IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 February 2025, accepted 6 April 2025, date of publication 16 April 2025, date of current version 2 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3561506

RESEARCH ARTICLE

Reducing Head Pose Estimation Data Set Bias With Synthetic Data

ROBERTO VALLE[1], JOSÉ M. BUENAPOSADA[2], AND LUIS BAUMELA[1]
[1]Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, 28660 Madrid, Spain
[2]Departamento de Informática y Estadística, Universidad Rey Juan Carlos, Móstoles, 28933 Madrid, Spain

Corresponding author: Roberto Valle (rvalle@fi.upm.es)

This work was supported by the Project PID2022-137581OB-I00 funded by MICIU/AEI/10.13039/501100011033 FEDER, UE.

The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Data set bias not only compromises the fairness, accuracy and effectiveness of trained models, but also leads to a lower performance in real-world scenarios compared to the evaluation results obtained with a specific data set. This issue is especially evident in the estimation of head pose, as current data sets suffer from a limited number of images, imbalanced data distributions, the high cost of annotation, and ethical concerns. Synthetic data offers a promising solution to address these challenges, but current semi-synthetic data sets fail to deliver satisfactory results, likely due to the limited realism of the generated faces and the heavily skewed pose distribution. In this paper, we report the existence of data set biases in the most widely used head pose estimation benchmarks, which lead to an optimistic estimation of model performance in real-world scenarios. To mitigate this issue, we create a synthetic image data set using a generative model with explicit control over the head pose. Our experiments demonstrate that incorporating our synthetic images leads to improved generalization and accuracy.

INDEX TERMS Synthetic data, head pose estimation, generative models, data set bias, ethics.

I. INTRODUCTION

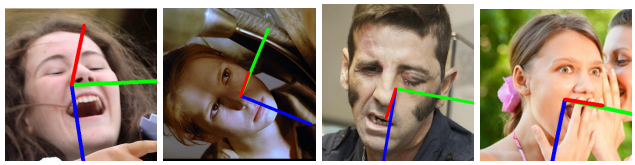
Deep learning models have become prevalent in several face analysis tasks, by significantly boosting the performance of different domains, such as face detection [1], [2], [3], face alignment [4], [5], [6], head pose estimation [7], [8], [9] or face recognition [10], [11], among others. However, the success of these models is highly dependent on the quantity, diversity, and quality of the images and the annotations used during training. Large, diverse, and well-annotated data sets are essential for developing robust deep models that allow them to generalize well under different conditions.

Head pose estimation (HPE) is a relevant pre-processing step, and it is often integrated into multi-task deep models in conjunction with other face analysis tasks [4], [12], [13]. By HPE, we mean predicting the relative orientation between the viewer and the target head [7]. This information is essential for applications, such as face recognition, gaze estimation, or driver monitoring [11], [14], since the performance of these systems is affected by the variability of facial appearance, and most of it depends on the orientation of the face.

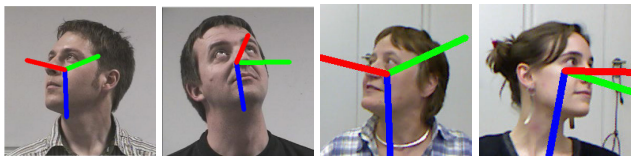
In this task, the demand for large-scale data sets that are accurately annotated poses significant challenges, as image collection is time-consuming, expensive, and must comply with data protection regulations [15]. The process of annotating images with precise head pose information adds further complexity due to the need for specialized equipment. As a result, popular HPE data sets often suffer from lack of variability [16], [17], imbalanced label distributions, composed, for example, of predominantly semifrontal faces [18], [19], and reliance on automatic labeling [19], [20], [21]. Models trained with them suffer from the so-called data set bias [22]. They do not accurately represent the real world and, as a consequence, do not generalize well across data sets [23], [24], [25], [26], [27].

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

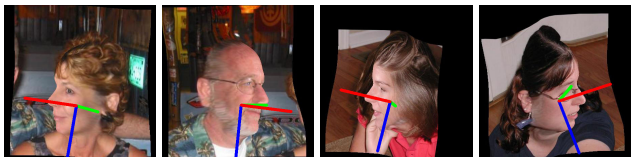
A solution to the aforementioned issues lies in the utilization of synthetic data. It refers to artificially generated images that mimic the real-world, but are created through algorithms rather than being directly captured with a camera [28]. In Figures 1a and 1b, we compare the distinct characteristics of real images in controlled and unrestricted scenarios, respectively, with synthetic ones in Figures 1c and 1d. The samples in Figure 1c are from the popular 300W-LP semi-synthetic data set [21], produced by warping images from 300W [29] with a 3D model [30]. 300W-LP is largely composed of profile views with poor realism. Hence, as shown in Sect. IV, models trained on them do not perform well against real images.



(a) Real in-the-wild images [19], [20]



(b) Real laboratory images [16], [17]



(c) Synthesized faces from real images [21]

(d) Our synthetic images

FIGURE 1. Qualitative comparison of HPE data sets acquired in realistic conditions (a) compared to face images acquired in controlled scenarios (b), semi-synthetic images simulated by fitting a 3D face model (c), and the synthetic faces generated in this work using a controllable generative model [31] (d).

In this work, we advocate for the use of generative models to create synthetic images for HPE, offering an efficient and cost-effective alternative, improving model robustness in handling varied scenarios, lighting conditions, head poses, or facial expressions (see Figure 1d). Unfortunately, many challenges arise when using generative models to create synthetic images. They can produce excellent results under ideal conditions, but they may encounter issues such as generating faces that appear artificial or distorted, particularly in near-profile viewpoints [32]. Thus, the quality of the generated images can vary significantly depending on the model's complexity and the training data distribution. Furthermore, generative models are also prone to leak information from the underlying training data, which is specifically of concern when human data is involved, since identity leakage poses a significant risk to personal privacy rights [23].

To the best of our knowledge, we are the first to address the presence of biases in HPE benchmarks, which mainly stem from varying acquisition conditions and label distributions across different data sets. To address this, we introduce a synthetic data set, HPGEN, using a controllable generative model that automates head pose annotations. As shown in our experiments, training a deep model for HPE that includes our synthetic images demonstrates better generalization capabilities across all major benchmarks. The HPGEN images and labels can be downloaded at https://dx.doi.org/10.21227/d31p-nt47.

II. RELATED WORK

In this section, we examine current state-of-the-art (SOTA) methods that generate images conditioned to the head pose, we analyze the data set bias problem for other facial analysis tasks, and review the accuracy of present literature for HPE.

