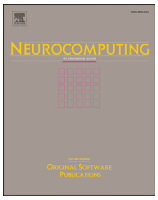
Neurocomputing 704 (2026) 134793

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Survey paper

Visual hand gesture recognition with deep learning: A comprehensive review of methods, datasets, challenges and future research directions

Konstantinos Foteinos [a], Manousos Linardakis [a], Panagiotis Radoglou-Grammatikis [b,c], Vasileios Argyriou [d], Panagiotis Sarigiannidis [b], Iraklis Varlamis [a], Georgios Th. Papadopoulos [a,*]

[a] Department of Informatics and Telematics, Harokopio University of Athens, Thiseos 70, Athens, GR 17676, Attiki, Greece
[b] Department of Electrical and Computer Engineering, University of Western Macedonia, Active Urban Planning Zone, Kozani, GR 50150, Kozani, Greece
[c] K3Y, Studentski district, Vitosha quarter, bl. 9, Sofia, BG 1700, Sofia City Province, Bulgaria
[d] Department of Networks and Digital Media, Kingston University, 53-57 High Street, Kingston upon Thames, KT1 1LQ, Surrey, United Kingdom

ARTICLE INFO

Communicated by C. Shan

Keywords:
Vision-based hand gesture recognition
Human-robot interaction
Human-computer interaction
Sign language
Deep learning
Gesture recognition taxonomy

ABSTRACT

The rapid evolution of deep learning (DL) models and the ever-increasing size of available datasets have raised the interest of the research community in the always-important field of ***visual hand gesture recognition*** (VHGR), and delivered a wide range of applications, such as sign language understanding and human-computer interaction using cameras. Despite the large volume of research works in the field, a structured and complete survey on VHGR is still missing, leaving researchers to navigate through hundreds of papers in order to find the current state-of-the-art (SOTA). The current survey aims to fill this gap by presenting a comprehensive overview of this computer vision field. With a systematic research methodology that identifies the SOTA works and a structured presentation of the various methods, datasets, and evaluation metrics, this review aims to constitute a useful guideline for researchers, helping them to propose improvements. Specifically, this survey focuses on four fundamental questions: what are the main VHGR aspects, what are the current SOTA methods, what comparative insights can be drawn across methods and tasks, and which challenges shape future research. Starting with the methodology used to locate the related literature, the survey identifies and organizes the key VHGR approaches in a taxonomy-based format, and presents the various dimensions that affect the final method choice, such as input modality, task type, and application domain. The SOTA techniques are grouped across three primary VHGR tasks: static, isolated dynamic and continuous gesture recognition. For each task, the architectural trends and learning strategies are listed. To support the experimental evaluation of future methods in the field, the study reviews commonly used datasets and presents the standard performance metrics. Our survey concludes by identifying the major challenges in VHGR, including both general computer vision issues and domain-specific obstacles, and outlines promising directions for future research.

1. Introduction

The recognition and interpretation of hand gestures [2] play a central role in advancing Human-Computer Interaction (HCI) and assistive communication systems. In recent years, deep learning-based vision-based hand gesture recognition (VHGR) has gained significant momentum as an active subfield of computer vision, building upon progress in related areas such as action [199] and facial expression recognition [194]. This has led to an increased focus on Hand Gesture Recognition (HGR) across various applications. For example, HGR is widely used in Human-Robot Interaction (HRI) in industrial settings (e.g., robotic arm control [124,171]), in healthcare (e.g., robot-assisted surgery [179]), and in the autonomous navigation of mobile robots such as Unmanned Aerial [240] and Ground [21,133] Vehicles (UAVs and UGVs). It also plays a key role in HCI, especially in VR/AR environments [6,43,246]. In addition, the large population of deaf and hard-of-hearing individuals has driven research into automatic Sign Language Recognition (SLR)

* Corresponding author.
Email addresses: kfoteinos@hua.gr (K. Foteinos), manouslinard@gmail.com (M. Linardakis), pradoglou@k3y.bg (P. Radoglou-Grammatikis), vasileios.argyriou@kingston.ac.uk (V. Argyriou), psarigiannidis@uowm.gr (P. Sarigiannidis), varlamis@hua.gr (I. Varlamis), G.Th.Papadopoulos@hua.gr (G.T. Papadopoulos).

https://doi.org/10.1016/j.neucom.2026.134793
Received 19 January 2026; Received in revised form 9 May 2026; Accepted 10 August 2026
Available online 12 August 2026

Table 1
Comparison of this survey with selected others. ML refers to traditional machine learning/computer vision techniques. The current survey mainly examines publications from the last four years, focusing on recent advancements regarding architectures, learning strategies, and related challenges, although older datasets that are relevant to this day have also been included.

Publication	Range	Papers	Input data	Models	Scope	Comments	Aspects that this survey aims to complement
[2]	2016–2020	–	Vision-based	DL	CSLR	Provides an in-depth comparison of existing approaches, loss functions and pretraining schemes for CSLR.	This survey expands the scope including non-CSLR methods and examines more recent methods.
[159]	2014–2020	98	Vision-based	DL & ML	General	Investigates the progress on hand gesture representation and Data Acquisition (DAQ) techniques as well as their performance and future research directions.	It examines relatively older methods. The current work aims to emphasize recent DL approaches and related challenges and future perspectives.
[8]	2014–2021	138	Sensor- & Vision-based	DL	SLR	Categorizes Continuous and Isolated SLR methods based on their architectures and discusses pose estimation for ISLR in detail.	This survey covers more topics including HCI & HRI, and highlights task-specific architectural patterns and challenges for CSLR.
[3]	2001–2021	649	Vision-based	DL & ML	SLR	Discusses relationships between research teams and DAQ techniques.	It includes relatively old methods. This survey emphasizes recently proposed ones in order to facilitate the evolution of the field.
[168]	2000–2020	571	Sensor- & Vision-based	DL & ML	General	Provides many statistical results and bibliometric charts.	It does not identify contemporary challenges or future research directions; this work seeks to bridge this gap.
[57]	1998–2020	–	Sensor- & Vision-based	DL & ML	SLR	Discusses in detail existing methods for SLR and provides multiple tables for quantitative comparison.	This survey analyzes newer approaches for SLR and expands taxonomy and comparison to more application domains, beyond SLR.
[108]	2010–2021	80	Vision-based	DL & ML	SLR	A Natural Language Processing (NLP) toolkit was utilized to filter publications. Insightful statistics are also provided.	Only a few methods are discussed and the current trends are not highlighted. On the contrary, this survey aims to identify trends and challenges.
[198]	2015–2023	–	Vision-based	DL	ISLR	Provides statistical charts and summary tables based on a four-fold taxonomy.	This review introduces additional taxonomy criteria and covers more tasks.
[206]	2014–2024	–	Sensor- & Vision-based	DL & ML	General	Considers a variety of modalities and examines methods for each one.	The current work aims to discuss further multi-modality fusion strategies and identify more challenges.
[84]	2018–2023	256	Sensor- & Vision-based	ML	General	Covers various aspects of HGR, including DAQ techniques, hand segmentation and feature extraction.	Focuses exclusively on traditional ML approaches, which are currently surpassed by DL ones. This survey, searching for recent advancements, focuses on DL approaches.
[7]	2015–2024	58	Sensor- & Vision-based	DL & ML	SLR	Examines various SLR methods and highlights the superiority of DL methods over older ones.	It does not identify challenges and future perspectives; this survey aims to cover this, going deeper into details.
[219]	2017–2024	–	Sensor- & Vision-based	DL & ML	SLR	Groups methods into RNN-based, CNN-based and transformers and discusses well-known benchmarks.	This survey expands the scope to non-SLU papers and seeks to provide methods comparison and future prospects.
[180]	2008–2024	–	Sensor- & Vision-based	DL & ML	General	Covers a wide spectrum of gesture-based interaction and related technologies.	The considered methods are relatively old. This survey aims to identify more recent advancements.
[134]	2018–2025	125	Vision-based	DL & ML	General	Discusses jointly hand gesture capturing and recognition techniques. Many challenges are highlighted, including data bias and privacy issues.	It covers a limited number of datasets for HGR and relatively old methods; this survey complements it by analyzing more datasets and recent advancements.
[58]	2002–2023	–	Sensor- & Vision-based	DL & ML	CDHGR	Focuses on online gesture recognition, discussing DAQ technologies and evaluation metrics, among others.	This survey covers more topics and provides comparison of state-of-the-art methods.
[115]	2010–2024	–	Sensor- & Vision-based	DL & ML	CSLR	Discusses in detail methods and datasets for CSLR.	It does not conduct methods comparison and the challenges are general; this survey highlights optimization difficulties in the context of CSLR.

systems, aiming to improve communication and inclusion for these communities [34,243].

Given the importance of the applications mentioned above, HGR has grown into a broad research field, generally divided into two main approaches: vision-based (VHGR) and sensor-based. VHGR relies on visual data such as RGB and depth images, as well as features derived from them, such as pose heatmaps or optical flow. Sensor-based HGR, on the other hand, uses data from devices such as wearable gloves or surface electromyography sensors, and even non-wearable sources like acoustic signals. Researchers have also explored multi-modal methods that combine inputs from different sources. In this study, we focus on VHGR, as sensor-based methods often involve high costs and complex setups, making them less practical for real-world use. Vision-based approaches better align with the goal of HGR: to enable more natural, unobtrusive, and accessible gesture-based interaction.

The great importance and wide scope of HGR have led researchers to conduct surveys focusing on different aspects of the field, as shown in Table 1. Compared to existing review papers, this survey provides an in-depth examination of the recent deep learning (DL) methods (published since 2021), covering the current state-of-the-art (SOTA). In contrast, other existing reviews either focus on older approaches (published since 1998) [57,84,168] or mainly cover traditional machine learning techniques [134,180,219]. While such works are valuable for understanding the fundamental methods and providing statistical insights, they are less suited for reflecting the recent evolution of the field. The present survey stands out by its broader and more up-to-date scope, including 238 papers published since 2021–an average of nearly 47 publications per year. By comparison, Adeyanju et al. [3] and Ojeda-Castelo et al. [168] analyze 649 and 571 papers, respectively, but span roughly 20 years of research, corresponding to only about 30 publications per year. Finally, the current survey focuses exclusively on VHGR. Including sensor-based approaches would significantly expand the scope and potentially confuse readers. Additionally, such methods generally offer lower user convenience and reduced applicability in real-world scenarios. Nevertheless,

the review maintains a broad and inclusive perspective, acknowledging that approaches from different domains may share common principles and can inform each other.

More specifically, the survey focuses on vision-based methods for three key VHGR tasks, which are discussed in detail in the following sections: static HGR (SHGR) [167,256], isolated dynamic HGR (IDHGR) [78,93,106,254,262], and continuous (dynamic) HGR (CDHGR or CHGR) [38,99,157]. SHGR involves recognizing gestures from single frames or spatial data only. IDHGR extends this by analyzing short, pre-segmented sequences to capture both spatial and temporal patterns. CDHGR goes a step further by detecting multiple gestures within long, unsegmented video sequences. Each task presents its own unique challenges. For example, CDHGR often relies on weak supervision, making optimization more difficult. At the same time, all tasks share common issues such as interference from irrelevant background movements or variations in lighting, viewpoint, and hand appearance.

VHGR has a long history, beginning with traditional computer vision techniques based on hand-crafted features and evolving into complex deep learning pipelines [160]. Early approaches typically followed a three-stage process: pre-processing, feature extraction, and classification. Pre-processing often involved segmenting the hand region using skin color or other visual cues. Feature extraction relied on descriptors like Histogram of Oriented Gradients (HOG), while classification was usually performed using models such as Support Vector Machines (SVMs) or Hidden Markov Models (HMMs), depending on the task. With the rise of DL, these hand-crafted features were gradually replaced by learned representations, which offer better generalization. Modern deep learning architectures have become increasingly deep and sophisticated, especially for dynamic HGR tasks. These models first capture short-term motion patterns and then aggregate them to understand long-term context [81,98–100,157]. Due to their superior performance, this study focuses primarily on DL-based methods.

The main contributions of this study can be summarized as follows:

1. A structured and comprehensive methodology for reviewing the VHGR literature;
2. A broad overview of VHGR, organized according to multiple taxonomy criteria;
3. A detailed analysis of recent methods, highlighting their key characteristics and innovations;
4. A synthesis of current SOTA approaches along with a discussion of the key challenges facing the field.

In particular, the key research questions that the current study aims to address are the following:

Q1 What are the principal aspects of VHGR? Specifically, what are its main tasks, gesture types, general architectures, learning strategies, and application domains?
Q2 What are the current SOTA methods for each category and the corresponding recent developments?
Q3 What are the theoretical and quantitative insights obtained through the comparative analysis of methods, both per-task and cross-task?
Q4 What are the current challenges and future research directions concerning the domain?

The structure of this paper aims to guide the reader through a comprehensive review of VHGR. Section 2 presents the literature review methodology, outlining the systematic approach used to collect and evaluate relevant works. Section 3 introduces a taxonomy of VHGR methods based on diverse criteria, offering a structured classification of existing approaches. Sections 4–6 delve into the three main VHGR tasks – static, isolated dynamic, and continuous gesture recognition – discussing representative methods and their underlying principles. Section 7 provides a unified comparison of methods for VHGR regardless of the task. Section 9 presents the main datasets used for the evaluation of the presented methods and Section 10 summarizes the evaluation metrics commonly used in the field. Finally, Section 11 outlines key challenges in VHGR and highlights current solutions, while Section 12 concludes the survey by reflecting on recent trends and emphasizing open challenges that shape the future of VHGR research.

2. Literature review methodology

The literature review was conducted following a systematic methodology designed to ensure a comprehensive and structured exploration of the VHGR field. The study scope was first defined, guiding the development of inclusion and exclusion criteria for selecting relevant publications. Subsequently, targeted queries were executed across major scientific databases to retrieve works meeting these criteria. Finally, the selected literature was thoroughly examined to extract key innovations, contributions, and methodological insights, which are detailed in the following sections.

2.1. Survey scope

This work sets out to map the evolving landscape of VHGR. By distilling the most impactful recent contributions, it seeks to clarify the field's current trajectory and highlight its most pressing research needs. The review is structured around five foundational pillars, which are crucial for answering the research questions, as mentioned in the previous section (Section 1):

- Advanced deep learning **methods** introduced in recent literature;
- Benchmark **datasets** employed across various VHGR tasks;
- Established **evaluation metrics** used to measure the effectiveness of the proposed solutions;
- Key **challenges** that continue to shape the research agenda;
- Innovative **solutions** and **future directions** designed to address current limitations.

2.2. Selection criteria

To ensure the quality, relevance, and completeness of this review, a well-defined set of inclusion and exclusion criteria was established prior to initiating the literature search. The selection process was guided by three key axes:

- **Relevance**: Only works that directly address VHGR were considered. This includes methods focused on hand gesture recognition and sign language recognition based solely on visual modalities such as RGB images, infrared (IR), optical flow, motion history, depth maps, and pose representations (e.g., skeleton landmarks and heatmaps derived either from RGB-based pose estimators or visual data acquisition devices). Studies centered on Human Action Recognition (HAR), sensor-based techniques (e.g., using electromyography or radar), or multimodal approaches combining visual and non-visual data were excluded to maintain a focused scope.
- **Up-to-date**: To capture the current state of the field, priority was given to publications from 2021 onward, especially those introducing new methods. Regarding datasets, no strict publication date was enforced; however, only those actively used by the community or released within the past four years were included. All surveyed review papers were published after 2021.
- **Impact**: To further emphasize seminal and influential works, publications were primarily sourced from highly cited venues (i.e., those with an h5-index of 90 or higher on Google Scholar), with a single exception for the IEEE International Conference on Automatic Face & Gesture Recognition, due to its strong relevance to VHGR. As this h5-index criterion may not be sufficient when applied as a standalone criterion, potentially significant works may be discarded. Therefore, the references of the collected publications

were examined to identify additional milestone or seminal works. Their selection was guided by their accuracy on common benchmarks, applicable only to methods, and by the number of citations. Furthermore, additional queries were submitted to identify highly cited publications, ensuring that the search included the majority of important works. Finally, to provide the theoretical context of the study, a small number of milestone papers were referenced, although they were not analyzed in detail in the main methods section.

2.3. Literature search

To identify publications meeting the predefined criteria, a comprehensive literature search was performed using specialized queries across established scientific databases, including *Scopus* and *Web of Science (WoS)*. *Google Scholar* was additionally consulted to capture potentially overlooked works from major venues, and *arXiv* was also used to find relevant high-impact pre-prints. Queries were constructed according to the inclusion/exclusion criteria, employing specific commands (e.g., `EXACTSRCTITLE`, `LIMIT-TO`) and Boolean operators (`AND`, `OR`, `NOT`) to refine results. The search targeted keywords in titles, abstracts, and index terms, including *gesture recognition, gesture classification, sign language recognition, sign language understanding, pointing recognition, human finger pointing, hand gestures, human gestures, gesture, hand gesture, and body gesture*. This process retrieved over 900 publications, of which 210 were relevant to VHGR. Additional queries on Google Scholar and arXiv, with duplicates removed, produced a final set of 278 publications, encompassing review papers, methods, and datasets.

2.4. Categorizing the identified articles

After applying the inclusion and exclusion criteria, the remaining publications were carefully reviewed for relevance. Those that did not meet the criteria were excluded from further analysis. The selected papers were then categorized into three distinct groups: *methods*, *datasets*, and *review papers*. Each category was analyzed separately, and the corresponding information was organized into three dedicated tables. Publications relevant to more than one category were included in each table as separate entries. This procedure resulted in 185 articles presenting methods, 72 presenting datasets, and 21 survey articles. Key information for each paper was systematically recorded in three tables with corresponding columns, as detailed below:

- **Methods:** For each article, we recorded the name of the method it proposes, the datasets used for evaluation, the evaluation metrics, a brief description of the architecture, the input modality, and the learning strategy.
- **Datasets:** For each dataset, we recorded its name, the task based on temporal characteristics (e.g., static, isolated dynamic, or continuous), the data modality, the total number of samples, the number of subjects, and, when applicable to CDHGR, the number of gesture instances. For Sign Language Understanding (SLU) datasets, we also noted the language, vocabulary sizes (both sign and text), and total duration (for video-based datasets). Additional attributes include the source of the dataset (e.g., laboratory, television, or web), frame rate, resolution, and the environmental conditions under which the data were recorded.
- **Surveys:** For survey papers, we provide a concise description of the content. Furthermore, the proposed taxonomy, the time range covered, and the scope of the survey were recorded. Although we examine reviews considering both sensor-based and vision-based HGR techniques, we focus only on content related to VHGR.

