base. (EffNetV2-S)	0.127 / 4.799	0.278 / 9.244	0.106 / 4.135	0.178 / 6.372	0.122 / 4.864
	Ours (ResNet50)	0.127 / 4.810	0.322 / 9.581	0.103 / 4.106	0.159 / 6.463	0.127 / 4.974
	Ours (RepVGG)	0.126 / 4.760	0.307 / 9.110	0.105 / 4.180	0.149 / 6.041	0.121 / 4.683
	Ours (EffNetV2-S)	0.124 / 4.636	0.309 / 8.620	0.102 / 4.047	0.145 / 5.897	0.123 / 4.761
	Ours (EffNetV2-S)*	0.124 / 4.408	0.349 / 7.666	0.098 / 3.916	0.132 / 5.362	0.116 / 4.434

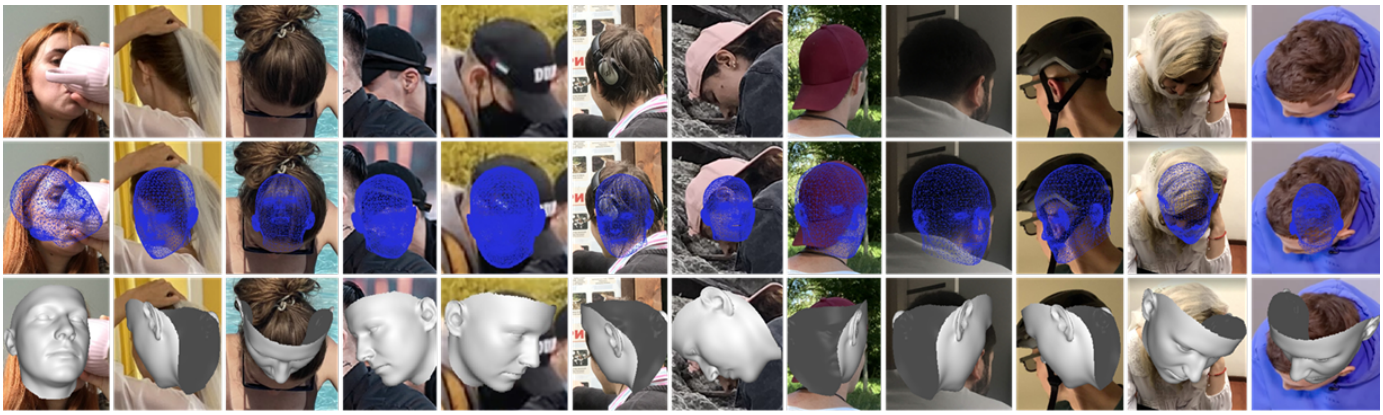


Fig. 13. Qualitative comparison results of full-range head pose estimation (HPE) between our method (*3rd line*) and DAD-3DNet [16] (*2nd line*). All head images are from DAD-3DHeads test-set (*1st line*), which never appeared during SSL training.

TABLE 6
Comparing our method with the baselines on the 6 categories of Pascal3D+ dataset with few annotations. All results are averaged.

Method	7 images		20 images		50 images	
	Med.↓	Acc_{30°↑	Med.↓	Acc_{30°↑	Med.↓	Acc_{30°↑
Res50-Gene	39.1	36.1	26.3	45.2	20.2	54.6
Res50-Spec	46.5	29.6	29.4	42.8	23.0	50.4
StarMap [95]	49.6	30.7	46.4	35.6	27.9	53.8
NeMo [96]	60.0	38.4	33.3	51.7	22.1	69.3
NVSM [71]	37.5	53.8	28.7	61.7	24.2	65.6
FisherMatch [36]	28.3	56.8	23.8	63.6	16.1	75.7
SAA [90]	26.0	57.4	22.0	64.7	15.2	78.1
FisherMatch+ [37]	25.5	58.0	21.4	65.1	14.8	78.9
Ours (Res50)	21.5	61.2	18.6	67.7	12.8	83.1
Full Sup.	8.1	89.6	8.1	89.6	8.1	89.6

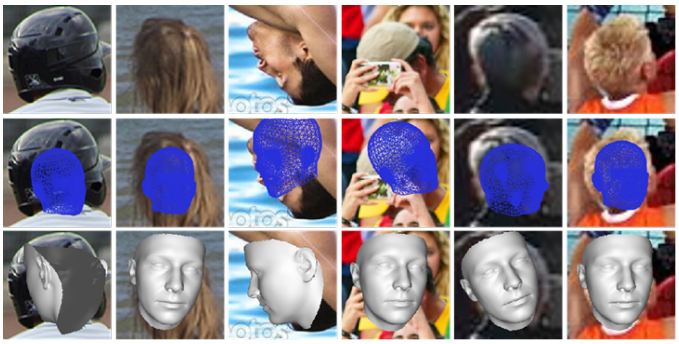


Fig. 14. Failure cases (*third line*) of our SemiUHPE on some wild challenging heads (*fisrt line*). These samples are also very hard for DAD-3DNet [16] (*second line*) to deal with.

gives a satisfactory estimation. More convincing qualitative results can be found in our project link.

Although the overall effect is impressive, our SemiUHPE may still fail on some cases of heavy blur, invisible face, severe occlusions, or atypical pose (like faces upside-down). Sometimes, there is more than one challenge, as shown in Fig. 14 of the third case (atypical pose + self-occlusion) and the last two cases (blurry + backward). Although we humans have strong prior knowledge to help infer the poses of challenging heads, the network cannot do this so far. We believe that in these situations, it is necessary for the model to rely on the context of the human body to identify its head pose, especially for the first, second and fourth cases in Fig. 14, where heads facing back and completely covered by other objects such as hats and hair.