A. CONTROLLABLE IMAGE GENERATION

Generative Adversarial Networks (GANs) and Generative Diffusion Models (GDMs) are two of the most prominent methodologies in the field of generative modeling. Recent advancements in computer vision have led to the development of sophisticated 3D-aware image generators [33], [34]. These models aim to generate images that not only look realistic but also adhere to the constraints of 3D structures.

Introduced by Goodfellow et al. [35], GANs consist of two networks, a generator and a discriminator. The generator creates synthetic images by learning the distribution of training samples, whereas the discriminator evaluates their authenticity by classifying whether the input samples stem from the training set or are produced by the generator. Both networks are trained in an adversarial manner, which drives the generator to produce increasingly realistic images over time. Nowadays, GANs have been extended incorporating additional geometric information, such as semantic maps or 3D models, into the generation process to handle 3D-aware image generation [36], [37], [38], [39]. Techniques like GANs conditioned on pose information have been developed to ensure that generated images maintain consistent and realistic 3D structures when viewed from different angles. However, they continue to struggle in accurately reconstructing 3D shapes with non-rigid entities such as human faces that exhibit non-rigid deformations and varied appearances [31], [32], [38].

GDMs represent an alternative approach to image generation [40]. Unlike GANs, which use an adversarial training setup, diffusion models rely on a process of iterative refinement. Starting from random noise, GDMs iteratively denoise the image through a sequence of steps, guided by learned

probability distributions. Analogously, diffusion models have been also adapted for 3D-aware image generation [41], [42], [43], [44], [45], [46]. These models condition the diffusion process to ensure that the generated images respect the underlying 3D structure.

The choice between GANs and GDMs depends on their respective strengths in controlling output images. In Table 1 we note that 3D-aware GANs [31], [32] are better suited for tasks requiring high quality image outputs with high diversity, but can suffer from training instability compared to 3D-aware GDMs [41], [43]. Given that our application (HPE) prioritizes the generation of realistic images properly annotated, we have opted to use GANs over GDMs. Specifically, OP3D [31], which proposes a novel feature map representation, regarded as ''OrthoPlanes'', aimed at enhancing the 3D awareness of the volume rendering stage (see Figure 2), which significantly improves how well the subjects' identity is maintained under different poses.

TABLE 1. Comparative of existing literature on 3D-aware image generation using FFHQ [18], which differentiates the quality and diversity of the generated images, and possible control of specific attributes on the generated images. Lower FID denotes better quality in the synthetic faces (see Sect. IV).

Method	Model	Control	FID (↓)
ZestGuide [42]	GDM	Prompt, Semantic map	22.08
ControlNet [44]	GDM	Prompt, Semantic map, Keypoints	15.27
GRAM-HD [36]	GAN	Euler angles, FoV, Identity	12.00
VersatileDiffusion [43]	GDM	Prompt	11.10
GMPI [38]	GAN	Euler angles, Identity	8.29
GLIGEN [41]	GDM	Prompt, Semantic map, Keypoints	5.61
SURF-GAN [37]	GAN	Euler angles, Identity	4.72
EG3D [32]	GAN	Euler angles, FoV, Identity	4.70
OP3D [31]	GAN	Euler angles, FoV, Identity	**4.01**

B. DATA SET BIAS

Data set bias [22] is a well-known problem in other facial analysis tasks, such as facial alignment [6] or facial recognition [22], [27]. Synthetic data can help mitigate it [28], but a model trained only on synthetic data suffers from a larger bias than the same model trained on real data [24]. According to [25], while deep models trained using synthetic data deliver reasonable accuracy, real images lead to superior results. Domain adaptation helps address this gap by adjusting a model trained on synthetic data to perform well on real-world images [26], [47]. Alternatively, Joshi et al. [23] proposes the combination of synthetic and real data to reduce biases caused by the unequal distributions often observed in real-world data sets.

HPE data sets bias will also lead to inaccurate or skewed estimations of head pose across different conditions. Adding synthetic data helps mitigate the bias in current HPE data sets, improving face diversity, augmenting underrepresented head poses, and reducing labeling inconsistencies (see Table 5).

C. HEAD-POSE ESTIMATION

HPE refers to the process of determining the relative orientation between the camera and the target head. It is usually parametrized using popular Euler angles. Recent SOTA literature shows remarkable results [7] (see Table 2). Most approaches follow the same evaluation protocol. They use a large semi-synthetic data set, 300W-LP [21], to train a deep model that is evaluated on real data sets, AFLW2000-3D [21] and Biwi [16], for unrestricted and laboratory conditions.

We also observe that the most successful methods, OpNet [48] and TRG [49], have incorporated additional training data sets, suggesting that models trained exclusively with 300W-LP semi-synthetic images struggle to generalize to different scenarios.

In this work, our main goal is to analyze the existence of HPE data set biases that negatively affect generalization. To address this, we introduce a synthetic data set, HPGEN, generated with a GAN with explicit control over head pose.

TABLE 2. Comparative of existing literature on HPE using the popular AFLW2000-3D and Biwi data sets [16], [21]. Lower GE denotes better head pose estimation (see Sect. IV).

Method	AFLW2000-3D					Biwi				
	MAE (↓)				GE (↓)	MAE (↓)				GE (↓)
	yaw	pitch	roll	mean		yaw	pitch	roll	mean	
HopeNet [9]	6.47	6.56	5.44	6.15	9.93	4.81	6.61	3.27	4.89	9.53
FSA-Net [8]	4.50	6.08	4.64	5.07	8.16	4.27	4.96	2.76	4.00	7.64
WHENet [50]	4.44	5.75	4.31	4.83	-	3.60	4.10	2.73	3.48	-
MFDNet [51]	4.30	5.16	3.69	4.38	-	3.40	4.68	2.77	3.62	-
img2pose [52]	3.42	5.03	3.27	3.91	6.41	4.56	3.54	3.24	3.78	7.10
MNN [13]	3.34	4.69	3.48	3.83	-	3.98	4.61	2.39	3.66	-
LPONet [12]	2.83	4.56	3.30	3.56	-	4.94	4.87	2.85	4.22	-
DSFNet [5]	**2.65**	4.28	2.82	3.25	-	-	-	-	-	-
CIT [4]	2.68	4.38	3.45	3.50	-	**3.01**	4.54	4.15	3.90	-
SRHP-Euler [53]	2.76	4.25	2.76	3.26	5.29	4.54	5.05	2.80	4.13	7.49
SRHP-6D [53]	2.85	4.59	3.04	3.49	5.37	4.58	4.65	2.71	3.98	7.30
OpNet [48]	2.79	**4.18**	**2.49**	**3.15**	**5.23**	3.66	4.61	2.44	3.57	7.01
TRG [49]	-	-	-	-	-	3.04	**3.44**	**1.78**	**2.75**	**5.35**

III. DATA SET GENERATION METHODOLOGY

A critical component in developing robust HPE models is the availability of data sets that generalize well under different conditions. Annotating a large-scale image data set poses significant challenges that can hinder the progress in this field. Joshi et al. [23] suggests that training a model with a combination of real and synthetic images yields outstanding results and helps mitigate biases.

