ISSN 2464-4617 (print)
ISSN 2464-4625 (online)
Computer Science Research Notes - CSRN
http://www.wscg.eu
WSCG 2025 Proceedings

HaGRIDv2: 1M Images for Static and Dynamic Hand Gesture Recognition

Anton Nuzhdin
SaluteDevices
anton.nuzhdin.hse@gmail.com

Alexander Nagaev
SaluteDevices
sashanagaev1111@gmail.com

Alexander Sautin
SaluteDevices
sasautin007@gmail.com

Alexander Kapitanov
SaluteDevices
kapitanovalexander@gmail.com

Karina Kvanchiani
SaluteDevices
karinakvanciani@gmail.com

ABSTRACT

This paper proposes the second version of the widespread image-based Hand Gesture Recognition dataset HaGRID – HaGRIDv2. We have added 15 new gestures to the existing 18, encompassing both conversational and control functions, including two-handed gestures. Building on the foundational concepts proposed by HaGRID's authors, we implemented the dynamic gesture recognition algorithm and further enhanced it by adding three new groups of manipulation gestures. The "no gesture" class was significantly expanded and redefined with a new semantic focus to include a diverse range of natural hand movements, which led to a 16-fold reduction in false positives on HaGRIDv2. The HaGRIDv2 dataset outperforms the original HaGRID in pre-training models for gesture-related tasks. Besides, we achieved the best generalization ability among gesture and hand detection datasets. Additionally, the second version of the dataset provides a diverse range of hand samples, which is crucial for fine-tuning modern diffusion models. By fine-tuning on HaGRIDv2, these models achieve improved outcomes in generating anatomically correct hand gesture images. HaGRIDv2, pre-trained models, and a dynamic gesture recognition algorithm are publicly available.

Keywords
Hand Gesture Recognition, Human-Computer Interaction, HGR Dataset

Figure 1: The 15 outlined in red new gesture classes added to HaGRID's 18 ones ("inv" stands for "inverted").

1 INTRODUCTION

Hand gestures, as natural and intuitive expressions, effectively reflect emotions and facilitate communication [16]. Their ability to convey messages quickly makes gestures invaluable for Human-Computer Interaction (HCI) [57]. Thus, the development of Hand Gesture Recognition (HGR) systems has the potential to significantly enhance user interfaces in various domains [27, 66] such as robotic control [54], driver assistance [52], and medicine [64, 69, 14] for touchless interaction. The proposed research aims to develop a comprehensive HGR system for video conferencing [2, 1] and home automation devices [5, 65, 17, 7, 12]. The system should enhance participants' communication, enable remote control of device functions [67, 26], allow the manipulation of various objects on the screen, and activate different platform features [15]. Considering the described application area, the system should enable intuitive operation through easy-to-demonstrate functional gestures, offer instant feedback, and efficiently operate on resource-constrained edge devices.

Neural networks have recently become the primary component of HGR systems [28, 47, 46] and action recognition in general [32, 60, 11]. Large datasets aligned with the system constraints described above are required to train a model resilient to real-world conditions. Based on the requirements described above, the appropriate dataset should include a range of gestures categorized as manipulative (e.g., clicking, swiping, or zooming the screen), control (e.g., taking screenshots), and conversational (e.g., expressing approval, disapproval, love, anger, regret) [59, 24]. The first category usually implies dynamic gestures, which involve motion and are therefore captured as videos. In contrast, the latter two categories consist of static gestures, in which a specific hand or finger pose remains fixed (e.g., giving a thumbs-up to convey approval). Thus, the dataset should encompass static and dynamic gestures (i.e., videos rather than images) to cover all three categories effectively. The HGR model should efficiently provide real-time inference on the CPU while remaining lightweight, as it should be integrated into the resource-constrained device. Compliance with such requirements significantly restricts the choice of neural network architectures capable of operating in the temporal dimension.

Most existing gesture datasets are limited in several ways: some overlook functional and commonsense gestures, others only cover static or dynamic gestures, and many lack sufficient diversity in subjects (i.e., the individuals performing the gestures). The most suitable dataset, HaGRID [23], includes various conversational and control gestures and proposes an algorithm for dynamic manipulative gesture recognition. The algorithm required static gestures to construct dynamic ones, satisfying the conditions of the lightweight model. The authors of [23] aimed to create the HGR system for home automation and video conferencing services. Therefore, all chosen gestures are straightforward and perform a specific function. However, HaGRID lacks support for essential device interaction gestures, such as clicking, zooming, taking screenshots, and cursor control. Besides, the extra “no gesture” class contains too homogeneous hands, provoking the model to predict false positives. In this regard, we decided to expand the dataset with new gestures to support the development of dynamic gestures, introduce a new “no gesture” class (see Fig. 4), and enhance the dynamic gesture recognition algorithm.

It was decided to expand the HaGRID dataset by adding new classes. This paper introduces HaGRIDv2, the second version of the HaGRID dataset, designed to enrich the functionality of HGR systems for video conferencing and home automation. Although HaGRIDv2 was developed primarily for gesture detection, we explore its applicability to other tasks, including full-frame gesture classification, hand detection, and text-to-image gesture generation (see Fig. 2). The contribution of this paper is four-fold:

- **Extended gesture classes.** HaGRIDv2 incorporates 15 new gesture classes (Fig. 1) performing control and conversational functions. Cross-dataset evaluation experiments confirmed that HaGRIDv2 demonstrates the best domain generalization ability among gesture detection datasets (see Tab. 5).
- **Improved “no gesture”.** The “no gesture” class has been improved compared to HaGRID by incorporating domain-specific natural hand positions, resulting in a 16-fold reduction in false positives (see Fig. 8).
- **Support for dynamic gestures.** We extend the dynamic gesture recognition algorithm [23] by developing swipes, clicks, zooms, drag-and-drops, and other manipulative gestures.
- **Multi-task experiments.** Experimental results show that HaGRIDv2 excels even beyond its main scope: it achieves top scores on standard benchmarks for hand detection (Tab. 6), significantly improves pre-training for various gesture-related tasks (Fig. 7), and effectively addresses anatomically incorrect gestures generated by diffusion models (Fig. 9).

The dataset and the extended algorithm for recognizing dynamic gestures are publicly available[1][2] under the modified Creative Commons CC-BY 4.0 license.

The paper is organized as follows: Sec. 2 reviews key gesture datasets regarding applicability to device control systems. Sec. 3 outlines the HaGRIDv2 collection process and its key characteristics. Sec. 4 introduces the dynamic gesture recognition algorithm trained solely on static images, and Sec. 5 presents baseline experiments across various tasks. Cross-dataset generalization is examined in Sec. 6, and Sec. 7 compares HaGRIDv2 and HaGRID in terms of pre-training performance, false positive triggering, and gesture generation quality. Ethical considerations and limitations are discussed in Sec. 8 and Sec. 9, respectively. Finally, Sec. 10 summarizes findings and future directions.

2 RELATED WORK

There are a variety of HGR datasets with gestures categorized based on their applications [59, 24]: sign language [53], control [68, 58, 9, 21, 23, 51, 62], conversational [10, 23, 4, 51, 42, 43, 39, 21] and manipulative gestures [47, 68, 41, 9, 34, 56, 6]. The proposed research aims to develop an HGR system for device control and video conferencing, where manipulative, control, and conversational gestures are essential. Therefore, the system should recognize both static and dynamic gestures, which are reviewed in this study. Sign language recognition datasets are excluded because their gestures are unsuitable for performing the described functions.