4.5 Ablation Studies

In this part, we give detailed studies for explaining the effect of our proposed three strategies. Then, we present the studies of dynamic thresholds changing and unsupervised convergence tendency, respectively.

Aspect-Ratio Invariant Cropping. We took FSA-Net [18] and 6DRepNet [20] using the naive cropping-resizing way for comparing. Then, we replaced them with aspect-ratio invariant cropping, and retrained new versions FSA-Net† and 6DRepNet†. We also listed results of the trivial supervised method. As shown in Table 8, both the original FSA-Net and 6DRepNet are significantly improved. When using

TABLE 7
Comparison of **3D head reconstruction** results on DAD-3DHeads *val-set*. The backbone of all models is MobileNet-w1 [97]. All four metrics are adopted from the paper [16]. Baselines Mean-Teacher [28] and FixMatch [30] used the original DAD-3DNet without adding the HPE sub-branch. The ratio 5% (1,892 labels), 10% (3,784 labels), 20% (7,568 labels) or 100% (37,840 labels) means labeled samples in DAD-3DHeads *train-set*.

Type	Method	NME↓				Z_5 Accuracy↑				Chamfer Distance↓				Pose Error↓			
		5%	10%	20%	100%	5%	10%	20%	100%	5%	10%	20%	100%	5%	10%	20%	100%
SL	DAD-3DNet [16]	7.861	4.766	4.025	2.467	0.919	0.929	0.938	0.954	3.957	3.468	3.247	2.982	0.327	0.270	0.233	0.149
SSL	Mean-Teacher [28]	5.599	4.969	4.860	—	0.926	0.928	0.936	—	3.459	3.320	3.383	—	0.291	0.277	0.241	—
	Mean-Teacher [28] + T_{CutOcc}	4.071	3.845	3.645	—	0.936	0.939	0.939	—	3.267	3.210	3.084	—	0.245	0.233	0.226	—
	FixMatch [30] + Fixed Thre.	4.460	3.790	3.543	—	0.935	0.938	0.939	—	3.248	3.127	3.075	—	0.252	0.230	0.225	—
	Ours (Geo. Dist. Filtering)	3.990	3.781	3.553	—	0.937	0.939	0.942	—	3.179	3.144	3.046	—	0.239	0.229	0.221	—
	Ours (Entropy Filtering)	3.833	3.748	3.490	—	0.941	0.941	0.947	—	3.152	3.055	3.027	—	0.228	0.217	0.206	—