For each publication, various metadata such as title, authors' names, venue, publication year, citation count, keywords, task, application domain, authors' motivation, challenges, and future directions were recorded. The most recent surveys have already been presented in Table 1 and discussed in the Introduction section. The methods and datasets are presented in an organized manner in the following Sections, and some bibliometrics are provided in Fig. 1.

3. Taxonomy of hand gesture recognition methods

To better understand the diversity and evolution of vision-based hand gesture recognition (VHGR) methods and answer the first research question (see Q1 in Section 1), this section introduces a structured taxonomy of existing approaches. It classifies related works according to complementary criteria that capture key aspects of VHGR systems—such as input type, data representation, and application domain—each further divided into subcategories. The overall taxonomy is illustrated in Fig. 2 and described in detail below.

3.1. Application domain

This criterion refers to the various domains in which VHGR solutions have been applied. The following categories are considered:

- **Sign Language Understanding** (SLU): Involves the automatic interpretation of gestures in either a single sign language (monolingual SLU [13,130]) or across multiple sign languages (multilingual SLU [162]).
- **Human-Computer Interaction** (HCI): Encompasses any form of interaction between users and computers. Numerous studies focus on HCI [167,239], with particular emphasis on VR/AR applications [6,43,246].
- **Human-Robot Interaction** (HRI): Includes gesture-based control of mobile robots such as UGVs [21,196] and UAVs [240,259], industrial robotic arms [124], assistive robots [40], and surgical robots [179].
- **Human-Machine Interaction** (HMI): Focuses on interfaces between humans and machines, especially in automotive contexts. VHGR enables recognition of traffic gestures (e.g., from police officers) [61,76,230] and supports the development of more intuitive in-vehicle user interfaces [68,69,146,160,178].

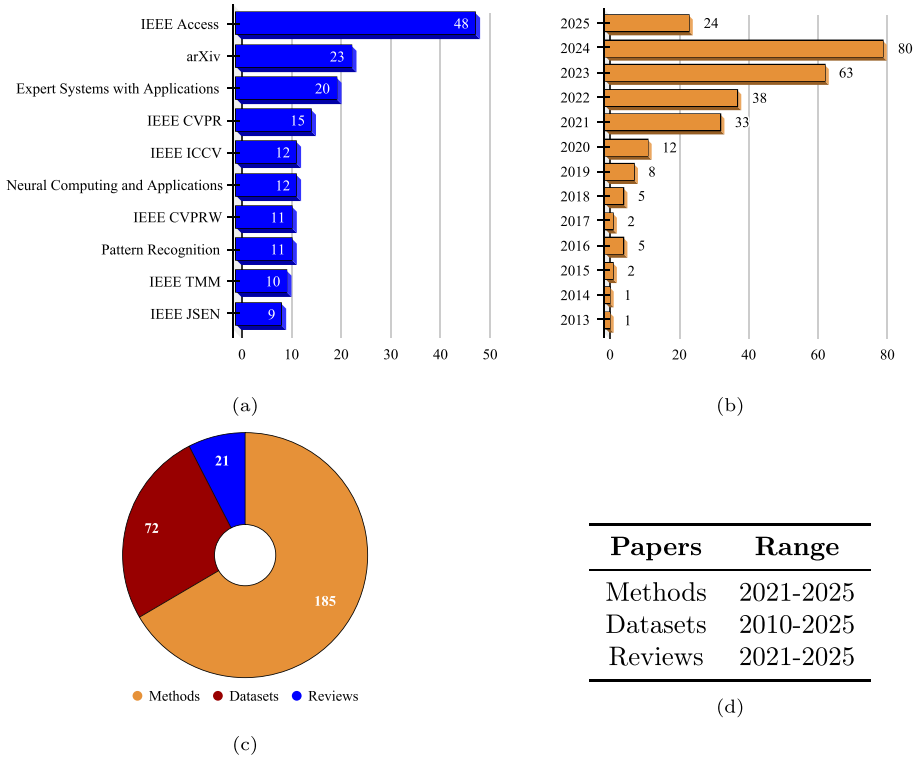
3.2. Task

Complementing the application domains, this criterion focuses on the nature of the gestures themselves, specifically their temporal characteristics. Gestures are broadly classified as either **static** or **dynamic**, depending on whether they involve a fixed hand posture or a sequence of movements over time. Dynamic gestures typically consist of three temporal phases: preparation, core (or nucleus), and retraction [160], which reflect the gesture's evolution from initiation to completion. The recognition of dynamic gestures can be further divided into **isolated** and **continuous**, considering only one gesture per data sequence and identifying gestures in a continuous manner, respectively.

3.3. Input

In continuation of the previous criteria, which focused on application contexts and gesture types, this criterion examines the nature and variety of visual input data used for VHGR. The choice of **input modality** directly influences the representation architecture and the deep learning models employed. While RGB frames remain the most commonly used modality, other forms of visual data have also been explored—such as depth maps (including RGB-like representations, such as HHA [106]), infrared (IR) images, skeleton joints (and derived features like bones and motion [50]), pose heatmaps [38,262], optical flow and motion history [154].

In terms of the **number of input modalities**, methods are classified as *unimodal*, when they use a single type of input data, or *multi-modal*, when they integrate two or more modalities. In the case of multiple modalities in the input, a **multi-modality fusion strategy** is chosen [78,222]. This could be:



Papers	Range
Methods	2021-2025
Datasets	2010-2025
Reviews	2021-2025

(d)

Fig. 1. Bibliometric statistics. Figure a) shows the distribution of the selected papers on the top-10 venues, in terms of number of publications. Almost 50 of the papers considered in the current survey have been published in IEEE Access, covering the whole spectrum of domains and tasks. Figure b) illustrates the number of publications & pre-prints per year. As the current survey focuses on recent advancements, most of the selected papers have been published since 2021. Methods and reviews published before 2021 were ignored. The number of methods, datasets and reviews (which may overlap) is shown in figure c). The publication range for each category is reported in table d).

- **Early (data-level) fusion**: Combines modalities at the input level before processing [65,226].
- **Late (score-level) fusion**: Fuses final prediction outputs from independent unimodal models [67,154,211].
- **Intermediate (feature-level) fusion**: Merges features extracted from separate modalities at a single processing level [217].
- **Multi-level fusion**: Enables interaction across modalities at multiple stages of processing [78,262].

Finally, based on the **type of input**, the methods are broadly categorized into two groups: *pose-based* (or skeleton-based) approaches, which exclusively process skeleton data, and *appearance-based* approaches, which rely on visual cues from RGB, depth, infrared, or similar modalities.

3.4. Type of information conveyed

Building upon the previous criterion, which focused on the nature of the input, this dimension considers the **meaning** or **intent** behind the gestures. Specifically, it examines the kind of information gestures are meant to transmit [19], resulting in the following categories (illustrated in Fig. 4):

- **Semaphoric** gestures convey specific commands or instructions and are commonly used in gesture-based interfaces [21,160,167].
- **Language** gestures function similarly to semaphoric ones but are exclusively used in the context of sign languages [13,202].
- **Deictic (pointing)** gestures are used to indicate objects, directions, or spatial targets [163].
- **Manipulative** gestures aim to interact with or mimic the manipulation of physical objects [66].

3.5. Camera point of view

Complementing the previous criteria on gesture content and input modalities, this dimension focuses on the spatial relationship between the camera and the performer [246]. The camera's perspective influences both the visual characteristics of the gesture and the recognition model design. Three primary viewpoints are considered (see Fig. 5):

- **First-person view (egocentric)**: The camera is mounted on the performer, capturing the gestures from their own perspective.
- **Second-person view**: The camera acts as the direct recipient of the gestures, typically simulating the perspective of a human observer or interacting agent.
- **Third-person view**: The camera observes the gestures from an external position, detached from the performer.

3.6. Architecture

Building on the previous criteria, which focused on the nature of input and camera perspective, this criterion concerns the architectural design of the gesture recognition system, regarding the modeling stages performed by the backbone. Specifically, it refers to how spatial and temporal information is processed and integrated to model gesture dynamics. Three main categories are identified:

- **Spatial** architectures consider only spatial data, such as isolated RGB frames, and feed them into corresponding DNNs (mainly 2DCNNs) to capture spatial information.
- **Temporal** architectures extract only temporal features, utilizing sequential models to process sequences of raw data, such as coordinates of skeleton joints.

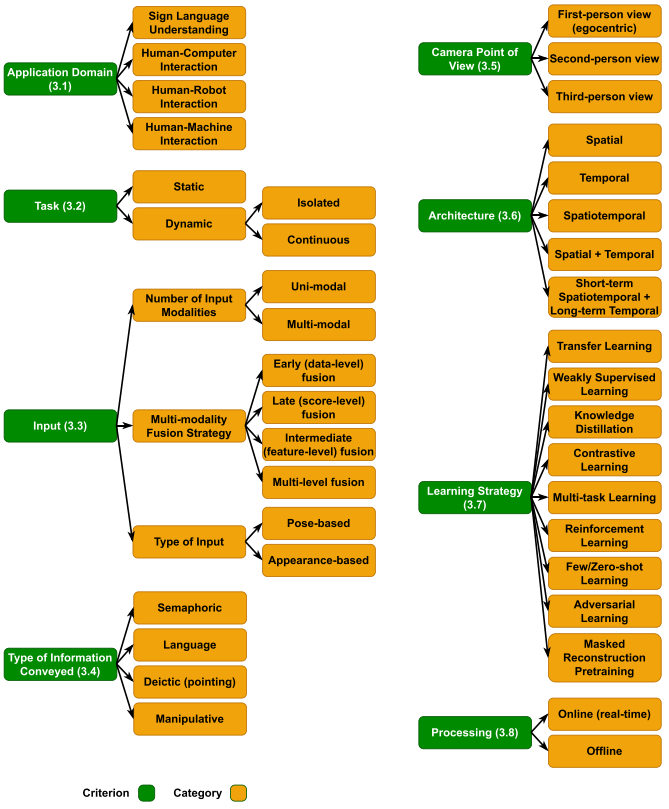
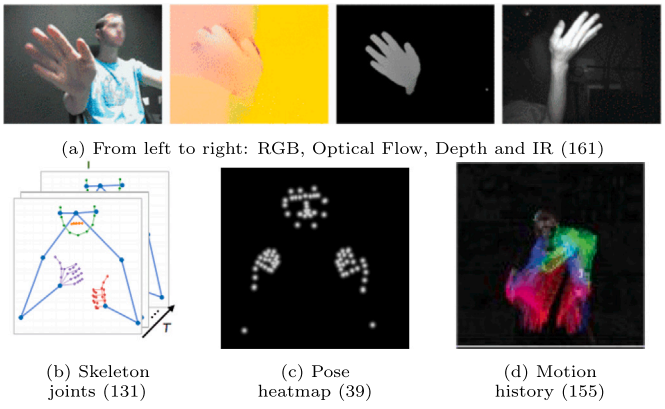


Fig. 2. Taxonomy of hand gesture recognition methods. A diverse set of criteria has been proposed by many works. This survey adds a few more, including "Architecture", considering it essential for a method-oriented presentation.



(a) From left to right: RGB, Optical Flow, Depth and IR (161)

(b) Skeleton joints (131)

(c) Pose heatmap (39)

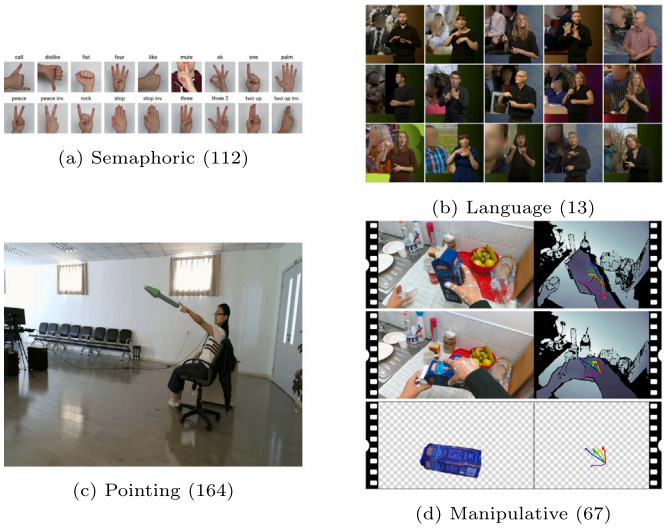
(d) Motion history (155)

Fig. 3. Modalities that are considered for VHGR.

- **Spatiotemporal** modeling architectures simultaneously analyze both the spatial and temporal domains, as in 3DCNNs.
- **Spatial + Temporal** pipelines first extract spatial features (e.g., using 2DCNNs) and then pass them through a sequential model (e.g., BiLSTM) to perform temporal modeling.
- **Short-term Spatiotemporal + Long-term Temporal** pipelines aim to capture local context from small parts of the overall sequence and then feed these spatiotemporal features into a temporal model to aggregate partial information and obtain a global vector representation.

3.7. Learning strategy

This criterion explores the different training methodologies adopted by researchers to develop VHGR models, often influenced by data availability, task complexity, and desired generalization. The main strategies identified include:



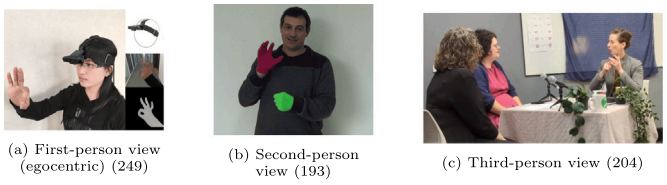
(a) Semaphoric (112)

(b) Language (13)

(c) Pointing (164)

(d) Manipulative (67)

Fig. 4. Main categories of hand gestures based on type of information conveyed.



(a) First-person view (egocentric) (249)

(b) Second-person view (193)

(c) Third-person view (204)

Fig. 5. Different camera points of view for VHGR.

- **Transfer Learning**, where models are pretrained on large-scale datasets (e.g., Jester [149]) and subsequently fine-tuned for specific downstream VHGR tasks [143].
- **Weakly Supervised Learning**, often used in CDHGR, where temporal annotations are coarse or unaligned, lacking exact gesture boundaries [234].
- **Knowledge Distillation**, in which one part of the network (often a teacher model) transfers knowledge to another (student model) to enhance feature extraction and efficiency [38,157,244].
- **Contrastive Learning**, used either in a CLIP-based framework to support zero-shot/few-shot recognition by leveraging textual descriptions [28,29,186], or to improve the discriminative capacity of learned representations [250,252].
- **Multi-task Learning**, where a single model is optimized to perform multiple related tasks simultaneously, encouraging knowledge sharing and improved generalization [130].
- **Reinforcement Learning**, applied to optimize decision-making processes, such as keyframe selection, by rewarding effective representations [62].
- **Few/Zero-shot Learning**, which enables the model to generalize to new gestures or languages with limited or no annotated examples, often leveraging semantic or textual guidance [28,29,186].
- **Adversarial Learning**, used as a data augmentation strategy [14] to improve robustness, as demonstrated in adversarial video synthesis networks [164].
- **Masked Reconstruction Pretraining**, inspired by BERT-style architectures, in which models are pretrained to reconstruct masked skeleton sequences, injecting structural prior knowledge (e.g., hand topology) [92,93,251].

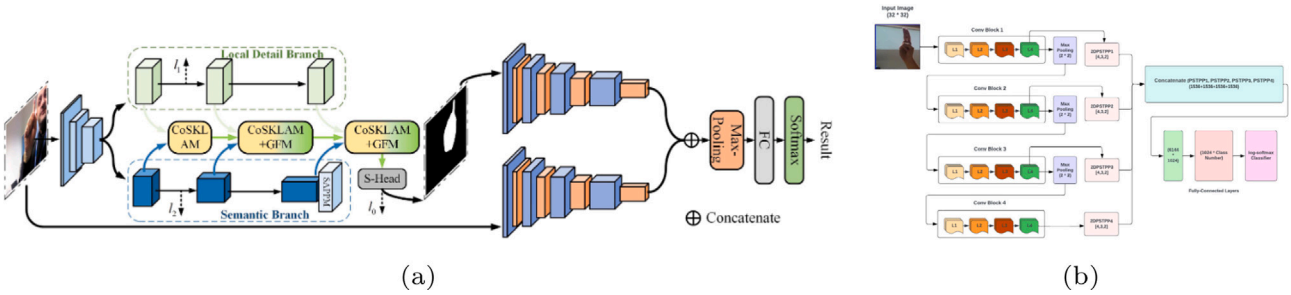


Fig. 6. Representative networks for SHGR: FGDSNet [256] (a) propose a dual-stream hand palm segmentation module followed by a classification model, while 2DPSTPP-Net [104] (b) incorporates a novel pooling mechanism within a custom network with dense connections.

3.8. Processing

Existing methods for isolated and continuous dynamic HGR can be further categorized based on whether the models have full access to the entire gesture(s) sequence or not. According to Molchanov et al. [160], two types of processing strategies apply:

- **Online (real-time) processing**: Online HGR models perform gesture recognition in a stream of input frames. In other words, these models take as input the gesture sequence recorded until a specific time. Thus, online gesture recognition is suitable for real-time applications, as the gesture is recognized before the retraction part, reducing the latency. Sliding windows and gesture detector/classifier pipelines are often employed [25,58].
- **Offline**: Offline processing means that the model treats the entire gesture sequence. This approach is limited to applications that do not require real-time recognition. Most of the existing methods for VHGR fall into this category, such as those using BiLSTM for temporal modeling [157].

The presented taxonomy of VHGR methods highlights the diversity of approaches developed to address different aspects of gesture understanding—from isolated dynamic gestures to sign language recognition and multimodal human-computer interaction. It provides a comprehensive foundation for analyzing current trends and identifying research gaps. Building on this overview, the following three sections delve deeper into specific VHGR categories, examining in detail the latest methods for Static (Section 4), Isolated Dynamic (Section 5), and Continuous Dynamic (Section 6) gesture recognition, respectively.

