

ReVA-HPE: Relative 6DoF Head Pose Estimation with Synthetic Pose Supervision and Personalized Virtual Anchors

Vasiliki Vasileiou^{1,2} Panagiotis P. Filntisis¹ Petros Maragos^{1,2} Kostas Daniilidis^{3,4}

¹HERON - Hellenic Robotics Center of Excellence, Athena Research Center, Greece

²National Technical University of Athens, School of Electrical and Computer Engineering, Greece

³Robotics Institute, Athena Research Center, Greece ⁴University of Pennsylvania, USA

v.vasileiou@athenarc.gr

Abstract

Monocular head pose estimation is usually formulated as absolute regression from a single face crop, yet many applications observe the same person over time and can supply an explicit anchor image. Recently, relative head pose estimation was proposed, which casts the problem as estimation of the relative geometric change between an anchor and a query image. However, existing methods still require the anchor image pose to be known at inference, significantly limiting their practical use. We first extend relative head pose estimation to full 6DoF by additionally predicting a depth-normalized relative translation, which composes with the anchor to recover metric query pose in the camera frame. To eliminate the need for a ground-truth anchor pose, we reconstruct a personalized avatar from a single unlabeled enrollment image and render a bank of virtual anchors with known poses. The bank provides robust rotation hypotheses through multi-anchor consensus, while the real enrollment image serves as the real anchor in the translation branch via landmark PnP. We also introduce an efficient variant that amortizes the bank cost once at enrollment, after which each query requires a single real-to-real relative prediction, with limited accuracy sacrifice. The relative pose estimator is trained exclusively on synthetic face renderings, without real pose supervision, pseudo-labels, domain adaptation, or fine-tuning on the evaluation sets. On real 6DoF benchmarks, ReVA-HPE achieves state-of-the-art rotation accuracy on BIWI and, given a near-frontal enrollment image, on FaMoS, while remaining competitive in metric translation despite relying solely on synthetic pose supervision. Project page and code: <https://vasilikivas.github.io/ReVA-HPE/>

1. Introduction

Head pose provides a compact geometric cue to a person’s attention and interaction with the surroundings. Reliable estimation of this cue supports applications including gaze anal-

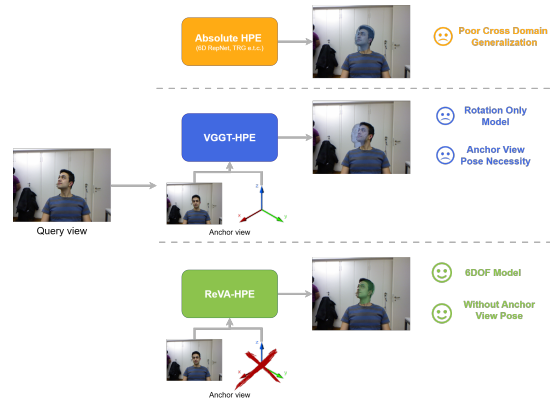


Figure 1. Absolute HPE predicts pose directly from the query and may generalize poorly across domains. VGGT-HPE uses relative prediction but requires a reference image with known pose and estimates rotation only. ReVA-HPE removes this requirement, using personalized virtual anchors for rotation and PnP on an unlabeled enrollment image for translation, and estimates full 6DoF head pose without any ground-truth pose.

ysis, driver monitoring, facial understanding, telepresence, sign-language systems, and human-robot interaction [1, 23].

Most learned head pose estimators focus on orientation, predicting yaw, pitch, and roll- or an equivalent rotation representation-directly from a single face crop [13, 25, 34, 37]. A smaller body of work targets full 6DoF pose, additionally recovering the metric translation of the head with respect to the camera [2, 5, 15]. This setting is considerably more challenging: rotation must remain robust to extreme poses and appearance variation, while translation also depends on camera intrinsics, crop geometry, subject scale, and the precise definition of the anatomical head origin.

Recently, VGGT-HPE [27] reformulated head pose estimation as relative pose prediction: given an anchor and a query image, the model estimates their relative geometric transformation and composes it with the anchor pose to recover the query pose. This image-matching formulation enabled VGGT-HPE to outperform leading absolute-regression methods on real-face benchmarks, including meth-

ods using real or mixed pose supervision, despite relying solely on synthetic pose supervision.

However, two limitations remained. First, VGGT-HPE was evaluated mainly as an orientation estimator. A practical 6DoF system also needs a well-defined relative translation target, correct metric decoding, and camera-frame translation evaluation. Second, its strongest protocol still assumed a known anchor pose. Even one ground-truth anchor pose per sequence is an oracle, yet the obvious fix is self-defeating: an absolute estimator accurate enough to supply that would already solve the problem that motivates the relative formulation. Replacing the ground truth anchor with a weaker absolute estimate injects a persistent offset into every composed query, while autoregressive anchoring drifts over time.

In this paper, we turn relative head pose estimation into a complete 6DoF system that does not require a ground-truth anchor pose, called Relative Virtual-Anchor Head Pose Estimation (ReVA-HPE) and summarized in Figure 1. We first extend the previous rotation-centered formulation to predict both relative rotation and relative translation, allowing the query head pose to be recovered metrically in the camera frame. We then remove the known-anchor-pose assumption by constructing, from a single unlabeled enrollment image, a personalized virtual-anchor bank with known poses. We estimate the query rotation by robustly combining predictions from the virtual anchors in the bank. Metric translation is obtained separately through a single PnP solve on the enrollment image. We also introduce an efficient variant that processes the virtual-anchor bank only once, during enrollment. After this one-time calibration, each new query requires a single network inference, with only a small loss in accuracy. The relative pose estimator is trained exclusively on procedural face renderings, without real pose supervision, pseudo-labels, or target-domain fine-tuning. Despite this restricted supervision, ReVA-HPE achieves state-of-the-art rotation accuracy on BIWI and, given a near-frontal enrollment image, on FaMoS, while remaining competitive in metric translation. Our contributions are:

- **Relative 6DoF head pose estimation.** We extend the previously proposed relative head pose rotation formulation to full 6DoF by learning relative translation together with relative rotation and decoding metric query pose in the camera frame.
- **Ground-truth-free virtual anchors.** We remove the ground-truth anchor pose requirement by replacing the labeled real anchor with a personalized virtual anchor set constructed from one unlabeled enrollment image.
- **Efficient enrollment calibration.** We introduce a lightweight variant that uses the virtual anchor bank once at enrollment to calibrate its rotation, combines it with a PnP-grounded translation estimate, and processes every subsequent query with a single real-to-real relative prediction.