Dynamic gesture datasets are typically annotated for action recognition, classifying entire video sequences.

[1] https://github.com/hukenovs/hagrid
[2] https://github.com/ai-forever/dynamic_gestures

Dataset	Samples	Classes	Subjects	Scenes	Resolution	Annotations	Annotation Method
				Gesture Detection			
LaRED, 2014 [21]	243,000	81	10	10	640×480	masks	automatically
OUHANDS, 2016 [42]	3,000	10	23	various	640×480	masks, boxes	automatically
HANDS, 2021 [51]	12,000	29	5	5	960×540	boxes	–
SHAPE, 2022 [4]	33,471	31+1	20	various	4128×3096	masks, boxes	manually
HaGRID, 2023 [23]	554,800	18+1	37,583	≥37,583	1920×1080	boxes, keypoints*	manually
HaGRIDv2, 2025 (ours)	1,086,158	33+1	65,977	≥65,977	1920×1080	boxes, keypoints*	mixed
				Gesture Classification			
HTU HGR, 2011 [55]	1,000	10	10	various	640×480	labels	manually
Kinect Leap, 2014 [39, 40]	1,400	10	14	–	640×480	labels	manually
Senz3D, 2015 [43, 44]	1,320	11	4	–	640×480	labels	manually
SIT-HANDS, 2023 [10]	4,200	14	10	various	1920×1080	labels	manually
				Hand Detection			
HAND, 2011 [45]	5,628	1	–	–	mixed	boxes	automatically
EgoHands, 2015 [8]	4,800	1	various	various	720×1280	masks, boxes	mixed
Human-Parts, 2019 [31]	14,962	3	–	various	mixed	boxes	automatically
TV-Hands, 2019 [50]	9,498	1	various	various	–	boxes	manually
ContactHands, 2020 [49]	20,516	1	–	various	–	boxes	manually
BodyHands, 2022 [48]	20,490	2	–	various	–	boxes	manually

Table 1: The main parameters of the most popular gesture datasets. "+1" in the third column means the dataset contains an extra class "no gesture". "-" in some columns means the information was not found. * – keypoints prepared by using the MediaPipe [38] hand model. Subjects refer to the distinct individuals performing the gestures, while scenes include variations in background, lighting, camera angle, and other environmental factors. Such diversity helps the model learn to generalize across different setups and reduces overfitting to specific conditions.

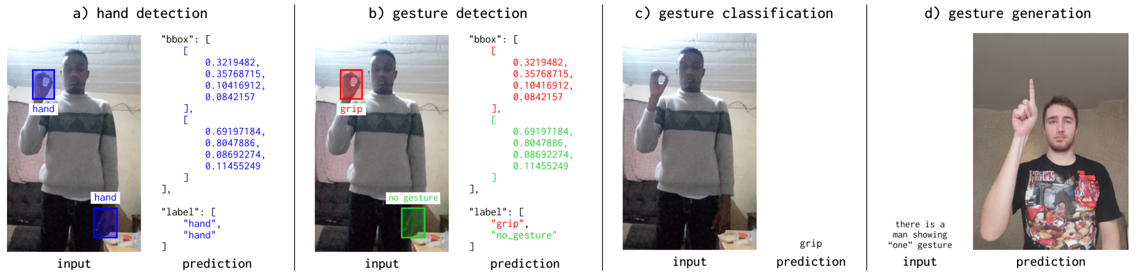

Figure 2: Different tasks addressed by HaGRIDv2. (a) The hand detector recognizes all hands by bounding boxes with the same label "hand"; (b) The gesture detector aims to predict a bounding box with a label for each hand on the image; (c) The gesture classifier produces a label for the entire image; (d) The gesture generator creates an image of a person showing a gesture according to the prompt.

In contrast, static gesture recognition can be achieved by solving various tasks, including gesture detection and classification (see Fig. 2 for difference), hand gesture keypoint estimation, and gesture segmentation. However, classification labels are impractical in multi-person frames, keypoints can stick together when the person is far away, and segmentation masks are excessive for gesture recognition and yield no meaningful benefit over detection-based approaches. While many existing detection datasets typically feature just one person per frame, bounding box annotations and augmentations such as mosaic used in YOLO [63] still enable training models to detect an unlimited number of gestures in a single image. Additionally, only third-person (i.e., captured from an external viewpoint rather than the user's) data are suitable for video conferences and device control.

An overview of datasets related to HGR is organized as follows. Sec. 2.1 discusses datasets relevant to the device control task, while Sec. 2.2 examines datasets related to gestures data domains not central to this research.

2.1 Devices Control

Static Gestures. Tab. 1 shows that only HaGRID [23], LaRED [21], OUHANDS [42], HANDS [51], and SHAPE [4] datasets are relevant for the required annotations. Each dataset has limitations that can affect its suitability for developing a reliable HGR system for real-world conditions. The LaRED [21] and the OUHANDS [42] datasets include images captured from close distances, making them unsuitable for training models intended to operate in larger environments. Besides, there is no currently access to the LaRED samples due to the outdated link. Also, OUHANDS [42] and HANDS [51], constructed with only 23 and 5 subjects, respectively, are inappropriate for developing a robust HGR system. In addition, even the most diverse datasets on classes lack gesture variety for control and conversational purposes. So, while the SHAPE [4] dataset lacks enough control gestures, the HaGRID [23] suffers from a deficiency of conversational ones, containing only "like" and "dislike" emotional gestures. Since emotional gestures are essential for video meetings, such an omission restricts the overall functionality of the HGR system.

Dynamic Gestures. Datasets such as [58, 34, 56, 62, 47, 6, 68, 41, 9] are the most relevant for dynamic gesture recognition in device interaction and video conferencing. There are only ChAirGest [56], Jester [41], and IPN Hand [9] datasets meeting the described requirements about the existence of functional gestures and third-person view. However, these datasets focus mainly on manipulative dynamic gestures and lack the necessary static conversational and control gestures. This gap highlights their insufficiency in covering the full spectrum of gestures needed for device control and video conferencing.

While a completely suitable dataset is unavailable, the HaGRID dataset includes both control and conversational gestures, allowing for the recognition of dynamic manipulative gestures as proposed in the algorithm referenced in [23]. Therefore, we introduced a number of substantial improvements and developed the second version of HaGRID, which includes new gesture classes and a more diverse and representative 'no gesture' category (see Fig. 4).

2.2 Other Tasks

HaGRIDv2's design, with one gesture per frame, supports full-frame classification, which is ideal for single-user interactions with personal devices. Additionally, HaGRIDv2 includes a wide variety of hand postures, including complex ones and natural hand positions, making it well-suited for hand detection tasks.

Hand Gesture Classification. All the static datasets for object detection reviewed below can also be used for image classification. In addition to them, there are NTU HGR [55], Kinect Leap [39], Senz3D [43], and SIT-HANDS [10] (see Tab. 1). However, these datasets also have disadvantages in solving the human-computer interaction problem. The NTU HGR [55], Senz3D [43], and Kinect Leap [39] datasets contain 100-140 samples per gesture class, which is not enough to build a sufficiently high-quality model (see ablation study in [23]). In addition, Senz3D and Kinect Leap scenes are homogeneous, and the gestures are too close to the camera. The SIT-HANDS [10] dataset was made heterogeneous in subjects, lighting conditions, and background. However, the dataset has only 2,800 training frames and contains a relatively small variety of gestures, severely limiting the system's functionality.