4. Methods for static VHGR

Static VHGR (SHGR) aims to identify gestures relying only on spatial data. As this approach neglects temporal information, it is only suitable for recognizing static hand gestures. For gesture classification, fine-tuned variants of well-known 2DCNN architectures are employed, such as VGG and DenseNet, or custom lightweight architectures are trained from scratch. When classification is combined with localization of gestures, variants of YOLO are adopted. Hand segmentation is often applied to eliminate the background, reducing the interference of gesture-irrelevant information. Hand or human detection is also utilized in some cases to normalize the spatial dimensions, especially for long-range recognition. Regarding application domains, SHGR methods are used in SLU, HCI, and HRI. However, a few more general-purpose models have also been proposed. Representative network structures and algorithmic pipelines for SHGR are illustrated in Fig. 6. The rest of the section analyzes representative methods in detail, highlights the current SOTA, and compares them, answering Q2 and Q3 (see Section 1).

4.1. Sign language understanding (SLU)

For SLU, the authors in [200] developed Gesture-CNN, a shallow 2DCNN to recognize Indian Sign Language (ISL) signs and slightly outperformed various versions of VGG on a self-collected dataset. The authors in [71] proposed the STN-ND-Net, a combination of a Spatial Transformer Network (STN) that improves robustness to spatial transformations, with a 2DCNN backbone that uses deconvolution operations to reduce the pixel-wise and channel-wise correlations and eliminate the information redundancy of colored images. The success of the vision transformer motivated the authors of SIGNFORMER [122] to adopt it for ISL recognition, while authors in [205] leveraged both skeleton and RGB data in their dual-stream feature extractor and an MLP to perform intermediate fusion. A YOLO variant, namely YOLOv4-CSP [9] has also been developed, to simultaneously detect and identify Turkish Sign Language gestures.

4.2. Human-computer interaction (HCI)

For gesture-based HCI the authors in [245] proposed a lightweight 2DCNN, named LHGR-Net, which combines a Multiscale Structure (MSS) that aims to make the model scale-invariant via atrous convolution with a Lightweight Attention Structure (LAS) based on CBAM. The authors in [104] proposed 2DPSTPP-Net, a 2DCNN with four stacked convolutional blocks whose features are pooled using max pooling and a novel method, then concatenated to form the final representation, tailored for HCI scenarios. The authors in [79] modified the YOLOv5 to make it more lightweight and developed a model that can be used for gesture-based user interfaces in smart sports stadiums, improving the experience of spectators and athletes. In Alam et al. [10], a pipeline for egocentric HGR and fingertip localization uses YOLOv2 for hand detection, followed by VGG16 for feature extraction and gesture classification. In Dang et al. [46], a lightweight HCI pipeline segments the hand using DeepLabV3 + or Unet, then classifies the cropped image with a 2DCNN like MobileNet or EfficientNet. Zhou et al. [256] introduced FGDSNet, a two-stage approach with dual-stream segmentation (local detail and semantic branches) and multi-level fusion for feature integration. The segmentation mask and original image are fed into a dual-stream classification module, where mid-level fusion combines features for final classification via an MLP.

4.3. Human-robot interaction (HRI)

HRI, particularly gesture-based mobile robot navigation, has attracted research interest. URGR [21] is a three-stage approach using YOLOv3 for human cropping, High-Quality Network (HQ-Net) for super-resolution to address long-range scenarios, and a classifier, Graph-Vision Transformer (GViT), for gesture recognition. The HQ-Net comprises convolutional and deconvolutional pathways. Features extracted by Canny Edge algorithm and a convolutional module followed by self-attention

Table 2
Static VHGR: comparison of method families and key insights.

Aspect	Sign Language Understanding (SLU)	Human-Computer Interaction (HCI)	Human-Robot Interaction (HRI)	General Purpose
Primary Function	Automated recognition of gesture alphabets, numeric signs, and static fingerspelling	Gesture-based control for VR/AR, smart environments, and device interaction	Intuitive mobile robots (UGVs/UAVs) guidance, robotic arm control, and Robotic Assisted Surgery (RAS)	Foundational feature extraction and gesture recognition across various contexts
Most suitable scenarios	Small- or medium-scale static sign vocabularies and fingerspelling, especially when gestures are captured under relatively controlled conditions	Real-time command interfaces for VR/AR, smart environments, and edge devices where low latency is more important than large vocabulary coverage	Long-range robot guidance and robotic-control scenarios where human/hand localization, scale normalization, or super-resolution is required	Baseline recognition, transfer learning, or general-purpose feature extraction when the application does not impose strong domain-specific constraints
Architectures	• Shallow 2DCNNs • 2DCNN variants (VGG, DenseNet) • Vision transformers (ViT) • YOLO variants	• Lightweight 2DCNNs • Hand palm segmentation • Multi-scale networks • YOLO variants	• Spatial normalization and super-resolution • Hybrid GCN-ViT classifiers	• Enhanced DenseNet • Multi-stream 2DCNNs • Pyramid feature aggregation • Multiple kernel sizes
Strengths	• High discriminative power for fine-grained details • Capturing of global semantic dependencies • Robust to spatial transformations	• Low-latency and real-time inference • Suitable for resource-constrained edge devices • Scale-invariance through (e.g., atrous) convolutions	• Effective ultra-range recognition at long distances (up to 25m) • Robustness to motion blur	• High generalization ability • Capturing of multiple levels of granularity • Rotation and scale invariance
Limitations	• High computational complexity for Transformer backbones • Sensitivity to visually similar signs • Requirement for large annotated datasets	• Difficulty with intricate background interference • Limited gesture vocabulary in lightweight models • High susceptibility to lighting	• High complexity of multi-stage pipelines • Dependency on accurate human/hand detection • Increased computational overhead for super-resolution	• Lack of domain-specific optimization • High number of parameters • Prone to overfitting on small, non-diverse datasets
Indicative Methods	Gesture-CNN [200], STN-ND-Net [71], SIGNFORMER [122]	LHGR-Net [245], 2DPSTPP-Net [104], FGDSNet [256]	URGR [21], DRCAM [196]	EDenseNet [216], SpAtNet [27], MMFANet [131]

Table 3
Performance of representative SHGR methods. Many approaches achieve high accuracy even with small model size, but the limited size and diversity of the datasets used may result in performance reduction during real-world deployment. Methods marked with star (*) perform hand palm segmentation before applying classification. "Architecture" refers to the structure of the classification model and "Params" to the total number of parameters. Results are reported as in the original publications and should be interpreted mainly within each benchmark dataset.

Method	Year	Modality	Architecture	Params	ASL-FS	HGR1	NUS-I	NUS-II	OUHANDS
FGDSNet [256] *	2024	RGB	Dual-stream 2DCNN	1.20M	–	94.80%	–	99.80%	97.80%
SpAtNet [27]	2024	RGB	Three-stream 2DCNN	3.80M	–	93.33%	–	–	–
[46] *	2023	RGB	EfficientNetB0	5.34M	–	98.15%	–	–	89.20%
LHGR-Net [245]	2023	RGB	2DCNN + lightweight attention	1.81M	–	93.15%	–	–	98.57%
2DPSTPP-Net [104]	2023	RGB	2DCNN	19.01M	–	–	100%	99.61%	–
DRCAM [196] *	2023	Depth	Attention-based 2DCNN	–	93.44%	–	–	–	–
EDenseNet [216]	2021	RGB	Modified DenseNet	7.77M	99.99%	–	–	99.30%	–

are fused with the latent representation. GViT utilizes two GCN layers that capture fine-grained details followed by a modified ViT. This setting enables effective recognition of gestures at a distance of up to 25m for UGV guidance. Similarly, Sahoo et al. [196] introduced a two-stage method based on depth data: hand segmentation followed by gesture classification using DRCAM, a 2DCNN with stacked Residual Channel Attention Modules (RCAMs) and dense interconnections between the cascaded blocks. Because of the attention mechanisms, the model focuses effectively on the most informative regions of the image.

4.4. General purpose

General-purpose models have also been developed. EDenseNet [216] extends DenseNet by adding convolutional layers between dense blocks. SpAtNet [27] is a lightweight 2DCNN with three streams: two Multi-scale Attentive Feature Fusion (MAFF) modules for coarse-grained features and one Interleaved Module for high-level representations; their fused features are fed to an MLP for classification. This approach, employing multiple kernel sizes, increases accuracy by capturing various levels of granularity. MMFANet [131] performs gesture detection and recognition using a 2DCNN backbone followed by the proposed feature aggregation pyramid network (FAPN) and task-specific regressor and classifier branches. The FAPN module aims to extract multiscale features through upsampling/downsampling operations and fusion of adjacent feature maps, while reducing the computational complexity by utilizing novel expanded convolutional blocks with different dilation rates combined with attention mechanisms that eliminate redundant information.

4.5. Comparison and critical summary of SHGR methods

Table 2 summarizes the main qualitative trade-offs among SHGR method families, while Table 3 reports the quantitative performance of representative methods. Most existing methods for SHGR are evaluated on different or custom datasets, which prevents a direct comparison of their performance. Despite their high accuracy and low parameter count – suggesting that SHGR might be a largely solved problem – the datasets used (ASL-FS [177], HGR1 [114], NUS-I [123], NUS-II [175], and OUHANDS [150]) remain relatively small (except for ASL-FS) and lack diversity. Consequently, model performance would likely degrade in real-world, uncontrolled environments. Among the existing methods, LHGR-Net [245] and FGDSNet [256] demonstrate strong results on the OUHANDS dataset with lightweight architectures, making them suitable for gesture-based HCI. On the other hand, EDenseNet [216] achieves high accuracy on ASL-FS but at a higher computational cost, making it better suited for static sign language recognition.

Combining Tables 2 and 3, we can draw the following conclusions (partially answering Q3 in Section 1):

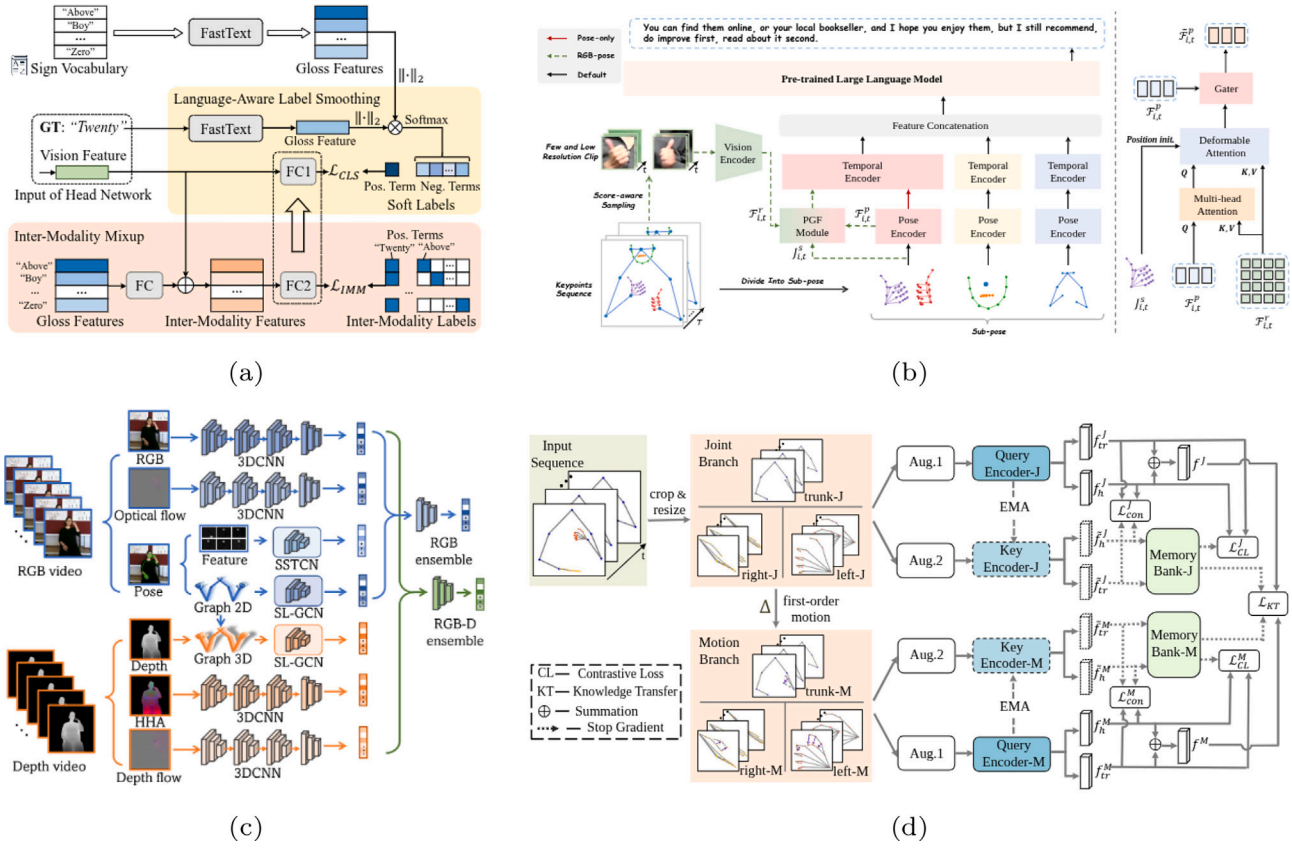


Fig. 7. Representative network structures and algorithmic pipelines for IDHGR: NLA-SLR [262] (a) introduces language-aware label smoothing to facilitate training, Uni-Sign [130] (b) proposes a multi-task framework incorporating a language model, SAM-SLRv2 [105] (c) deploys several modalities for accurate recognition, and [252] (d) applies consistency constraints in a self-supervised manner.

- Direct comparison of recognition accuracy should be interpreted cautiously, since methods are often evaluated on different datasets and not all works report computational cost, preprocessing overhead, or protocol-level details. Therefore, the quantitative comparison in Table 3 is intended mainly as a benchmark-oriented summary rather than a fully controlled comparison across methods.
- The RGB modality remains dominant, while depth information has rarely been leveraged [196] and has been associated with lower accuracy. Only a few works have used multiple modalities, as standalone RGB-based approaches already achieve sufficient performance. Since these multimodal approaches have been evaluated on custom datasets, they are not shown in the table.
- Applying background removal through hand segmentation [46, 196,256] does not appear to significantly improve accuracy compared to end-to-end learning approaches.
- Slightly modified general-purpose classifiers [46,216] show competitive performance compared to models specifically tailored to HGR.
- More recent works use a smaller number of parameters, revealing a shift toward resource-efficient recognition [27,256].
- Newer publications also consider multi-stream approaches [27, 256] to enrich the extracted features without substantially increasing computational cost.

5. Methods for isolated dynamic VHGR

Unlike SHGR approaches, which fail to capture temporal dynamics and struggle with dynamic gestures, recent advances in **Isolated Dynamic Vision-based Hand Gesture Recognition (IDHGR)** have focused on recognizing individual gesture instances from spatiotemporal inputs such as RGB video sequences or skeleton joint trajectories.

When gestures are part of a sign language, the task is referred to as Isolated Sign Language Recognition (ISLR). A notable distinction is that ISLR models frequently leverage both manual features (e.g., hand movements) and non-manual features (e.g., facial expressions, upper body motion), which can significantly enhance recognition accuracy. As a result, many ISLR approaches go beyond hand-centric analysis to incorporate richer contextual cues.

In the following subsections, we present IDHGR methods grouped by input modality as (a) **unimodal appearance-based**, (b) **unimodal pose-based**, and (c) **multimodal** approaches. Then, we conclude with a summary of key innovations and SOTA performance; this aims to answer Q2 and Q3 (see Section 1). Representative network structures and algorithmic pipelines for IDHGR are illustrated in Fig. 7.

5.1. Unimodal appearance-based methods

Unimodal appearance-based methods rely on a single input type—such as RGB, optical flow, infrared (IR), motion history, depth, or depth flow—to classify hand gestures. As shown in Fig. 3, these representations resemble color images, which allows for the application of similar computer vision techniques. These methods can be further categorized based on the architecture of the feature extractor, which typically follows one of three main pipelines: (1) spatiotemporal, (2) spatial + temporal, and (3) short-term spatiotemporal + long-term temporal backbones. The following paragraphs explain each of these architectures in more detail.

5.1.1. Spatiotemporal

Spatiotemporal backbones perform spatial and temporal modeling jointly, typically using 3D CNNs or (2 + 1)D CNNs—either custom-built or adapted from existing models.

In Wang et al. [228], the authors introduced a custom network, (2 + 1)D-SLR, composed of five stacked (2 + 1)D convolutional blocks (i.e., a 2D convolution followed by a 1D convolution). This structure reduces the computational overhead compared to traditional 3D convolutions, making the model more practical for deployment. To enhance temporal modeling, authors in [26] proposed a trainable key frame selection method based on error correction. VGG16 features are used to identify key frames, which are then preprocessed and fed into a custom 3DCNN for classification.

In Wang et al. [231], a dual-stage multi-stream pipeline for Chinese ISLR was developed. An adapted EfficientDet model—enhanced with cross-level Bi-FPN connections and a dual-channel spatial attention module (DCSAM)—detects both hands. The full image and cropped hand patches are passed through three parallel 3D-ResNet-18 models (following the P3D scheme), with their features fused and classified via an MLP. Authors in [82] introduced TS3C-Net, a modified 3D ConvNeXt that adds an extra 3 × 3×3 convolutional branch to each block, enhancing spatiotemporal representation. Finally, Tang et al. [218] addressed limitations in fixed-pipeline designs by combining features from both ResC3D and ConvLSTM using the Selective Spatiotemporal (SeST) module. The fused features are passed through fully connected and softmax layers for classification.

5.1.2. Spatial + Temporal

These backbones first extract spatial features from each frame, typically using 2D CNNs or vision transformers, and then model the temporal dynamics with sequential modules such as GRUs or LSTMs. The resulting sequence embeddings are aggregated, and a classifier (e.g., MLP) outputs the final prediction.

In Hax et al. [85], a method tailored for HCI tasks using only depth data was proposed. The method combined Inception-v3 which extracted spatial features with an LSTM for temporal modeling. For ISLR, authors in [223] introduced a pipeline to recognize Indian sign language words used by deaf farmers. They combined GoogleNet for spatial encoding with a BiLSTM for sequence modeling.