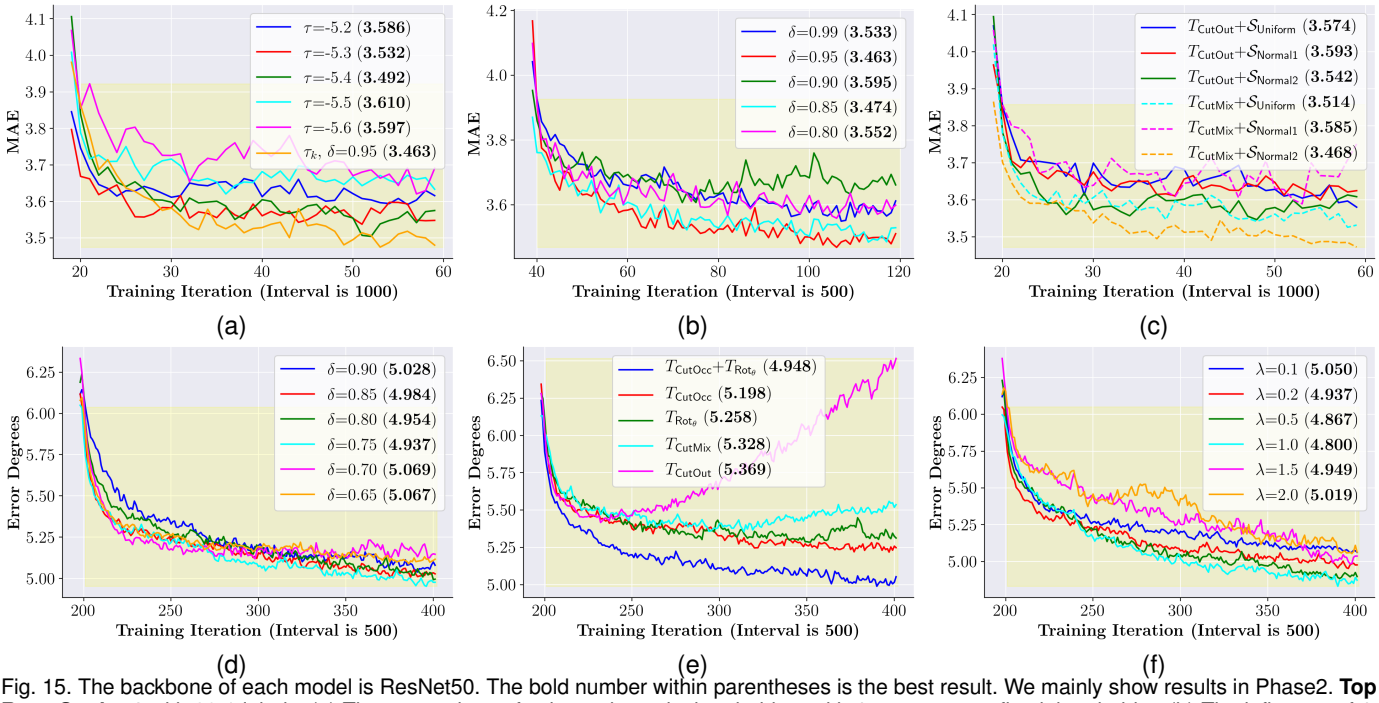


Fig. 15. The backbone of each model is ResNet50. The bold number within parentheses is the best result. We mainly show results in Phase2. **Top Row: Setting1** with 20% labels. (a) The comparison of using a dynamic threshold τ_k with $\delta=0.95$ or pre-fixed threshold τ. (b) The influence of δ. (c) The effect of different sampling ways. **Bottom Row: Setting3**. (d) The influence of δ with $\lambda=0.2$. (e) The effect of different augmentations with $\lambda=0.2$ and $\delta=0.75$. (f) The influence of λ with $\delta=0.75$.

TABLE 8
Euler angles errors on AFLW2000. Models are trained on 300W-LP with different input cropping ways and pose rotation representations.

Method	Backbone	Cropping	Rot-Rep	Pitch	Yaw	Roll	MAE
FSA-Net [18]	ResNet50	Naive	Euler angles	6.08	4.50	4.64	5.07
FSA-Net† [18]	ResNet50	Ours	Euler angles	5.42	4.01	3.75	4.39
6DRepNet [20]	RepVGG	Naive	trivial matrix	4.91	3.63	3.37	3.97
6DRepNet† [20]	RepVGG	Ours	trivial matrix	4.58	3.04	2.86	3.49
Supervised	ResNet50	Ours	matrix Fisher	4.58	3.20	2.95	3.58
Supervised	RepVGG	Ours	matrix Fisher	4.50	3.18	2.81	3.50
Supervised	EffNetV2-S	Ours	matrix Fisher	4.40	2.99	2.75	3.38

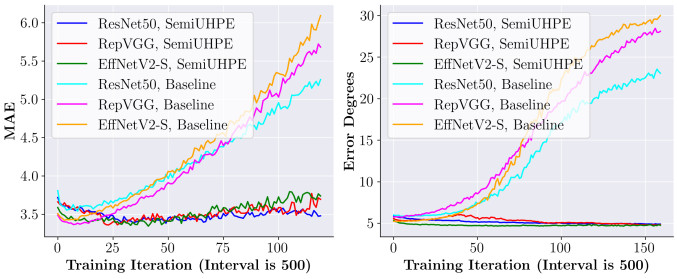


Fig. 16. Testing results of the baseline FisherMatch and our proposed SemiUHPE in **Setting2** (a) and **Setting3** (b). Only training steps in Phase2 are shown for clearer comparison.

RepVGG and new cropping, 6DRepNet† can even compete with the supervised method equipped an advanced matrix Fisher representation. These prove the superiority of this simple yet efficient size-invariant preprocessing.

Dynamic Entropy-based Filtering. We designed three groups of experiments for explaining the effectiveness of dynamic filtering strategy. Firstly, in Fig. 15a, we searched for the optimal pre-fixed threshold τ used by FisherMatch. Nonetheless, our dynamic threshold τ_k with $\delta=0.95$ got the best result. We also searched for the optimal δ for different unlabeled datasets. Then, in Fig. 15b, we got the optimal $\delta=0.95$, which is a large ratio due to that both labeled and unlabeled data are in 300W-LP. Finally, in Fig. 15d, we got a lower optimal $\delta=0.75$, which is caused by more difficult and noisy heads in the unlabeled COCOHead. Usually, an ideal δ should equal to $1-\|\mathcal{D}^u_{ood}\|/\|\mathcal{D}^u\|$. But we cannot obtain the accurate ratio of $\mathcal{D}^u_{ood}$ in $\mathcal{D}^u$. We thus estimated a suitable δ by ablation studies. For example, the optimal δ of COCOHead is 0.75 in **Setting2** and **Setting3**, which means δ may be dataset-related yet not sensitive to task settings.

Head-Oriented Strong Augmentations. We conducted two groups of experiments for showing the effectiveness

TABLE 9
The HPE results on the *val-set* of dataset DAD-3DHeads by using different unlabeled datasets.

Method	Unlabeled Dataset	$\|\mathbf{I}-\mathbf{R}_1\mathbf{R}_2^T\|_F \downarrow$	Angle error (degree)↓
DAD-3DNet [16]	—	0.130	5.456
SemiUHPE (ours)	WiderFace	0.127	5.300$_{(\ 2.9\%\downarrow)}$
	CrowdHuman	0.125	5.226$_{(\ 4.2\%\downarrow)}$
	OpenImageV6	0.123	5.174$_{(\ 5.2\%\downarrow)}$
	COCOHead	0.116	4.800$_{(12.0\%\downarrow)}$

TABLE 10
Euler angles errors on AFLW2000. Models are trained on 300W-LP with different ratios of label.