Hand Detection. Among the hand detection datasets there are HAND [45], EgoHands [8], HumanParts [31], TV-Hands [50], ContactHands [49], BodyHands [48]. Hands detection datasets may differ in the annotation type: some use horizontally aligned bounding boxes, while others use oriented bounding boxes. Since most well-known detection models work with horizontally aligned bounding boxes, this type is preferred, so we did not conduct experiments on HAND [45], TV-Hands [50], and ContactHands [49] for simplicity.

Datasets vary in the number of people, subjects, resolution, and scenes. EgoHands [8] is intended for first-person gesture detection and segmentation across various backgrounds, while BodyHands [48] and HumanParts [31] offer 14,000 and 15,000 diverse third-person samples, respectively. Due to their heterogeneity, these datasets are valuable benchmarks for hand detection, but they focus only on natural hand postures and do not include specific gestures or complex finger positions. By addressing some key limitations in existing datasets, HaGRIDv2 offers improved support for these areas.

3 HAGRIDV2 DATASET

The HaGRIDv2 dataset follows the same design as HaGRID, featuring 33 static gestures (shown in Fig. 1) along with additional "no gesture" samples to prevent false positives. The dataset is used to train gesture recognition models, enabling the development of HCI and device control systems.

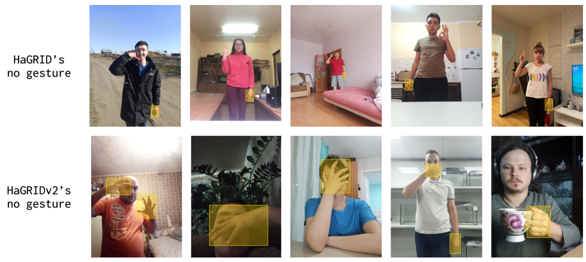

Figure 4: Samples of the "no gesture" class in HaGRID and HaGRIDv2 datasets.

The HaGRIDv2 dataset differs from the HaGRID with the following key updates:

- We added "holy", "heart" (in two variations, see Fig. 1), "middle finger", and "gun" as emotional, conversational gestures used during the conversation, and "three3" for extra functions.
- We expanded the range of control gestures by one-handed "thumb index", "grip", "point", "pinkie", and "grabbing", and two-handed "thumb index2", "timeout", "take photo", "xsign".
- Some of added static gesture classes were designed to enable the extension of the dynamic gesture recognition algorithm by developing such gestures as "drag and drop", "click", "zoom in", "zoom out" and new variations of swipes.

  In total, 15 new static gesture classes and four groups of dynamic gestures (swipes, drag-and-drops, zooms, clicks) were developed to cover various device functions (see Fig. 1).

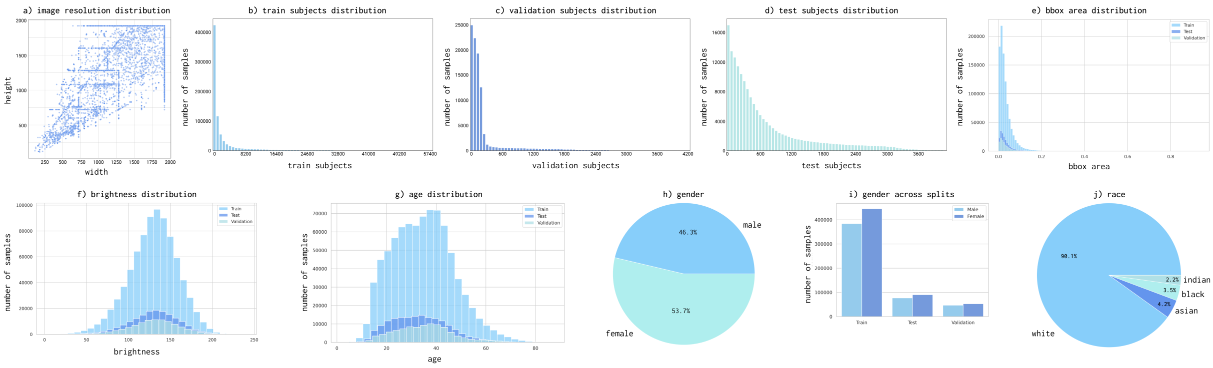

Figure 3: The key statistics of HaGRIDv2. (a) Image resolution distribution showing the scatter of image dimensions; (bâd) Distribution of subjects in the training, validation, and test sets, with the x-axis indicating the number of subjects who recorded the number of samples shown on the y-axis; (e) Bounding box area distribution; (f) Brightness distribution; (g-i) Age and gender distributions of subjects, received automatically by MiVOLO [29] neural network; (j) Racial distribution of subjects, received automatically by FairFace [25] neural network.

- The HaGRIDv2's "no gesture" class was explicitly designed to address false positives by including a broader range of natural hand positions (e.g., relaxed hands near the face, holding a cup, natural gesticulation), whereas HaGRID is limited to a single relaxed hanging hand posture (see Fig. 4). We studied popular hand positions in setups involving device interaction and video conferencing platforms, identified the most frequent gestures, and specifically collected these natural movements.

3.1 Dataset Creating Pipeline

The data creation pipeline almost followed the one proposed by the original HaGRID authors [23] to maintain the consistency of HaGRIDv2's data distribution. Also, such a pipeline allows us to collect heterogeneous samples in large volumes. We used the same crowdsourcing platforms, such as Yandex.Toloka[3] and ABC Elementary[4] and instructions for crowdworkers through mining, validation, and filtration steps, described in Dataset Creating Pipeline in [23]. In the mining stage, crowdworkers captured photos with a specified gesture under controlled conditions. In the validation stage, validators review images on the crowdsource platforms to ensure the gesture is performed correctly, and only properly executed photos are retained. Finally, the filtration stage aimed to remove ethically sensitive images, including those featuring individuals under 18.

Annotation. We have decided to replace HaGRID's manual box annotation with an automated one due to its time- and labor-intensive nature. A substantial size of HaGRID is enough to train a robust hand detector for automated annotations, as demonstrated by HaGRID's authors in their ablation study [23]. We have also implemented crowd moderation for quality assurance.

[3] https://toloka.yandex.ru/
[4] https://elementary.activebc.ru/

The annotation process consisted of two stages: hand detection and distinguishing gesturing hands from "no gesture". We trained the YOLOv10x [63] detector on the HaGRID, previously reducing all gesture classes to one class "hand". Each bounding box was assigned either the corresponding gesture label or "no gesture," leveraging the fact that the gesture name for each image is known from the data collection stage. For images featuring one-handed gestures, the hanging hand is identified as the one that is always below the hand with the gesture (see Fig. 5a). For two-handed gestures, we obtain joint boxes by combining two hand boxes, making a box from the upper left edge of the left hand to the lower edge of the right hand (see Fig. 5b). Fig. 5c shows the exception of the "xsign" gesture, where two boxes are merged to a square, the side of which is equal to the distance between the extreme points of the boxes.

3.2 Dataset Characteristics

The HaGRIDv2 dataset is an extension of the widely used HGR dataset HaGRID. Adding 531,358 samples divided into 15 new gesture classes and the extra "no gesture" class to the original HaGRID, the received combination contains over a million primarily Full HD RGB images (see Fig. 1 and Fig. 3a). Since the presented research aims to build a system for home automation devices and conference control, added classes are related to these domains. Thus, each gesture is intended to perform a specific associative function. A special "no gesture" class encompasses domain-specific hand postures common for the mentioned applications, such as hands near the face, relaxed, or holding objects. Added static gestures contributed to an extension of the dynamic gesture recognition algorithm presented in [23]. There is the opportunity to recognize such dynamic gestures as "zoom", "click", and others.