In this context, the challenges inherent in the development of a data set for HPE highlight the need for innovative solutions, such as the use of synthetic data with a controllable generative model. As illustrated in Figure 2, OP3D [31] produces high-resolution synthetic faces (*i.e.*, 512 × 512 pixels) conditioned on a specific head pose input denoted by yaw and pitch angles. OP3D uses the styleGAN-based backbone [18], which yields excellent results in data-driven generative image modeling (see Table 1). In this network, a random latent code z sampled from a normal distribution defines the identity of the subject. This latent vector z is processed by the mapping network to obtain the style vector w, used to control a distinct visual attribute (*e.g.*, pose, gender, identity) disentangled within the synthesis network. Finally, the generated 2D

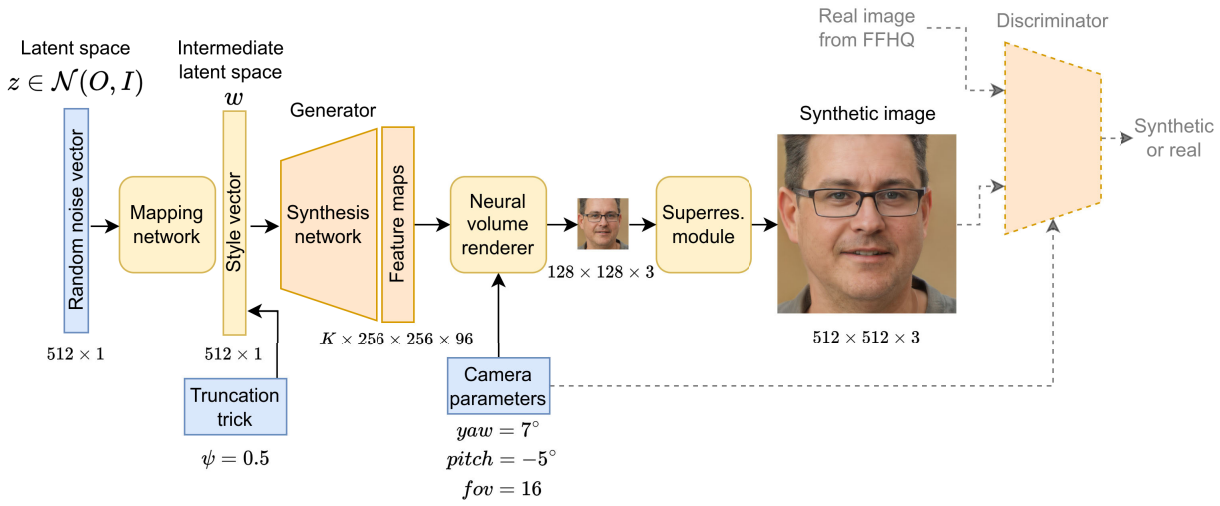


FIGURE 2. The controllable generative model, OP3D [31] requires as input the random latent code z, truncation trick value ψ, and camera parameters based on *yaw*, *pitch* and *fov*, colored with blue boxes respectively. The final discriminator with dashed lines and higher transparency is used exclusively during training.

feature maps are converted to a 3D representation used as input for a neural volume renderer that simulates how each face would look in a 3D space. This formulation regarded as "OrthoPlanes" uses 3 orthogonal groups aligned with the axes of feature planes, each with K parallel planes with a resolution of $256 \times 256 \times 96$. In this work, we use the OP3D model based on $K = 12$ planes trained on FFHQ [18].

Although generative models excel at producing highly realistic synthetic face images, they are also prone to leak information from the training data set [23]. This is of particular concern when human data is involved, as leaks of identity information infringe the right to personal privacy. The implementation of data protection laws [15], *e.g.*, the *General Data Protection Regulation* (GDPR), defines strict rules for processing data that can reveal identity information. As specified by Feng et al. [54], identity leaks become less likely when using data bases with high diversity and a large number of identities for training, which alleviates model overfitting. FFHQ consists of 70K identities acquired from Flickr without explicit consent, which bypasses the legal requirements of the individuals at the mercy of a permissive Creative Commons license.[1] We note that such information leakage related to the human identities within the training set is of concern, so we use the standard truncation trick [18] that defines how far the generated images can be from the "average face image" (and how diverse the output can be). We use a $\psi = 0.5$ that tends to produce averaged identities at expense of deteriorate image realism and diversity.

As a result, we generate a synthetic data set that consists of a total of 400K images generated using 4K different identities. Each identity is consistent across multiple views. A total of 100 views properly balanced by yaw angle are generated, by adjusting the yaw angle from $-90°$ to $90°$, the pitch angle from -60° to 60°, and the field of view (FoV), which refers to the extent or angle of the scene that is visible through the camera lens or sensor, between 5° and 25°. In this way, a pitch range up to $\pm 60°$ guarantees that the gimbal lock problem [53] can be ignored, which occurs when the pitch angle is $\pm 90°$, using a yaw-pitch-roll configuration [19]. We denote as HPGEN the synthetic data set generated. Figure 3 shows several samples acquired from two different identities.

In Table 3, we assess that HPGEN stands out as the most balanced among the standard HPE data sets. Regarding yaw and pitch distributions, it maintains a consistent percentage of samples across all orientations, addressing limitations found in other data sets that exhibit uneven sample distributions.

As with most HPE data sets in the literature (see Table 3), it is necessary to annotate the position of the face in our HPGEN images. So, we annotate each image with the bounding box provided by the RetinaNet [1] face detector trained on the popular WIDER [2] data set. In case of several detections in an image, we only choose the most significant in terms of confidence and/or size.

IV. EXPERIMENTS

Traditionally, HPE techniques use public image sets obtained in laboratory conditions. In this context, Biwi [16] is the most popular data set. It contains 15677 frames from 24 videos of 20 subjects in a controlled environment, where people sit in front of a Kinect sensor and freely turn their heads.