2. Related Work

2.1. Monocular Head Pose and 6DoF Face Pose

Early head-pose systems relied on facial correspondences, 3D templates, and perspective pose solvers. Modern approaches predominantly learn an absolute mapping from a face crop to Euler angles or a rotation representation. HopeNet combines classification and regression [25], FSA-Net aggregates fine-grained facial structure [33], WHENet targets wide-range orientation [37], 6DRepNet uses a continuous 6D rotation representation with geodesic supervision [13], and TokenHPE learns transformer orientation tokens [34]. These methods are effective for orientation but learn a dataset-specific absolute reference from training images.

Fewer methods evaluate complete 6DoF head pose. img2pose directly predicts face pose and detection [2], PerspNet reconstructs facial geometry under perspective projection and recovers 6DoF pose [15], and TRG explicitly couples head translation, rotation, and facial geometry [5]. Landmark PnP remains a strong geometric baseline when calibrated intrinsics and reliable 2D-3D correspondences are available [9, 17]. Our method differs in predicting both rotation and translation *relative to a visible anchor*, then using virtual anchors to remove the need for a GT pose.

Recent methods address robustness and uncertainty in complementary ways. DSFNet [19] fuses image and 3D representation spaces for occlusion-robust dense face alignment, while HeadDiff [31] models uncertainty in head rotation using a diffusion process. OPAL [6] studies rotation representations across different pose ranges, and SemiUHPE [36] uses semi-supervised learning to improve unconstrained head-pose estimation. These methods primarily target rotation or facial alignment, whereas ReVA-HPE estimates metric 6DoF pose relative to an explicit visual anchor and removes the need for a labeled anchor.

2.2. Relative Pose & Geometry Foundation Models

Relative transformations are standard in visual odometry, SLAM, tracking, and reference-based object pose estimation because correspondences across observations constrain displacement directly [26, 30]. Recent geometry foundation models generalize this idea across broad image domains. CroCo learns cross-view completion [32], DUST3R predicts point maps without known calibration [29], MAST3R grounds feature matching in 3D [18], and VGGT jointly predicts cameras and dense scene geometry from one or more images [28]. VGGT-HPE [27] adapts VGGT’s camera branch to paired head images with LoRA. In this work we keep the relative image-pair formulation, extend it to explicit relative translation, and address the test-time anchor problem by constructing RGB-derived anchor sources from a single enrollment image.

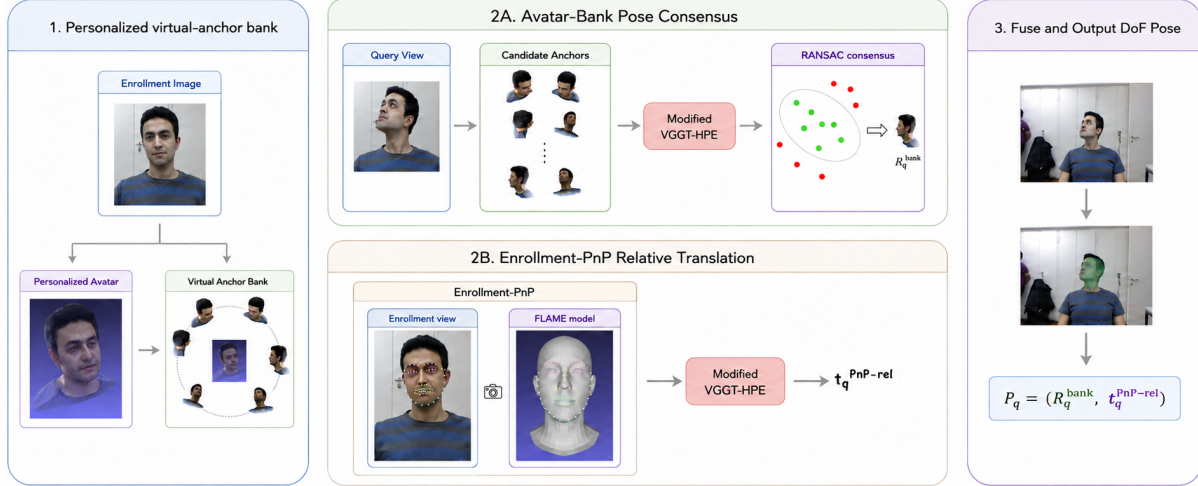


Figure 2. **ReVA-HPE overview.** From a single unlabeled enrollment image, a personalized Gaussian avatar is reconstructed and rendered at M known poses to form the virtual anchor bank. **(2A)** For each query, the relative 6DoF model compares the query against all bank anchors, producing M rotation candidates that are fused by pose-space RANSAC consensus into $\hat{R}_q^{\text{bank}}$. **(2B)** In parallel, 68 landmarks detected on the enrollment image ground a metric PnP solve. The relative model then predicts the enrollment-to-query translation update, yielding $\hat{t}_q^{\text{PnP-rel}}$. Modified VGGT-HPE denotes the frozen VGGT camera branch adapted with LoRA and trained to predict relative rotation together with the depth-normalized translation update $(\delta_x, \delta_y, \delta_z)$ from an ordered image pair. The two components are fused into the final 6DoF pose $\hat{P}_q = (\hat{R}_q^{\text{bank}}, \hat{t}_q^{\text{PnP-rel}})$ without any ground-truth anchor pose at inference.

2.3. Synthetic Pose Supervision

Synthetic supervision in head pose estimation commonly relies on pose-expanded real images. For example, 300W-LP [38] uses 3D face profiling to generate large-pose views from real photographs and is widely used to train absolute estimators such as 6DRepNet [13] and TokenHPE [34]. Procedural face models provide a different form of supervision: FLAME [20] allows direct control over identity, expression, viewpoint, and pose without deriving training examples from real images. VGGT-HPE [27] showed that relative rotation learned from such procedural renderings transfers effectively to real faces. We build on this setting, extending the relative target to full 6DoF while removing the need for a known anchor pose. The task-specific adaptation of our relative estimator uses no real head-pose images, pose annotations, pseudo-labels, or target-domain fine-tuning.

