ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11461406

AIRGLOVE: EXPLORING EGOCENTRIC 3D HAND TRACKING AND APPEARANCE GENERALIZATION FOR SENSING GLOVES

Wenhui Cui[1,2*], *Ziyi Kou*[1], *Chuan Qin*[1], *Ergys Ristani*[1], *Li Guan*[1]

[1] Meta Reality Labs Research
[2] Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California

ABSTRACT

Sensing gloves have become important tools for teleoperation and robotic policy learning as they are able to provide rich signals like speed, acceleration and tactile feedback. A common approach to track gloved hands is to directly use the sensor signals (e.g., angular velocity, gravity orientation) to estimate 3D hand poses. However, sensor-based tracking can be restrictive in practice as the accuracy is often impacted by sensor signal and calibration quality. Recent advances in vision-based approaches have achieved strong performance on human hands via large-scale pre-training, but their performance on gloved hands with distinct visual appearances remains underexplored. In this work, we present the first systematic evaluation of vision-based hand tracking models on gloved hands under both zero-shot and fine-tuning setups. Our analysis shows that existing bare-hand models suffer from substantial performance degradation on sensing gloves due to large appearance gap between bare-hand and glove designs. We therefore propose *AirGlove*, which leverages existing gloves to generalize the learned glove representations towards new gloves with limited data. Experiments with multiple sensing gloves show that AirGlove effectively generalizes the hand pose models to new glove designs and achieves a significant performance boost over the compared schemes.

***Index Terms*—** 3D Hand Tracking, Adversarial Learning

1. INTRODUCTION

Sensing gloves have recently become essential in robotics research, supporting diverse applications such as teleoperation [1, 2] and human-to-robot imitation learning [3, 4]. To track human hands with gloves, common approaches mainly rely on on-glove sensors like inertial measurement units (IMUs) [5]. However, these solutions often suffer from instability in real-world applications as sensors often require sophisticated calibration processes and can degrade in quality over time [6]. Recently, significant progress has been made in vision-based pose tracking by leveraging large-scale human hand pose datasets [7] and advanced models trained on diverse hand data [8, 9]. With egocentric cameras, the vision-based models can provide stable and fine-grained hand pose estimation that accurately derives finger joints and wrist locations even under challenging conditions like in occlusion or with skin appearance variations [8, 9, 10]. While such models may be applied for sensing gloves, it remains unknown whether and how much the appearance discrepancies between human hands and gloves can lead to the performance degradation. Therefore, in this work, we investigate the potential domain discrepancy between human hands and sensing gloves in the appearance-level, which is relatively unexplored despite its critical impact on glove tracking.

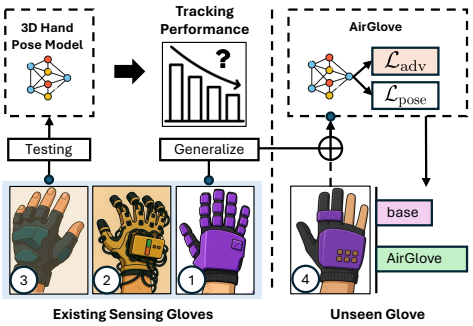

Fig. 1. *Overview of Our Study*. **(Left)** We quantitatively explore the potential degradation of vision-based hand tracking models on sensing gloves. **(Right)** We propose an appearance-invariant representation learning framework for glove generalization, which leverages adversarial learning on existing sensing glove data to enhance gloved hand tracking performance on unseen glove designs.

We further illustrate the studied problem in Fig. 1. Sensing gloves encompass a wide variety of designs tailored to different applications. One possible solution of adapting vision-based hand models to gloves is to collect glove-specific datasets with annotated hand poses [11, 12], and then finetune the models with the data. However, the correlation between the model improvement and the data scale of sensing gloves is unclear due to the lack of sufficient glove data. Moreover,

*Work done during internship at Meta.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:58:20 UTC from IEEE Xplore. Restrictions apply.

the data collection is time-consuming or even impractical in the long run given the fact that new glove designs with unique appearances are emerging.