Nowadays, HPE progress has shifted towards evaluations that involve more realistic and challenging situations by using images acquired in-the-wild, that include extreme rotations, exaggerated facial expressions, arbitrary illumination, blurriness, partial occlusions, etc. A common methodology for performing cross-data set experiments is based on using

[1] https://exposing.ai/ffhq/

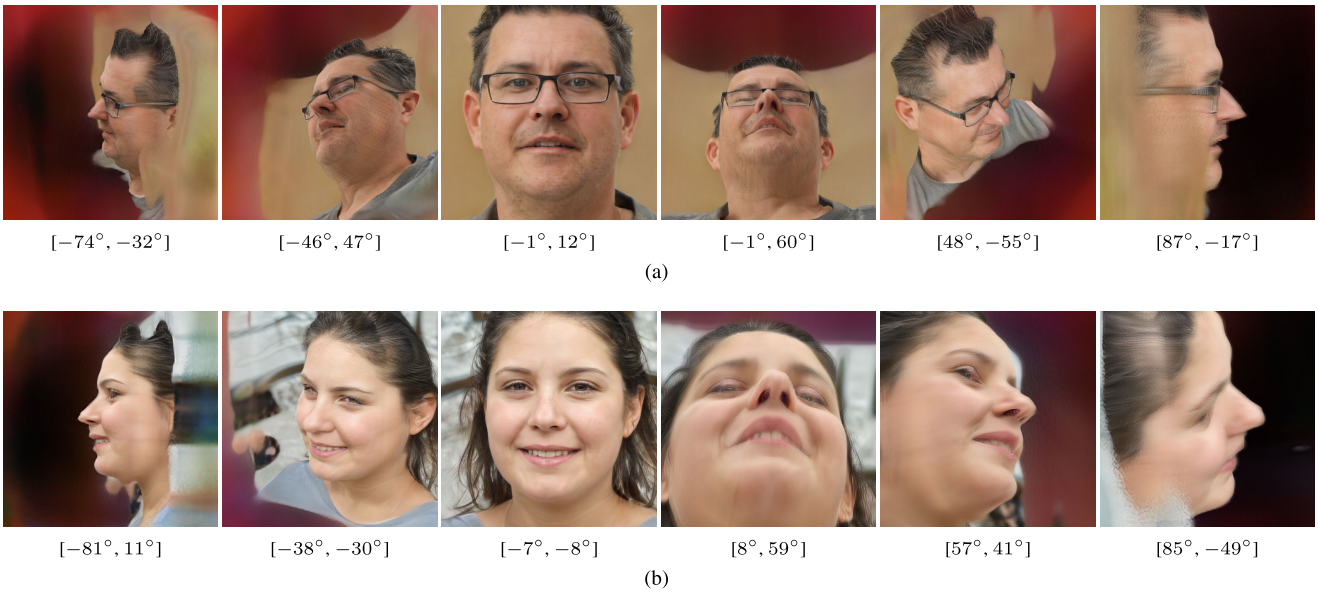


FIGURE 3. Representative face samples of HPGEN related to 2 different identities in each row. The text below represents yaw and pitch angles in degrees required to control the output. We note that identity remains consistent across various perspectives, although the realism of near-frontal face images appears to be higher.

TABLE 3. Comparison among HPGEN and the most popular HPE data sets. We show the head pose distribution of each data set computing the percentage of faces (%) around each orientation. For example, 15° includes all faces between 0° and 30°.

Database	# images	# subjects	yaw (%)						pitch (%)			
			-75°	-45°	-15°	15°	45°	75°	-45°	-15°	15°	45°
AFLW	20954	24140	7.62	12.04	30.04	29.12	10.01	6.84	2.11	65.49	31.67	0.66
DAD-3D	42152	42152	2.47	8.58	35.32	42.28	8.41	1.67	2.61	64.92	30.30	1.66
300W-LP	61225	3837	18.76	18.57	12.31	12.81	18.69	18.83	3.19	75.18	21.07	0.52
AFLW2000-3D	1969	1969	7.26	9.90	35.70	30.87	9.39	6.85	3.55	65.10	26.86	2.79
Biwi	13219	20	2.93	20.34	40.34	22.67	12.14	1.54	8.22	22.77	58.41	10.56
HPGEN	400000	4000	16.64	16.63	16.39	16.84	16.68	16.79	25.07	24.51	25.30	24.67

300W-LP and AFLW2000-3D data sets [21]. The former consists of 61225 synthetic images that expand 300W [29] through extreme orientations, while the latter is a reannotated subset of AFLW [19]. Both include automatically annotated Euler angles generated by fitting a 3DMM to each face. The most popular protocol uses 300W-LP as training set and both AFLW2000-3D and Biwi as test sets. The standard evaluation protocol [9] discards faces with yaw outside the range ±99°.

The second popular benchmark is AFLW [19]. It provides a collection of 25993 faces in 21997 in-the-wild images, with head pose ranging between ±120° for yaw and ±90° for pitch and roll angles. These Euler angles have been computed assuming the structure of a mean 3D face and applying the POSIT algorithm [55] with the manually labeled landmarks. We discard some images with reported annotation errors and follow the same train and test protocol as [13].

Finally, DAD-3D [20] is a newly released in-the-wild data set balanced over a wide range of poses, face expressions, and occlusions. It consists of 42152 and 2746 annotated images as train and test sets respectively. DAD-3D also includes the rotation matrix used to accurately fit the 3DMM to each face.

A. METRICS

In the context of generative deep models, the Fréchet Inception Distance (FID) [56] and the Kernel Inception Distance (KID) [57] are used to evaluate the quality and diversity of the generated images. The FID in Eq. (1) measures the similarity between extracted features of the real and generated data sets as multivariate Gaussian distributions,

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \quad (1)$$

being μ_r/μ_g the means, and Σ_r/Σ_g the covariances of features from FFHQ [18] and HPGEN respectively.

The KID is based on the Maximum Mean Discrepancy, which measures the difference between distributions using kernel methods. It does not assume any specific distribution (*e.g.*, Gaussian) for the data. Lower scores FID and KID indicate a greater similarity between real and generated images. KID has the advantage of having an unbiased estimator and can be more stable for smaller sample sizes.

The Identity Similarity (ID) [58] measures, using the Arcface [10] cosine similarity, how well the images generated in HPGEN preserve the identity of the subjects under different poses. Higher ID scores indicate better identity preservation in the generated images.

In the context of HPE, the most popular metric for quantifying error is the Mean Absolute Error (MAE). Measures the head pose resorting to the Euler angles representation, which is dependent on the order in which yaw, pitch, and roll rotations are applied.