Type	Method	Backbone	2%	5%	10%	20%	All
SL	Sup.	TinyViT-22M	4.116	3.734	3.677	3.512	3.455
	Sup.	EffNetV2-S	4.009	3.678	3.517	3.444	3.379
SSL	Base.	TinyViT-22M	4.031	3.691	3.464	3.415	—
	Base.	EffNetV2-S	3.991	3.596	3.448	3.372	—
	Ours	TinyViT-22M	3.914	3.647	3.514	3.434	—
	Ours	EffNetV2-S	3.835	3.526	3.377	3.348	—

of new strong augmentations. In Fig. 15c, we can see that CutMix is always better than CutOut. The best sampling distribution is $\mathcal{S}_{\text{Normal2}}$ for its reasonable concentration of occlusion generation. In Fig. 15e, we kept using $\mathcal{S}_{\text{Normal2}}$ and observed the same effect of CutOut and CutMix. When combining them together, the new T_{CutOcc} can further reduce HPE errors. Independently, the proposed rotation consistency augmentation T_{Rot_θ} can also improve performance. Finally, when applying both T_{CutOcc} and T_{Rot_θ}, we achieved the best result with a remarkable promotion.

Influence of Unsupervised Loss Weight. As shown in Fig. 15f, we selected the unsupervised loss weight λ from the list (0.1, 0.2, 0.5, 1.0, 1.5, 2.0). Our method performed best when setting $\lambda = 1.0$, which indicates that it does not require careful adjustments of the unsupervised part weight and has a robust performance about this hyper-parameter.

Convergence Curve of Baseline FisherMatch. As shown in Fig. 16, we plotted convergence curves of the baseline FisherMatch as well as our proposed SemiUHPE. It is obvious that whether it is in **Setting2** (see Fig. 16a) or **Setting3** (see Fig. 16b), the baseline method always begins to collapse after quickly reaching an optimal point with using either backbone. We assume that this is caused by the domain differences between labeled and unlabeled datasets. Our proposed SemiUHPE can significantly alleviate this problem and converge relatively smoothly.

4.6 Other Optional Setups

4.6.1 Choosing Different Unlabeled Datasets.

As discussed in the previous content, in addition to COCOHead [32], there are also other alternative unlabeled datasets such as WiderFace [33], CrowdHuman [34] and OpenImageV6 [31] that contain many wild heads. A natural doubt is, will applying these similar substances yield better results than using COCOHead. To answer this question, we followed the steps of generating COCOHead, processed original WiderFace and CrowdHuman datasets, and obtained the corresponding unlabeled head sets. Specifically, we removed samples with head size smaller than 25, 30 and 30 pixels in WiderFace, CrowdHuman and OpenImageV6. Then, we got about 62K, 163K and 165K heads, respectively. Then, we followed **Setting3** to implement the similar **DAD-WiderFace**, **DAD-CrowdHuman** and and **DAD-OpenImage** experiments. Considering that both CrowdHuman and OpenImageV6 have about 2× samples than COCOHead (∼74K) or WiderFace, we adjusted their iterations in Phase2 from 100K into 200K. Without loss of generality, all compared models used the ResNet50 [87] as their network backbones.

As shown in Table 9, although our method SemiUHPE using either unlabeled dataset can surpass the supervised DAD-3DNet [16], it has significant discrepancies when applying different unlabeled datasets. For the angle error, our method can improve DAD-3DNet by 12.0% when using COCOHead, which is much more prominent than using WiderFace (2.9%), CrowdHuman (4.2%) or OpenImageV6 (5.2%). This is understandable because most human heads in these datasets are face-visible, especially the WiderFace originally built for face detection task. Besides, CrowdHuman and OpenImageV6 have many harmful head bounding box annotations which are illegal or unrecognizable. In summary, COCOHead is the most suitable unlabeled choice. In practical applications, if we want to pursue higher performance, we can also choose to merge these datasets for co-training (such as the last row in Table 4).

4.6.2 Choosing Transformer-based Backbones

Although we have chosen three different backbones including ResNet50 [87], RepVGG [88] and EfficientNetV2-S [89] to conduct extensive experiments, these networks are all based on CNNs, which have been challenged in recent years by transformer-based alternatives [98]–[100]. Actually, the fully supervised method TokenHPE [58], [101] has initially revealed the great potential of using transformer networks (*e.g.*, the ViT-Base/16 [98]) to deal with the HPE task. Therefore, we also considered whether we could use a more superior transformer-based backbone to boost SemiUHPE. After considering various aspects, we decided to adopt TinyViT-22M [102] as a trial. On the one hand, TinyViT-22M has similar parameters to our used CNN-based networks, and on the other hand, its pre-trained model has higher classification accuracy yet smaller calculation amount than the original ViT-Base/16. This means it is more likely to achieve better HPE results.

Specifically, with using TinyViT-22M as the backbone, we followed the **Setting1** and implemented all three methods including Sup., Base and SemiUHPE. As shown in Table 10, although TinyViT-22M has a similar top-1 accuracy with EffNetV2-S on ImageNet (84.8% vs. 84.9%), its performance can be much worse than EffNetV2-S when trained with full or semi-supervision. As a reasonable reference, in Table 3, the transformer-based TokenHPE [58] also failed to achieve lower HPE errors than many CNN-based counterparts. And our reproduced supervised method using the TinyViT-22M backbone and matrix Fisher representation can get quite low results (MAE=3.46) if comparing with methods in Table 3. In particular, when using TinyViT-22M for SSL training, our method performed worse than the baseline method when the labeling rate is 10% or 20%. This anomaly is very different from the case when using other backbones. We suspect that this is because the transformer structure needs to preprocess the input image into smaller patches,

which conflicts with the operations such as CutOut [77] and CutMix [79] used in the strong augmentations on unlabeled images. Nevertheless, we envision that these troubles may be alleviated by applying data augmentation strategies specifically designed for vision transformer families [103]–[105], which can be considered as a trivial extension of our work. On the other hand, it has been widely proven that vision transformers require a large amount of labeled data to train in order to unleash their scaling laws, such as for 3D face reconstruction [106] and 3D hand reconstruction [107]. But as we have revealed, this does not seem to benefit HPE tasks in SSL settings.