2.4. Personalized Avatars as Virtual References

Recent 3D Gaussian-splatting methods have enabled controllable 3D human-head avatars and high-quality novel-view synthesis [16]. Single-image or few-image avatar models reconstruct personalized human heads for view generation or animation, including FaceLift [21], LAM [12], and SEGA [10]. These works primarily target visual fidelity, editability, and controllable rendering.

In contrast, the present work uses a personalized avatar as a calibration device for relative pose estimation. The reconstructed avatar renders identity-consistent references with known synthetic poses, which provide multiple anchor-

conditioned rotation hypotheses. Metric translation is not taken from the avatar; it is grounded by the real enrollment image through PnP. Importantly, FaceLift is trained entirely on synthetic data and is used off-the-shelf, without adaptation to the evaluation sets.

3. Method

We build on VGGT-HPE [27], which estimates head pose by predicting the geometric change between an anchor and a query image. Two components are missing for a practical 6DoF system. First, the original relative prediction covers only rotation. We add an explicit depth-normalized translation target, allowing us to recover metric camera-frame translation once the translation of the anchor is known. Second, VGGT-HPE requires the ground-truth pose of the anchor image at test time. We remove this requirement by replacing the labeled anchor with a single unlabeled enrollment image. From this image, we construct a personalized bank of virtual anchors for rotation and use landmark-based PnP to obtain the metric translation of the enrollment anchor. The full pipeline is illustrated in Figure 2.

3.1. VGGT-HPE Preliminaries

VGGT-HPE [27] reframes head pose estimation as relative prediction. Instead of regressing an absolute pose from a single crop, it receives an anchor image and a query image and predicts the geometric change between them, which is then composed with the anchor image ground-truth pose. This replaces the implicit canonical frame learned by abso-

lute regressors with an explicit visual reference. We keep the pairwise VGGT-HPE architecture and relative-rotation composition, and extend its camera target with explicit depth-normalized translation.

Pose convention. We work in the RGB camera frame throughout. A head pose is

$$\mathbf{P} = (\mathbf{R}, \mathbf{t}), \quad \mathbf{R} \in \text{SO}(3), \quad \mathbf{t} = [x, y, z]^\top \in \mathbb{R}^3, \quad (1)$$

where $\mathbf{R}$ maps canonical head coordinates into the camera frame and $\mathbf{t}$ locates the canonical head origin. The model always receives an ordered pair: an anchor I_a with known pose $\mathbf{P}_a = (\mathbf{R}_a, \mathbf{t}_a)$ and a query I_q , and predicts the geometric change from anchor to query.

Backbone and adaptation. We reuse VGGT’s multi-view camera branch [28], keeping the backbone frozen and disabling the depth, point-map, and tracking heads. Rank-8 LoRA adapters with scale 16 [14] are inserted into the attention and MLP projections of the frame and global transformer blocks, adapting a small number of parameters to facial geometry while preserving the pretrained geometric prior.

Relative rotation. From the adapted pair, the model predicts the anchor-to-query rotation

$$\Delta \mathbf{R}_{q \leftarrow a} = \mathbf{R}_q \mathbf{R}_a^\top, \quad (2)$$

and recovers the query orientation by composing it with the anchor pose,

$$\widehat{\mathbf{R}}_q = \widehat{\Delta \mathbf{R}_{q \leftarrow a}} \mathbf{R}_a. \quad (3)$$

3.2. Relative 6DoF Pose Learning

3.2.1. Depth-Normalized Relative Translation

Translation is less direct than rotation. A homogeneous relative-pose translation couples translation to the anchor rotation; with a distant head, small rotation errors can induce large apparent translation errors. Directly regressing $\mathbf{t}_q - \mathbf{t}_a$ is also strongly scale-dependent. We therefore predict a depth-normalized displacement,

$$\delta_x = \frac{x_q - x_a}{z_a}, \quad \delta_y = \frac{y_q - y_a}{z_a}, \quad \delta_z = \log \frac{z_q}{z_a}. \quad (4)$$

The lateral components are normalized by anchor depth, while the log-depth ratio makes forward and backward motion approximately symmetric.

At inference, once an anchor translation $\mathbf{t}_a = [x_a, y_a, z_a]^\top$ is available, the metric query translation is decoded as

$$\widehat{z}_q = z_a \exp(\widehat{\delta}_z), \quad (5)$$

$$\widehat{x}_q = x_a + \widehat{z}_q \widehat{\delta}_x, \quad \widehat{y}_q = y_a + \widehat{z}_q \widehat{\delta}_y. \quad (6)$$

The output $(\widehat{\Delta \mathbf{R}_{q \leftarrow a}}, \widehat{\delta}_x, \widehat{\delta}_y, \widehat{\delta}_z)$ forms our relative 6DoF representation. Its translation component is not the translation block of a homogeneous $\text{SE}(3)$ product; Eqs. (5)-(6) define the verified decoder used by all reported experiments.

Branch-wise composition. Under Eqs. (3)-(6), the anchor rotation enters only the rotation composition, while the anchor translation enters only the translation decoder. This separation is important in the ground-truth-free setting: the two anchor components may be estimated from different measurements as long as they are expressed in the same RGB camera coordinate frame. We use this property in Sec. 3.3: personalized virtual anchors provide robust rotation hypotheses, while the real enrollment image provides a metric translation gauge.

3.2.2. Training Objective

VGGT emits camera predictions at K intermediate stages. We supervise every stage with exponentially decayed weights,

$$\mathcal{L}_{\text{cam}} = \frac{1}{K} \sum_{k=1}^K \gamma^{K-k} \left(\lambda_R \mathcal{L}_R^{(k)} + \lambda_T \mathcal{L}_T^{(k)} + \lambda_F \mathcal{L}_F^{(k)} \right), \quad (7)$$

where $\mathcal{L}_R$ is an ℓ_1 quaternion loss, $\mathcal{L}_T = \|\widehat{\delta} - \delta\|_1$, and $\mathcal{L}_F$ supervises the inter-frame focal-length ratio in log-tangent space. We use $\lambda_R = 1$, $\lambda_T = 1$, and $\lambda_F = 0.5$. Joint camera supervision is retained because removing translation and field-of-view supervision degrades cross-domain rotational accuracy in the preliminary ablation [27].