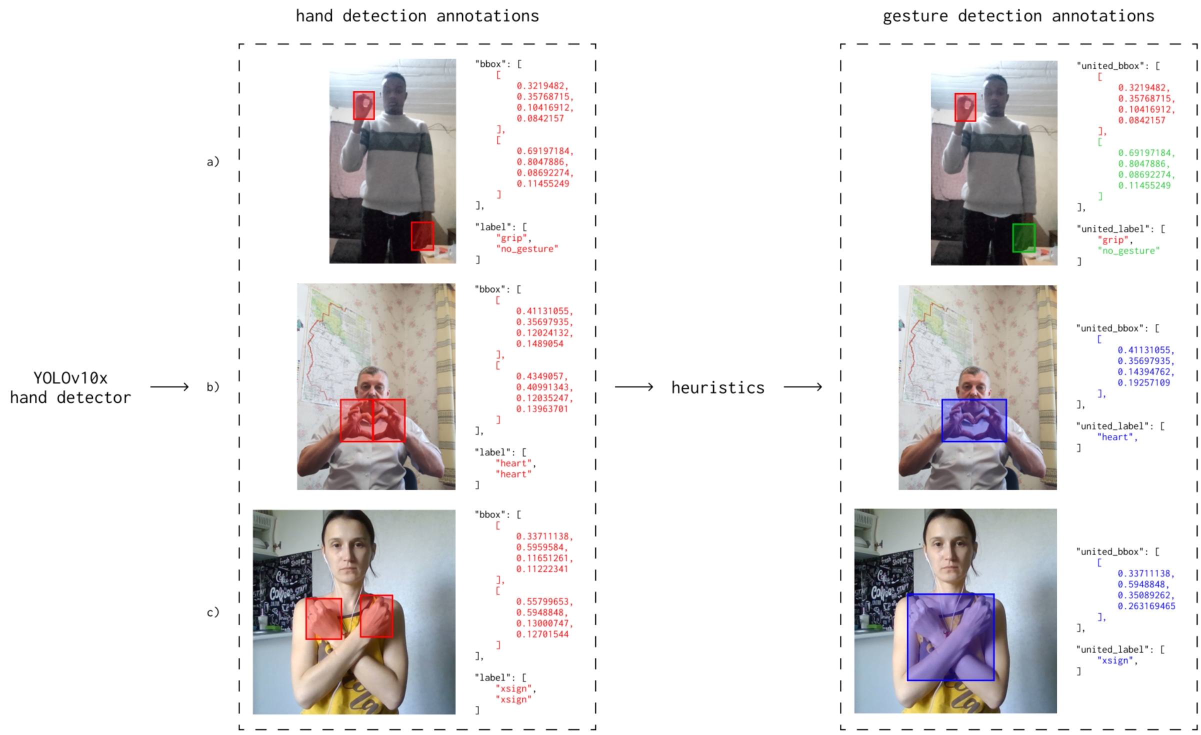

Figure 5: The pipeline for automatic image annotation. a) The higher hand on the one-handed gesture image is marked as the gesticulating hand; b) Two predicted boxes for the two-handed gesture are merged into a single box. c) A two-handed "xsign" gesture is marked by a square bounding box, received by stretching a vertical line equal to the distance between two boxes.

Content. New samples were recorded by 28,394 unique crowdworkers, each located in their own scene. Fig. 3e-g shows the age, gender, and race distributions of the subjects, calculated for all HaGRIDv2's 65,977 subjects. We preserve the HaGRID distribution and ensure domain-specific relevance by collecting samples in realistic indoor conditions with varying lighting conditions and subject-to-camera distances. The mean and standard deviation of the pixel values of the HaGRIDv2 images for the RGB channels are [0.54, 0.5, 0.474] and [0.234, 0.235, 0.231], respectively. These statistics were calculated after normalizing the original 8-bit RGB images with pixel values in the range [0â255].

Annotations. HaGRIDv2 includes bounding box annotations for all hands. Each image has one or two boxes for one-handed gestures: one for the gesturing hand and one for the non-gesturing hand if it is in the frame (see Fig. 5a). Images with two-handed gestures strictly correspond to two boxes for each hand. Moving from the hand detection task to gesture detection, an additional box encompassing both hands is added to two-handed gesture images (see Fig. 5b-c). Bounding box annotations are proposed in COCO [33] format with normalized relative coordinates.

Splitting. The dataset was divided into training (76%), validation (9%), and testing (15%) sets, recorded by 57,656, 4,209, and 4,114 subjects, respectively. Note that training, validation, and test sets of original HaGRID are the subsets of corresponding HaGRIDv2 sets. Fig. 3b-d illustrate the subjects' distribution across three sets with improved heterogeneity in the test and validation sets compared to the training one. Sets are balanced in age, gender, brightness, and race due to randomness.

In addition, we provide hashed user IDs for researchers to split the dataset on their own. Also, such automatically received meta information as age, gender, and race for each subject, and keypoints obtained by MediaPipe [38] for each hand have also been supplied. Since the dataset is large, we also provided the lightweight version, with all images resized to 512 pixels on the shortest side.

4 DYNAMIC GESTURE RECOGNITION ALGORITHM

The original HaGRID [23] authors proposed a dynamic gesture recognition algorithm that allows recognition based solely on static gestures, eliminating the need for video-based training and temporal dimension processing. This paper presents a novel approach based on such a logic with extended functionality.

Algorithm. Fig. 6 shows the pipeline of dynamic gesture recognition. Note that the algorithm processes each

ISSN 2464-4617 (print)
ISSN 2464-4625 (online)

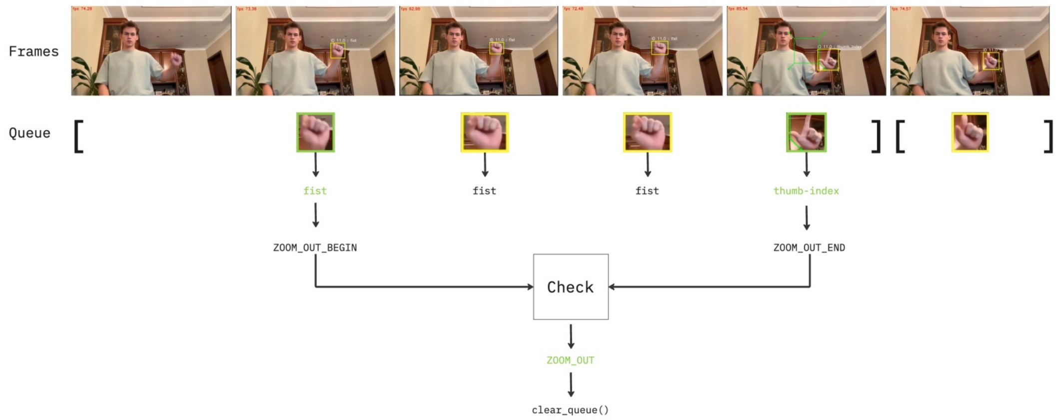

Figure 6: The algorithm for recognizing dynamic gestures, exemplified by the "zoom out" gesture.

frame independently during inference without analyzing the entire sequence from start to end. We detect hands in each frame with a lightweight RFB [3] model, ensuring faster inference. Further, the received crops are classified into gestures utilizing a single Residual Block [18]. To accurately identify the boundaries of the gesture sequence, we employ a sliding queue that stores recognized gestures from the last n frames, which is experimentally set to 30. The choice of 30 frames for the sliding queue was experimentally determined to balance the need for temporal accuracy and computational efficiency. This value was found to provide a sufficient window to accurately capture gesture sequences while avoiding excessive overlap between consecutive gestures. Each dynamic gesture is defined by specific static gestures marking its start and end. Remarkably, we maintain a separate queue for every detected hand, enabling the system to detect multiple dynamic gestures simultaneously. The algorithm monitors the queue for the starting gesture and triggers the classification of the dynamic gesture upon detecting the corresponding closing gesture. Additionally, we implement checks on each dynamic gesture's duration, distance, and orientation to ensure accurate classification, shown and described in Tab. 2.