Alternatively, the Geodesic Error (GE) [59],

$$\text{GE} = \frac{1}{N}\sum_{i=1}^{N}(\cos^{-1}\left(\frac{\text{Tr}(\hat{\mathbf{R}}_i\mathbf{R}_i^T)-1}{2}\right)), \tag{2}$$

where $\hat{\mathbf{R}}_i$, $\mathbf{R}_i$ represent the ground truth and predicted rotation matrices respectively for i-th image, directly compares the rotation matrices by computing their geodesic distance. It does not make any errors in the presence of the gimbal lock, that in HPE usually occurs when the pitch $\approx \pm 90°$ [53].

B. IMPLEMENTATION DETAILS

We use the same architecture for all the HPE experiments, an EfficientNet-B4 [60] pre-trained on ImageNet. We fine-tune this model using a SGD optimizer with an initial learning rate $\alpha = 10^{-3}$, which is halved when the validation loss plateaus for 5 epochs. We shuffle each training data set, and split it into 90% train and 10% validation, to determine the meta-parameters of the learning process. In 300W-LP and HPGEN we split the validation data according to different identities.

We also augment our training data by applying to each sample the following random operations: in plane rotation between $\pm 5°$, scaling by $\pm 10\%$, translation by $\pm 20\%$ of the bounding box size, mirroring the face image horizontally and changing color multiplying each HSV channel by a random value between $\pm 20\%$.

In AFLW/DAD-3D/HPGEN we crop faces using the provided bounding box ground-truth. In 300W-LP/AFLW2000-3D we use the rectangle enclosing the annotated landmarks. In Biwi, the bounding boxes provided by the MTCNN face detector [3] to make fair comparisons with the previous literature [8], [51], [52], [53] (see Table 2).

At runtime each model estimates the head pose on the Biwi videos at a mean rate of 12 FPS, using an NVIDIA GeForce RTX 3080 Ti (12GB) GPU, PyTorch Lightning and OpenCV libraries. Code and trained models are available at https://github.com/pcr-upm/access25_headpose.

C. IMAGE GENERATION RESULTS

As mentioned in Sect. III, the controllable generative model OP3D [31] has been trained using FFHQ [18]. So, we perform several experiments to assess the reliability of the generated face images compared to those in FFHQ.

In Table 4 we compare the quality of our HPGEN synthetic images with the real images from FFHQ. HPGEN always yields results that fall short compared to those reported in the original OP3D study. The ID metric suggests that the faces of HPGEN maintain consistency of the identity of subjects, although not as effectively as described in [31], *i.e.*, 0.73 vs 0.417 ID.

TABLE 4. Comparison between FID/KID/ID metrics computed by using HPGEN images against the results provided by OP3D [31]. HPGEN (yaw/pitch ∈ [-30°, 30°]) represents a subset of HPGEN images with face orientation in ±30°.

Database \ Metric	FID (↓)	KID (↓)	ID (↑)
OP3D [31]	4.01	1.23	0.73
HPGEN (yaw/pitch ∈ [-30°, 30°])	6.05	3.50	0.685
HPGEN	73.35	47.47	0.417

However, the metrics FID and KID indicate that the HPGEN images exhibit a greater divergence from the real images, 4.01 vs. 73.35 FID. The key reason stems from the head pose distribution in the original FFHQ data set. In FFHQ, 60% of the images have a yaw angle between $\pm 25°$, and 96% of them are between $\pm 45°$ [24]. This bias in the orientation distribution means that the models trained on FFHQ perform better on images with similar yaw angles, but struggle against extreme head poses. In HPGEN the data are balanced properly according to the yaw angle, which ranges between $\pm 90°$.

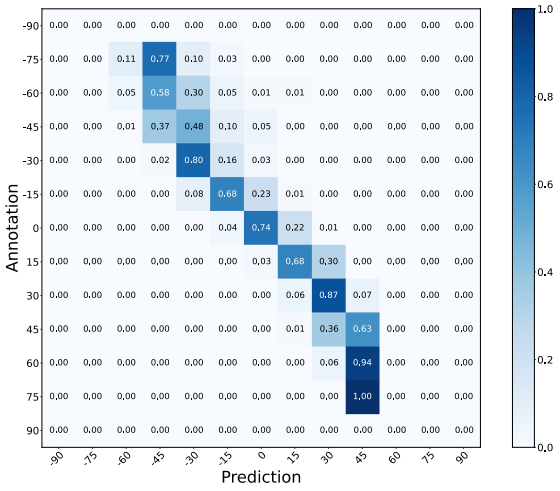
To assess this hypothesis, we consider a subset of HPGEN containing only images with a head orientation within $\pm 30°$. When computing the FID, KID and ID metrics again, we get a noticeable reduction in FID/KID metrics, and an increase in consistency ID (see Table 4). In Figure 3 we qualitatively verify that images with extreme poses exhibit artifacts around the face that compromise visual quality.

At this point, we also analyze whether the images in HPGEN with a head pose close to profile are good, not only in terms of realism, but also in terms of whether the applied control to generate them has been successful. In Figure 4a we show the confusion matrix of a deep model trained with HPGEN and evaluated with Biwi, where we discretize each angle in steps of 15°. Note that the classification error increases as the head pose approaches the profile view, which indicates that the model struggles with handling poses that are underrepresented in FFHQ.

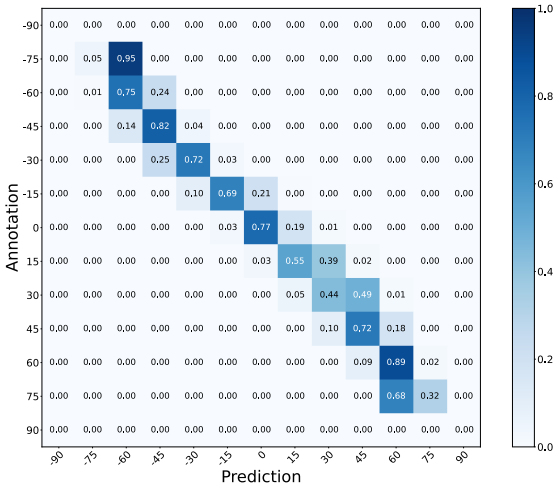
In the end, we reannotate the entire HPGEN data set using the DAD-3DNet model [20] to determine if the error in the near-profile cases is due to the facial appearance quality, or is attributed to the original yaw/pitch control labels. In Figure 4b we replace the previous HPGEN labels with the reannotated angles, thus reducing the classification error in the near-profile samples. For example, considering only those Biwi samples with an absolute yaw angle greater than 60°, we significantly reduce the classification error from 98.50% to 49.75%. This means that the images generated in near-profile poses do not agree fully with the orientation control signal. We speculate that this is caused by the lack of training images in those angles in FFHQ.