3.2.3. Controlled Synthetic Training Data

The head-pose network is trained *only* on procedural FLAME renders [20]: 250 identities - 200 with synthetic hairstyles, 50 bald - rendered under two HDR environment maps, ten expressions, and five viewpoints per expression, giving 25,000 base images across 54 skin-tone textures. Pair members share identity and illumination but may differ in expression and viewpoint.

Compared with VGGT-HPE, we use a more controlled geometric protocol: a fixed list of 10^6 predefined anchor-query pairs instead of online random sampling, fixed-focal copies of the data, and a unified pair crop with margin 0.5 and analytically updated intrinsics. Appearance augmentation is retained; geometry-changing randomness from focal variation and independent crop perturbations is removed. Only LoRA parameters are optimized with AdamW and mixed precision, the VGGT backbone remains frozen. Examples of our synthetic data can be found in the Suppl. Mat.

3.3. Removing the Ground-Truth Anchor Pose

The relative predictor requires an anchor pose to recover an absolute query pose. The GT-anchor oracle provides the ground-truth pose of the enrollment anchor from benchmark annotations, but such a protocol is not deployable. We instead assume a single *unlabeled* enrollment image I_e for each subject. A personalized virtual-anchor bank provides query-rotation hypotheses from anchors with known synthetic poses, while enrollment PnP estimates the metric

translation of the real enrollment anchor. Our final method therefore obtains query rotation through virtual-anchor consensus and query translation through enrollment-to-query relative decoding.

3.3.1. Avatar-Bank Pose Consensus

Anchor poses are known by construction, as each anchor is rendered from a viewpoint sampled over fixed front-hemisphere directions in canonical head coordinates. The avatar is registered to this frame at enrollment by triangulating 68 landmarks across calibrated rendered views and aligning them to the enrollment FLAME fit with a RANSAC-robust similarity transform. The Suppl. Mat. quantifies the alignment residual. For each query I_q , we compare the query with every rendered virtual anchor:

$$(\widehat{\Delta \mathbf{R}}_{q \leftarrow m}^{(m)}, \widehat{\delta}_{q \leftarrow m}^{(m)}) = f_\theta(A_m, I_q). \quad (8)$$

Each virtual anchor gives one absolute 6DoF candidate,

$$\widehat{\mathbf{R}}_q^{(m)} = \widehat{\Delta \mathbf{R}}_{q \leftarrow m}^{(m)} \mathbf{R}_m, \quad \widehat{\mathbf{t}}_q^{(m)} = \text{Decode}(\widehat{\delta}_{q \leftarrow m}^{(m)}, \mathbf{t}_m). \quad (9)$$

The bank candidates are fused by pose-consensus RANSAC. Each candidate i is used as a seed. For every other candidate j , we compute a rotation residual and a translation residual,

$$r_R^{(i,j)} = d_R(\widehat{\mathbf{R}}_q^{(j)}, \widehat{\mathbf{R}}_q^{(i)}), \quad r_t^{(i,j)} = \|\widehat{\mathbf{t}}_q^{(j)} - \widehat{\mathbf{t}}_q^{(i)}\|_2, \quad (10)$$

where d_R is geodesic rotation distance. Candidate j is an inlier to seed i when

$$r_R^{(i,j)} \leq \tau_R, \quad r_t^{(i,j)} \leq \tau_t. \quad (11)$$

We use $\tau_R = 10^\circ$, $\tau_t = 50 \text{ mm}$, and require at least three inliers. These values are fixed globally and are not retuned on BIWI or FaMoS. A synthetic-only sensitivity analysis, including corrupted anchor hypotheses, shows that this operating point lies within a stable region spanning approximately $\tau_R = 7.5^\circ - 15^\circ$ and $\tau_t \geq 50 \text{ mm}$, full results are provided in the supplementary material. In the final ReVA-HPE method, τ_t is used only to determine the internally consistent bank hypotheses contributing to rotation. Each seed i is scored first by its inlier count $|\mathcal{I}_i|$. Ties are broken by the mean normalized residual

$$s_i = \frac{1}{|\mathcal{I}_i|} \sum_{j \in \mathcal{I}_i} \left(\frac{r_R^{(i,j)}}{\tau_R} + \frac{r_t^{(i,j)}}{\tau_t} \right). \quad (12)$$

Let $\mathcal{I}^*$ be the best consensus set. The bank rotation is the projected mean of the inlier rotations,

$$\widehat{\mathbf{R}}_q^{\text{bank}} = \Pi_{\text{SO}(3)} \left(\frac{1}{|\mathcal{I}^*|} \sum_{j \in \mathcal{I}^*} \widehat{\mathbf{R}}_q^{(j)} \right), \quad (13)$$

where $\Pi_{\text{SO}(3)}$ denotes SVD projection onto $\text{SO}(3)$ [22].

3.3.2. Enrollment-PnP Relative Translation

For translation, the enrollment image I_e serves as the real anchor. It fixes the subject under the target camera and provides a metric camera ray through landmark PnP. We detect 68 2D landmarks in I_e , pair them with the landmarks of the mean-neutral FLAME head, with zero shape and expression parameters at the model's metric scale, and solve an EPnP-RANSAC problem using the RGB intrinsics to obtain the enrollment translation:

$$\widehat{\mathbf{t}}_e^{\text{PnP}} = [\widehat{x}_e^{\text{PnP}}, \widehat{y}_e^{\text{PnP}}, \widehat{z}_e^{\text{PnP}}]^\top. \quad (14)$$

The query is then paired with the enrollment anchor:

$$(\widehat{\Delta \mathbf{R}}_{q \leftarrow e}, \widehat{\delta}_{q \leftarrow e}) = f_\theta(I_e, I_q). \quad (15)$$

For the final method, we use only the predicted relative translation update from this real-to-real pair. The query translation is decoded using the enrollment-PnP translation:

$$\widehat{\mathbf{t}}_q^{\text{PnP-rel}} = \text{Decode}(\widehat{\delta}_{q \leftarrow e}, \widehat{\mathbf{t}}_e^{\text{PnP}}), \quad (16)$$

where Decode denotes Eqs. (5)-(6). This branch requires landmark detection only once, at enrollment, and provides the translation component through one real-to-real relative prediction per query.