To reduce inference time, a lightweight ISLR model, called TIM-SLR has been proposed in [229]. The model uses a 2DCNN for per-frame feature extraction, followed by a Time Interaction Module (TIM)—a zero-parameter unit that fuses frame features into a compact representation. Finally, authors in [54] proposed a two-stage transformer-based model that extracts spatial features per frame, while a causally masked transformer processes the sequence, ensuring that only past and present frame features contribute at each step.

5.1.3. Short-term spatiotemporal + Long-term temporal

The methods in this category follow a two-stage architecture: the first stage captures short-term spatiotemporal patterns from segments of the input sequence, while the second stage aggregates these features to form a final representation of the entire sequence.

EgoFormer [178] is a model for egocentric dynamic gesture recognition optimized for edge devices, which has been introduced in the context of autonomous driving. EgoFormer combines a custom 3DCNN with a transformer module and an MLP for final classification. For long-range gesture-based UGV control, authors in [24] proposed a pipeline where a lightweight 2DCNN extracts per-frame features, which are used to select key frames. After cropping the human region, the selected frames are passed through a SlowFast network, followed by a transformer and an MLP.

To address data scarcity in sign language understanding, Bilge et al. [28] proposed a zero-shot learning framework. Sign videos are represented using BERT-encoded textual descriptions combined with handcrafted attributes. This representation guides a dual-stream backbone composed of short-term spatiotemporal modules (e.g., I3D) and long-term models (e.g., BiLSTM).

5.1.4. Other

Some appearance-based methods fall outside the previously discussed categories but offer valuable contributions, especially in ISLR. For example, the method proposed in [144] for Indian ISLR compresses an entire frame sequence into a single image, which is then processed by a MobileNet. A second MobileNet extracts per-frame features, which are passed through a two-layer LSTM. The outputs of both streams are concatenated and refined using a custom shallow 2DCNN. A different approach was presented in [121], where the authors introduced a dual-stream ISLR architecture combining Inception V4 with a novel convolutional module called stem. Liu et al. [141] proposed a dual-stream model with keyframe selection. A 2DCNN and attention mechanism filter out less informative frames. The remaining keyframes are processed by two branches: the Hand stream, which crops and analyzes hand regions with attention blocks, and the Full stream, which processes the entire frame. The fused features are used for final classification.

5.2. Unimodal pose-based methods

Pose-based methods, also known as skeleton-based approaches, rely on structured motion data—such as skeleton graphs and pose heatmaps—to classify gestures. These methods often incorporate various forms of pose information, including joint coordinates, bones, motion trajectories, and joint angles, sometimes in combination [50]. They also explore joint subsets related to specific body regions [172], often adopting multi-stream architectures to handle these diverse inputs.

Given the natural graph structure of skeletal data, many methods employ Graph Convolutional Networks (GCNs) originally developed for HAR. In the sections that follow, we group the approaches based on the structure of their backbone, following the same organization as in the previous subsection. Notably, pose information alone is often insufficient for accurate sign recognition [161]. Consequently, many models integrate additional cues—typically from RGB inputs—using 3DCNNs and perform score-level fusion to enhance performance.

5.2.1. Spatiotemporal

Pose-based methods that incorporate spatiotemporal representations simultaneously process spatial and temporal data to improve SLU and HCI.

In SLU, pose-based methods model dynamic hand and body movements for gesture recognition. SKIM [132] combines dynamic GCNs and temporal convolutional networks (TCNs) to analyze global and local motion, while MC-LSTM [169] fuses multiple streams (face, hand, body) via temporal convolutions and GCNs. P3D [126] employs transformer-based part-wise and whole-body encoders, and HWGAT [172] enhances key body-part focus with hierarchical attention. TMS-Net [50] introduces a six-stream GCN for skeleton, bone, and motion data with inter-stream fusion. To tackle data limitations, SLGTformer [212] and VSCT [30] use transformers and few-shot learning, respectively. Dual-stream approaches such as [205] and MULTISTAGE-CGAR [155] integrate joint and motion data through temporal convolutions and attention. Other notable advances include SL-TSSI-DenseNet [125], which converts joint coordinates into image-like matrices (termed Tree Structure Skeleton Image, TSSI) for DenseNet processing; STST [249], a two-stage transformer for local-global spatiotemporal features; and DSTA-SLR [101], which employs dynamic adjacency matrices and multi-stream fusion for improved recognition accuracy. Dual-stream architectures such as STGCN-LSTM [80] combine ST-GCN and self-attention ConvLSTM for capturing global and local features.

In HCI, pose-based methods address challenges such as data overfitting and the need for self-supervised learning frameworks. MAE-based frameworks [103] use a masked autoencoder to reconstruct skeleton graphs, focusing on motion and spatial relationships. 3sISTGCN [211] proposes a three-stream model to process joints, bones, and motions,

applying modified ST-GCN layers and late fusion for improved classification. Another approach, DG-STA [6], tackles class-incremental learning in HCI by training models on non-overlapping tasks without access to old training data. DSTSA-GCN [41] addresses temporal expression limitations by stacking three stages of graph convolutions and multiscale temporal convolutions. In another approach, TD-GCN [137] introduces a pose-based method for IDHGR. Unlike traditional methods that use fixed skeleton topologies, TD-GCN employs temporal-sensitive topology learning to enhance feature quality. First, graph convolution is used to extract high-level spatiotemporal representations, which are then used to calculate temporal and channel-independent adjacency matrices. These matrices help model the skeleton, and the features are fused using the proposed TCF module. Finally, authors in [96] proposed ST-CGNet, a dual-stream architecture comprising a 3DCNN and a convolutional LSTM module to capture different levels of information. Attention mechanisms and an advanced fusion strategy are utilized to efficiently incorporate the features of each stream.

5.2.2. Spatial + Temporal

Many models use a two-step approach, where the spatial domain is first processed by a spatial perception module to extract spatial features. These features are then passed into a sequential model for temporal aggregation, resulting in a combined spatial-temporal representation of the input sequence. For instance, GCN-BERT [221] utilizes a two-layer GCN for spatial modeling, followed by BERT for temporal modeling.

Other general-purpose methods adopt a multi-stream, two-stage pose-based approach. In Narayan et al. [165], three streams, each containing three 1D convolutional layers, are combined with a parallel module of average pooling layers. The extracted features are then concatenated and input into an LSTM, followed by a fully connected layer and softmax for classification. Similarly, STr-GCN [209] employs a two-stage model where a GCN spatial feature extractor is first applied to obtain intra-frame features and then passes the features through a transformer encoder for modeling inter-frame temporal relationships. The final classification scores are generated by an MLP.

5.2.3. Short-term spatiotemporal + Long-term temporal

In this approach, the backbone first extracts features from smaller segments of the skeleton sequence to capture short-term spatial-temporal relationships. Then, long-term temporal modeling is applied to aggregate global context across the entire sequence.

For example, gesture recognition in autonomous driving has become an important research area. In Guo et al. [76], the authors propose a pose-based approach called MG-GCT for recognizing police traffic gestures. The model uses two graph convolution streams: one to process skeleton joints and another to analyze joint motions. The features extracted from both streams are fused and passed through a transformer encoder for temporal modeling, with the final classification performed by an MLP. Authors in [170] proposed a dual-stream backbone processing both hand and whole body keypoints using ST-GCN models. Cross-attention is then applied to focus on the most important features, which are subsequently passed through an LSTM to model long-term dependencies.

5.2.4. Temporal

Some pose-based methods bypass spatial modeling and focus directly on sequential processing. Typically, keypoints are slightly preprocessed—often by extracting geometric features such as bone angles—and then passed into sequential models like RNNs or TCNs to produce a vector representation of the input sequence.

In Qi et al. [179], the authors proposed a pose-based method for surgical robot teleoperation. Preprocessed skeleton joints are fed into an LSTM, and the resulting features are passed to an MLP for final classification. The model was evaluated on a custom dataset. Similarly, in [195], a cross-domain gesture recognition approach was developed using geometric features (e.g., bone angles), which are processed by a TCN. The resulting features are summarized into a sequence descriptor, and a linear classifier predicts the final class. To improve signer independence and reduce overfitting, the SwC GR-MMixer [140] incorporates GRU units for temporal modeling and a shifted window concatenation strategy to enhance sequence representation.

To address data scarcity, the authors in [164] introduced AVSN, an adversarial learning-based data augmentation framework. The method jointly optimizes augmentation parameters and the HGR classifier. A BiLSTM serves as the discriminator, while an MLP acts as the generator. The generator produces challenging augmented data that exploits weaknesses in the discriminator, which in turn learns to correctly classify gestures.

5.2.5. Other

Several works explore alternative architectures or skeleton representations beyond the main categories. In Abdullahi et al. [1], a Fourier CNN with BiLSTM extracts spatiotemporal features from 3D joints for ISLR, using a trainable binary mask for feature selection. Similarly, Alamri et al. [11] applies feature thresholding based on variance and importance, followed by deep CNN classification. To overcome color-coded map limitations, Kishore et al. [116] introduces the Joint Motion Affinity Map (JMAM) with a multi-stream attention-based 2DCNN for feature extraction and recognition.

Key frame selection is addressed in [62] via FS-Net, a reinforcement learning agent that selects informative frames for DD-Net classification. Li et al. [129] proposes a lightweight variant merging attention-based and autoencoder streams. Finally, the Hierarchical Self-Attention Network (HAN) [138] employs modular attention across finger joints (J-Att, F-Att, T-Att, Fusion-Att) for efficient spatial-temporal modeling, achieving strong accuracy with just 0.05 GFLOPs.

5.3. Multimodal methods

Multimodal approaches combine different input types (e.g., RGB, Depth, Skeleton, Text) to improve gesture recognition performance by leveraging complementary information. IDHGR multimodal methods do not seem to follow any trend, likely due to the wide spectrum of alternative options (combinations of streams, fusion techniques, modalities, etc.), so we group them by application domain.

5.3.1. General purpose

In Li et al. [128], a dual-stream 3DCNN with average fusion was proposed for RGBD gesture recognition, using a memory bank to retain discriminative features (z) while filtering redundant ones (v) by maximizing mutual information with class labels. RAAR3DNet [254] employed Neural Architecture Search (NAS) with Dynamic Static Attention (DSA) modules and a modified I3D backbone to emphasize motion and hand regions, supported by an auxiliary HeatmapNet and late RGBD fusion.

To better exploit modality interactions, Zhou et al. [255] introduced a decoupling-recoupling framework combining independent spatial and temporal modeling with mid-level fusion through a shared MLP-AE module. Multi-level fusion was further explored in [78], integrating 3D CNNs with a Convolutional Transformer Fusion Block (CTFB) that fuses intermediate cross-stream features.

To address data scarcity, several works explored zero-shot learning by incorporating textual gesture descriptions encoded by pretrained language models (mainly BERT). In Rastgoo et al. [186], a dual-stream model combined a vision transformer for spatial perception and an LSTM-MLP for temporal modeling, learning visual features aligned with textual embeddings. Similarly, Rastgoo et al. [187] fused skeleton-derived geometric features with RGB-C3D-LSTM-AE representations into a shared semantic space for classification. To reduce computational costs, Garg et al. [69] replaced the transformer's self-attention mechanism with convolutional layers and a gated feedforward network, resulting in the lightweight ConvMixFormer architecture.

Knowledge distillation and contrastive learning were applied in [226], where three 3DCNNs (RGB, Depth, Optical Flow) were fused with Coordinate Attention, using I3D as teacher and ResNet as student, while integrating CLIP for zero-shot recognition. Finally, Ma et al. [145] and Yu et al. [242] replaced standard convolutions with 3D Central Difference Convolution (3D-CDC), combining hybrid residual blocks, transformers, and NAS-based optimization for enhanced spatiotemporal modeling.

5.3.2. Sign language understanding (SLU)

Recent works on ISLR have focused on multimodal, pose-centric architectures combining RGB and skeleton data. SignBERT [93] employs a spatial GCN with temporal and hand chirality encoding, feeding features into a transformer encoder and a hand-model-aware decoder for self-supervised pretraining, with parallel RGB predictions fused at inference. Its enhanced variant, SignBERT+ [92], adds spatial-temporal position encoding and an improved RGB-pose fusion module. Similarly, MASA [250] uses the same GCN + Transformer backbone with contrastive learning on pose reconstruction, while BEST [251] embeds pose triplets with BERT and applies late fusion with RGB data.

To improve feature invariance, Zhao et al. [252] applies a contrastive approach enforcing consistency among hand, trunk, and motion features, using the same GCN + Transformer backbone with late fusion of pose and RGB predictions. SAM-SLR [106] fuses six streams (RGB, depth, skeleton, optical flow, pose heatmaps, HHA), while SAM-SLR-v2 [105] adds a 3D hand graph stream and learnable fusion weights.

VTN-HC/PF [48] processes hand-cropped images via a 2DCNN, transformer, and MLP for final predictions, with PF incorporating pose-based geometrical features. Motion History Images are used in [154], analyzed by a 3DCNN with attention, and in a dual-stream setup combined with 2DCNN RGB features via weighted summation.

To tackle data scarcity, ZSSLR [264] uses a CLIP-based zero-shot framework, while [29] proposes a few-shot cross-lingual method combining two 3DCNNs and a pose estimator with a Temporal Feature Aggregation Module. NLA-SLR [262] targets semantically similar signs with a multistream S3D backbone and language-aware label smoothing. MSNN [148] uses five I3D streams plus ST-GCN, and hDNN-SLR [182] combines a 3DCNN, attention-based BiLSTM, and a VAE-based module. Finally, Uni-Sign [130] proposes a unified multimodal architecture with three pose encoders and an RGB encoder fused via attention, followed by temporal encoding and LLM-based pretraining for ISLR and CSLR.

5.3.3. Human-computer interaction (HCI)

Motivated by the difficulty of recognizing fine-grained gestures in HCI applications, authors in [248] proposed HandSense, a method relying on depth and RGB data. Two 3DCNNs are deployed to extract features for each modality. These features are fused and fed into an SVM to obtain the final prediction. Authors in [112] proposed an approach relying on RGBD Dynamic Images (RDDIs). These images are obtained by combining RGB and estimated depth data. The resulting RDDI is fed into a modified Xception, resulting from removing the top four Xception modules, in order to extract meaningful features. These features are finally used to perform the classification.

5.3.4. Other

Inspired by the success of transformers, GestFormer [67] introduces a lightweight transformer-based framework for gesture-driven HRI. A 2DCNN extracts per-frame features, processed by PoolFormer layers (replacing attention) and two custom blocks for multiscale pooling/WT and token mixing. Each modality is processed independently, and outputs are averaged. In Gao et al. [65], a hybrid 3DCNN/LSTM model fuses RGB, depth, and skeleton data (from an enhanced pose estimator) via weighted sum. The fused input feeds a spatiotemporal model combining 2D/3D convolutions and ConvLSTM layers. Authors in [59] emphasize two major issues: information redundancy and the difficulty of recognizing similar gestures. Their approach leverages a CLIP encoder to provide qualitative embeddings and enhances the feature extractor through label smoothing as well as a semantic filter to incorporate multimodal data.

5.4. Comparison and critical summary of IDHGR methods

Table 4 provides the qualitative comparison of the main IDHGR method families, while Table 5 reports their quantitative performance on the most commonly used benchmarks. The performance of IDHGR methods in the most commonly used benchmarks (IsoGD [225], SHREC [210], DHG [49], MSASL [109], WLASL [127], CSL500 [102] and AUTSL [208]) is shown in Table 5. Since the results are collected from the original publications rather than reproduced under a unified experimental setup, they should be interpreted cautiously. Direct comparison is most meaningful within the same benchmark dataset and reported protocol; across datasets, modalities, and preprocessing pipelines, differences in data difficulty, pose-estimation overhead, additional training data, and fusion strategies may affect the reported performance. For ISLR, Uni-Sign [130] surpasses relevant methods in terms of recognition accuracy, but requires powerful hardware to execute quickly, due to its large size (592.1M parameters, most of them contained in the language model). Thus, smaller but less accurate models like NLA-SLR [262], and SignBERT+ [93,252] are more reasonable solutions for real-world applications. Even though the multimodal SAM-SLR [106] and SAM-SLR-v2 [105] are very accurate models for ISLR, they require a large number of modalities that are not fused effectively (using late fusion), which may be suboptimal. Pose-based methods for HCI and HRI including TD-GCN [137], STST [249] and DSTSA-GCN [41] achieve high recognition rates with relatively low computational cost. However, the computational complexity of pose estimation and the effect of its incorrect predictions in more challenging environments (e.g., fast movements) may reduce their overall performance. Appearance-based methods are not affected by off-the-shelf pose estimators but they comprise many more parameters (up to 251.4M [242]).

Overall, the main insights arising from Tables 4 and 5, resolving Q3 (see Section 1), are the following (note that most of the purely appearance-based methods are evaluated on either custom or lesser-known datasets, so only a few of them are included in the table):

- Keyframe selection [62,141] does not appear to introduce additional errors due to failures, unlike pose estimators, which may degrade accuracy. However, only a few works have leveraged its potential to reduce computational cost, probably due to the complexity of training this additional component.
- A few works consider human body regions and process each of them independently [130,138,141,148,170,172,252]. However, it remains unclear whether the independent processing of different human regions improves accuracy, as observed in some works [130,252], or not, as reported in others [148].
- NAS is deployed in older approaches [242,254], surpassing their contemporary counterparts at the time, around 2021. However, although such methods explore a wide spectrum of possible architectures, they use older backbones, such as I3D [254] instead of more advanced models like S3D [262], or limited training strategies, such as [242]. This leads to inferior accuracy compared to more recent methods.
- Unimodal pose-based approaches still incorporate derived skeleton modalities, such as motion and bones [101,132,211], without any significant impact on recognition accuracy.
- Having been thoroughly explored, the ST-GCN scheme dominates pose-based IDHGR [132,211] and achieves superior performance [41,137]. However, spatiotemporal transformers are emerging as competitive alternatives in terms of accuracy [249].
- Researchers seldom rely solely on temporal modeling [140], aiming instead to fully exploit the spatial domain.
- In general, pure transformers have not yet been applied for appearance-based feature extraction. Instead, only 3DCNN backbones, either custom models or models originally proposed for

Table 4
Isolated dynamic VHGR: comparison of appearance-based, pose-based, and multimodal method families.