Motivated by these limitations, we conduct a systematic study on the performance of vision-based hand tracking models across diverse sensing gloves. To mitigate data collection efforts for new glove designs, we introduce **AirGlove**, an **A**ppearance-**I**nvariant **R**epresentation Learning framework that leverages existing sensing gloves to learn appearance-invariant glove representations and generalizes to new gloves. However, two technical challenges exist to achieve the above goals. The first is *insufficient glove data* as there is currently no large-scale and real-world glove datasets that cover various glove applications. The second is *the lack of effective design for the adversarial learning method*. While the classical adversarial framework [13] can be applied, this optimization involves two competing objectives that are inherently unstable and prone to convergence issues [14, 15].

To address the first challenge, we create a multi-sensing glove dataset with millions of video frames and Optitrack-based 3D pose labels [11] to adapt and evaluate vision-based models. To address the second challenge, we design an energy-based adversarial learning strategy for AirGlove by equalizing the probability of hidden glove representations across all glove appearances, thus encouraging the model to learn appearance-agnostic representations while maintaining robust pose estimation performance. To the best of our knowledge, our work is the first to explicitly investigate and mitigate the visual appearance discrepancies between sensing gloves and human hands for vision-based hand tracking models. Experiments with multiple sensing gloves demonstrate substantial performance degradation of vision-based models and the significant enhancement of them by AirGlove.

2. PROBLEM DEFINITION

Definition 1. Vision-Based Gloved Hand Tracking: Considering a user wearing a sensing glove on their hands and an egocentric vision camera on their heads (e.g., from head-worn smart glasses or head-mount display), the user then performs various hand activities. We denote the recorded videos as $\mathcal{X}_g = \{x_{g,t}\}_{t=1}^{T}$ where $x_{g,t}$ represents t^{th} frame and g the specific sensing glove. The goal of a vision-based hand tracking model is to take $x_{g,t}$ as input and predict the hand skeleton as a set of 3D landmarks. In particular, we denote the 21 predicted landmarks as $p_{g,t} = \{p_{g,t}^k\}_{k=1}^{21}$, where each $p_{g,t}^k \in \mathbb{R}^3$ represents a 3D coordinate vector.

Definition 2. Hand Pose Ground-Truth: similar to model predictions, we represent the hand pose labels as a set of 3D landmarks [16]. Formally, the hand pose at time t and for glove g is denoted as $q_{g,t} = \{q_k\}_{k=1}^{21}$. We illustrate more details of the ground-truth collection in Section 3.

Definition 3. Glove Appearance Generalization: Given a new sensing glove with limited pose labels, our goal is to enhance vision-based models that learn appearance-invariant representations from existing sensing gloves so that the representations can be generalized to new gloves with unseen appearance. Specifically, we formulate the problem as an adversarial optimization task where an adversarial classifier mitigates the appearance gaps of different gloves in the latent features, which can be mathematically denoted as:

$$\mathcal{M}^* = \arg\min_{\mathcal{M}} \; \mathbb{E}_{(x,p)\sim\mathcal{X}} \left[\mathcal{L}_{\text{pose}}(\mathcal{M}(x), p) + \mathcal{L}_{\text{adv}}\right], \quad (1)$$

where $\mathcal{M}$ denotes the vision-based model. More detailed explanation about the terms can be found in Section 4.1 and 4.2.

Dataset	# Sessions (Train)	# Sessions (Eval)	# Frames (Train)	# Frames (Eval)
IMU-Glove	28	6	858,059	446,200
PS-Glove	4	1	98,962	49,673
MoCap-Glove	13	3	224,875	160,561
Haptic-Glove	5	1	247,479	158,946
Bare-hand	5	3	232,104	56,883

Table 1. Summary of Multi. Sensing Glove Dataset.

3. VISION-BASED GLOVE TRACKING

In this section, we detail the setup for evaluating vision-based hand tracking models on sensing gloves.

Multi. Sensing Glove Dataset. We adopt four representative types of sensing gloves to cover diverse use cases across major AR/VR and robotics applications. The prototypes of the gloves are shown in Fig. 1. These include ① *IMU-Glove* equipped with inertial sensors for capturing motion dynamics, ② *Haptic-Glove* integrating actuators such as motors or vibrotactile units to provide force and haptic feedback, ③ *Pressure-sensing (PS)-Glove* with distributed pressure sensors for measuring contact interactions, and ④ *MoCap-Glove* with optical markers for precise motion capture and annotation. We used Quest 3 to capture multi-view egocentric videos at 60 Hz and an optical motion capture system to obtain ground-truth hand poses [11]. In addition, we collect a bare-hand dataset to establish a baseline for the hand tracking models of interest. Each participant contributes one session, during which they are instructed to perform a variety of hand motions (e.g., swipe, pinch, free-form). For each glove, we randomly split the train/test data at the session level and show the summary of the data in Table 1.