3.3.3. Final Selected Method

We refer to the combination of avatar-bank rotation and enrollment-PnP relative translation as **ReVA-HPE**. The final query pose is

$$\widehat{\mathbf{P}}_q = (\widehat{\mathbf{R}}_q^{\text{bank}}, \widehat{\mathbf{t}}_q^{\text{PnP-rel}}) \quad (17)$$

where each component is sourced from its most reliable estimate: the bank for rotation and enrollment PnP for metric translation. Bank translation is used only internally during pose-consensus RANSAC.

4. Experiments

Datasets and Protocols. The BIWI Kinect Head Pose Database [8] contains 24 calibrated RGB-D recordings of 20 individuals with frame-level 6DoF head poses. We treat each recording as a separate subject. For every subject, the first available RGB frame is fixed as enrollment before any score is computed and removed from the query set, leaving 15,654 queries. We additionally evaluate on the FAMOS [3] test set to assess transfer beyond BIWI. We use only camera index 0 throughout the evaluation. Frame-level 6DoF labels are obtained from the provided ground-truth registrations: for each subject, we rigidly align its identity-specific FLAME-topology template to each registered frame using Procrustes analysis, and use the resulting transformation as the head pose in the dataset camera frame. We follow the

Method	Yaw ↓	Pitch ↓	Roll ↓	MAE _r ↓	T _x ↓	T _y ↓	T _z ↓	MAE _t ↓	Pose Supervision
<i>Reported numbers</i>									
Direct 6DoF Regress [15]	16.49	14.03	5.81	12.11	62.36	85.01	366.52	171.30	R
Refined Pix2Pose [15, 24]	5.75	5.06	11.23	7.35	16.82	21.30	255.36	97.83	R
MICA [39]	5.40	7.17	3.80	5.46	9.32	13.66	60.13	27.70	M
JMLR [11]	6.31	6.17	3.72	5.40	8.66	7.27	32.63	16.19	M
PerspNet [15]	3.10	3.37	2.38	2.95	4.15	6.43	46.69	19.09	R
TRG [5]	3.28	3.52	<u>1.87</u>	<u>2.89</u>	8.41	7.38	<u>27.13</u>	<u>14.31</u>	R
TRG* [5]	3.04	<u>3.44</u>	1.78	2.75	<u>7.83</u>	<u>6.99</u>	24.07	12.97	M
OPAL [6]	4.54	5.05	2.80	4.13	—	—	—	—	M
<i>Reproduced under shared MTCNN protocol</i>									
6DRepNet [13]	3.74	4.95	3.04	3.91	—	—	—	—	M
TokenHPE-v1 [34]	5.57	6.23	3.79	5.20	—	—	—	—	M
TRG* [5]	4.58	7.18	3.68	5.15	13.03	13.59	119.52	48.71	M
SemiUHPE [36]	3.30	3.71	<u>2.60</u>	3.20	—	—	—	—	R
OPAL [6]	3.74	7.23	3.25	4.74	—	—	—	—	M
VGGT-HPE-Abs [27]	4.90	7.01	3.53	5.15	—	—	—	—	S
PnP-enrollment + Rel.	2.27	4.19	3.41	3.29	8.43	11.14	31.74	17.10	S
ReVA-HPE (ours, first frame)	<u>2.20</u>	<u>2.94</u>	2.47	<u>2.53</u>	8.43	11.14	31.74	17.10	S
ReVA-HPE (ours, $\alpha \leq 15^\circ$)	2.01	2.84	2.61	2.49	<u>8.77</u>	<u>11.73</u>	<u>36.34</u>	<u>18.94</u>	S
<i>Methods using a GT anchor pose</i>									
VGGT-HPE [†] [27]	2.23	3.07	3.38	2.89	47.32	20.93	62.21	43.48	S
ReVA-HPE-GT [†] (ours)	1.60	1.92	2.72	2.08	7.53	6.31	14.86	9.57	S

[†] Requires the ground-truth pose of the enrollment image at test time.

Table 1. BIWI comparison. Rotation is in degrees and translation in millimeters, lower is better. Reported values follow [5], reproduced results use the shared MTCNN protocol. “First frame” fixes each subject’s first available frame as enrollment, without pose-based selection. **Bold** and underlined denote the best and second-best results within each non-oracle block. The GT-anchor block is ranked independently, with its best values in **bold**. S, R, and M denote synthetic, real, and mixed pose supervision.

same protocol as BIWI, using the first available frame for enrollment and the remaining frames as queries.

Metrics. As earlier works, for orientation, we report yaw, pitch, and roll mean absolute error (MAE) and their average, MAE_r. For translation, we report the per-axis mean absolute errors and their arithmetic mean, MAE_t, in millimeters.

Implementation Details. Following the literature [13, 34], we detect face crops with MTCNN [35] (We also include an ablation study using RetinaFace[7] in the Suppl. Mat.) FaceLift [21] is used with its released pretrained model trained only on synthetic data to create the personalized avatar. We initialize the frozen backbone from the released general-purpose VGGT checkpoint and adapt its camera branch through LoRA. The MTCNN and the FAN-style landmark detector [4] use released checkpoints pretrained on real face data. We use EPnP [17] inside 300-iteration RANSAC [9], with a 5-pixel threshold, confidence 0.999, and at least six inliers, followed by iterative inlier refinement. We additionally evaluate the enrollment EPnP-RANSAC re-projection threshold using synthetic landmark projections with controlled noise and outliers. The fixed 5-pixel threshold lies at the tighter end of a stable 5–10 pixel region, the complete synthetic sensitivity analysis is provided in the Suppl. Mat. Neither the pose-consensus nor EPnP sensitivity analysis uses BIWI or FaMoS labels for parameter selection.

Baselines. We compare against head pose estimation methods spanning both rotation-only and 6DoF approaches:

6DRepNet [13], OPAL [6], SemiUHPE [36], PerspNet [15], and TRG [5]. All baselines are trained on real-world data or mixed regimes (e.g., real images expanded via synthetic pose profiling such as 300W-LP [38], ARKitFace [15], or combinations thereof). We also evaluate **PnP-enrollment+Rel.**, which uses the same relative network but directly uses the enrollment-PnP rotation, without virtual-anchor consensus. This serves as the direct ablation of the bank’s contribution to rotation.