Aspect	Unimodal appearance-based methods	Unimodal pose-based methods	Multimodal methods
Primary function	Direct spatiotemporal feature extraction from raw pixel intensities	Geometric modeling of skeletal joint relationships and trajectories	Integration of hierarchical features from cross-modal data streams
Most suitable scenarios	Trimmed RGB/depth videos where hand shape, texture, and scene context are informative and the computational budget allows video backbones	Privacy-sensitive, low-light, or background-cluttered settings where reliable skeletons are available and geometric motion is sufficient	High-accuracy ISLR or RGB-D HCI/HRI scenarios where complementary cues can justify the added sensing, synchronization, and fusion cost
Input data types	• RGB • Depth • Infrared (IR) • Optical Flow (OF) • Motion history	• 2D/3D skeletal joints • Bones • Joint angles • Motion vectors	• RGB + Skeleton • Depth + Pose • Visual + textual/semantic guidance
Architectures	• 3DCNN • (2 + 1)D CNN • ConvLSTM • 2DCNN + RNN • Video Transformers	• Graph Convolutional Networks (GCNs) • Temporal Convolutional Networks (TCNs) • Transformers	• Multi-stream ensembles • Cross-attention fusion modules • Hybrid GCN-CNNs
Learning strategy	• Transfer learning • Contrastive learning • Supervised pretraining	• Masked reconstruction • Self-Supervised Learning (SSL) • Adversarial data augmentation	• Knowledge distillation • Multi-task learning • Zero-shot/Few-shot learning
Strengths	• Capturing of fine-grained hand shapes, texture and environmental context	• Computationally efficient • Privacy-preserving • Lighting-invariant	• Highest accuracy • Robust to occlusion • Handling of visually similar gestures
Limitations	• High computational cost • Sensitive to background clutter and lighting	• Dependency on pose estimator accuracy • Lack of textural detail	• Extreme complexity • Large parameter footprint • Difficulty in synchronization
Indicative methods	(2 + 1)D-SLR [228], TS3C-Net [82], EgoFormer [178], TIM-SLR [24,229]	TD-GCN [137], STST [249], DSTSA-GCN [41]	SAM-SLR [106], Uni-Sign [130], NLA-SLR [128,262]

action recognition, have been used to extract features from related modalities [105,106,148,218,262].
- Incorporating multiple modalities and cues, such as face and body regions, under a multi-task, high-capacity framework significantly improves accuracy but increases computational cost [130].
- Numerous approaches have achieved more than 90% accuracy on datasets with small corpora, such as SHREC [210] and DHG [49]. More complicated datasets, especially for ISLR, that comprise large vocabularies and only a few samples per class remain highly challenging, such as MSASL [109] and WLASL [127].

6. Methods for continuous dynamic VHGR

In contrast to IDHGR, where each sample contains at most one gesture, Continuous Dynamic Hand Gesture Recognition (CDHGR) deals with unsegmented gesture sequences. The goal is to recognize gestures from untrimmed videos and generate the corresponding gesture sequence. CDHGR is typically treated as a **weakly supervised task**, as the precise temporal boundaries of gestures are not provided; the model learns to identify them implicitly, often using Connectionist Temporal Classification (CTC) loss [72,234].

While CDHGR has been explored in domains like autonomous driving [61,230], most research focuses on Sign Language Understanding (SLU), where it is commonly referred to as **Continuous Sign Language Recognition (CSLR)** or sign2gloss. CSLR aims to produce a sequence of glosses[1] from video input. Due to the importance of non-manual features (e.g., facial expressions and upper-body movements), many approaches also incorporate these regions to improve recognition accuracy. Some CSLR methods are further extended to support gloss-based Sign Language Translation (SLT) by adding a gloss-to-text module [38,73,263].

[1] Glosses are atomic lexical units in sign language, analogous to words in a spoken language.

In the following, we review recent advances in CDHGR, grouped by input modality. We highlight key architectures, learning strategies, and benchmark results as required by Q2 and Q3 (see Section 1). Representative network structures and algorithmic pipelines for CSLR are illustrated in Fig. 8.

6.1. RGB-based

Most SOTA RGB-based CDHGR models follow a **three-stage pipeline** architecture: (1) a spatial model (commonly a 2DCNN or a transformer) to extract frame-wise features; (2) a short-term temporal model (mainly 1DCNN or TCN) for local temporal modeling and gloss-level feature extraction; and (3) a contextual module (e.g., BiLSTM) for long-term sequence modeling and final gesture sequence prediction [81,90,91,97–100,157,244]. This standard architecture was introduced in [42]. This decoupled spatial-temporal structure has been favored over unified spatiotemporal models like 3DCNNs due to the latter's inability to define clear temporal boundaries [2].

Many RGB-based CDHGR methods build on the canonical 2DCNN + 1DCNN + BiLSTM architecture by introducing auxiliary losses and multi-task objectives. For example, VAC-CSLR [157] augments the visual and temporal modules with a Visual Enhancement (VE) loss and a Visual Alignment (VA) distillation loss to accelerate convergence and improve representation quality. Similarly, SMKD [81] adds two auxiliary classifiers—one after the 1DCNN and one after the BiLSTM—to perform synchronous gloss segmentation alongside recognition, using a three-stage training schedule to mitigate gradient "spikes". C^2SLR [261] enforces Spatial Attention Consistency (SAC) by focusing the visual module on pose-informed regions, and Sentence Embedding Consistency (SEC) to align the visual and sequential representations of the same gloss.

To better capture local temporal patterns, several works replace or enhance the standard 1DCNN. Temporal Lift Pooling (TLP) [98] adopts a learned pooling scheme inspired by the lifting transform, eliminating handcrafted kernels. mLTSF-Net [237] couples content-aware feature

Table 5
Performance of IDHGR methods on various datasets. Notably, accuracy depends on dataset difficulty; for example, methods with accuracy near 100% on AUTSL do not exhibit similar performance on more challenging datasets with thousands of classes, such as MSASL and WLASL. Top-3 accuracies are shown in bold. KS, R and NAS stand for keyframe selection, regions and Neural Architecture Search, respectively. Colors correspond to architecture type: Spatiotemporal, (Short-term) Spatiotemporal + (Long-term) Temporal, Spatial + Temporal, Temporal and other. Methods marked with "†" have used additional training data. Results are reported as in the original publications; therefore, comparisons should be made primarily within the same benchmark dataset and reported protocol [94,156].

Method	Year	KS	R	NAS	Fusion	Backbone	Params	IsoGD	SHREC		DHG		MSASL1000		WLASL2000		CSL500	AUTSL
									14	28	14	28	PI	PC	PI	PC		
Unimodal appearance-based																		
SeST [218]	2021	✗	✗	✗	Mid-level	ResC3D & ConvLSTM	–	60.27%	–	–	–	–	–	–	–	–	–	–
[141]	2024	✓	✓	✗	Mid-level	Other	–	–	–	–	–	–	–	–	–	–	**98.87%**	92.53%
Unimodal pose-based																		
STGCN-LSTM [80]	2025	✗	✗	✗	Mid-level	ST-GCN & ConvLSTM	–	–	–	–	–	–	–	–	–	–	95.2%	–
DSTSA-GCN [41]	2025	✗	✗	✗	–	GCN/TCN	8.08M	–	**97.74%**	**95.37%**	**95.04%**	**93.57%**	–	–	–	–	–	–
TMS-Net [50]	2024	✗	✗	✗	Multi-level	GCN/TCN	–	–	–	–	–	–	–	–	56.4%	–	–	–
SKIM [132]	2024	✗	✗	✗	Late	GCN/TCN	–	–	–	–	–	–	–	–	57.73%	55.40%	–	–
DSTA-SLR [101]	2024	✗	✗	✗	Late	GCN/TCN	7.4M	–	–	–	–	–	65.74%	62.31%	53.68%	51.17%	98.15%	–
STST [249]	2024	✗	✗	✗	–	Transformer	–	–	**97.62%**	**95.83%**	**94.82%**	**93.18%**	–	–	–	–	–	–
HWGAT [172]	2024	✗	✓	✗	–	Transformer & GCN	–	–	–	–	–	–	–	–	48.49%	–	–	**95.80%**
[103]	2024	✗	✗	✗	–	MAE + ST-GCN	–	–	94.1%	90.0%	–	–	–	–	–	–	–	–
TD-GCN [137]	2024	✗	✗	✗	Late	GCN/TCN	1.36M	–	**97.02%**	**95.36%**	93.9%	91.4%	–	–	–	–	–	–
SL-TSSI-DenseNet [125]	2023	✗	✗	✗	–	TSSI + DenseNet	7.2M	–	–	–	–	–	–	–	–	–	–	93.13%
3sISTGCN [211]	2022	✗	✗	✗	Late	Modified ST-GCN	7.1M	–	96.7%	94.9%	93.7%	91.2%	–	–	–	–	–	–
[170]	2025	✗	✓	✗	Mid-level	ST-GCN + Attention+LSTM	–	–	–	–	–	–	–	–	–	–	–	86.43%
SBI-DHGR [165]	2023	✗	✗	✗	Mid-level	Multistream 1DCNN + LSTM	28M	–	97%	92.36%	**94.64%**	91.79%	–	–	–	–	–	–
STr-GCN [209]	2023	✗	✗	✗	–	GCN + Transformer	–	–	93.39%	89.20%	–	–	–	–	–	–	–	–
SwC GR-MMixer [140]	2024	✗	✗	✗	–	GRU	–	–	–	–	–	–	–	–	–	–	**98.54%**	–
HAN-2S [138]	2025	✗	✓	✗	Multi-level	Transformer	0.94M	–	95%	92.86%	92.71%	89.15%	–	–	–	–	–	–
[129]	2024	✗	✗	✗	Late	AE & Transformer	0.21M	–	96.9%	94.17%	94.21%	**92.11%**	–	–	–	–	–	–
[156]	2023	✗	✗	✗	Mid-level	GCN & MLP	–	–	97.01%	92.78%	92.00%	88.78%	–	–	–	–	–	–
DRL-DDNet [62]	2023	✓	✗	✗	–	Other	–	–	96.5%	93.8%	–	–	–	–	–	–	–	–
Multimodal																		
MSNN [148]	2024	✗	✓	✗	Late	I3D & ST-GCN streams	–	–	–	–	–	–	50.59%	–	47.26%	–	–	–
NLA-SLR [262]	2023	✗	✗	✗	Multi-level	Multi-stream S3D variant	–	–	–	–	–	–	**73.80%**	**70.95%**	**61.26%**	**58.31%**	–	–
[128]	2023	✗	✗	✗	Late	Multi-stream Res3D	–	**74.08%**	–	–	–	–	–	–	–	–	–	–
CTFB [78]	2023	✗	✗	✗	Multi-level	Dual-stream I3D	–	**69.02%**	–	–	–	–	–	–	–	–	–	–
SlowFast + CDC [242]	2021	✗	✗	✓	Multi-level	Multi-stream SlowFast	251.4M	66.23%	–	–	–	–	–	–	–	–	–	–
[94]	2021	✗	✗	✗	Late	ST-GCN & Res3D	–	–	–	–	–	–	69.39%	66.54%	51.39%	48.75%	98.3%	–
RAAR3DNet [254]	2021	✗	✗	✓	Late	I3D with attention	–	66.62%	–	–	–	–	–	–	–	–	–	–
SAM-SLR-v2 [105]	2021	✗	✗	✗	Late	3DCNN & GCN/TCN	–	–	–	–	–	–	–	–	**59.39%**	**56.63%**	**99.00%**	**98.53%**
SAM-SLR [106]	2021	✗	✗	✗	Late	3DCNN & GCN/TCN	19.2M	–	–	–	–	–	–	–	58.73%	55.93%	–	**98.53%**
Uni-Sign [130] †	2025	✗	✓	✗	Mid-level	Various encoders + LLM	592.1M	–	–	–	–	–	**78.16%**	**76.97%**	**63.52%**	**61.32%**	–	–
MASA [250]	2024	✗	✗	✗	Late	GCN + Transformer & I3D	–	–	–	–	–	–	–	–	55.77%	53.13%	–	–
[252]	2024	✗	✓	✗	Late	GCN + Transformer & I3D	–	–	–	–	–	–	**74.18%**	**72.18%**	58.06%	55.66%	97.8%	–
BEST [251]	2023	✗	✗	✗	Late	GCN + Transformer & I3D	–	–	–	–	–	–	71.21%	68.24%	54.59%	52.12%	97.7%	–
SignBERT+ [92]	2023	✗	✗	✗	Late	GCN + Transformer & I3D	–	–	–	–	–	–	73.71%	70.77%	55.59%	53.33%	97.8%	–
De + Recouple [255]	2022	✗	✗	✗	Multi-level	Other	–	**66.79%**	–	–	–	–	–	–	–	–	–	
SignBERT [93]	2021	✗	✗	✗	Late	GCN + Transformer & I3D	–	–	–	–	–	–	71.24%	67.96%	54.69%	52.08%	97.6%	–

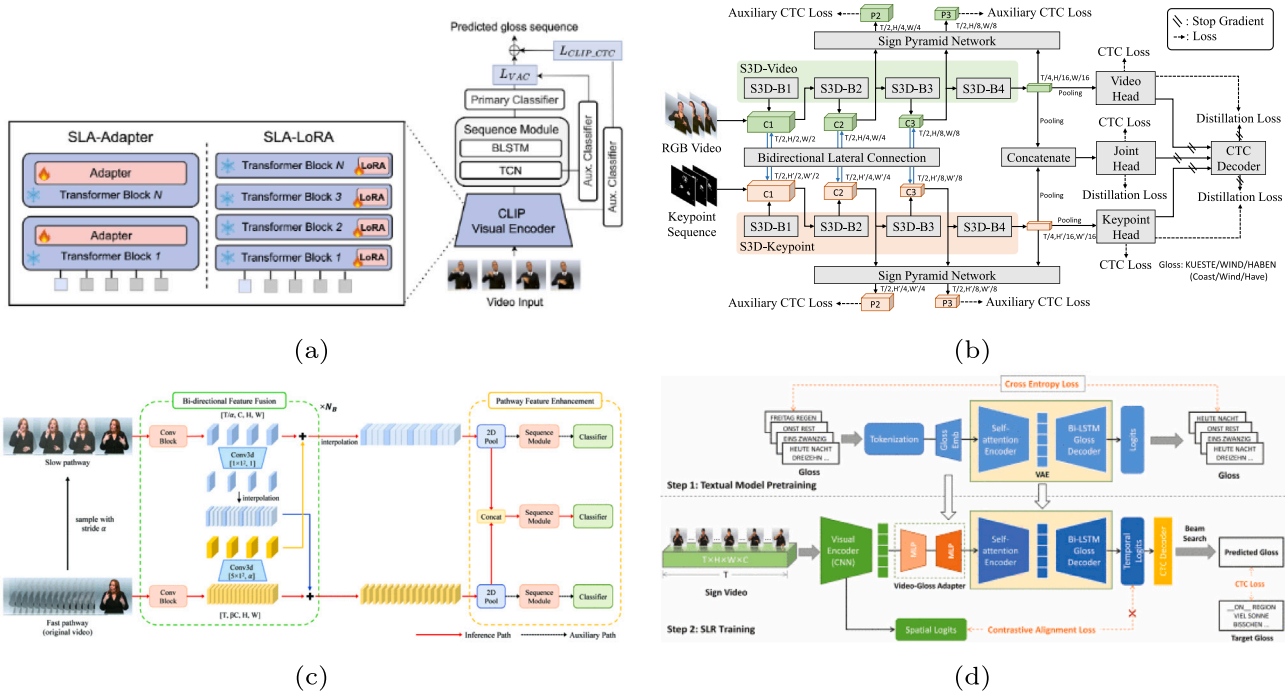


Fig. 8. Representative network structures and algorithmic pipelines for CSLR: CLIP-SLA [15] (a) enhances spatial perception using a foundation model, TwoStream-SLR [38] (b) proposes an end-to-end 3DCNN for multimodal CSLR with an advanced supervision scheme, SlowFastSign [5] (c) considers various frame rates to increase performance, and CVT-SLR [253] (d) uses a generative model to produce glosses instead of the standard decoder.

selection with position-aware convolutions to fuse adjacent-frame similarities at multiple scales. HS-I3D [236] extends I3D by extracting intermediate embeddings and deploying a U-Net-style decoder that upsamples temporally, enabling simultaneous sign detection and boundary localization.

Other methods inject lightweight attention and correlation blocks into the spatial encoder. CorrNet [99] adds parallel correlation and identification modules after each 2DCNN stage to model inter-frame trajectories and highlight informative regions; CorrNet+ [97] further replaces its correlation operator with a more efficient variant and inserts an attention layer. SEN-CSLR [100] introduces self-emphasizing spatial (SSEN) and temporal (TSEN) modules to reweight feature maps and keyframes with minimal overhead. STNet [233] applied frame-wise loss to the 2DCNN backbone output to enhance feature extraction and inserted multiple Spatial Resonance Modules to facilitate inter-frame modeling.

Many recent approaches replace 2D CNNs with transformer-based or hybrid CNN/attention modules. CLIP-SLA variants [15] use a frozen CLIP visual encoder with LoRA or MLP adapters and an auxiliary classifier for early supervision. Swin-MSTP [16] combines a Swin transformer with an MS-TCN using TLP and dilated convolutions, followed by a BiLSTM for global context.

TCNet [142] augments a CNN backbone with trajectory and correlation modules, while SlowFastSign [5] modifies SlowFast-101 with bidirectional fusion and auxiliary temporal classifiers. To better model spatial correlations, SignGraph [63] treats frame patches as nodes in a dynamic graph, alternating between Local Sign Graph (LSG) and Temporal Sign Graph (TSG) modules; its multiscale variant, MultiSignGraph, further improves spatiotemporal context.

CORE [247] proposed a different approach aiming to leverage the relations between regions within an input frame, deploying an off-the-shelf pose estimator to detect fine-grained and contextual cues; the corresponding frame regions are fed into a 2DCNN and the extracted features are fused using a GNN to obtain the final embedding. These per-frame representations pass through a BiLSTM to perform global modeling. SD-Net [64] enhanced temporal modeling by using three parallel branches. Each branch consists of a two-stage temporal context module that captures both frame-wise and gloss-wise features. Contrastive loss is applied between two of the branches, based on the proposed Entity and Template scheme. Authors in [70] formulated CSLR as a cross-modal alignment task using a diffusion model to generate the glosses aiming to mitigate the weak supervision, achieving superior performance. DSTEN-CSLR [241] replaced the BiLSTM with a custom contextual model, consisting of memory and feed-forward blocks. Furthermore, the 2DCNN backbone was enhanced by adding a Spatial-Channel Attention (SCA) mechanism, supervised by ground-truth heatmaps, provided by an off-the-shelf pose estimator.