D. HEAD-POSE ESTIMATION RESULTS

The most realistic methodology for evaluating HPE is a cross-data set procedure, in which the deep model is



(a) Using head pose labels required to control OP3D [31]



(b) Using head pose reannotated with a model trained on [20]

FIGURE 4. Confusion matrices obtained using HPGEN and Biwi data sets for training and evaluating respectively. We compare results using the original head pose labels required to control the generative model output (a), against the reannotated head pose using the DAD-3DNet model [20].

trained with a data set and evaluated with a different one. In Table 5 we perform several cross-data set experiments to evaluate the quality of current benchmarks, including AFLW, DAD-3D, 300W-LP/AFLW200-3D and Biwi data sets. We also compare the generalization capability of these models compared to "All", which is a model that combines AFLW, DAD-3D and 300W-LP. "All" can mitigate bias by using the diversity of combined data sets. The key finding of this experiment is that we always achieve the best result when training with the train subset of the same data base, *e.g.*, 6.85 and 5.97 in AFLW and DAD-3D respectively. These results demonstrate the existence of a significant data set bias in most benchmarks. Unexpectedly, the model trained with DAD-3D also achieves the lowest GE on AFLW2000-3D and Biwi, while the standard benchmark relies on 300W-LP [21], which confirms that a model trained with 300W-LP synthesized images struggle to generalize to different scenarios.

TABLE 5. Cross-data set experiment using the most popular HPE data sets. Note that we use the Geodesic Error (GE) as metric. "All" denotes a model trained with AFLW, DAD-3D and 300W-LP training subsets. "Avg" represents the average GE.

Train \ Test	AFLW	DAD-3D	AFLW2000-3D	Biwi	HPGEN	Avg
AFLW	6.85	9.08	6.21	11.26	7.55	8.19
DAD-3D	8.89	5.97	5.67	7.78	5.14	6.69
300W-LP	10.69	12.44	6.46	11.64	6.15	9.47
All	7.23	6.03	5.10	10.40	5.94	6.94
HPGEN	12.47	13.19	9.24	9.90	1.56	9.27
All+HPGEN	7.00	5.86	5.09	8.66	1.53	5.62

To measure the reliability of the synthetic images in HPGEN, we compare the generalization capability of a deep model trained on "All" standard HPE data sets compared to All+HPGEN, that also includes the synthetic data. As stated previously, synthetic data can also contain biases since we achieve the best result, 1.56 GE in the test subset of HPGEN, training with its training subset. Another key finding is that by computing the mean GE of all data sets we achieve the best result combining the All+HPGEN benchmarks, 5.62 GE. Similar to Wood et al. [25], we confirm that models trained only on synthetic data demonstrate competitive performance, but combining real and synthetic images we achieve the best results.

Moreover, we compare 300W-LP results in Table 5 with previous literature using AFLW2000-3D and Biwi. In Table 2 we note that models trained only with 300W-LP struggle to generalize to different scenarios. Certain methods [13], [50], [52] tend to perform well in laboratory data, but do not generalize in images involving more challenging situations, while other methods [4], [5], [53] produce superior results in AFLW2000-3D images compared to those obtained in Biwi. Similarly, our model provides a GE of 6.46 and 11.64 in AFLW2000-3D and Biwi respectively. We also note a GE reduction from 11.64 to 9.92 aligning the reference systems as in [48] and [53]. Even so, our goal was not to exceed the SOTA, but rather to demonstrate that incorporating synthetic images improves the generalization and reduces the data set bias.

Although not strictly comparable with SOTA [48], [49] due to the differences in the additional data used for training, we analyze our best model All+HPGEN combining real and synthetic data. First, we establish the new SOTA in AFLW2000-3D and we achieve a significant GE relative reduction of 2.68% over OpNet [48], from 5.23 to 5.09. In Biwi, we also obtain a competitive GE, 8.66, which is further reduced to 7.04 by aligning the reference systems as in [53]. This ranks second, surpassed only by TRG [49], 5.35, which incorporates additional training data in laboratory conditions such as ARKitFace [17].

In the end, we visually compare the six worst All+HPGEN predictions from the test data sets in Figure 5. By focusing on the top errors, we can clearly identify the specific conditions