Gesture type	Checks for correctness		
	Distance	Duration	Orientation
Swipes	✓	✓	up / down / left / right
Clicks	✗	✓	non-moving
Drag-n-Drop	✗	✓	✗
Zooms (one-handed)	✗	✓	non-moving
Zoom (two-handed)	✗	✗	✗

Table 2: Checks for dynamic gesture recognition. Distance is the literal distance between the centers of the bounding boxes for the gesture's start and end positions on the screen. Duration measures the number of frames between these positions. Orientation determines the gesture's direction: left, right, up, down, or non-moving (the hand remains stationary in one screen location throughout the gesture).

New Dynamic Gestures. The algorithm supports four categories of gestures with their variations: swipes, zooms, clicks, and drag-and-drops. These variations allow for associating different functionalities with each gesture, enhancing the system's versatility.

Based on such an approach, the system is predictable and lightweight, with 276,292 parameters and 106.14 Million Floating Point Operations Per Second (MFLOPs) for the detector and 102,605 parameters and 6.9 MFLOPs for the classifier, and runs efficiently on standard CPU hardware. We obtained 89 Frames Per Second (FPS) on the Apple M1 CPU, using only a single CPU core for inference. Additionally, it is easily expandable with new custom gestures, requiring only static images for training.

5 BASE EXPERIMENTS

5.1 Experimental Setup

The base experiments are divided into three groups: gesture detection, gesture classification, and hand detection. These experiments are designed to showcase the dataset's ability to generalize across popular models for different tasks. We resized the images' maximum side to 224 and padded the result with zeros to a square. The full-frame gesture classification is based on 33 main classes without the "no gesture" class, as each image contains one of the target gestures. For performance evaluation, we used the F1-score and Mean Average Precision (mAP) metrics for classification and detection tasks, respectively. All models were trained on a single Tesla H100 with 80GB for 100 epochs with a batch size 128. Early stopping was triggered after 10 epochs without the metric increasing by at least 0.01.

Hand and Gesture Detection. Three detection architectures, SSDLite MobileNetV3 Large [35], YOLOv10n, and YOLOv10x [63], were employed to ensure that on HaGRIDv2 it is possible to train a robust gesture detector. We use the Stochastic Gradient

Model	Optimizer	Weight Decay	Learning Rate	Scheduler	Scheduler' Params.
ResNet	SGD	1^{-4}	1^{-1}	ReduceLROnPlateau	mode: min, factor: 0.1
MobileNetV3	SGD	5^{-4}	5^{-3}	StepLR	step size: 30, gamma: 0.1
VitB16	SGD	5^{-4}	5^{-3}	CosineAnnealingLR	T max: 8
ConvNext	AdamW	5^{-2}	4^{-3}	CosineAnnealingLR, LinearLR	T max: 8, factor: 0.001
SSDLite	SGD	5^{-4}	1^{-4}	StepLR	step size: 30, gamma: 0.1
YOLOv10	SGD	5^{-4}	1^{-2}	LambdaLR	sinusoidal function

Table 3: Training hyperparameters.

Model	Model size (MB)	Parameters (M)	Inference time (ms)	Metrics	
				F1-score	mAP
Gesture Detection					
SSDLite MobileNetV3 Large [35]	10.3	2.7	26.3	–	72.7
YOLOv10n [63]	10.33	2.7	85.67	–	88.2
YOLOv10x [63]	121.5	31.6	1145.6	–	**89.4**
Full-Frame Classification					
ResNet18 [19]	42.7	11.2	37.8	98.3	–
ResNet152 [19]	222.65	58.2	226.94	**98.6**	–
MobileNetV3 Small [20]	6	1.6	5.1	86.7	–
MobileNetV3 Large [20]	16.3	4.2	11.34	93.4	–
ViTB16 (pretrained) [13]	327.4	85.8	350.9	91.7	–
ConvNeXt Base [36]	334.2	87.6	320	96.4	–
Hand Detection					
YOLOv10n [63]	10.3	2.7	75.8	–	87.9
YOLOv10x [63]	120.8	31.6	1150.7	–	**88.8**

Table 4: Models training results on the HaGRIDv2.

Descent (SGD) optimizer with an initial learning rate of 0.01 for YOLOs and 0.0001 for SSDLite. The YOLO models employed default augmentations such as mosaic, hsv, and horizontal flips, while SSD trained without any modifications. Using the same setup, the YOLOv10x model was used to train for the hand detection task.

Gesture Classification. In addition to ResNet-18, ResNet-152, MobileNetV3 Small, MobileNetV3 Large, and pre-trained on ImageNet ViTB16 utilized in [23], we also employed ConvNext [36] as a full-frame gesture classifier. The AdamW optimizer [37] with a specified initial learning rate, weight decay, and scheduler for each architecture was used (see Tab. 3).

5.2 Results

Tab. 4 presents the evaluation metrics on the HaGRIDv2 test subset. The metrics are remarkably high, demonstrating the dataset's effectiveness in training robust models. To ensure the model's ability to work in real-life conditions, we provide a demo of gesture classification and detection models in our repository.

6 CROSS-DATASET EVALUATION

Experimental Setup. This section covers cross-dataset evaluation for hand and gesture detection[5]. We utilized the same setup across all experiments, employing YOLOv10n as a detector, the hyperparameters and augmentations described in Sec. 5, and mAP as a detection metric.

[5] Cross-dataset evaluation for classification was excluded due to the lack of accessible datasets with sufficient overlapping gestures, making the comparison non-informative.

6.1 Gesture Detection

Datasets. We were limited to the HANDS and OUHANDS datasets in the gesture detection task since we could not access the LaRED and SHAPE datasets. These datasets intersect only in 5 classes — namely “fist,” “one,” “palm,” “peace,” and “three” — with HaGRIDv2. Thus, we left only samples with these overlapping gestures, reducing three training, validation, and testing sets. The original HaGRID dataset was excluded from these experiments, as its content is entirely subsumed within HaGRIDv2.

Results. Tab. 5 indicates the HaGRIDv2's complexity, as evidenced by the lowest test average mAP. Furthermore, HaGRIDv2 exhibits superior domain generalization, supported by the highest train average mAP. Notably, HaGRIDv2 is the only dataset that consistently achieves valuable metrics across tests on other datasets, underscoring its value in training robust models.