5 Conclusion and Discussion

In this paper, we aim to address the unconstrained head pose estimation task on less-touched wild head images. Due to the lack of corresponding labels, we turn to semi-supervised learning techniques. Based on empirically effective frameworks, we propose the dynamic entropy-based filtering for gradually updating thresholds and head-oriented strong augmentations for better enforcing consistency training. By combining the proposed aspect-ratio invariant cropping, our method can achieve optimal HPE performance quantitatively and qualitatively on omnidirectional wild heads. We also demonstrate the scalability and versatility of SemiUHPE on generic object rotation regression and 3D head reconstruction. We expect that our work will greatly inspire related downstream applications.

Last but not least, although our method has achieved stunning results in estimating the pose of wild head, there are still many aspects worth exploring in depth. We summarize the possible perspectives for future research as follows.

- **Domain gaps among datasets.** Strictly speaking, there are conspicuous differences between labeled and unlabeled datasets we used in this paper. For example, the labeled 300W-LP is artificially synthesized through the face profiling algorithm with artifacts, while the unlabeled COCOHead is collected in the wild containing realistic and natural samples. We may utilize domain adaptation strategies to alleviate this problem.
- **Combination with vision-language models.** Recently, large vision-language models (VLMs) and multimodal LLMs (mLLMs) have demonstrated strong generalizable visual reasoning abilities by aligning image and text inputs. Beyond using text descriptions to mitigate data scarcity [108] or to improve robustness in HPE [109], a more concrete strategy is to cast unconstrained HPE into a visual question answering (VQA) task. For example, both the cropped head image and its reconstructed 3D mesh can be provided to an mLLM, together with a query about the plausibility of the estimated pose. If the mLLM detects a severe error, the system can re-estimate after small perturbations or reject the prediction. Our preliminary test with Gemini 2.5 Flash shows that it can reliably flag most obvious errors, suggesting a promising future direction.
- **More accurate head pose estimation.** Our method still cannot address some hard cases shown in Fig. 14, where most facial features are missing due to severe occlusion or backward orientation. A potential remedy is to leverage contextual cues beyond the cropped head. For instance, the orientation of the upper body [6], [110] can provide a strong prior: if the torso clearly faces away from the camera, the head pose should be consistent with a back-of-head orientation. Similarly, surrounding information such as hair-dominated regions with absent facial features [53], [111] can help detect backward heads where the yaw absolute angle must exceed $90°$. Integrating such priors may effectively reduce catastrophic errors that cannot be solved by single-frame appearance cues alone.
- **Video-level HPE extension.** While our SemiUHPE current focuses on single-frame, applying it to video sequences may lead to temporal jitter or abrupt errors, especially for backward-facing heads. A promising direction is to incorporate temporal consistency constraints, ensuring smooth pose transitions across consecutive frames. Another possibility is to exploit short-term temporal windows, where neighboring frames provide complementary cues to disambiguate extreme poses. Such extensions could further enhance the robustness of SemiUHPE in real-world video applications.
- **Video-based SemiUHPE.** Another promising extension is to move from single-frame SemiUHPE to video-based semi-supervised head pose estimation. By redesigning the backbone to process short frame sequences, the model could implicitly capture temporal head-motion cues and output smooth 3D pose trajectories. A teacher–student framework built on annotated video datasets (e.g., BIWI [21], CMU Panoptic [112]) together with large-scale unlabeled online videos would provide a natural baseline. This direction also parallels recent advances in sequential visuomotor imitation learning, where temporal visual observations are mapped to continuous 6DoF trajectories, suggesting that frameworks such as ACT [113] or Diffusion Policy [114] could be adapted for head pose estimation.
- **Impact on downstream tasks.** Our SemiUHPE also has implications for broader head-related tasks. For 3D head generation, it can help diagnose and re-balance pose distributions in training datasets, mitigating bias caused by underrepresented large poses. For talking-head generation, robust estimation of extreme angles may reduce jittering artifacts in synthesized videos. For face-swapping, accurate pose priors can guide the alignment of wrapped facial textures, alleviating failures under large yaw angles. These extensions highlight the potential of SemiUHPE as a general building block for robust head-driven applications.

References

[1] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *TPAMI*, vol. 31, no. 4, pp. 607–626, 2008.

[2] A. F. Abate, C. Bisogni, A. Castiglione, and M. Nappi, "Head pose estimation: An extensive survey on recent techniques and applications," *PR*, vol. 127, p. 108591, 2022.

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

[4] K. Ahuja, D. Kim, F. Xhakaj, V. Varga, A. Xie, S. Zhang, J. E. Townsend, C. Harrison, A. Ogan, and Y. Agarwal, "Edusense: Practical classroom sensing at scale," *UbiComp*, vol. 3, no. 3, pp. 1–26, 2019.

[5] K. Ahuja, D. Shah, S. Pareddy, F. Xhakaj, A. Ogan, Y. Agarwal, and C. Harrison, "Classroom digital twins with instrumentation-free gaze tracking," in *CHI*, 2021, pp. 1–9.

[6] S. Nonaka, S. Nobuhara, and K. Nishino, "Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination," in *CVPR*, 2022, pp. 2192–2201.

[7] N. Zapata, G. Pérez, L. Bonilla, P. Núñez, P. Bachiller, and P. Bustos, "Guessing human intentions to avoid dangerous situations in caregiving robots," in *IROS*, 2024.