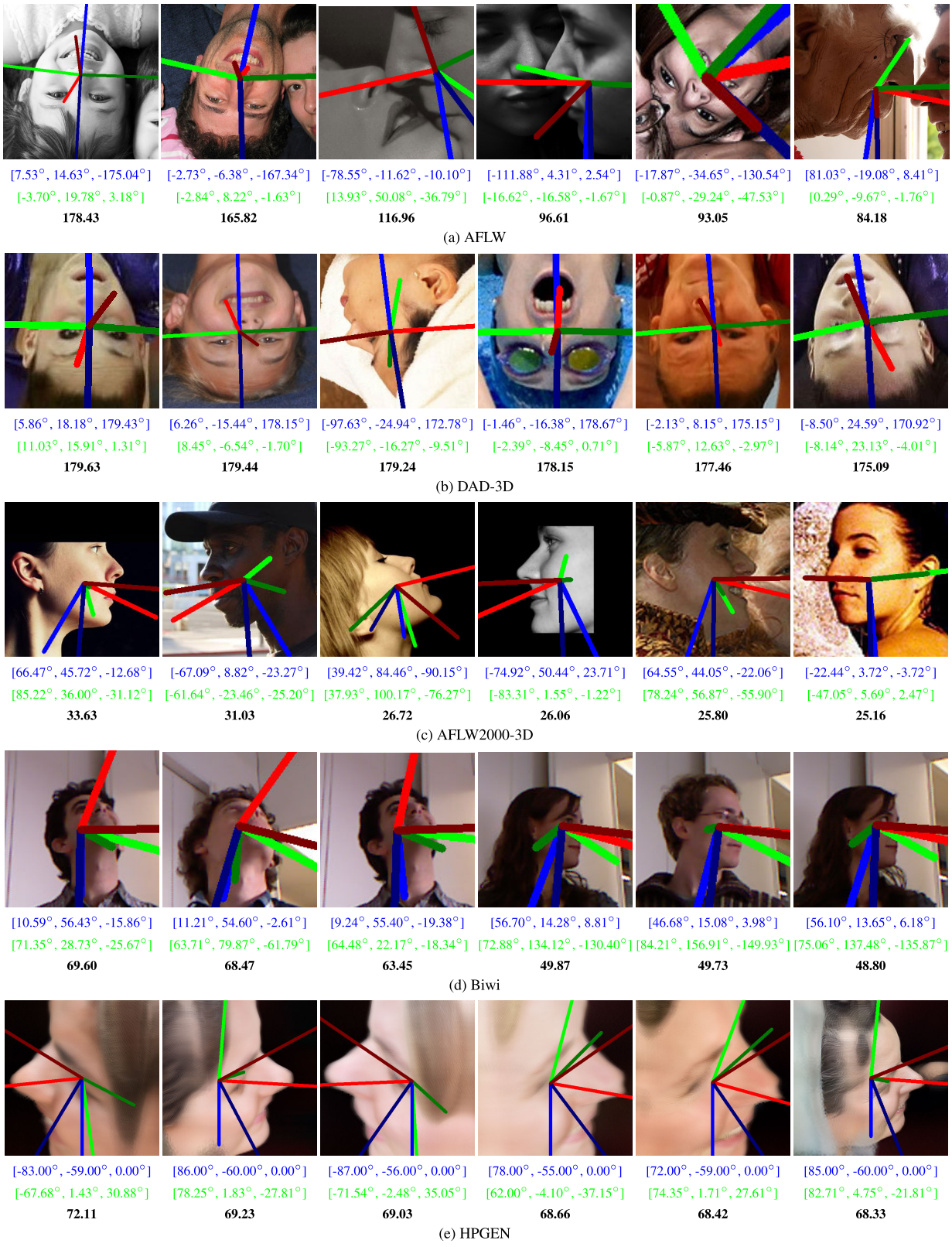


FIGURE 5. Top-6 errors using the All+HPGEN model in AFLW, DAD-3D, AFLW2000-3D, Biwi and HPGEN test data sets. The blue, green and black text below represent Euler angles ground truth, prediction and Geodesic error respectively.

or scenarios where the model fails most often. We note that results are still unsatisfactory when two faces appear together and both are within the cropped bounding box (see Figure 5a), faces are turned upside (see Figure 5b), certain facial regions are hidden/occluded (see Figure 5c), heads are tilted sharply with significant yaw/pitch angles (see Figure 5d), or bounding boxes are captured in a close-up view (see Figure 5e).

V. CONCLUSION

In this work, we propose the addition of synthetic face images provided by a controllable generative model as a solution to mitigate the problem of data set bias in the context of HPE. To this end, we generate a comprehensive synthetic data set, HPGEN, with annotated head poses, properly balanced covering yaw and pitch angles from -90° to 90° and -60° to 60° respectively.

By combining the synthetic face images with several well-known real image sets, we train a regressor that demonstrates superior generalization capabilities. The All+HPGEN model achieves excellent results across all benchmarks, highlighting the effectiveness of incorporating synthetic faces to alleviate bias and improve performance. Even so, this model tends to perform best when faces are in near-frontal positions (within ±45°), since we utilize a pre-trained controllable generative model that produces more realistic images in this range.

In future work, we plan to apply *domain adaptation* to further enhance the performance of models trained on synthetic data across in-the-wild evaluation data sets. Additionally, we aim to train the controllable generative model by exclusively using ethical images with explicit permissions, ensuring compliance with ethical standards and allowing us to further enhance the realism of faces.

ACKNOWLEDGMENT

The authors thank Íñigo Sanz for generating the HPGEN data set. José M. Buenaposada and Luis Baumela are members of the Madrid Ellis Unit, funded by the Regional Government of Madrid.

REFERENCES

[1] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[2] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5525–5533.

[3] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[4] Y. Li, G. Tan, and C. Gou, "Cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose," *Int. J. Comput. Vis.*, vol. 132, no. 4, pp. 1242–1257, Apr. 2024.

[5] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 4531–4540.

[6] R. Valle, J. M. Buenaposada, A. Valdés, and L. Baumela, "Face alignment using a 3D deeply-initialized ensemble of regression trees," *Comput. Vis. Image Understand.*, vol. 189, Dec. 2019, Art. no. 102846.

[7] A. F. Abate, C. Bisogni, A. Castiglione, and M. Nappi, "Head pose estimation: An extensive survey on recent techniques and applications," *Pattern Recognit.*, vol. 127, Jul. 2022, Art. no. 108591.

[8] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1087–1096.

[9] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2155–215509.

[10] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022.

[11] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, S. Yan, and J. Feng, "Towards pose invariant face recognition in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2207–2216.

[12] Y. Kim, J.-H. Roh, and S. Kim, "Facial landmark, head pose, and occlusion analysis using multitask stacked hourglass," *IEEE Access*, vol. 11, pp. 30970–30981, 2023.

[13] R. Valle, J. M. Buenaposada, and L. Baumela, "Multi-task head pose estimation in-the-Wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2874–2881, Aug. 2021.

[14] D. Kang and D. Kang, "Head pose-aware regression for pupil localization from a-Pillar cameras," *IEEE Access*, vol. 12, pp. 11083–11094, 2024.

[15] S. Mittal, K. Thakral, R. Singh, M. Vatsa, T. Glaser, C. C. Ferrer, and T. Hassner, "On responsible machine learning datasets with fairness, privacy, and regulatory norms," *arXiv:2310.15848*, 2023.

[16] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 437–458, Feb. 2013.

[17] Y. Kao, B. Pan, M. Xu, J. Lyu, X. Zhu, Y. Chang, X. Li, and Z. Lei, "Toward 3D face reconstruction in perspective projection: Estimating 6DoF face pose from monocular image," *IEEE Trans. Image Process.*, vol. 32, pp. 3080–3091, 2023.

[18] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4217–4228, Dec. 2021.

[19] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2144–2151.

[20] T. Martyniuk, O. Kupyn, Y. Kurlyak, I. Krashenyi, J. Matas, and V. Sharmanska, "DAD-3DHeads: A large-scale dense, accurate and diverse dataset for 3D head alignment from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20910–20920.

[21] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3D total solution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 78–92, Jan. 2019.

[22] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. CVPR*, Jun. 2011, pp. 1521–1528.

[23] I. Joshi, M. Grimmer, C. Rathgeb, C. Busch, F. Bremond, and A. Dantcheva, "Synthetic data in human analysis: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, pp. 4957–4976, Jul. 2024.

[24] M. Huber, A. T. Luu, F. Boutros, A. Kuijper, and N. Damer, "Bias and diversity in synthetic-based face recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 6203–6214.