Vision-Based Bare-Hand Models. We adopt two vision-based bare-hand models that are directly compatible with our motion capture system annotations: (i) **MEgATrack** [16]: a real-time egocentric hand-tracking framework that consists of a ResNet-based pose regressor to regress 3D hand keypoints. [7]. (ii) **UmeTrack** [17]: a multi-view hand tracking framework designed for VR that takes egocentric wide-FOV camera inputs and predicts 3D hand pose in world coordinates. Both are trained on over two million bare-hand egocentric visual data and generalize well across hand sizes and skin tones.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:58:20 UTC from IEEE Xplore. Restrictions apply.

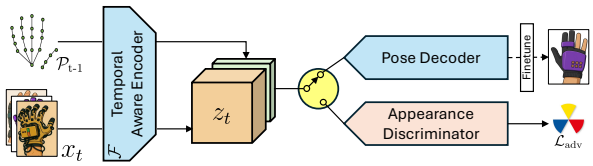

Fig. 2. *Overview of AirGlove.* The temporal-aware encoder extracts visual representations from egocentric videos, followed by the 3D pose decoder for pose estimation. The adversarial appearance discriminator iteratively regulates the glove representations to derive appearance-invariant features.

4. GLOVE APPEARANCE GENERALIZATION

Given the exploration in Section 3 on vision-based hand tracking models, we further enhance their glove tracking performance when limited annotated data is available. A simple solution is to fine-tune existing bare-hand models on the new glove data. However, when the available data scale is small, the models risk overfitting and fail to capture sufficient hand pose diversity. In addition, continuously collecting annotated datasets for every new glove design is impractical, as glove designs evolve rapidly with varying visual appearances, and obtaining 3D pose annotations is a resource-intensive process.

We propose AirGlove that consists of two modules: i) a temporal-aware deep visual network and ii) an adversarial appearance-invariant discriminator. In particular, AirGlove is optimized on data collected from existing gloves, denoted as $\mathcal{G} = \{g_k\}$, and expected to generalize to a new glove, denoted as $\hat{g}$. We show the overview of our proposed AirGlove framework in Fig. 2 and introduce each module below.

4.1. Temporal-Aware Deep Visual Network (TADV-Net)

We follow [16] to build our visual encoder by leveraging the video frame $x_{g,t}$ and the hand pose $q_{g,t-1}$ in the previous timestamp to provide contextual motion priors. We formulate the encoding process as $z_{g,t} = \mathcal{F}_{\text{enc}}(x_t, q_{g,t-1})$, where $z_{g,t} \in \mathbb{R}^d$ denotes the encoded d-dimensional representations. Note that during the inference stage, we replace $q_{g,t-1}$ with the $p_{g,t-1}$ to keep the auto-regressive predictions by the model.

The pose decoder is a multi-stacked MLP that converts $z_{g,t}$ to both 2D and 1D heatmaps for hand keypoints and depth, respectively. Then a 3D hand pose loss is applied as

$$\mathcal{L}_{\text{pose}} = \frac{1}{K}\sum_{k=1}^{K} \|\hat{h}_k^{\text{2D}} - h_k^{\text{2D}}\|_2^2 + \|\hat{h}_k^{\text{1D}} - h_k^{\text{1D}}\|_2^2 \qquad (2)$$

where K is the total number of keypoints, $h^{\text{1D}}/h^{\text{2D}}$ represent the 1D/2D heatmaps. $\hat{h}_k$ denotes the predictions.