Some models integrate linguistic embeddings with visual features to leverage gloss co-occurrence and language structure. SignBERT-HKSL [260] combines a (3 + 2+1)D ResNet with BERT and BiLSTM, pretraining on masked inputs and fine-tuning iteratively on masked outputs. C^2ST [244] replaces the CTC decoder with a Conditional Gloss Decoder (CGD) to model gloss dependencies and aligns visual and gloss embeddings via cross-modal matching loss. GPGN [75] similarly applies a cross-modal matching loss between BERT-derived gloss features and PDC-TFE visual embeddings, followed by a BiGRU for sequence modeling.

CVT-SLR [253] pretrains an asymmetric VAE on a pseudo gloss-to-gloss task, then connects it to a visual encoder with a video-gloss adapter and contrastive alignment loss. Fully transformer-based PiSLTRc [238] uses content-aware neighborhood gathering and position-informed attention for end-to-end CSLR, while [90] introduces prior-aware cross-modality augmentation, training a transformer to reconstruct masked gloss sequences and generate pseudo video-text pairs via guided gloss substitutions and insertions.

Finally, data augmentation and cross-lingual strategies help alleviate limited annotations. Fangyun Wei et al. [234] build a “dictionary” by using a pretrained CDHGR model to automatically segment continuous videos into isolated signs, then map German and Chinese sign corpora into a shared embedding space. CSGC [184] maintains a Gloss Prototype

memory that stores class prototypes updated during training; a gloss contrastive loss then enforces cross-sentence consistency.

6.2. Pose-based

Pose-based (or skeleton-based) approaches for CDHGR rely exclusively on body pose information, typically in the form of 2D/3D skeleton joints or pose heatmaps. These methods aim to model human motion while discarding appearance-based noise. However, due to the rapid motions (especially in the CSLR context) pose estimators often fail, making solely pose-based methods frequently inaccurate.

Several works exploit structured keypoint information by dividing the body into semantic regions such as the face, hands, and upper body. MSKA-SLR [73] processes each region independently via attention-enhanced temporal convolutions across four branches. The resulting features are fused to decode gloss sequences with CTC loss supervision, complemented by frame-level self-distillation. Similarly, CoSign [107] employs group-specific GCNs for each subset and introduces regularization to preserve inter-stream consistency. Their dual-stream variant integrates joint positions and motion features using multi-level fusion to capture both static and dynamic cues.

Temporal localization of gestures has also been explored in the literature. TY-Net [52] draws inspiration from YOLO to detect gesture instances in time by predicting their boundaries and centers, analogous to spatial bounding boxes.

Beyond joint coordinates, pose heatmaps serve as an alternative input modality. The multi-cue framework in [258] combines spatial and temporal reasoning by localizing facial and hand regions via a pose estimator and subsequently encoding them with a temporal module, before decoding gloss sequences with a CTC head. The FA-STGCN model [61] further improves robustness to motion scale and frequency variability. It normalizes skeleton inputs and applies alternating spatial and temporal GCNs. Spatial modules learn latent joint relationships, while temporal processing is split into global and local paths—extracting coarse-grained and fine-grained motion patterns, respectively.

Finally, in the context of autonomous driving, Wang et al. [230] developed a two-stage pose-based system to classify road intersection gestures. The first stage extracts geometric features and encodes them via LSTM layers. The second stage uses a dual-stream setup, where one branch processes geometric descriptors and the other uses a ConvLSTM to handle co-occurrence matrices. Features are fused with directional orientation indicators for final classification.

6.3. Multimodal

Multimodal CDHGR approaches combine inputs such as RGB, pose heatmaps, and depth to overcome unimodal limitations and capture complementary information.

A key example is TwoStream-SLR [38], a dual-stream architecture for CSLR and SLT. It processes RGB and pose heatmaps in parallel via the first four S3D blocks in each branch, with bidirectional lateral connections for multi-level information exchange. Sign Pyramid Networks and auxiliary heads (video, keypoint, joint) provide additional supervision, while self-distillation minimizes KL-divergence across outputs. The same framework has been applied to ISLR [263] using a sliding window to extract sign boundaries and construct training dictionaries.

Other CSLR models have adopted multimodal ISLR backbones. Uni-Sign [130] can be used directly for CSLR, while SignBERT+ [92] can be extended by appending temporal pooling and a CTC decoder to generate gloss sequences.

Beyond sign language, general-purpose CDHGR systems have explored multimodality across broader visual sensor inputs. For example, Lu et al. [143] propose a framework that segments RGB, depth, infrared, and optical flow inputs using a shared U-Net and processes them via separate 3D DenseNet-43 backbones. Extracted features are fused through Canonical Correlation Analysis to align shared representations across modalities.

6.4. Comparison and critical summary of CDHGR methods

Table 6 provides a qualitative horizontal comparison of CDHGR method families, while Table 7 reports the computational cost and Word Error Rate (WER) of CSLR methods on three benchmark datasets. Since these results are compiled from the original publications, comparisons should be interpreted mainly within the same benchmark dataset and reported protocol. Comparing other CDHGR methods (e.g., for HRI) is infeasible, as they are evaluated on different datasets and use other metrics. Leveraging Diffusion Models (DMs) as in [70] yields a substantial performance gain without dramatically increasing parameters. SlowFastSign [5] achieves strong accuracy with a moderate number of parameters, while TwoStream-SLR [38] has slightly higher WER, requires twice as many parameters, and involves a more complex training setup, though WER improves with additional training data [234]. C^2ST [244] is accurate but contains 187.8M parameters, which may limit real-world deployment. For resource-constrained scenarios, the pose-based CoSign-2s [107] requires only 30.1 GFLOPs, though pose estimation cost is not included. DSTEN-CSLR [241] is a lightweight RGB-based model but less accurate. Overall, a trade-off between accuracy and computational cost persists.

Summarizing, the key insights that can be derived from Tables 6 and 7, resolving Q3 (see Section 1), are the following:

- Only a few works, most of them pose-based, process regions corresponding to human body parts independently by applying separate networks to each region [73,107,247,258], as this usually requires pose estimation. This technique does not seem to increase performance in RGB-based CSLR [247] and may introduce additional errors due to pose-estimation failures.
- The three-stage paradigm, namely the 2DCNN + 1DCNN + BiLSTM pipeline [157], remains dominant even with modifications. However, such methods are often outperformed by specific single-stage approaches [38,234], which achieve lower WER due to advanced training strategies and the integration of multiple modalities.
- ResNet [157] is the most commonly used spatial-perception baseline.
- Considering spatiotemporal modeling in the early stages instead of frame-wise feature extraction may offer slight performance gains for RGB-based CSLR [5].
- RGB-based CSLR can also benefit from adopting visual foundation models for spatial perception [15], owing to their large capacity and extensive pretraining.
- The two-stage scheme for temporal modeling, namely short-term followed by long-term modeling, is rarely replaced by generative models [70,130,253] or single sequential modules [73,142].
- Diffusion models for gloss generation are examined in a single work [70], which achieves superior performance and overcomes some limitations of the standard weakly supervised paradigm.

7. Summary of SOTA methods

After reviewing numerous recent works, this section focuses on SOTA methods across tasks, summarizing the field and comparing key methods in terms of architecture, deployment, strengths, and limitations in order to address Q2 and Q3 (see Section 1). Table 8 highlights each method's task, application domain, input modality, architecture, and major strengths and weaknesses.

One insight from the table is that SLU dominates other fields, likely due to its complexity, which requires advanced visual cues (including facial expressions), extensive temporal modeling for dynamic gestures, and classifiers capable of distinguishing highly similar gestures. Consequently, few SOTA methods exist for static SLR, as they struggle to capture hand movements. General-purpose IDHGR methods, used as baselines for tasks like SLU, HRI, and HCI, have also been widely studied;

Table 6
Continuous dynamic VHGR: horizontal comparison of RGB-based, pose-based, and multimodal method families.

Aspect	RGB-Based Methods	Pose-Based Methods	Multimodal Methods
Primary function	Mapping of unsegmented pixel sequences directly to gloss/command ones	Conversion of skeletal joint coordinate sequences into labels	Fusion of complementary signals (e.g., RGB + pose) to resolve ambiguities
Most suitable scenarios	CSLR and continuous command recognition when RGB video is available and the model must exploit hand shape, facial cues, body context, and scene information	Resource-constrained or privacy-sensitive continuous recognition where pose extraction is reliable and appearance information is less critical	High-performance CSLR/SLT settings where RGB and pose/depth cues can be fused to handle co-articulation, ambiguous signs, and weak temporal supervision
Data representation	• Raw pixel values	• 2D/3D joints • Bones • Motion vectors • Pose heatmaps	• Heterogeneous data embeddings (e.g., RGB + skeleton + depth)
Temporal modeling	• 1D-TCN • BiLSTM • Multi-scale temporal perception modules	• ST-GCNs • Temporal decoupling • Transformers	• Cross-modal attention • Lateral connections • Distillation across streams
Neural network architectures	• 2DCNN + 1DCNN + BiLSTM • 3DCNNs (I3D, Res3D) • Transformers (Swin-MSTP, CLIP-SLA)	• Graph Convolutional Networks (ST-GCN, CoSign) • 1DCNN/TCN + BiLSTM	• Dual-stream encoders (TwoStream-SLR) • Unified frameworks with LLM decoders (Uni-Sign)
Strengths	• Rich visual context • Capturing of facial/body cues • No need for off-the-shelf estimators	• Invariant to illumination and background • Low computational cost • Privacy-safe approaches	• Mitigation of individual sensor noise • Superior accuracy • Robust to co-articulation
Limitations	• Susceptible to motion blur, occlusions, and cluttered backgrounds • High GFLOP rate	• Reliance entirely on pose estimation quality • Ignoring of appearance cues	• Complex training • High number of network parameters • Inference latency
Indicative methods	SlowFastSign [5], CorrNet+ [97], CLIP-SLA [15,70]	CoSign [107], FA-STGCN [61], MSKA-SLR [73]	TwoStream-SLR [38], Uni-Sign [130], SignBERT+ [92,234]

Table 7
Comparison of CSLR methods in terms of computational cost and accuracy (Word Error Rate, WER) on three benchmarks: CSL-Daily [257], Phoenix [60], and Phoenix T [34]. "Params" denotes the number of *all* parameters, both trainable and frozen. Methods marked with "†" have utilized additional training data and "‡" enhance the vanilla spatial perception model with custom modules (e.g., correlation modules [97,99] and CBAM [261]). "Regions" refers to whether the methods consider the body regions (e.g., grouping keypoints accordingly [73]) or not. Reported GFLOPs and parameters follow the original publications; preprocessing overhead, such as pose estimation, may not be included unless explicitly stated. The top-3 metrics are shown with bold.

		Architectural Elements				Performance				
Method	Year	Regions	Stages	Backbone	Decoder	GFLOPs	Params	CSL-Daily	Phoenix	Phoenix-T
RGB-based										
[70]	2025	✗	2	ResNet	DM	–	86.64M	**23.7%**	**16.8%**	**17.2%**
CORE [247]	2025	✓	3	ResNet	GNN + BiLSTM	–	–	31.1%	21.7%	21.6%
SD-Net [64]	2025	✗	–	ResNet	Hybrid	–	–	27.9%	–	20.3%
STNet [233]	2025	✗	3	ResNet	1DCNN + BiLSTM	–	–	27.2%	19.3%	19.8%
CSGC [184]	2024	✗	3	ResNet	1DCNN + BiLSTM	–	–	26.7%	19%	19.5%
GPGN [75]	2024	✗	3	ResNet	TCN + BiGRU	–	–	30%	20.4%	20.5%
CTCA [74]	2023	✗	2	ResNet	TCN & BiLSTM	–	–	29.4%	20.1%	20.3%
CVT-SLR [253]	2023	✗	2	ResNet	VAE	–	–	–	20.1%	20.3%
TLP [98]	2022	✗	3	ResNet	TLP + BiLSTM	2950.5	48M	–	20.8%	21.2%
SMKD [81]	2021	✗	3	ResNet	1DCNN + BiLSTM	1032.3	20.3M	–	21%	22.4%
VAC-CSLR [157]	2021	✗	3	ResNet	1DCNN + BiLSTM	1120.4	22.3M	–	22.3%	–
DSTEN-CSLR [241]‡	2025	✗	3	ResNet	TCN + FFN & Memory	366.3	30.19M	28.1%	20.0%	20.3%
TCNet [142]‡	2024	✗	2	ResNet	BiLSTM	932.1	19.3M	29.3%	18.9%	19.4%
CorrNet+ [97]‡	2024	✗	3	ResNet	1DCNN + BiLSTM	–	–	28.2%	18.2%	19.1%
CorrNet [99]‡	2023	✗	3	ResNet	1DCNN + BiLSTM	1035.4	20.5M	30.1%	19.4%	20.5%
SEN-CSLR [100]‡	2023	✗	3	ResNet	1DCNN + BiLSTM	1144.2	23.1M	30.7%	21%	20.7%
C^2SLR [261]‡	2024	✗	3	VGG	TCN + Transformer	1124.6	32.6M	–	20.4%	20.4%
SLA-Adapter [15]	2025	✗	3	CLIP ViT	TCN + BiLSTM	–	–	25.8%	18.8%	19.5%
SLA-LoRA [15]	2025	✗	3	CLIP ViT	TCN + BiLSTM	–	–	25.8%	19.3%	19.4%
Swin-MSTP [16]	2025	✗	3	SwinT	TLP + BiLSTM	–	–	27.1%	18.7%	19.7%
C^2ST [244]	2023	✗	3	SwinT	1DCNN + BiLSTM	1635	187.8M	25.8%	**17.7%**	18.9%
SlowFastSign [5]	2024	✗	3	SlowFast	1DCNN + BiLSTM	–	52.5M	**24.9%**	18.3%	**18.7%**
MultiSignGraph [63]	2024	✗	3	ST-GCN	1DCNN + BiLSTM	–	–	26.4%	19.1%	19.1%
Pose-based										
MSKA-SLR [73]	2025	✓	2	MHA	TCN	–	–	27.1%	21.2%	19.8%
CoSign-2s [107]	2023	✓	3	GNN	1DCNN + BiLSTM	30.1	28.2M	27.2%	20.1%	20.1%
STMC [258]	2022	✓	3	VGG	TCN + BiLSTM	1742.1	40.71M	–	20.7%	21%
Multimodal										
Uni-Sign [130]†	2025	✓	3	Various	Temporal encoders + LLM	–	592.1M	26%	–	–
[234]†	2023	✗	1		– Dual-stream S3D –	–	105.2M	–	**16.7%**	**18.5%**
TwoStream-SLR [38]	2022	✗	1		– Dual-stream S3D –	–	105.2M	**25.3%**	18.8%	19.3%

Table 8
Comparison of current SOTA methods. They are compared in terms of task, application domain (referred to as "Domain" for the sake of brevity), input modality (or modalities) and general architecture. The strengths of each method or the key innovations introduced as well as some notable limitations of them, are also mentioned.

Method	Task	Domain	Modality	Architecture	Strengths or Key Innovation	Limitations
FGDSNet [256]	SHGR	HRI	RGB	Dual-stream hand segmentation model + dual-stream 2DCNN + mid-level fusion	The dual-stream classifier addresses any failures of the task-specific segmentation model.	The segmentation model and the 2DCNN backbones must be trained separately.
LHGR-Net [245]	SHGR	HCI	RGB	Atrous convolutional layers followed by depth separable convolutions and attention.	Suitable for resource-constrained deployment.	Very small capacity.
2DPSTPP-Net [104]	SHGR	HCI	RGB	2DCNN with custom pooling layers and residual connections	Spatial pyramid pooling and residual connections obtain robust representations.	The model is relatively large for SHGR.
Gesture-CNN [200]	SHGR	SLU	RGB	2DCNN consisting of four convolutional layers	The simplicity of the architecture enables easy implementation and training.	Relatively large (67M parameters).
SAM-SLR [106], SAM-SLR-v2 [105]	IDHGR	SLU	RGBD, Pose, HHA, OF, DF	Six/seven streams followed by late fusion	Incorporated a large number of modalities, achieving very high accuracy.	Insufficient fusion and computationally expensive.
SignBERT [93], SignBERT+ [92] (ISLR)	IDHGR	SLU	RGB, Pose	Spatial GCN + Transformer & 3DCNN + late fusion	The model gains hand topology prior through SSL pre-training (masked hand tokens reconstruction, inspired by BERT)	Needs extensive SSL pretraining, the processing and incorporation of the RGB modality may be suboptimal
SignBERT+ [92] (CSLR)	CDHGR	SLU	RGB, Pose	Spatial GCN + Transformer & 3DCNN + intermediate fusion + BiLSTM		
BEST [251]	IDHGR	SLU	RGB, Pose	VAE + Spatial GCN + Transformer & 3DCNN + late fusion		
NLA-SLR [262]	IDHGR	SLU	RGB, Pose	Four-stream network based on S3D with two frame rates and multi-level fusion	Novel architecture performing multimodality fusion efficiently. Language-aware label smoothing facilitates training. Robust to visually similar signs recognition.	Is affected by the performance of word encoders
Uni-Sign [130]	IDHGR, CDHGR	SLU	RGB, Pose	A pretrained LLM process features extracted by pose, visual and temporal encoders	Provides a unified framework for multiple SLU tasks, increasing the recognition accuracy.	Requires additional training data
TD-GCN [137]	IDHGR	–	Pose	ST-GCN variant	Facilitates temporal modeling through non-fixed time-dependent topology	Depends on the pose estimation performance.
SBI-DHGR [165]	IDHGR	–	Pose	Multi-stream 1DCNN + BiLSTM followed by late fusion	Utilizes features of various levels of granularity.	Very large compared to other pose-based methods
DSTSA-GCN [41] (4s)	IDHGR	–	Pose	Four stream GCN + TCN	Group channel-wise and temporal-wise GCNs with dynamic graph topologies improve the performance.	The static part of the hand topology is initialized randomly, lacking prior.
SlowFast + NAS [242]	IDHGR	–	RGBD, OF	Multi-stream 3DCNN with 3D-CDC and multi-level fusion	Enhanced temporal modeling through central difference convolution and optimal hyperparameter selection through NAS.	Needs substantial computational resources (251.4M params, 232.1GFLOPs) compared to similar methods
RAAR3DNet [254]	IDHGR	–	RGBD	Dual stream I3D with adoptable cells, spatial and temporal attention	NAS improves the architecture. The spatial attention mechanism is supervised by an off-the-shelf pose estimator.	It relies on the correctness of the pose estimator
CTFB [78]	IDHGR	–	RGBD	Dual-stream 3DCNN + Multi-level fusion with transformer-based modules	Performs efficient multimodality fusion via multiple attention modules placed in different layers.	The scaling to more than two modalities is not straightforward
[70]	CDHGR	SLU	RGB	Visual encoder followed by a Diffusion Model (DM).	Using a DM for gloss generation makes it very accurate.	Increased complexity.
CLIP-SLA [15]	CDHGR	SLU	RGB	CLIP visual encoder (with MLP adapters or LoRA) + TCN + BiLSTM	Early supervision by adding an extra classifier accelerates training, while the utilization of a foundation model enhances the feature extraction.	The utilization of CLIP implies high computational cost.
Swin-MSTP [16]	CDHGR	SLU	RGB	SwinT + 1DCNN/TLP	Deploying a multiscale TCN, the model is able to recognize continuous signs of varying time durations.	Computationally expensive.
C2ST [244]	CDHGR	SLU	RGB	SwinT + 1DCNN+BiLSTM	CGD loss (that avoids the conditional independence assumption of CTC) and self-distillation increase the performance.	The vision transformer as well as the language model utilized during training increase the computational complexity.
CoSign-2s [107]	CDHGR	SLU	Pose	ST-GCN variant	The computational cost is negligible, thus CoSign-2s perfectly balances efficient execution and accurate recognition.	The performance depends on the attributes of the pose estimator.
[234]	CDHGR	SLU	RGB, Pose	TwoStream-SLR	Utilization of extra training data via dictionary construction.	Needs additional training data.
TwoStream-SLR [38]	CDHGR	SLU	RGB, Pose	Dual-stream 3DCNN with bidirectional lateral connections	Extra supervision by auxiliary losses and self-distillation applied to additional prediction heads facilitate training. The multi-level fusion incorporates efficiently the two modalities.	Increased complexity due to the use of two streams and an off-the-shelf pose estimator. Very large.

the similarity of gestures in HRI and HCI allows the same methods to serve both domains.