[25] E. Wood, T. Baltrusaitis, C. Hewitt, S. Dziadzio, T. J. Cashman, and J. Shotton, "Fake it till you make it: Face analysis in the wild using synthetic data alone," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3661–3671.

[26] S. Basak, P. Corcoran, F. Khan, R. Mcdonnell, and M. Schukat, "Learning 3D head pose from synthetic data: A semi-supervised approach," *IEEE Access*, vol. 9, pp. 37557–37573, 2021.

[27] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2261–2268.

[28] C. M. de Melo, A. Torralba, L. Guibas, J. DiCarlo, R. Chellappa, and J. Hodgins, "Next-generation deep learning based on simulators and synthetic data," *Trends Cognit. Sci.*, vol. 26, no. 2, pp. 174–187, Feb. 2022.

[29] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image Vis. Comput.*, vol. 47, pp. 3–18, Mar. 2016.

[30] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 146–155.

[31] H. He, Z. Yang, S. Li, B. Dai, and W. Wu, "OrthoPlanes: A novel representation for better 3D-awareness of GANs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 22939–22950.

[32] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. de Mello, O. Gallo, L. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, "Efficient geometry-aware 3D generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16102–16112.

[33] F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, L. Liu, A. Kortylewski, C. Theobalt, and E. Xing, "Multimodal image synthesis and editing: The generative AI era," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15098–15119, Dec. 2023.

[34] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10850–10869, Sep. 2023.

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020.

[36] J. Xiang, J. Yang, Y. Deng, and X. Tong, "GRAM-HD: 3D-consistent image generation at high resolution with generative radiance manifolds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 2195–2205.

[37] J.-G. Kwak, Y. Li, D. Yoon, D. Kim, D. K. Han, and H. Ko, "Injecting 3D perception of controllable NeRF-GAN into StyleGAN for editable portrait image synthesis," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2022, pp. 236–253.

[38] X. Zhao, F. Ma, D. Güera, Z. Ren, A. G. Schwing, and A. Colburn, "Generative multiplane images: Making a 2D GAN 3D-aware," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2022, pp. 18–35.

[39] F. Yin, Y. Zhang, X. Wang, T. Wang, X. Li, Y. Gong, Y. Fan, X. Cun, Y. Shan, C. Öztireli, and Y. Yang, "3D GAN inversion with facial symmetry prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 342–351.

[40] J. Ho, A. N. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Jan. 2020.

[41] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "GLIGEN: Open-set grounded text-to-image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22511–22521.

[42] G. Couairon, M. Careil, M. Cord, S. Lathuilière, and J. Verbeek, "Zero-shot spatial layout conditioning for text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 2174–2183.

[43] X. Xu, Z. Wang, E. Zhang, K. Wang, and H. Shi, "Versatile diffusion: Text, images and variations all in one diffusion model," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7720–7731.

[44] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3813–3824.

[45] S. Zhao, D. Chen, Y. Chen, J. Bao, S. Hao, Y. Lu, and K. K. Wong, "Uni-ControlNet: All-in-one control to text-to-image diffusion models," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Jan. 2023.

[46] T. Kirschstein, S. Giebenhain, and M. Nießner, "DiffusionAvatars: Deferred diffusion for high-fidelity 3D head avatars," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 5481–5492.

[47] F. Kuhnke and J. Ostermann, "Deep head pose estimation using synthetic images and partial adversarial domain adaption for continuous label spaces," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10163–10172.

[48] M. Welter, "On the power of data augmentation for head pose estimation," Jul. 2024, *arXiv:2407.05357*.

[49] S.-H. Chun and J. Y. Chang, "6DoF head pose estimation through explicit bidirectional interaction with face geometry," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2024, pp. 146–163.

[50] Y. Zhou and J. Gregson, "WHENet: Real-time fine-grained estimation for wide range head pose," in *Proc. Brit. Mach. Vis. Conf.*, Jan. 2020.

[51] 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," *IEEE Trans. Multimedia*, vol. 24, pp. 2449–2460, 2022.

[52] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner, "img2pose: Face alignment and detection via 6DoF, face pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7613–7623.

[53] A. Cobo, R. Valle, J. M. Buenaposada, and L. Baumela, "On the representation and methodology for wide and short range head pose estimation," *Pattern Recognit.*, vol. 149, May 2024, Art. no. 110263.

[54] Q. Feng, C. Guo, F. Benitez-Quiroz, and A. Martinez, "When do GANs replicate? On the choice of dataset size," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6681–6690.

[55] D. F. Dementhon and L. S. Davis, "Model-based object pose in 25 lines of code," *Int. J. Comput. Vis.*, vol. 15, nos. 1–2, pp. 123–141, Jun. 1995.

[56] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Jan. 2017, pp. 6626–6637.

[57] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2018.

[58] F. P. Papantoniou, A. Lattas, S. Moschoglou, J. Deng, B. Kainz, and S. Zafeiriou, "Arc2Face: A foundation model of human faces," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 241–261.

[59] D. Q. Huynh, "Metrics for 3D rotations: Comparison and analysis," *J. Math. Imag. Vis.*, vol. 35, no. 2, pp. 155–164, Oct. 2009.

[60] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2019, pp. 6105–6114.

ROBERTO VALLE received the integrated B.S. and M.S. degrees from the Universidad Rey Juan Carlos (URJC), in 2013, and the Ph.D. degree in artificial intelligence from the Universidad Politécnica de Madrid (UPM), in 2020. Since 2022, he has been an Assistant Professor with UPM and a member of the Computer Perception Group. His research interests include computer vision and deep learning algorithms for facial analysis problems.

JOSÉ M. BUENAPOSADA received the integrated B.S. and M.S. and the Ph.D. degrees in computer science from the Universidad Politécnica de Madrid (UPM), in 1999 and 2005, respectively. From 2008 to 2018, he was an Associate Professor with the Universidad Rey Juan Carlos (URJC), Spain, where he has been a Professor of computer science, since July 2023. He is a member of the GAVAB Research Group, URJC, and the Computer Perception Group, UPM. His research interests include image alignment, face image analysis, object detection, and efficient computer vision.

LUIS BAUMELA received the integrated B.S. and M.S. and the Ph.D. degrees in computer science from the Universidad Politécnica de Madrid, in 1989 and 1995, respectively. From 1989 to 1992, he was an Engineer with Telefónica's Research and Development Laboratories. From 1997 to 2016, he was an Associate Professor. Since 2016, he has been a Professor of computer science with the ETSI Informáticos, Universidad Politécnica de Madrid, where he leads the Computer Perception Group. His research interest includes the development of ethical vision systems that look at people.