4.1. Quantitative Results

Tables 1 and 2 present quantitative results on BIWI and FaMoS datasets.

BIWI. We report results from the original publications for context, together with results reproduced on the same test frames using a shared preprocessing pipeline. For the reproduced comparison, we use MTCNN because it follows the established BIWI evaluation protocol of 6DRepNet and subsequent work [13, 27]. In the Suppl. Mat., we verify that this choice does not favor ReVA-HPE by repeating the evaluation with RetinaFace and showing that the detector choice minimally changes method results.

Our default first-frame protocol fixes each subject’s first available frame as the enrollment image, without pose-based selection. We additionally report a guided protocol using an enrollment image with $\alpha = \max(|\text{yaw}|, |\text{pitch}|, |\text{roll}|) \leq 15^\circ$. Ground-truth angles are used only to construct this diagnostic protocol and are never supplied to ReVA-HPE. In

Method	Yaw ↓	Pitch ↓	Roll ↓	MAE _r ↓	T _x ↓	T _y ↓	T _z ↓	MAE _t ↓	Pose Supervision
<i>Reproduced under shared MTCNN protocol</i>									
6DRepNet [13]	4.54	8.03	5.36	5.98	—	—	—	—	M
TokenHPE-v1 [34]	8.62	5.69	4.46	6.26	—	—	—	—	M
TRG* [5]	6.18	6.26	4.32	5.59	7.52	5.06	91.40	34.66	M
SemiUHPE [36]	1.98	<u>3.65</u>	<u>2.94</u>	<u>2.86</u>	—	—	—	—	R
OPAL [6]	8.33	5.38	5.14	6.28	—	—	—	—	M
VGGT-HPE-Abs [27]	34.41	10.60	9.73	18.29	27.14	9.15	85.10	40.46	S
PnP-enrollment + Rel.	9.32	35.08	15.62	20.01	20.73	11.82	<u>54.83</u>	<u>29.13</u>	S
ReVA-HPE (ours, first frame)	<u>1.34</u>	6.69	3.31	3.78	20.73	11.82	<u>54.83</u>	<u>29.13</u>	S
ReVA-HPE (ours, $\alpha \leq 15^\circ$)	1.21	3.39	2.80	2.47	<u>11.18</u>	<u>7.49</u>	49.15	22.61	S
<i>Methods using a GT anchor pose</i>									
VGGT-HPE-Rel [†] [27]	2.93	4.91	5.47	4.44	24.85	15.00	33.68	24.51	S
ReVA-HPE-GT [†] (ours)	1.38	1.77	2.70	1.95	12.43	3.68	13.49	9.87	S

[†] Requires the ground-truth pose of the enrollment image at test time.

Table 2. FaMoS comparison under the shared MTCNN protocol. Rotation is in degrees and translation in millimeters, lower is better. “First frame” uses each subject’s first available frame as enrollment, without pose-based selection. In the non-oracle block, **bold** and underlined denote the best and second-best results. The GT-anchor block is ranked independently, with its best values in **bold**. S, R, and M denote synthetic, real, and mixed pose supervision.

practice, this condition can be encouraged simply by asking the user to face the camera while providing the enrollment image. A detailed enrollment-pose ablation across all angle ranges is presented later.

Under the shared protocol, the default first-frame ReVA-HPE obtains 2.53° , outperforming all reproduced competing baselines, the next best is SemiUHPE at 3.20° . With the guided $\alpha \leq 15^\circ$ enrollment protocol, ReVA-HPE improves further to 2.49° . Both results are also numerically below the best separately reported result, TRG* at 2.75° , although these cross-protocol numbers are provided only for context.

The comparison with PnP-enrollment+Rel. shows where this improvement comes from. Both variants use the same relative pose estimator and the same translation branch. The difference is how query rotation is obtained: PnP-enrollment+Rel. relies on a single PnP estimate of the enrollment rotation, whereas ReVA-HPE combines evidence from the virtual-anchor bank. This reduces rotation error from 3.29° to 2.53° , showing that the bank provides a more reliable query-rotation estimate. Their translation results are identical by design for the same enrollment image because both use the same enrollment-PnP translation and relative translation decoder.

Translation remains the more difficult part of the problem. The separately reported TRG* result, which uses mixed pose supervision, has lower translation error than ReVA-HPE, although this comparison is not protocol-matched. Under the shared protocol, ReVA-HPE remains competitive without real pose supervision. Finally, the GT-anchor variant reduces rotation error further to 2.08° . This result measures the remaining cost of estimating the anchor automatically and indicates that improved enrollment calibration could further improve performance.

FaMoS. FaMoS is particularly challenging because camera 0 is positioned below the subject. This introduces a

strong pitch bias in the PnP estimate obtained from the enrollment image. As a result, PnP-enrollment+Rel. reaches a pitch error of 35.08° and an overall rotation error of 20.01° . ReVA-HPE avoids using this PnP rotation and instead estimates query rotation by combining predictions from multiple virtual anchors. Under the default first-frame protocol, this reduces the rotation error to 3.78° , outperforming all reproduced non-oracle baselines except SemiUHPE, which achieves 2.86° . Under the guided $\alpha \leq 15^\circ$ protocol defined above, ReVA-HPE improves further to 2.47° . As shown later in Table 5, SemiUHPE obtains 3.00° on exactly the same evaluation cases. ReVA-HPE therefore achieves the best rotation accuracy in this guided protocol while using only synthetic pose supervision for its task-specific adaptation. The same table provides a detailed analysis across the remaining enrollment-pose ranges.

The benefit is also clear for translation. Under the default first-frame protocol, ReVA-HPE achieves an MAE_t of 29.13 mm, compared with 34.66 mm for TRG and 40.46 mm for VGGT-HPE-Abs. Under the guided protocol, its translation error decreases further to 22.61 mm.

Finally, the ground-truth enrollment-pose oracle reaches 1.95° , showing that the relative predictor generalizes strongly to FaMoS and that much of the remaining rotation error comes from estimating the enrollment reference automatically.

Qualitative results of full 6DoF pose on BIWI and FaMoS visualizing both rotation and metric translation for all compared methods are included in the Suppl. Mat.