4.2. Adversarial Appearance-Invariant Discriminator

The design of the Adversarial Appearance-Invariant Discriminator (**AAID**) is based on the assumption that each glove with a unique appearance can be considered as a semantic class. Given the representation $z_{g,t}$ of a particular frame, AAID should assign equal probabilities to all gloves if $z_{g,t}$ is not biased to any glove appearance. Specifically, we construct a convolutional classifier θ_{cls} that takes $z_{g,t}$ as input and outputs $|\mathcal{G}|$ dimensions as predicted classes. To suppress appearance information, we define the adversarial loss $\mathcal{L}_{\text{adv}}$ as the Kullback–Leibler (KL) [18] divergence between the classifier output and a uniform prior U. Minimizing the divergence can be formulated as $\mathcal{L}_{\text{adv}}$:

$$D_{\text{KL}}(U \parallel \theta_{\text{cls}}(z)) = -\frac{1}{|\mathcal{G}|}\sum_{c=1}^{|\mathcal{G}|} \log \theta_{\text{cls}}(z) - \log |\mathcal{G}| \qquad (3)$$

where $\theta_{\text{cls}}(z)$ is the output of the classifier. The total optimization objective for AirGlove is defined as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pose}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}$ where $\lambda_{\text{adv}} \in [0, 1]$.

We adopt an alternating optimization strategy to jointly optimize pose estimation and disentangle glove-specific appearances. The process alternates between two phases: (i) freeze TADV-Net and update θ_{cls} by minimizing the multi-class cross-entropy $\mathcal{L}_{\text{cls}}$; (ii) freeze θ_{cls} and update TADV-Net by minimizing $\mathcal{L}_{\text{total}}$. These two phases alternate every $E = 3$ epochs. By ensuring that AAID adapts to the evolving visual representations while TADV-Net improves its ability to challenge the discriminator in return, AirGlove effectively learns appearance-invariant glove representations.

Eval Data	Model	MKPE (↓)	MKPE.T (↓)	F-MKPE (↓)	F-MKPE.T (↓)
Bare-hand	MEgATrack	16.677	8.408	21.162	16.325
	UmeTrack	15.150	7.970	18.574	15.142
IMU-Glove	MEgATrack	33.938	13.009	40.049	25.837
	UmeTrack	18.559	10.273	22.823	20.286
PS-Glove	MEgATrack	73.979	19.401	78.751	39.514
	UmeTrack	54.601	13.804	62.398	27.350
MoCap-Glove	MEgATrack	21.487	11.976	27.046	22.967
	UmeTrack	57.452	12.099	62.953	23.146
Haptic-Glove	MEgATrack	63.509	16.232	66.466	32.151
	UmeTrack	60.525	14.840	66.473	29.190

Table 2. *Comparison of pose tracking results (mm) on bare-hand and sensing gloves.* Both evaluated models show substantial performance degradation when applied to gloves.

5. EVALUATION

We conduct extensive experiments on the collected datasets to answer the following research questions: **Q1**) How much does the pose tracking performance degrade when applying the vision-based bare-hand models directly to sensing gloves? **Q2**) Given a new glove with limited annotated data, how effectively can AirGlove enhance the glove tracking performance? **Q3**) How well does AirGlove learn appearance-invariant representations from existing sensing gloves?

Evaluating Bare-hand Models on Gloves (Q1). We follow [16] to adopt *Mean Keypoint Position Error* (MKPE) and

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:58:20 UTC from IEEE Xplore. Restrictions apply.

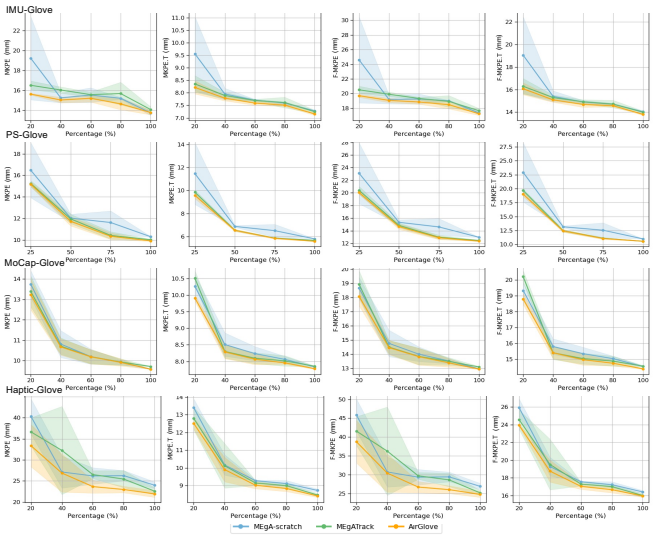

Fig. 3. Evaluation of AirGlove on sensing glove datasets. AirGlove achieves superior tracking performance compared to all baselines across different sensing gloves (row-wise) and evaluation metrics (column-wise). Best viewed in color.