6.2 Hand Detection

Datasets. We compare HaGRIDv2 with datasets designed explicitly for hand detection to assess its ability to solve this task. Since HaGRIDv2 was annotated with horizontal bounding boxes, we only compared it with similarly annotated BodyHands [48], HumanParts [31], and EgoHands [8] datasets. We also included the original HaGRID [23] dataset to test the impact of more quantity and variety in gestures in the HaGRIDv2. The training, validation, and test sets of HaGRIDv2 are designed to prevent any overlap with the HaGRID subsets, eliminating the risk of data leakage and ensuring the integrity of the comparison.

Trained / Tested	[42]	[51]	HaGRIDv2	Train avg. mAP (↑)
OUHANDS [42]	65	0.07	4.25	2.16
HANDS [51]	0.0	65.4	**7**	3.5
HaGRIDv2 (ours)	**67**	**66.3**	88.1	**66.65**
Test avg. mAP (↓)	33.5	33.19	**5.63**	

Table 5: Cross-dataset evaluation in the gesture detection task. mAP was computed for each pair of datasets and averaged separately for training and testing. Higher train average mAP indicates greater model robustness, while lower test average mAP reflects that the dataset serves as both a strong benchmark and a challenging test, offering a comprehensive evaluation of model performance. Diagonal values were excluded from the averages to ensure unbiased comparison and assess generalization.

Results. Although neither version of HaGRID was initially designed for hand detection, and their samples are generally simpler compared to other datasets, Tab. 6 demonstrates that models trained on HaGRIDv2 generalize better to other distributions. In particular, the model trained on HaGRIDv2 shows superior performance when evaluated on external datasets, indicating a stronger ability to recognize hands under varying conditions. Additionally, HaGRIDv2 achieves a higher mAP on HaGRIDv2 than HaGRID itself, suggesting that increasing the diversity of gesture classes improves the model's ability to generalize older gestures.

7 HAGRID VS HAGRIDV2

7.1 Pre-train Impact

Experimental Setup. We compared HaGRID and HaGRIDv2 as pre-training datasets to demonstrate that almost 2× increase in samples and class diversity in HaGRIDv2 consistently yields more reasonable results. ResNet18 for gesture classification and YOLOv10n for gesture detection were employed, applying the same hyperparameters and metrics from base experiments in Sec. 5.

Datasets. HANDS and OUHANDS were utilized to fine-tune detectors on their training sets with further assessment on their test sets. The pre-trained classifiers were fine-tuned on the Kinect Leap, Senz3D, OUHANDS, and HANDS training sets, while HTU HGR and SIT-HANDS were unavailable due to the invalid link. For the Kinect Leap and Senz3D datasets, we performed our own test split by user identity, as the authors provided no original split.

Results. Fig. 7 shows that models pre-trained on HaGRIDv2 consistently outperformed those pre-trained on HaGRID, improving model generalization and proving its indispensability for pre-training.

7.2 False Positive Triggering

Experimental Setup. We conducted a check for false positive triggering to assess the impact of diversifying

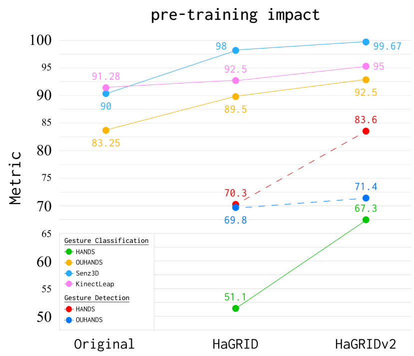

Figure 7: Impact of pre-training on gesture classification and detection across HaGRID and HaGRIDv2. "Original" metrics are sourced from the respective dataset papers; missing values indicate metrics not reported by the authors.

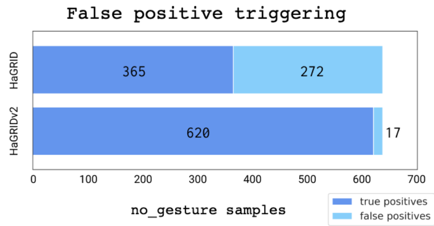

Figure 8: Comparing the false positives on the "no gesture" class for HaGRID and HaGRIDv2 datasets.

the "no gesture" class (see Fig. 4). We trained two YOLOv10n detectors: one on the original HaGRID dataset and the other on HaGRIDv2. Each trained model was assessed on the "no gesture" samples from the HaGRIDv2 test set to evaluate the number of false positives.

Results. The HaGRIDv2-trained model produces 16 times fewer false positive errors than the HaGRID-trained model, which is especially important for production-level HGR systems (see Fig. 8). The mAP metrics were 57 for HaGRID and 72.9 for HaGRIDv2, highlighting the usefulness of the new "no gesture" configuration.

7.3 Gesture Generation

Experimental Setup. This section aims to demonstrate that adding new gestures improves the quality of generating images of people showing gestures. We fine-tuned two Stable Diffusion 2.1 models using the LoRA [22] (Low-Rank Adaptation) from the diffusers library [61] on the HaGRID and HaGRIDv2 datasets. Each image in the datasets was annotated with an automatically generated description using BLIP-2 [30], following the format: *"{blip_caption} showing {gesture_name} gesture"*. During sample generation, we used prompts like *"there is a person showing {gesture_name} gesture"*. To conduct a fair comparison, we also compared the HaGRIDv2 fine-tuned model with the original Stable Diffusion 2.1 without fine-tuning.

Trained / Tested	EgoHands	BodyHands	Human-Parts	HaGRID	HaGRIDv2	Train avg. mAP (↑)
EgoHands [8]	75.4	0.9	1.16	3.54	4.15	2.4
BodyHands [48]	14	35.2	26.4	49.4	37.3	31.7
Human-Parts [31]	35.7	**21.8**	50.5	58.7	53.3	42.4
HaGRID [23]	39.2	19.7	**35**	86.7	**73.9**	42
HaGRIDv2 (ours)	**39.4**	17.3	34.9	**87.2**	87.9	**44.7**
Test avg. mAP (↓)	32	**14.9**	24.4	49.7	42.2	

Table 6: Similar cross-dataset evaluation as in Tab. 5 for hand detection task.

Evaluation. To compare HaGRID and HaGRIDv2, we evaluated the generation quality of 18 gestures from the original HaGRID. For the original diffusion model, we tested 26 gestures from HaGRIDv2, excluding "inverted" and "thumb_index" gestures due to their specificity, and kept only one version of each gesture (e.g., keeping "heart" and removing "heart2"). Each model generated three images per gesture, resulting in 6 images per gesture. We conducted a Subjective Benchmark Scoring (SBS) study to evaluate their quality. Three independent crowdworkers compared each pair of images. Notably, the crowdworkers were unaware of which model had generated each image, ensuring an unbiased evaluation.

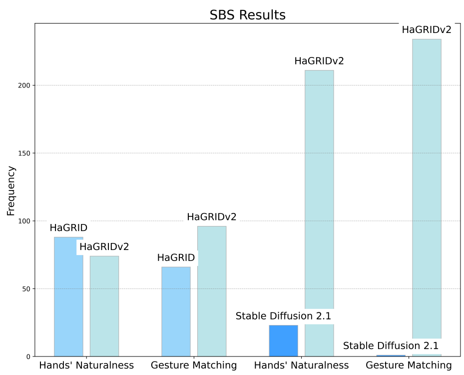

Figure 9: The SBS results compare Stable Diffusion 2.1 fine-tuned on the HaGRID and HaGRIDv2 datasets, as well as a comparison between HaGRIDv2 and the original Stable Diffusion 2.1 model.