Depth, pose, and other non-RGB modalities are used less frequently and typically in combination with RGB streams. This is because they either cannot represent fine-grained gestures (depth, optical flow) or perform poorly in dynamic recognition (pose estimators often fail on blurry images). RGB-based models—even older or simpler ones—already achieve strong results on SHGR datasets [167], so incorporating pose can incur significant computational overhead without substantial accuracy gains. In CSLR, rapid hand movements degrade pose estimator performance, making pose usable only alongside RGB [38] and increasing complexity. Thus, pose-based approaches are mostly limited to IDHGR, with some exceptions [107], where incorporating multiple modalities can notably improve performance [105,106], even with late fusion.

Regarding architecture, SHGR's simplicity allows even small 2DCNNs [200,216] to achieve very high accuracy (≥95%). However, the limited size and diversity of datasets such as those proposed in [123, 150,175,220] can lead to overfitting, reducing real-world performance. In contrast, IDHGR and especially ISLR rely on more complex models. ST-GCN variants [137,211] and GCN + Transformer pipelines [92,93,251,252] mainly process skeleton data, often combined with RGB 3DCNNs (e.g., I3D, Res3D) to compensate for pose estimation errors. While many studies focus on improving pose-based streams, RGB modalities are less explored. SAM-SLR [106] and SAM-SLRv2 [105] propose advanced separate processing for each modality, representing current SOTA for IDHGR. RGBD models [78,128,254] typically use dual-stream 3DCNNs, increasing computational cost. CSLR demands even more complex architectures to capture short-term (gloss-wise) and long-term (global) context. The 2DCNN + 1DCNN + BiLSTM pipeline remains common, with relatively fewer parameters than SHGR models (e.g., Gesture-CNN [200] 67M vs. large CSLR models 48M [98]). Vision transformers increasingly replace 2DCNN backbones [15,16,244] for their attention-based focus on critical regions. TwoStream-SLR [38] employs a dual-stream 3DCNN with multimodal fusion and pose heatmaps (instead of skeleton joints) alongside RGB frames, achieving high accuracy.

As mentioned earlier, SHGR is a relatively easy task, so researchers have not proposed any notable learning strategies. Pose-based methods for ISLR often [92,93,251,252] insert hand topology priors through SSL. NAS is also utilized to determine the optimal structure of 3DCNNs. Models for CSLR face serious optimization difficulties compared to other methods due to their complexity. Thus, extra supervision is provided via additional losses attached to various layers of the models [15,157]. Leveraging additional training data combined with accurate architectures during training [234] increases performance; therefore, methods marked as using additional data should not be compared one-to-one with methods trained only on the original benchmark data.

In summary, LHGR-Net [245] and stand-alone pose-based methods [107,129,137] are suitable for resource-constrained use cases (assuming that the cost of pose estimation is negligible, which is true especially when the keypoints are extracted by the deployed camera), balancing the computational complexity with recognition accuracy. TwoStream-SLR [38], C^2ST [70,244] achieve superior performance for CSLR but come with complex training settings and large computational costs. NLA-SLR [262] and SAM-SLRv2 [105] surpass most of the related methods, but they require substantial computational resources. Uni-Sign [130] introduced a novel approach by combining various SLU tasks into a single sequence generation task. Although it achieves significant performance and unifies a variety of tasks providing a shared framework, the computational complexity makes real-world deployment difficult, and the requirement of training data with detailed annotations prevents its adoption in other sign languages.

8. Gesture-based interaction paradigms

Having investigated the different algorithmic perspectives of Visual Hand Gesture Recognition (VHGR) methods and their particular characteristics in Sections 3–7, this section outlines how automatically recognized gestures are practically used in real-world interactive systems. In particular, the principal and most widely used gesture-based interaction paradigms, enabled by the use of VHGR methods, fall into the following closely related areas, which are described below: a) Human-Computer Interaction (HCI), with particular emphasis on immersive Virtual and Augmented Reality (VR/AR), b) Human-Robot Interaction (HRI), and c) Human-Machine Interaction (HMI). The following figure (Fig. 9) illustrates indicative applications of gesture-based interaction.

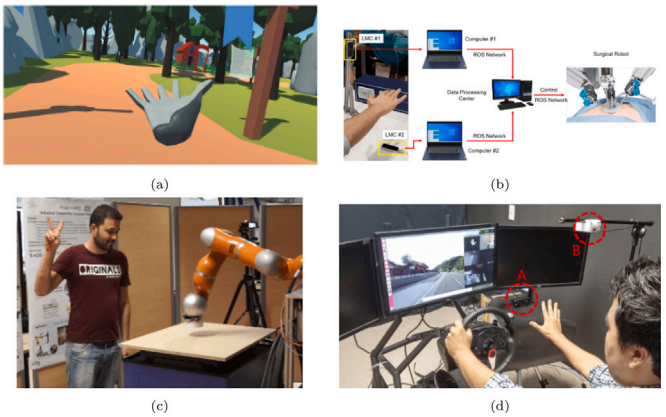


Fig. 9. Indicative works for gesture-based interaction schemes: a) locomotion within a virtual environment by identifying the palm orientation [33], b) surgical robot teleoperation via gestures [179], c) robotic arm control in a teach-by-demonstration paradigm [152] and d) car manipulation [160].

8.1. Gesture-based interaction in HCI: VR/AR

Gesture-based interfaces have been expanded in immersive VR/AR applications, replacing controllers and more cumbersome devices to enable more convenient, user-friendly, and intuitive interaction. In this context, the main gesture-based interaction paradigms are as follows:

- **Selection and manipulation**: Authors in [31] introduce the so-called "Go-Go" and "HOMER" techniques for arm-extension-based virtual object selection. Additionally, Mine et al. [158] show that body-relative interaction techniques exploiting proprioception can improve both speed and accuracy compared to world-relative ones, a principle that motivates the prevalence of pinch and grab metaphors in current Head-Mounted Displays (HMDs). The survey [18] provides an overview of 3D selection techniques, while [153] performs a similar task for direct object manipulation, jointly establishing an evaluation framework for gesture-based interfaces. Authors in [110] propose a framework termed DirectionQ for multiple object selection, considering three gesture commands: pinch to start the operation, release to stop the selection and hand movement in the desired direction to indicate one of the selected objects.
- **Mid-air gesture vocabularies for VR/AR**: Authors in [176] perform an elicitation study for HMD-AR, deriving a user-defined gesture set and concluding that one-handed (or unimanual) semaphoric commands are strongly preferred over two-handed (also referred to as bimanual) ones. Authors in [32] propose FingARtips, which comprises a gesture-based direct-manipulation AR interface based on a fingertip-tracked pinch metaphor, and [188] show that combining static HGR with HMM-based dynamic HGR is sufficient to drive AR menu systems. Interaction paradigms for immersive VR/AR on consumer hardware (Microsoft HoloLens 2, Meta Quest) are summarized in [87], reporting that practitioners converge on a small core gesture vocabulary (pinch, point, grab

and air-tap) providing a minimum set of classes that models should support.

- **Locomotion and bimanual interaction**: Locomotion refers to navigation through the virtual environment. An indicative work is [33], which examines three settings for gesture-based locomotion: palm orientation, index-finger direction, and gaze steering. User feedback showed that these techniques are comparable to conventional interfaces, such as controllers, in terms of preference. Additionally, Cutler et al. [44] shows that bimanual direct manipulation on the Responsive Workbench substantially outperforms unimanual alternatives, anticipating the bimanual interaction techniques prevalent in consumer HMDs.
- **Ergonomics and the gorilla-arm problem**: A persistent concern in VR/AR is the so-called "gorilla arm" effect that arises when users hold their arms aloft to perform mid-air gestures. Authors in [88] introduce the "Consumed Endurance" metric to quantify shoulder fatigue and show that subtle changes in elbow elevation drastically extend the comfortable interaction time. Consumed Endurance is widely used in evaluations of new gesture interfaces and motivates a class of micro-gesture interaction styles performed close to the body (e.g., [37,214]).
- **Multimodal gesture-gaze coupling**: An active research trend involves the combination of gestures with gaze. In particular, authors in [173] introduce the gaze + pinch paradigm, in which gaze performs targeting and a discrete pinch gesture confirms the action; the combination greatly mitigates the Heisenberg effect and the precision cost of mid-air pointing, and is the default interaction style of products, such as the Apple Vision Pro.

8.2. Gesture-based interaction in HRI

In the context of HRI, gestures serve a dual purpose, comprising both explicit commands issued to a robot and implicit social cues that ground joint activity. Indicative gesture-based HRI paradigms are described below:

- **Industrial and collaborative robotics**: Authors in [136] summarize gesture-based HRI in manufacturing and articulate the trade-off between static-pose commands (low cognitive load, reliable recognition) and dynamic-trajectory gestures (greater expressivity, higher recognition cost and latency). Furthermore, authors in [201] study the complementary problem of robot-emitted instructional gestures, identifying the hand configurations that humans most readily decode. Authors in [152] propose static sign language gestures for defining the trajectory of a robotic arm in a teach-by-demonstration scheme.
- **Egocentric and shared-perspective gestures**: Other approaches propose on-board egocentric HGR systems to facilitate collaboration. In particular, human-robot handovers [215] show that the success of these short interactions depends less on grip force or trajectory accuracy than on the timely interpretation of subtle hand cues observed from the robot's first-person viewpoint. From the recognition side, egocentric datasets and pipelines, such as EgoHands [22] and recent in-the-wild hand-object reconstruction methods [36], provide the visual foundation that these scenarios require.
- **Legibility and social acceptability**: Recent findings suggest that gesture sets used in public-facing HRI should map onto culturally established symbolic actions, both to ease user learning and to make the interaction legible to bystanders. In particular, authors in [53] formalize legibility as the property of motion that allows an observer to infer the actor's intent quickly and confidently, explicitly distinguishing it from raw predictability and arguing that legibility drives the perceived naturalness of HRI. In parallel, the social acceptability literature [4,118,190] consistently shows that users self-censor gestures they perceive as conspicuous, regardless of how accurately the system recognizes them.

8.3. Gesture-based interaction in HMI

Beyond computers and robots, gesture interfaces increasingly mediate interaction with everyday machines, including vehicles, smart appliances, Internet of Things (IoT), public displays and wearable devices. In this context, the main gesture-based interaction paradigms are analyzed below:

- **Automotive in-cabin interaction**: In human-car interaction, mid-air gestures enable drivers to control infotainment, increasing safety and convenience. In particular, authors in [151] report that gesture commands combined with auditory feedback reduce visual demand compared with touch screens, while authors in [174] characterize the practical recognition challenges (variable lighting, partial occlusion, restricted gesture amplitudes) that in-cabin VHGR systems must address. Various datasets have been proposed for in-vehicle gesture interaction, including Briareo [146], nvGesture [160], and EgoDriving [178].
- **Smart-home, IoT and public displays**: For consumer electronics and smart-home control, works on imaginary interfaces [77], Skinput [83] and ShoeSense [20] show that gestures can be sensed on, around and through the body without dedicated displays, enabling pervasive control of IoT devices. For public displays, authors in [224] study how passers-by discover that mid-air gestures are accepted, identifying the gesture immediacy problem and showing that visible cues can dramatically increase the rate of spontaneous interaction.
- **Wearable computing**: Authors in [37,214] demonstrate that subtle, unimanual micro-gestures performed in the user's resting hand posture are well suited to socially acceptable always-on interaction.
- **Discoverability and feedback**: Across HMI scenarios, mid-air gestures suffer from being invisible affordances. Strategies developed, such as dynamic guides [23] and progressive disclosure, are increasingly adopted as overlays in AR/HMI to scaffold gesture learning.

9. Datasets

The performance and generalizability of gesture recognition models are heavily influenced by the quality, diversity, and scale of the datasets used during training and evaluation. Over the years, a wide variety of datasets have been developed to support different HGR tasks, ranging from static and isolated gestures to continuous sign language recognition in real-world settings. These datasets vary widely in terms of the number of gestures or glosses, the types of modalities captured (e.g., RGB, depth, pose, IR), the presence of signer diversity, and the complexity of the sequences. In the following subsections, we categorize and review the most prominent datasets according to the specific recognition task they target: (i) Isolated and Continuous Sign Language Recognition (IDSLR and CSLR, respectively), (ii) Semaphoric Gestures (both static and dynamic), and (iii) Other Gesture Recognition tasks. This organization aims to provide a clear overview of the resources available for each task type and facilitate their appropriate use in future research. Indicative RGB samples from selected representative datasets are shown in Fig. 10, covering sign language recognition and semaphoric HGR datasets. Some frames have been slightly cropped for visualization purposes.

9.1. Isolated and continuous SLR

The datasets that have been developed to support video-based Sign Language Recognition (SLR) can be broadly categorized into those designed for IDSLR and CSLR. The former datasets typically focus on individual signs, recorded in controlled lab environments with a fixed vocabulary, making them well-suited for classification-based models. In contrast, continuous datasets present real-world sequences of signs, often extracted from broadcast material or spontaneous signing, requiring temporal segmentation and sequence modeling approaches. Additionally, some datasets support Sign Language Translation (SLT) by

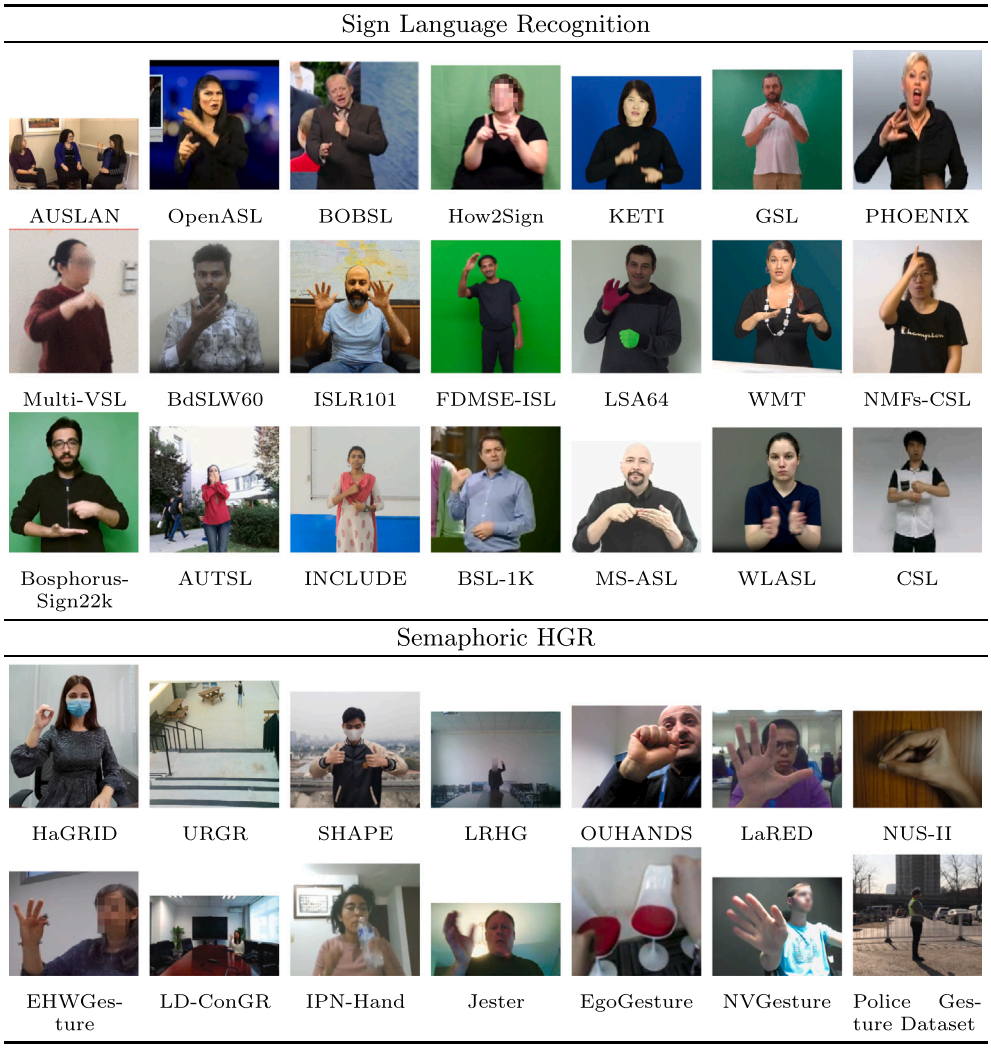


Fig. 10. Indicative RGB samples from selected representative datasets for sign language recognition and semaphoric HGR.

providing sentence-level gloss annotations and parallel spoken language transcriptions.