Fingertip MKPE (F-MKPE), together with their transformed variants (MKPE.T and F-MKPE.T) as the evaluation metrics. We show the evaluation results in Table 2. We observe that both MEgATrack and UmeTrack show substantially degraded performance on sensing gloves compared to bare-hands. The findings highlight the significant gap between sensing gloves and bare-hands in vision-based methods, emphasizing the challenge with appearance discrepancies.

Evaluating AirGlove in Fine-tuning Settings (Q2). To answer Q2, we take MEgATrack as the baseline to compare with AirGlove as they share the same visual encoder backbone. We train both MEgATrack and AirGlove on the union of three sensing glove datasets, leaving one glove out as the held-out set. Given a new target glove, we fine-tune both MEgATrack and AirGlove on increasing proportions of the training data from 20% to 100%. For each proportion, we randomly sample the data 5 times and report the averaged results with standard deviation. We also propose **MEgA-scratch** that is trained from scratch on the same data proportion of the target glove. We show the quantitative evaluation results in Fig. 3. We observe that AirGlove consistently outperforms all compared schemes at every data proportion and across all metrics. We attribute the generalization performance of AirGlove to its adversarial learning strategy, which effectively learns pose representations that are invariant to appearances.

We qualitatively evaluate the tracking performance of AirGlove by rendering 3D hand meshes based on the predicted hand poses [19]. For Haptic-Glove as the unseen glove, we compare reconstructed meshes from AirGlove and MEgATrack. As shown in Fig. 4, AirGlove yields hand poses more closely aligned with ground truth than the baseline.

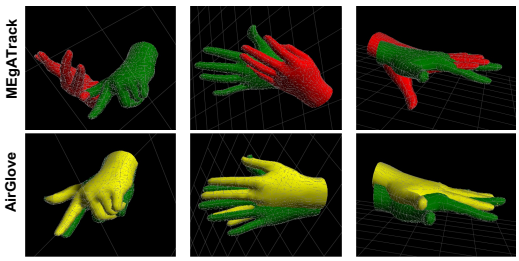

Fig. 4. *Visualization for AirGlove.* Compared to the baseline MEgATrack (red), AirGlove (yellow) generates predictions that are better aligned with the ground-truth (green).

	F1-Score (↓)	Accuracy (↓)
Fig. (a)	0.93	0.95
Fig. (b)	**0.32**	**0.43**

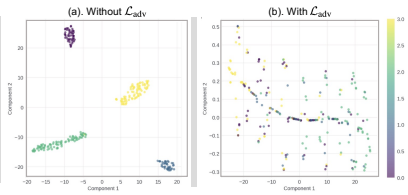

Fig. 5. *Glove classification results and t-SNE visualization of glove representations.* With $\mathcal{L}_{\text{adv}}$ vs. without, the model learns features that cannot differentiate glove appearances, leading to pose representations without appearance bias.

Ablation Study (Q3) We train AirGlove on the union of all glove datasets, once with $\mathcal{L}_{\text{adv}}$ enabled and once without it. After training, we freeze the pose encoder and evaluate the adversarial classifier θ_{cls} on held-out samples to measure classification *Accuracy* and *F1-score* [20]. We also extract the outputs from θ_{cls} and randomly select a subset of samples for visualization using t-SNE [21]. Fig. 5 shows both quantitative and qualitative results. We observe that enabling $\mathcal{L}_{\text{adv}}$ significantly degrades the classification performance, which is aligned with the visualization where the feature clusters for each glove collapses into a mixed distribution. These results confirm that $\mathcal{L}_{\text{adv}}$ effectively suppresses appearance-specific information in representations learned from the pose encoder, thus enhances generalization under glove appearance gaps.

6. CONCLUSION

In this work, we investigated the hand tracking problem for sensing gloves, where appearance variations introduced by diverse glove designs severely degrade the performance of vision-based bare-hand models. To mitigate the issue, we introduced AirGlove that learns appearance-invariant representations, enabling generalizable glove tracking without the need for large-scale training data on a new glove design. Experimental results demonstrate that AirGlove can enhance the generalization of gloved hand tracking on new glove designs, which consistently outperforms comparing schemes.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:58:20 UTC from IEEE Xplore. Restrictions apply.