[8] P. T. Singamaneni, P. Bachiller-Burgos, L. J. Manso, A. Garrell, A. Sanfeliu, A. Spalanzani, and R. Alami, "A survey on socially aware robot navigation: Taxonomy and future challenges," *IJRR*, p. 02783649241230562, 2024.

[9] 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," *TPAMI*, vol. 41, no. 1, pp. 121–135, 2017.

[10] C.-Y. Wu, Q. Xu, and U. Neumann, "Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry," in *3DV*. IEEE, 2021, pp. 453–463.

[11] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *CVPR*, 2016, pp. 146–155.

[12] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3d dense face alignment," in *ECCV*. Springer, 2020, pp. 152–168.

[13] R. Valle, J. M. Buenaposada, and L. Baumela, "Multi-task head pose estimation in-the-wild," *TPAMI*, vol. 43, no. 8, pp. 2874–2881, 2020.

[14] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner, "img2pose: Face alignment and detection via 6dof, face pose estimation," in *CVPR*, 2021, pp. 7617–7627.

[15] Y. Yu, K. A. F. Mora, and J.-M. Odobez, "Headfusion: 360° head pose tracking combining 3d morphable model and 3d reconstruction," *TPAMI*, vol. 40, no. 11, pp. 2653–2667, 2018.

[16] 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 *CVPR*, 2022, pp. 20 942–20 952.

[17] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *CVPRW*, 2018, pp. 2074–2083.

[18] 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 *CVPR*, 2019, pp. 1087–1096.

[19] Y. Zhou and J. Gregson, "Whenet: Real-time fine-grained estimation for wide range head pose," in *BMVC*, 2020.

[20] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, "6d rotation representation for unconstrained head pose estimation," in *ICIP*. IEEE, 2022, pp. 2496–2500.

[21] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3d face analysis," *IJCV*, vol. 101, pp. 437–458, 2013.

[22] J. Gu, X. Yang, S. De Mello, and J. Kautz, "Dynamic facial analysis: From bayesian filtering to recurrent neural network," in *CVPR*, 2017, pp. 1548–1557.

[23] F. Kuhnke and J. Ostermann, "Deep head pose estimation using synthetic images and partial adversarial domain adaption for continuous label spaces," in *CVPR*, 2019, pp. 10 164–10 173.

[24] Y. Wang, W. Zhou, and J. Zhou, "2dheadpose: A simple and effective annotation method for the head pose in rgb images and its dataset," *NN*, vol. 160, pp. 50–62, 2023.

[25] L. Zeng, L. Chen, W. Bao, Z. Li, Y. Xu, J. Yuan, and N. K. Kalantari, "3d-aware facial landmark detection via multi-view consistent training on synthetic data," in *CVPR*, 2023, pp. 12 747–12 758.

[26] T. T. Dao, D. H. Vu, C. Pham, and A. Tran, "Efhq: Multi-purpose extremepose-face-hq dataset," in *CVPR*, 2024, pp. 22 605–22 615.

[27] S. An, H. Xu, Y. Shi, G. Song, U. Y. Ogras, and L. Luo, "Panohead: Geometry-aware 3d full-head synthesis in 360deg," in *CVPR*, 2023, pp. 20 950–20 959.

[28] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *NeurIPS*, vol. 30, 2017.

[29] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *NeurIPS*, vol. 32, 2019.

[30] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *NeurIPS*, vol. 33, pp. 596–608, 2020.

[31] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *NeurIPS*, vol. 33, pp. 6256–6268, 2020.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[33] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *CVPR*, 2016, pp. 5525–5533.

[34] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "Crowdhuman: A benchmark for detecting human in a crowd," *arXiv preprint arXiv:1805.00123*, 2018.

[35] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *IJCV*, vol. 128, no. 7, pp. 1956–1981, 2020.

[36] Y. Yin, Y. Cai, H. Wang, and B. Chen, "Fishermatch: Semi-supervised rotation regression via entropy-based filtering," in *CVPR*, 2022, pp. 11 164–11 173.

[37] Y. Yin, J. Lyu, Y. Wang, H. Liu, H. Wang, and B. Chen, "Towards robust probabilistic modeling on so (3) via rotation laplace distribution," *TPAMI*, 2025.

[38] J. Levinson, C. Esteves, K. Chen, N. Snavely, A. Kanazawa, A. Rostamizadeh, and A. Makadia, "An analysis of svd for deep rotation estimation," *NeurIPS*, vol. 33, pp. 22 554–22 565, 2020.

[39] D. Mohlin, J. Sullivan, and G. Bianchi, "Probabilistic orientation estimation with matrix fisher distributions," *NeurIPS*, vol. 33, pp. 4884–4893, 2020.

[40] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *NeurIPS*, vol. 34, pp. 18 408–18 419, 2021.

[41] I. Nassar, S. Herath, E. Abbasnejad, W. Buntine, and G. Haffari, "All labels are not created equal: Enhancing semi-supervision via label grouping and co-training," in *CVPR*, 2021, pp. 7241–7250.

[42] J. Wu, H. Yang, T. Gan, N. Ding, F. Jiang, and L. Nie, "Chmatch: Contrastive hierarchical matching and robust adaptive threshold boosted semi-supervised learning," in *CVPR*, 2023, pp. 15 762–15 772.

[43] J. Kim, J. Jang, S. Seo, J. Jeong, J. Na, and N. Kwak, "Mum: Mix image tiles and unmix feature tiles for semi-supervised object detection," in *CVPR*, 2022, pp. 14 512–14 521.