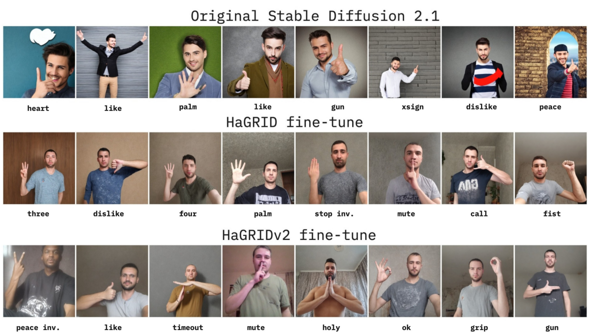

Figure 10: Examples of generated images with three Stable Diffusion 2.1 models: original, fine-tuned on HaGRID, and fine-tuned on HaGRIDv2.

Results. Fig. 9 shows that the model fine-tuned on HaGRIDv2 generates more accurate and natural-looking gestures, thanks to the broader variety of classes it was trained on. The accuracy here refers to the improved alignment between the generated gestures and the target gesture classes, as well as the higher visual naturalness of the outputs. Compared to the model fine-tuned on the original HaGRID or other datasets, the one trained on HaGRIDv2 more effectively replicates subtle motion patterns and hand configurations, resulting in gestures that are both semantically and visually closer to the ground truth. However, the original HaGRID achieved slightly better anatomical accuracy due to its simpler distribution and lack of complex hand postures. Additionally, the comparison with the original Stable Diffusion 2.1 model indicated its limited ability to generate recognizable gestures, with the anatomical accuracy of the hands also being inferior (see Fig. 10).

8 ETHICAL CONSIDERATION

Dataset Creation. HaGRIDv2's samples contain personal information, so crowdworkers must consent to collect, process, and publish their photos. We comply with Russia's Federal Law "On Personal Data" (27.07.2006 N152), ensuring legal data handling. For ethical reasons, images of children were excluded from the dataset. After validation, we justified each rejected photo and allowed crowdworkers to challenge rejections. As the HaGRID dataset is part of HaGRIDv2, we ensured that HaGRID adheres to the described ethical requirements by contacting its authors.

Biases. Utilizing only Russian crowdsourcing platforms can lead to an imbalance in the racial diversity of workers. We tried to minimize this gap and covered the most frequently identified races – Caucasian, Negroid, and Mongoloid. Even though there is an imbalance, trained on the HaGRIDv2 neural network can accurately recognize gestures from underrepresented racial groups.

Possible Misuse. There is the risk of misusing the datasets with faces to improve surveillance systems, profile individuals based on race, create deepfakes, and contribute to identity theft. We release the dataset under a public license for non-commercial use in research purposes, acknowledging the potential risk of its misuse for unlawful activities. We used anonymized user hash IDs in the dataset annotation to preserve crowdworkers privacy and enable the ability to split HaGRIDv2 by its users.

9 LIMITATIONS

Dataset. As noted in Sec. 8, the dataset is biased towards the white race, potentially compromising the algorithm's robustness and performance in diverse scenarios. Additionally, the Gaussian age distribution, limited to individuals 18 years and older, may lead to inaccurate predictions for children and seniors. The "no gesture" class expansion is specified to poses typical for device interactions, potentially leading to false positives in other scenarios. Furthermore, predominantly home-based scenes may lead to incorrect operation when people wear outdoor clothing and gloves in different weather conditions.

Dynamic Gesture Recognition Algorithm. The deterministic nature of the algorithm reduces its robustness, as it demands users to perform gestures with exact precision. This inflexibility can result in recognition inconsistencies, especially in real-world scenarios where slight variations in gesture execution are common, ultimately limiting the algorithm's reliability and user experience.

Generative Models. While the HaGRIDv2 dataset aids in training models to generate people displaying gestures (see Sec. 7.3), it has limitations. The focus of the dataset on specific gestures and the limited variety of hands in natural positions restrict the ability of models to generate anatomically accurate hands in free poses, which may limit broader applicability. Additionally, since the model was fine-tuned solely on samples from the limited distribution of HaGRIDv2, it tends to generate images that look quite similar, reducing overall variety.

10 CONCLUSION

This paper introduces HaGRIDv2, an enhanced version of HaGRID, which become the largest and most diverse gesture dataset for HGR systems. Including a new "no gesture" class significantly strengthens the system's robustness and adaptability to real-world conditions, paving the way for more reliable gesture-based interaction technologies. The HaGRIDv2 can also be used for robust hand generation by diffusion models. We also present a dynamic gesture recognition algorithm that identifies various manipulative gestures for device control. The dataset, pre-trained models, and the dynamic gesture recognition algorithm will be published in our repository.

REFERENCES

[1] Raise your hand with gesture detection in google meet. https://workspaceupdates.googleblog.com/2023/11/google-meet-hand-raise-gesture-detection.html.

[2] Zoom gesture recognition. https://support.zoom.com/hc/en/article?id=zm_kb&sysparm_article=KB0067842.

[3] Ultra-light-fast-generic-face-detector. https://github.com/Linzaer/Ultra-Light-Fast-Generic-Face-Detector-1MB, 2019.

[4] SHAPE Dataset. https://users.soict.hust.edu.vn/linhdt/dataset/, 2021.

[5] Bayan Alabdullah, Hira Ansar, Naif Almudawi, Abdulwahab Alazeb, Abdullah Alshahrani, Saud Alotaibi, and Ahmad Jalal. Smart home automation-based hand gesture recognition using feature fusion and recurrent neural network. *Sensors*, 23:7523, 08 2023.

[6] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, Jeff Kusnitz, Michael Debole, Steve Esser, Tobi Delbruck, Myron Flickner, and Dharmendra Modha. A low power, fully event-based gesture recognition system. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7388–7397, 2017.

[7] P. N. Arathi, S. Arthika, S. Ponmithra, K. Srinivasan, and V. Rukkumani. Gesture based home automation system. In *2017 International Conference on Nextgen Electronic Technologies: Silicon to Software (ICNETS2)*, pages 198–201, 2017.

[8] Sven Bambach, Stefan Lee, David Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[9] Gibran Benitez-Garcia, Jesus Olivares-Mercado, Gabriel Sanchez-Perez, and Keiji Yanai. Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition. In *25th International Conference on Pattern Recognition, ICPR 2020, Milan, Italy, Jan 10–15, 2021*, pages 4340–4347. IEEE, 2021.

[10] Dayananda Kumar N. C, K.V Suresh, Sheela S. J, and Dinesh R. Sit-hands: A novel rgb-d dataset for static hand gesture recognition using kinect sensor. In *2023 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES)*, pages 1–6, 2023.

[11] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[12] Smit Desai and Apurva Desai. Human computer interaction through hand gestures for home automation using microsoft kinect. In Nilesh Modi, Pramode Verma, and Bhushan Trivedi, editors, *Proceedings of International Conference on Communication and Networks*, pages 19–29, Singapore, 2017. Springer Singapore.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

[14] H. Pallab Jyoti Dutta, M. K. Bhuyan, Debanga Raj Neog, Karl Fredric MacDorman, and Rabul Hussain Laskar. Patient assistance system based on hand gesture recognition. *IEEE Transactions on Instrumentation and Measurement*, 72:1–13, 2023.

[15] Nesrine Fourati and Emmanuel Marilly. Gestures for natural interaction with video. In *Visual Information Processing and Communication III*, volume 8305, pages 170–181. SPIE, 2012.

[16] Susan Goldin-Meadow. The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, 3(11):419–429, 1999.