4.2. Ablation Studies

Bank size and efficient inference. Table 3 compares direct inference and VA-fast. Direct-1r evaluates M anchors at one radius, Direct-5r evaluates $5M$ candidates at five radii, and VA-fast uses the bank once at enrollment followed by one

real-to-real prediction per query. Accuracy largely saturates at $M = 20$: increasing Direct-5r to $M = 50$ changes MAE_r from 2.53° to 2.51° , while model time increases from 9.774 to 30.877 s. We therefore use $M = 20$. Avatar reconstruction and bank rendering are one-time enrollment costs.

Protocol	$M = 5$			$M = 10$		
	MAE_r	Time	Spd.	MAE_r	Time	Spd.
Direct-1r	4.30	0.683	—	2.92	1.165	—
Direct-5r	3.23	2.472	—	2.79	4.902	—
VA-fast	4.53	0.137	5.0/18.0	3.08	0.117	10.0/41.9

Protocol	$M = 20$			$M = 50$		
	MAE_r	Time	Spd.	MAE_r	Time	Spd.
Direct-1r	2.69	2.185	—	2.58	5.259	—
Direct-5r	2.53	9.774	—	2.51	30.877	—
VA-fast	2.85	0.109	20.0/89.7	2.97	0.105	50.1/294.1

Table 3. MAE_r is in degrees and time is model-only seconds per query. VA-fast speedups are relative to Direct-1r/Direct-5r.

Enrollment sensitivity. Tables 5 and 4 study complementary sources of enrollment error. Across the diagnostic pose bins, ReVA-HPE rotation degrades as the enrollment becomes more off-axis, gradually on BIWI and more strongly on FaMoS. SemiUHPE is query-only and is therefore unchanged by enrollment selection: ReVA-HPE is strongest for near-frontal enrollment, while SemiUHPE becomes stronger in the off-axis bins. Table 4(a) applies each image perturbation before face detection and reruns the complete pipeline, including FaceLift reconstruction, landmark processing, PnP, virtual-bank generation, consensus, and query inference. Downsampling has little effect, blur causes moderate degradation, and partial occlusion is the most damaging condition, particularly for metric translation. Panels (b–c) separately isolate the enrollment-PnP branch by perturbing only the 68 enrollment landmarks: mild jitter and random landmark removal are tolerated, whereas larger coordinate errors primarily corrupt the estimated depth, bank rotation remains unchanged by construction. The Suppl. Mat. provides the complete protocols, GT-use policy, all perturbation severities, coverage and failure accounting, consensus-inlier statistics, and the full landmark-jitter and removal results.

(a) Full-pipeline degradation			(b) Landmark jitter			(c) Landmark removal		
Condition	$\text{MAE}_R \downarrow$	$\text{MAE}_t \downarrow$	Condition	$T_z \downarrow$	$\text{MAE}_t \downarrow$	Condition	$T_z \downarrow$	$\text{MAE}_t \downarrow$
Clean	2.53	17.10	Clean	31.74	17.10	Clean	31.74	17.10
Blur, $\sigma = 2$	3.22	18.64	$\sigma = 1$ px	31.96	17.21	10%	32.52	17.41
Downsample 50%	2.77	15.97	$\sigma = 2$ px	31.92	17.40	25%	32.59	17.44
Occlusion 12.5%	4.46	38.46	$\sigma = 4$ px	45.55	22.35	40%	33.83	17.97

Table 4. Sensitivity to enrollment-image degradation and landmark errors on BIWI.

5. Limitations and Discussion

Our results lead to three main observations. First, synthetic pose supervision transfers well in the relative formulation: ReVA-HPE shows substantially less cross-dataset degradation than the matched synthetic absolute baseline. Domain shift nevertheless remains, as performance still depends on the dataset, anchor geometry, avatar quality, and camera

Dataset	Enrollment pose	Coverage	ReVA-HPE	SemiUHPE	ReVA-HPE
			$\text{MAE}_r \downarrow$	$\text{MAE}_r \downarrow$	$\text{MAE}_t \downarrow$
BIWI	$\alpha \leq 15^\circ$	99.85%	2.49	3.20	18.94
	$15^\circ < \alpha \leq 30^\circ$	99.85%	3.71	3.20	18.62
	$30^\circ < \alpha \leq 45^\circ$	99.85%	4.85	3.20	19.31
	$\alpha > 45^\circ$	99.85%	4.80	3.20	21.77
FaMoS	$\alpha \leq 15^\circ$	84.57%	2.47	3.00	22.61
	$15^\circ < \alpha \leq 30^\circ$	73.94%	6.02	2.39	26.83
	$30^\circ < \alpha \leq 45^\circ$	55.32%	10.04	2.40	26.59
	$\alpha > 45^\circ$	0%	—	—	—

Table 5. Sensitivity to enrollment pose on BIWI and FaMoS under the shared MTCNN protocol.

calibration. Second, the virtual-anchor bank remains effective even though no rendered view perfectly matches the query. Consensus across virtual anchors suppresses inconsistent hypotheses and improves rotation accuracy. Third, increasing the bank size helps up to $M=20$, after which performance plateaus or declines as additional virtual anchors extend farther from the frontal pose range.

Metric translation still depends on PnP applied to the real anchor image. This provides a simple and effective metric reference, but makes translation sensitive to landmark quality and enrollment pose. Extreme orientations can degrade landmark localization and the resulting PnP estimate. The real anchor should therefore preferably have a near-neutral pose, although moderate off-axis poses are well tolerated.

The main remaining limitation is the dependence on avatar and calibration quality. Errors in identity, hair, silhouette, or unseen profile geometry can weaken anchor consensus. Moreover, avatar reconstruction, landmark-based calibration, and relative pose estimation are separate stages and are not jointly optimized. Future work could optimize the avatar directly for geometric matching and exploit temporal continuity in video, since the current method treats each query independently.

6. Conclusion

We presented ReVA-HPE, which extends relative head pose estimation to complete metric 6DoF without requiring a ground-truth anchor pose at inference. The pose estimator uses only 25,000 procedural FLAME renderings, without real-image pose supervision, pseudo-labels, or target-domain fine-tuning. At inference, a personalized virtual-anchor bank provides robust rotation hypotheses through multi-anchor consensus, while a single PnP solve on the enrollment image grounds metric translation. On BIWI, ReVA-HPE achieves 2.53° rotation MAE and outperforms all baselines reproduced under the same protocol, including methods using real or mixed pose supervision.