[44] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *CVPR*, 2023, pp. 7236–7246.

[45] R. Xie, C. Wang, W. Zeng, and Y. Wang, "An empirical study of the collapsing problem in semi-supervised 2d human pose estimation," in *ICCV*, 2021, pp. 11 240–11 249.

[46] L. Huang, Y. Li, H. Tian, Y. Yang, X. Li, W. Deng, and J. Ye, "Semi-supervised 2d human pose estimation driven by position inconsistency pseudo label correction module," in *CVPR*, 2023, pp. 693–703.

[47] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *CVPR*, 2014, pp. 1867–1874.

[48] 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 *ICCV*, 2017, pp. 1021–1030.

[49] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *CVPR*, 2020, pp. 5203–5212.

[50] G. Cantarini, F. F. Tomenotti, N. Noceti, and F. Odone, "Hhp-net: A light heteroscedastic neural network for head pose estimation with uncertainty," in *WACV*, 2022, pp. 3521–3530.

[51] H. Zhang, M. Wang, Y. Liu, and Y. Yuan, "Fdn: Feature decoupling network for head pose estimation," in *AAAI*, vol. 34, no. 07, 2020, pp. 12 789–12 796.

[52] X. Geng, X. Qian, Z. Huo, and Y. Zhang, "Head pose estimation based on multivariate label distribution," *TPAMI*, vol. 44, no. 4, pp. 1974–1991, 2020.

[53] H. Zhou, F. Jiang, and H. Lu, "Directmhp: Direct 2d multi-person head pose estimation with full-range angles," *arXiv preprint arXiv:2302.01110*, 2023.

[54] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee, "Quatnet: Quaternion-based head pose estimation with multiregression loss," *TMM*, vol. 21, no. 4, pp. 1035–1046, 2018.

[55] Z. Cao, Z. Chu, D. Liu, and Y. Chen, "A vector-based representation to enhance head pose estimation," in *WACV*, 2021, pp. 1188–1197.

[56] H. Liu, S. Fang, Z. Zhang, D. Li, K. Lin, and J. Wang, "Mfdnet: Collaborative poses perception and matrix fisher distribution for head pose estimation," *TMM*, vol. 24, pp. 2449–2460, 2021.

[57] D. Dai, W. Wong, and Z. Chen, "Rankpose: Learning generalised feature with rank supervision for head pose estimation," in *BMVC*, 2020.

[58] C. Zhang, H. Liu, Y. Deng, B. Xie, and Y. Li, "Tokenhpe: Learning orientation tokens for efficient head pose estimation via transformers," in *CVPR*, 2023, pp. 8897–8906.

[59] Z. Ruan, C. Zou, L. Wu, G. Wu, and L. Wang, "Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction," *TIP*, vol. 30, pp. 5793–5806, 2021.

[60] Y. Kao, B. Pan, M. Xu, J. Lyu, X. Zhu, Y. Chang, X. Li, and Z. Lei, "Towards 3d face reconstruction in perspective projection: Estimating 6dof face pose from monocular image," *TIP*, 2023.

[61] H. Li, B. Wang, Y. Cheng, M. Kankanhalli, and R. T. Tan, "Dsfnet: Dual space fusion network for occlusion-robust 3d dense face alignment," in *CVPR*, 2023, pp. 4531–4540.

[62] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *CVPR*, 2018, pp. 4119–4128.

[63] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," *NeurIPS*, vol. 31, 2018.

[64] I. Nassar, M. Hayat, E. Abbasnejad, H. Rezatofighi, and G. Haffari, "Protocon: Pseudo-label refinement via online clustering and prototypical consistency for efficient semi-supervised learning," in *CVPR*, 2023, pp. 11 641–11 650.

[65] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *ICLR*, 2016.

[66] Z. Hu, Z. Yang, X. Hu, and R. Nevatia, "Simple: Similar pseudo label exploitation for semi-supervised classification," in *CVPR*, 2021, pp. 15 099–15 108.

[67] F. Yang, K. Wu, S. Zhang, G. Jiang, Y. Liu, F. Zheng, W. Zhang, C. Wang, and L. Zeng, "Class-aware contrastive semi-supervised learning," in *CVPR*, 2022, pp. 14 421–14 430.

[68] Y. Chen, X. Tan, B. Zhao, Z. Chen, R. Song, J. Liang, and X. Lu, "Boosting semi-supervised learning by exploiting all unlabeled data," in *CVPR*, 2023, pp. 7548–7557.

[69] G. Wang, F. Manhardt, J. Shao, X. Ji, N. Navab, and F. Tombari, "Self6d: Self-supervised monocular 6d object pose estimation," in *ECCV*. Springer, 2020, pp. 108–125.

[70] G. Wang, F. Manhardt, X. Liu, X. Ji, and F. Tombari, "Occlusion-aware self-supervised monocular 6d object pose estimation," *TPAMI*, 2021.

[71] A. Wang, S. Mei, A. L. Yuille, and A. Kortylewski, "Neural view synthesis and matching for semi-supervised few-shot learning of 3d pose," *NeurIPS*, vol. 34, pp. 7207–7219, 2021.

[72] Y. Yin, Y. Wang, H. Wang, and B. Chen, "A laplace-inspired distribution on so (3) for probabilistic rotation estimation," in *ICLR*, 2023.

[73] F. Kuhnke, S. Ihler, and J. Ostermann, "Relative pose consistency for semi-supervised head pose estimation," in *FG*. IEEE, 2021, pp. 01–08.