[17] Ayaka Hatori and Hiroyuki Kobayashi. A preliminary study of iot-device control using gestures recognition. In *2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pages 976–979, 2017.

[18] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[19] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[20] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. *CoRR*, abs/1905.02244, 2019.

[21] Yuan-Sheng Hsiao, Jordi Sanchez-Riera, Tekoing Lim, Kai-Lung Hua, and Wen-Huang Cheng. Lared: a large rgb-d extensible hand gesture dataset. In *ACM SIGMM Conference on Multimedia Systems*, 2014.

[22] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

[23] Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, and Andrei Makhliarchuk. Hagrid – hand gesture recognition image dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4572–4581, January 2024.

[24] Maria Karam and M. C. Schraefel. A taxonomy of gestures in human computer interactions. 2005.

[25] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

[26] Sanmukh Kaur, Anuranjana, and Nivedita Nair. Electronic device control using hand gesture recognition system for differently abled. In *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 371–375, 2018.

[27] Rafiqul Zaman Khan and Noor Ibraheem. Hand gesture recognition: A literature review. *International Journal of Artificial Intelligence & Applications (IJAIA)*, 3:161–174, 08 2012.

[28] Okan Köpüklü, Ahmet Gunduz, Neslihan Kose, and Gerhard Rigoll. Real-time hand gesture detection and classification using convolutional neural networks. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8, 2019.

[29] Maksim Kuprashevich and Irina Tolstykh. Mivolo: Multi-input transformer for age and gender estimation. 2023.

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023.

[31] Xiaojie Li, Lu Yang, Qing Song, and Fuqiang Zhou. Detector-in-detector: Multi-level analysis for human-parts, 2019.

[32] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7082–7092, 2018.

[33] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

[34] Li Liu and Ling Shao. Learning discriminative representations from rgb-d video data. pages 1493–1500, 08 2013.

[35] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *Computer Vision - ECCV 2016*, pages 21–37. Springer International Publishing, 2016.

[36] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[38] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019.

[39] G. Marin, F. Dominio, and P. Zanuttigh. Hand gesture recognition with leap motion and kinect devices. In *IEEE International Conference on Image Processing (ICIP)*, Paris, France, 2014.

[40] G. Marin, F. Dominio, and P. Zanuttigh. Hand gesture recognition with jointly calibrated leap motion and depth sensor. *Multimedia Tools and Applications*, 2015.

[41] J. Materzynska, G. Berger, I. Bax, and R. Memisevic. The jester dataset: A large-scale video dataset of human gestures. *IEEE/CVF (ICCVW)*, 2019.

[42] Matti Matilainen, Pekka Sangi, Jukka Holappa, and Olli Silvén. Ouhands database for hand detection and pose recognition. *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–5, 2016.

[43] A. Memo, L. Minto, and P. Zanuttigh. Exploiting silhouette descriptors and synthetic data for hand gesture recognition. *STAG: Smart Tools Apps for Graphics*, 2015.

[44] A. Memo and P. Zanuttigh. Head-mounted gesture controlled interface for human-computer interaction. *Multimedia Tools and Applications*, 2017.

[45] Arpit Mittal, Andrew Zisserman, and Philip H. S. Torr. Hand detection using multiple proposals. In *British Machine Vision Conference*, 2011.

[46] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kautz. Hand gesture recognition with 3d convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–7, 2015.

[47] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4207–4215, 2016.

[48] Supreeth Narasimhaswamy, Thanh Nguyen, Mingzhen Huang, and Minh Hoai. Whose hands are these? hand detection and hand-body association in the wild. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4879–4889, 2022.

[49] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai. Detecting hands and recognizing physical contact in the wild. In *Advances in Neural Information Processing Systems*, 2020.

[50] Supreeth Narasimhaswamy, Zhengwei Wei, Yang Wang, Justin Zhang, and Minh Hoai. Contextual attention for hand detection in the wild. 04 2019.

[51] C. Nuzzi, S. Pasinetti, R. Pagani, G. Coffetti, and G. Sansoni. Hands: an rgb-d dataset of static hand-gestures for human-robot interaction. *Data in Brief*, 35:106791, 2021.

[52] Francisco Parada-Loira, Elisardo GonzÃ¡lez-Agulla, and JosÃ© L. Alba-Castro. Hand gestures to control infotainment equipment in cars. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pages 1–6, 2014.

[53] Nicolas Pugeault and Richard Bowden. Spelling it out: Real-time asl fingerspelling recognition. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1114–1119, 2011.

[54] Jagdish Lal Raheja, Radhey Shyam, Umesh Kumar, and P. Bhanu Prasad. Real-time robotic hand control using hand gestures. In *2010 Second International Conference on Machine Learning and Computing*, pages 12–16, 2010.

[55] Zhou Ren, Junsong Yuan, and Zhengyou Zhang. Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. *MM'11 - Proceedings of the 2011 ACM Multimedia Conference and Co-Located Workshops*, pages 1093–1096, 11 2011.

[56] Simon Ruffieux, Denis Lalanne, and Elena Mugellini. Chairgest: A challenge for multimodal mid-air gesture recognition for close hci. *ICMI 2013 - Proceedings of the 2013 ACM International Conference on Multimodal Interaction*, pages 483–488, 12 2013.

[57] Ram Pratap Sharma and Gyanendra K. Verma. Human computer interaction using hand gesture. *Procedia Computer Science*, 54:721–727, 2015. Eleventh International Conference on Communication Networks, ICCN 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Data Mining and Warehousing, ICDMW 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Image and Signal Processing, ICISP 2015, August 21-23, 2015, Bangalore, India.

[58] Yale Song, David Demirdjian, and Randall Davis. Tracking body and hands for gesture recognition: Natops aircraft handling signals database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 500–506, 2011.

[59] Helman Stern, Juan Wachs, and Yaeledan. Designing hand gesture vocabularies for natural interaction by combining psycho-physiological and recognition factors. *International Journal of Semantic Computing*, 02, 11 2011.

[60] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[61] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

[62] Jun Wan, Stan Z. Li, Yibing Zhao, Shuai Zhou, Isabelle Guyon, and Sergio Escalera. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 761–769, 2016.

[63] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*, 2024.

[64] Huihui Wang, Bo Ru, Xin Miao, Qin Gao, Masood Habib, Long Liu, and Sen Qiu. Mems devices-based hand gesture recognition via wearable computing. *Micromachines*, 14, 04 2023.

[65] Chih-Hsiang Yang, Yi-Nan Lin, Sheng-Kuan Wang, Victor Shen, Yi-Chih Tung, Frank Shen, and Chun-Hsiang Huang. Smart control of home appliances using hand gesture recognition in an iot-enabled system. *Applied Artificial Intelligence*, 37, 02 2023.

[66] Mais Yasen and Shaidah S. Jusoh. A systematic review on hand gesture recognition techniques, challenges and applications. *PeerJ Computer Science*, 5, 2019.

[67] Ilya Zakharkin, Arman Tsaturyan, Miguel Altamirano, Jonathan Tirado, and Dzmitry Tsetserukou. Zoomtouch: Multi-user remote robot control in zoom by dnn-based gesture recognition. 11 2020.

[68] Y. Zhang, C. Cao, J. Cheng, and H. Lu. Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia (T-MM)*, 20(5):1038–1050, 2018.

[69] Hongyang Zhao, Shuangquan Wang, Gang Zhou, and Daqing Zhang. Gesture-enabled remote control for healthcare. In *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 392–401, 2017.