Acknowledgments



Funded by
the European Union

This project is funded by the European Union under Horizon Europe (grant No. 101136568 – HERON).

References

- [1] Andrea F. Abate, Carmen Bisogni, Aniello Castiglione, and Michele Nappi. Head pose estimation: An extensive survey on recent techniques and applications. *Pattern Recognition*, 127:108591, 2022. 1
- [2] Vítor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7613–7623, 2021. 1, 2
- [3] Timo Bolkart, Tianye Li, and Michael J. Black. Instant multi-view head capture through learnable registration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 768–779, 2023. 5
- [4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1021–1030, 2017. 6
- [5] Sungho Chun and Ju Yong Chang. 6dof head pose estimation through explicit bidirectional interaction with face geometry. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXIII*, page 146–163, Berlin, Heidelberg, 2024. Springer-Verlag. 1, 2, 6, 7
- [6] Alejandro Cobo, Roberto Valle, José M. Buenaposada, and Luis Baumela. On the representation and methodology for wide and short range head pose estimation. *Pattern Recognition*, 149:110263, 2024. 2, 6, 7
- [7] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5202–5211, Los Alamitos, CA, USA, 2020. IEEE Computer Society. 6
- [8] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc van Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101: 437 – 458, 2012. 5
- [9] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, pages 726–740, 1987. 2, 6
- [10] Chen Guo, Zhuo Su, Jian Wang, Shuang Li, Xu Chang, Zhaohu Li, Yang Zhao, Guidong Wang, and Ru-Qi Huang. SEGA: Drivable 3d gaussian head avatar from a single image. *ArXiv*, abs/2504.14373, 2025. 3
- [11] J. Guo, Jin-Ke Yu, Alexandros Lattas, and Jian-Kang Deng. Perspective reconstruction of human faces by joint mesh and landmark regression. In *ECCV Workshops*, 2022. 6
- [12] Yi-Sheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Wei-Hao Yuan, Zi-Long Dong, and Liefeng Bo. LAM: Large avatar model for one-shot animatable gaussian head. *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, 2025. 3
- [13] Thorsten Hempel, Ahmed A. Abdelrahman, and Ayoub Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2496–2500, 2022. 1, 2, 3, 6, 7
- [14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 4
- [15] Yueying Kao, Bowen Pan, Miao Xu, Jiangjing Lyu, Xiangyu Zhu, Yuanzhang Chang, Xiaobo Li, and Zhen Lei. Toward 3d face reconstruction in perspective projection: Estimating 6dof face pose from monocular image. *IEEE Transactions on Image Processing*, 32:3080–3091, 2023. 1, 2, 6
- [16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 2023. 3
- [17] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An accurate o(n) solution to the pnp problem. *Int. J. Comput. Vision*, 81(2):155–166, 2009. 2, 6
- [18] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with MAST3R. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXII*, page 71–91, Berlin, Heidelberg, 2024. Springer-Verlag. 2
- [19] Heyuan Li, Bo Wang, Yu Cheng, Mohan Kankanhalli, and Robby T. Tan. DSFNet: Dual space fusion network for occlusion-robust 3d dense face alignment. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4531–4540, 2023. 2
- [20] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (TOG)*, 36:1 – 17, 2017. 3, 4
- [21] Weijie Lyu, Yi Zhou, Ming-Hsuan Yang, and Zhixin Shu. FaceLift: Learning generalizable single image 3d face reconstruction from synthetic heads. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12691–12701, 2025. 3, 6
- [22] Maher Moakher. Means and averaging in the group of rotations. *SIAM Journal on Matrix Analysis and Applications*, 24(1):1–16, 2002. 5
- [23] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4): 607–626, 2009. 1
- [24] Kiru Park, Timothy Patten, and Markus Vincze. Pix2Pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7667–7676, 2019. 6
- [25] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2155–215509, 2018. 1, 2
- [26] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual slam for monocular, stereo, and rgb-d cameras. In *Advances in*

- Neural Information Processing Systems*, pages 16558–16569. Curran Associates, Inc., 2021. [2](#)
- [27] Vasiliki Vasileiou, Panagiotis P Filntisis, Petros Maragos, and Kostas Daniilidis. VGGT-HPE: Reframing head pose estimation as relative pose prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5464–5473, 2026. [1](#), [2](#), [3](#), [4](#), [6](#), [7](#)
- [28] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5306, 2025. [2](#), [4](#)
- [29] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUST3R: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024. [2](#)
- [30] Wenshan Wang, Yaoyu Hu, and Sebastian A. Scherer. TartanVO: A generalizable learning-based vo. In *Conference on Robot Learning*, 2020. [2](#)
- [31] Yaoxing Wang, Hao Liu, Yaowei Feng, Zhendong Li, Xiangjuan Wu, and Congcong Zhu. HeadDiff: Exploring rotation uncertainty with diffusion models for head pose estimation. *IEEE Transactions on Image Processing*, 33:1868–1882, 2024. [2](#)
- [32] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Bréger, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. CroCo: Self-supervised pre-training for 3d vision tasks by cross-view completion. *ArXiv*, abs/2210.10716, 2022. [2](#)
- [33] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. FSA-Net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1087–1096, 2019. [2](#)
- [34] Cheng Zhang, Hai Liu, Yongjian Deng, Bochen Xie, and Youfu Li. TokenHPE: Learning orientation tokens for efficient head pose estimation via transformers. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8897–8906, 2023. [1](#), [2](#), [3](#), [6](#), [7](#)
- [35] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. [6](#)
- [36] Huayi Zhou, Fei Jiang, Jin Yuan, Yong Rui, Hongtao Lu, and Kui Jia. Semi-supervised unconstrained head pose estimation in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3082–3099, 2026. [2](#), [6](#), [7](#)
- [37] Yijun Zhou and James Gregson. WHENet: Real-time fine-grained estimation for wide range head pose. *ArXiv*, abs/2005.10353, 2020. [1](#), [2](#)
- [38] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face Alignment Across Large Poses: A 3D Solution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, Los Alamitos, CA, USA, 2016. IEEE Computer Society. [3](#), [6](#)
- [39] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, page 250–269, Berlin, Heidelberg, 2022. Springer-Verlag. [6](#)