[74] F. Kuhnke and J. Ostermann, "Domain adaptation for head pose estimation using relative pose consistency," *TBIOM*, vol. 5, no. 3, pp. 348–359, 2023.

[75] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *CVPR*, 2017, pp. 212–220.

[76] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019, pp. 4690–4699.

[77] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[78] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.

[79] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *ICCV*, 2019, pp. 6023–6032.

[80] H. Rhodin, M. Salzmann, and P. Fua, "Unsupervised geometry-aware representation for 3d human pose estimation," in *ECCV*, 2018, pp. 750–767.

[81] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *ToG*, vol. 36, no. 6, pp. 194–1, 2017.

[82] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "Learning to regress 3d face shape and expression from an image without 3d supervision," in *CVPR*, 2019, pp. 7763–7772.

[83] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *AVSS*. IEEE, 2009, pp. 296–301.

[84] L. Yang, Q. Song, Z. Wang, M. Hu, and C. Liu, "Hier r-cnn: Instance-level human parts detection and a new benchmark," *TIP*, pp. 39–54, 2020.

[85] H. Zhou, F. Jiang, and H. Lu, "Body-part joint detection and association via extended object representation," in *ICME*. IEEE, 2023, pp. 168–173.

[86] H. Zhou, F. Jiang, J. Si, Y. Ding, and H. Lu, "Bpjdet: Extended object representation for generic body-part joint detection," *TPAMI*, 2024.

[87] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[88] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in *CVPR*, 2021, pp. 13 733–13 742.

[89] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *ICML*. PMLR, 2021, pp. 10 096–10 106.

[90] G. Gui, Z. Zhao, L. Qi, L. Zhou, L. Wang, and Y. Shi, "Enhancing sample utilization through sample adaptive augmentation in semi-supervised learning," in *CVPR*, 2023, pp. 15 880–15 889.

[91] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond pascal: A benchmark for 3d object detection in the wild," in *WACV*. IEEE, 2014, pp. 75–82.

[92] Y. Li, G. Tan, and C. Gou, "Cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose," *IJCV*, pp. 1–16, 2023.

[93] Y. Wang, H. Liu, Y. Feng, Z. Li, X. Wu, and C. Zhu, "Headdiff: Exploring rotation uncertainty with diffusion models for head pose estimation," *TIP*, 2024.

[94] A. Cobo, R. Valle, J. M. Buenaposada, and L. Baumela, "On the representation and methodology for wide and short range head pose estimation," *PR*, vol. 149, p. 110263, 2024.

[95] X. Zhou, A. Karpur, L. Luo, and Q. Huang, "Starmap for category-agnostic keypoint and viewpoint estimation," in *ECCV*, 2018, pp. 318–334.

[96] A. Wang, A. Kortylewski, and A. Yuille, "Nemo: Neural mesh models of contrastive features for robust 3d pose estimation," in *ICLR*, 2021.

[97] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[98] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020.

[99] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10 012–10 022.

[100] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *CVPR*, 2022, pp. 12 009–12 019.

[101] H. Liu, C. Zhang, Y. Deng, T. Liu, Z. Zhang, and Y.-F. Li, "Orientation cues-aware facial relationship representation for head pose estimation via transformer," *TIP*, vol. 32, pp. 6289–6302, 2023.

[102] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, and L. Yuan, "Tinyvit: Fast pretraining distillation for small vision transformers," in *ECCV*. Springer, 2022, pp. 68–85.

[103] J.-N. Chen, S. Sun, J. He, P. H. Torr, A. Yuille, and S. Bai, "Transmix: Attend to mix for vision transformers," in *CVPR*, 2022, pp. 12 135–12 144.

[104] J. Liu, B. Liu, H. Zhou, H. Li, and Y. Liu, "Tokenmix: Rethinking image mixing for data augmentation in vision transformers," in *ECCV*. Springer, 2022, pp. 455–471.
[105] R. Fang, P. Gao, A. Zhou, Y. Cai, S. Liu, J. Dai, and H. Li, "Feataug-detr: Enriching one-to-many matching for detrs with feature augmentation," *TPAMI*, 2024.
[106] T. Zhang, X. Chu, Y. Liu, L. Lin, Z. Yang, Z. Xu, C. Cao, F. Yu, C. Zhou, C. Yuan *et al.*, "Accurate 3d face reconstruction with facial component tokens," in *ICCV*, 2023, pp. 9033–9042.
[107] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, "Reconstructing hands in 3d with transformers," in *CVPR*, 2024, pp. 9826–9836.
[108] Y. Wang, Q. Yu, L. Lin, Z. Li, and H. Liu, "Language-driven ordinal learning for imbalanced head pose estimation," in *ICASSP*. IEEE, 2024, pp. 4495–4499.
[109] Y. Tian, T. Shao, T. Demizu, X. Wu, and H.-T. Wu, "Hpe-cogvlm: New head pose grounding task exploration on vision language model," *arXiv preprint arXiv:2406.01914*, 2024.
[110] H. Zhou, F. Jiang, J. Si, and H. Lu, "Joint multi-person body detection and orientation estimation via one unified embedding," *arXiv preprint arXiv:2210.15586*, 2022.
[111] C. Nakatani, H. Kawashima, and N. Ukita, "Interaction-aware joint attention estimation using people attributes," in *ICCV*, 2023, pp. 10 224–10 233.
[112] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *ICCV*, 2015, pp. 3334–3342.
[113] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *RSS*, 2023.
[114] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *IJRR*, p. 02783649241273668, 2023.