7. REFERENCES

[1] Noah Becker, Kyrylo Sovailo, Chunyao Zhu, Erik Gattung, Kay Hansel, Tim Schneider, Yaonan Zhu, Yasuhisa Hasegawa, and Jan Peters, "Integrating and evaluating visuo-tactile sensing with haptic feedback for teleoperated robot manipulation," 2024.

[2] Manuel Caeiro-Rodríguez, Iván Otero-González, Fernando A Mikic-Fonte, and Martín Llamas-Nistal, "A systematic review of commercial smart gloves: Current status and applications," *Sensors*, vol. 21, no. 8, pp. 2667, 2021.

[3] Qingtao Liu, Qi Ye, Zhengnan Sun, Yu Cui, Gaofeng Li, and Jiming Chen, "Masked visual-tactile pre-training for robot manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 13859–13875.

[4] Elmar Rueckert, Rudolf Lioutikov, Roberto Calandra, Marius Schmidt, Philipp Beckerle, and Jan Peters, "Low-cost sensor glove with force feedback for learning from demonstrations using probabilistic trajectory representations," 2015.

[5] Arvin Tashakori, Zenan Jiang, Amir Servati, Saeid Soltanian, Harishkumar Narayana, Katherine Le, Caroline Nakayama, Chieh-ling Yang, Z Jane Wang, Janice J Eng, et al., "Capturing complex hand movements and object interactions using machine learning-powered stretchable smart textile gloves," *Nature Machine Intelligence*, vol. 6, no. 1, pp. 106–118, 2024.

[6] Anna Filipowska, Wojciech Filipowski, Paweł Raif, Marcin Pieniażek, Julia Bodak, Piotr Ferst, Kamil Pilarski, Szymon Sieciński, Rafał Jan Doniec, Julia Mieszczanin, et al., "Machine learning-based gesture recognition glove: design and implementation," *Sensors (Basel, Switzerland)*, vol. 24, no. 18, pp. 6157, 2024.

[7] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin, "Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 12999–13008.

[8] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik, "Reconstructing hands in 3d with transformers," 2023.

[9] A. Prakash et al., "3d hand pose estimation in everyday egocentric images," *arXiv preprint arXiv:2404.09308*, 2024.

[10] Wencan Cheng, Hao Tang, Luc Van Gool, and Jong Hwan Ko, "Handdiff: 3d hand pose estimation with diffusion on image-point cloud," 2024.

[11] Gergely Nagymáté and Rita M Kiss, "Application of optitrack motion capture systems in human movement analysis," 2018.

[12] Gyeongsik Moon, Shoou i Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee, "Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image," 2020.

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," 2014.

[14] Maciej Wiatrak, Stefano V. Albrecht, and Andrew Nystrom, "Stabilizing generative adversarial networks: A survey," 2020.

[15] Wanying Wang et al., "Exploring gradient explosion in generative adversarial imitation learning: A probabilistic perspective," 2023.

[16] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D. Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, Asaf Nitzan, Gang Dong, Yuting Ye, Lingling Tao, Chengde Wan, and Robert Wang, "Megatrack: monochrome egocentric articulated hand-tracking for virtual reality," *ACM Trans. Graph.*, vol. 39, no. 4, Aug. 2020.

[17] Shangchen Han, Po-Chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, Randi Cabezas, Luan Tran, Muzaffer Akbay, Tsz-Ho Yu, Cem Keskin, and Robert Wang, "Umetrack: Unified multi-view end-to-end hand tracking for vr," in *SIGGRAPH Asia 2022 Conference Papers*. Nov. 2022, SA '22, p. 1–9, ACM.

[18] Solomon Kullback and Richard A Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[19] Sasha Salter et al., "emg2pose: A large and diverse benchmark for surface electromyographic hand pose estimation," 2024.

[20] Jasmina Dj Novaković, Alempije Veljović, Siniša S Ilić, Željko Papić, and Milica Tomović, "Evaluation of classification models in machine learning," *Theory and Applications of Mathematics & Computer Science*, vol. 7, no. 1, pp. 39, 2017.

[21] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:58:20 UTC from IEEE Xplore. Restrictions apply.