The RWTH-PHOENIX-Weather 2014 [60] and its extended version PHOENIX-2014-T1 [34] are benchmark datasets offering sentence-level annotations in German Sign Language (DGS) captured from television weather forecasts. A semi-automated annotation protocol was proposed, leveraging existing methods for speech recognition and accelerating the process. The signers perform gestures in a constrained environment (i.e., wearing dark clothes in front of a static gray background), probably limiting the generalization of models trained on it. On the contrary, more recent efforts such as BOBSL [13] and OpenASL [203] have significantly expanded the signer base and vocabulary, with hours of British (BSL) and American Sign Language (ASL) footage collected from TV and web sources, respectively. The more unconstrained environment and the increased number of signers improve the real-world performance of HGR models. The lab-recorded KETI [117] contains samples captured from two camera angles (i.e., front and side), increasing diversity. The aim of the dataset is to interpret Korean sign language gestures that are performed in emergency situations using skeleton and RGB data, reducing the need for professionals. Another multimodal dataset collected in a laboratory is the How2Sign [55] for ASL recognition and translation. The entire dataset was captured in the so-called Green Screen studio with a single RGB camera and depth sensor, while a subset of it was captured in the Panoptic studio using multiple cameras from various viewpoints, enabling accurate 3D pose estimation. Pose data and text transcripts are also considered in TVB-HKSL-News [166]. The samples were collected from TV news with subtitles and Optical Character Recognition was applied to extract the text sequence, while an off-the-shelf pose estimator was utilized to obtain the human keypoints. The RGB-based CSL-Daily [257] also provides glosses and textual transcriptions, aiding both CSLR and SLT tasks. A sign dictionary (referred to as SignDict) was also constructed to be used for other tasks, including sign spotting and segmentation. The Auslan Daily Dataset [202] consists of Australian Sign Language (AUSLAN) gestures performed by humans in a television series, TV news and web videos. These samples may contain multiple persons interacting with each other, which makes the dataset suitable for in-the-wild scenarios.

Other datasets for CSLR do not provide text translations. ChicagoFSWild+ [204] was collected from web videos, including ASL signs performed by 260 signers making this dataset the largest in terms of number of subjects. HKSL [260] was captured in a laboratory environment, providing RGBD data.

Datasets for IDSLR are larger than datasets for CSLR, as they do not require complex annotation, making data collection easier. Many of these datasets, such as CSL [102], GSL iso. Adaloglou et al. [2], BosphorusSign22k [265], AUTSL [208], KArSL [207] and FDMSE-ISL [172], include pose or depth modalities, enriching multimodal learning. All of them, except for the AUTSL, were collected in a laboratory

Table 9
Overview of the characteristics of each dataset related to SLU. The sign vocabulary (Sign Voc.) corresponds to the number of distinct classes included in each dataset.

Name	Year	Task	Modality	Samples	Signers	Sign Voc.	Language	Duration	Source
TVB-HKSL-News [166]	2024	CSLR/SLT	RGB, Pose	7000	2	6515	HKSL	16.07h	TV
Auslan-Daily [202]	2023	CSLR/SLT	RGB	25,000	67	–	AUSLAN	–	TV, Web
OpenASL [203]	2022	CSLR/SLT	RGB	98,417	220	–	ASL	288h	Web
CSL-Daily [257]	2021	CSLR/SLT	RGB	20,654	10	2000	CSL	23.27h	Lab
BOBSL [13]	2021	CSLR/SLT	RGB	1962	39	2281	BSL	1467h	TV
How2Sign (green studio) [55]	2021	CSLR/SLT	RGBD, Pose	2529	11	–	ASL	79.12h	Lab
Corpus FinSL [197]	2020	CSLR/SLT	RGB	343	21	–	FinSL	14.37h	–
KETI [117]	2019	CSLR/SLT	RGB, Pose	14,672	14	524	KSL	27.99h	Lab
GSL SD/SI [2]	2019	CSLR/SLT	RGBD	10,295	7	310	GSL	9.59h	Lab
PHOENIX-2014-T1 [34]	2018	CSLR/SLT	RGB	8257	9	1066	DGS	–	TV
PHOENIX-2014 [60]	2014	CSLR/SLT	RGB	–	9	1558	DGS	10.73h	TV
HKSL [260]	2021	CSLR	RGBD	2400	6	55	HKSL	–	Lab
ChicagoFSWild+ [204]	2019	CSLR	RGB	55,232	260	26	ASL	–	Web
Multi-VSL [51]	2025	IDSLR	RGB, Pose	84,764	30	1000	Vietnamese	–	Lab
BdSLW60 [193]	2025	IDSLR	RGB	9307	18	60	Bangla	–	Lab
ISLR101 [183]	2025	IDSLR	RGB, Pose	4614	10	101	Iranian	–	–
ISW-1000 [189]	2025	IDSLR	RGB, Pose, OF	10,000	10	1000	CSL	–	Lab
FDMSE-ISL [172]	2024	IDSLR	RGBD	40,033	20	2002	ISL	36h	Lab
LSA64 [192]	2023	IDSLR	RGB	3200	10	64	LSA	1.9h	Lab
Signsuisse Lexicon [162]	2023	IDSLR	RGB	18,221	–	–	3 (Mult)	–	Web
CSLD [229]	2022	IDSLR	RGB	120,000	30	400	CSL	66.8h	Lab
NCSL [228]	2022	IDSLR	RGB	90,000	30	300	CSL	50.1h	Lab
KArSL [207]	2021	IDSLR	RGBD, Pose	75,300	3	502	ArSL	–	Lab
NMFs-CSL [95]	2021	IDSLR	RGB	32,010	10	1.067	CSL	24.4h	Lab
BosphorusSign22k [265]	2020	IDSLR	RGBD, Pose	22,542	6	744	TSL	19h	Lab
AUTSL [208]	2020	IDSLR	RGBD, Pose	38,336	43	226	TSL	21h	Lab
INCLUDE [213]	2020	IDSLR	RGB	4287	7	263	ISL	3h	Lab
IISL2020 [120]	2020	IDSLR	RGB	12,100	16	11	ISL	7h	Lab
RKS-PERSIANSIGN [185]	2020	IDSLR	RGB	10,000	10	100	Persian	–	Lab
BSL-1K [12]	2020	IDSLR	RGB	280,000	40	1064	BSL	1060h	TV
MSASL [109]	2019	IDSLR	RGB	25,513	222	1000	ASL	24.65h	Web
WLASL [127]	2019	IDSLR	RGB	21,083	119	2000	ASL	14.11h	Web
GSL isol. [2]	2019	IDSLR	RGBD	40,785	7	310	GSL	6.44h	Lab
SMILE [56]	2018	IDSLR	RGB	–	30	100	DSGS	–	Lab
ASL Dataset [19]	2018	IDSLR	Pose	1200	20	30	ASL	–	Lab
CSL [102]	2018	IDSLR	RGBD, Pose	125,000	50	500	CSL	–	Lab

with fixed background. Datasets that do not include such modalities can be extended to them by using off-the-shelf DL pose and depth estimators. The ASL Dataset [19] is the only one consisting of pose sequences exclusively. MS-ASL [109] and WLASL [127] are two RGB-based datasets for isolated ASL recognition collected from publicly available web videos of the deaf community, including educational resources and ASL tutorials and performed by 222 and 119 subjects respectively, increasing the diversity. However, they contain relatively few samples compared to large-scale datasets such as BSL-1K [12], CSLD [229] and NCSL [228]. For isolated Irish Sign Language (ISL) recognition, only laboratory datasets have been proposed (e.g., INCLUDE [213], IISL2020 [120] and FDMSE-ISL [172]), making the collection of an in-the-wild dataset necessary. For Vietnamese SLR, authors in [51] collected a large-scale multi-view dataset, considering three different points of view, enhancing diversity. Other sign languages have also been considered including Argentinian (e.g., LSA64 [192]), Persian (e.g., RKS-PERSIANSIGN [185]) and Iranian (e.g., ISLR101 [183]). The rarity of some sign languages, such as those mentioned above, poses obstacles to finding relevant annotators, exacerbating the data scarcity challenge, as it will be discussed later (Table 9).

9.2. Semaphoric HGR

Semaphoric HGR research is supported by a diverse range of datasets categorized into three main types based on task formulation: SHGR, IDHGR and CDHGR. These datasets vary in their application domains (e.g., HCI, HMI, HRI), modalities (e.g., RGB, depth, pose, infrared), and in the number of gesture classes, subjects, and gesture samples they include. The statistics of datasets for semaphoric HGR are provided in Tables 10–12, regarding the task for which they are used (i.e., static, isolated dynamic and continuous HGR, respectively).

SHGR datasets target gesture recognition from single frames, making them easier to process due to the lack of temporal complexity. Older datasets like HGR-1 [114], NUS-I [123], NUS-II [175], and OUHANDS [150] contain few samples, limiting deep learning training, yet remain benchmarks for HCI due to their simplicity. LaRED [89] offers 243K RGBD samples across 81 gestures, but only from 10 subjects, restricting subject-independent recognition. HaGRID [111] is much larger and diverse, with nearly twice as many full-resolution RGB frames from 35K subjects in varied scenes, including bounding boxes for hand detection. HaGRIDv2 [167] expands this with 15 additional classes, 66K subjects, and over 1M samples. SHAPE [45] has a similar number of gestures (all two-handed) but fewer samples. LRHG [259] and MD-UHGRD [240] target UAV-based gesture recognition, with MD-UHGRD offering more samples but fewer classes, captured from varying distances and angles to enhance viewpoint robustness. URGR [21] focuses on ultra-range gestures for UGVs, with 6 robot-command gestures performed by 16 signers at distances from 2 to 25 meters.

IDHGR datasets contain gestures performed in isolation, with start and end points explicitly segmented. Most of the existing datasets (e.g., SKIG [139], NVIDIA (aka nvGesture) [160], DHG [49] and SHREC17 [210]) for IDHGR are relatively small with a few thousand samples and support multimodal settings by providing depth maps, IR frames, optical flow or skeleton joints, in addition to the RGB modality, aiming to enhance the accuracy of the models. Jester-20bn [149], proposed for gesture-based HCI, is the largest dataset for (non-SLU) IDHGR containing 150K videos performed by 1376 subjects. The remarkable

Table 10
Datasets for semaphoric SHGR.

Name	Year	Domain	Modality	Samples	Subjects	Classes	Resolution	Distance
HaGRID [111]	2024	HCI	RGB	552,992	34,730	19	1920 × 1080	0.5-4m
HaGRIDv2 [167]	2024	HCI	RGB	1,086,158	65,977	34	1920 × 1080	0.5-4m
MD-UHGRD [240]	2024	HRI	RGB	20,000	–	19	–	3-10m
URGR [21]	2024	HRI	RGB	347,483	16	6	640 × 480	1-25m
SHAPE [45]	2022	–	RGB	34,226	20	32	4128 × 3096	2-5m
LRHG [259]	2021	HRI	RGB	4320	8	10	640 × 480	1-7m
OUHANDS [150]	2016	HCI	RGBD	3150	23	10	640 × 480	≤ 1m
HGR-1 [114]	2014	–	RGB	899	12	27	174 × 131–640 × 480	≤ 1m
LaRED [89]	2014	–	RGBD	243,000	10	81	640 × 480	–
NUS-II [175]	2013	–	RGB	2000	40	10	160 × 120	≤ 1m
ASL-FS [177]	2011	–	RGBD	48K	4	24	–	–
NUS-I [123]	2010	–	RGB	240	–	10	160 × 120	≤ 1m

Table 11
Datasets for semaphoric IDHGR.

Name	Year	Domain	Modality	Samples	Subjects	Classes	Frames
EHWGesture [17]	2025	–	RGBD	1100	25	11	3.3M
EgoDriving [178]	2024	HMI	RGB	400	–	29	70,276
DATE [47]	2024	HCI	RGB	13,500	22	27	–
ZJUGesture (isol.) [239]	2023	HCI	RGB	9892	60	9	–
LD-ConGR (isol.) [135]	2022	HCI	RGBD	44,887	30	10	–
IPN Hand (isol.) [25]	2021	HCI	RGB	4218	50	13	–
TCG [235]	2020	HMI	Pose	250	5	4	839,350
Jester [149]	2019	HCI	RGB	148,092	1376	27	5,331,312
Briareo [146]	2019	HMI	RGBD, Pose, IR	1440	40	12	–
EgoGesture (isol.) [246]	2018	HCI	RGBD	24,161	50	83	2,953,224
SHREC17-14/28 [210]	2017	–	Depth, Pose	2800	28	14/28	–
nvGesture [160]	2016	HMI	RGBD, IR, OF	1532	20	25	–
ChaLearn IsoGD [225]	2016	–	RGBD	47,933	21	249	–
DHG-14/28 [49]	2016	–	Depth, Pose	2800	20	14/28	–
SKIG [139]	2013	–	RGBD	1080	6	10	–

Table 12
Datasets for semaphoric CDHGR.

Name	Year	Domain	Modality	Samples	Instances	Subjects	Classes
ZJUGesture (cont.) [239]	2023	HCI	RGB	–	9892	60	9
LD-ConGR (cont.) [135]	2022	HCI	RGBD	542	44,887	30	10
IPN Hand (cont.) [25]	2021	HCI	RGB	200	4218	50	13
Police Gesture Dataset [86]	2020	HMI	RGB	21	3354	–	8
EgoGesture (cont.) [246]	2018	HCI	RGBD	2081	24,161	50	83
ChaLearn ConGD [225]	2016	–	RGBD	22,535	47,933	21	249

diversity and size of this dataset make it ideal for pretraining DL models. Other datasets for HCI include IPN-Hand [25], ZJUGesture [239] and DATE [47], comprising 4218, 9892 and 13,500 samples respectively. LD-ConGR [135], ChaLearn IsoGD [225] and EgoGesture [246] contain a sufficient amount of data, constituting adequate baselines for IDHGR. EgoDriving [178], Briareo [146] and nvGesture [160] have been proposed for human-car interaction, mainly in a multimodal setting, including gestures that correspond to car functionalities (e.g., CW rotation and audio manipulation[2]). TCG [235] was also proposed for human-car interaction, but for recognizing traffic police gestures in autonomous driving scenarios relying only on pose data. EHWGesture [17] is the first dataset for recognizing clinical gestures and providing annotations for many other tasks, including gesture detection & triggering.

CDHGR datasets challenge models to identify and segment gestures from continuous video streams. Many of the datasets mentioned earlier (e.g., ChaLearn [225], IPN-Hand [25], LD-ConGR [135], EgoGesture [246] and ZJUGesture [239]) have been obtained by annotating the temporal boundaries of gestures within continuous sequences; thus these datasets have protocols for both IDHGR and CDHGR. Therefore, we will only discuss the Police Gesture Dataset [86] as the previous analysis could also be applied here. This dataset was introduced for recognizing police traffic gestures, similarly to TCG [235], containing a small number of samples (21 RGB videos) and 3354 gesture instances.

Each dataset supports specific sub-tasks within HGR research: static datasets aid in gesture classification, isolated dynamic datasets support temporal gesture modeling and classification, and continuous datasets allow for joint gesture spotting and classification—crucial for real-time and interactive applications. This diversity in dataset structure, scale, and modality reflects the evolving demands and challenges of semaphoric gesture recognition across interactive domains.

9.3. Other

In addition to standard semaphoric gesture datasets, several specialized datasets have been introduced to support research on more specific gesture types. The First-Person Hand Action (FPHA) dataset [66], published in 2018, contains 1175 video sequences captured from a first-person perspective by six participants. It focuses on isolated manipulative hand gestures across 45 action classes, offering RGB, depth, and skeleton data, and is particularly valuable for studies involving egocentric interaction. Another notable example is the DP Dataset [163], released in 2023, which, to the best of our knowledge, is the only

[2] https://youtu.be/NJmk1DUyyB8.

publicly available benchmark specifically designed for pointing gesture recognition. Collected from 33 participants in two indoor environments, the dataset comprises approximately 2.8 million frames and 6335 annotated gesture instances. It includes detailed annotations such as gesture timings and pointing directions, enabling both binary gesture classification (gesture vs. non-gesture) and directional estimation tasks. Finally, the SocialGesture [35] focuses on the recognition of natural gestures in multi-person environments. It comprises 9889 in-the-wild videos containing 42,533 instances and provides fine-grained annotations including bounding boxes, relations (for example, a subject is pointing to a target) and Visual Question Answering (VQA) details.

9.4. Datasets comparison

After examining the existing datasets for VHGR with significant impact on the current research, in this subsection we aim to determine pros and cons of the most important ones as well as the use cases that they fit. For gesture-based HCI, HaGRID [111] and its versions [167] seem to be a reliable solution, containing an enormous number of samples corresponding to 19 distinct classes (that can be mapped to computer commands). Given that these gestures are also static, even simple architectures - like the ones discussed in Sections 4 & 7 - can achieve sufficient performance. Furthermore, this dataset could also be used for pretraining, enhancing the backbone's expression capability. However, its size implies that training takes a long time, which constitutes a deterrent in some cases, compared to other datasets for HCI like OUHANDS [150]. For gesture-based HRI, URGR [21] is the only one for UGV guidance and considers various recognition ranges, which may be crucial for real-world scenarios. Even though it contains a large number of samples, the small number of classes may limit the commands given by gestures and, therefore, the functionalities of the entire gesture-based interface. Regarding SLU datasets, for ASL recognition, OpenASL [203] is the largest one, including almost 100K RGB videos (288h in total) collected from the web. MSASL [109] comprises almost the same number of subjects, but it is much smaller. Furthermore, OpenASL was proposed for CSLR whereas MSASL was proposed for IDSLR, thus it may be more suitable for real-world use cases, where the temporal boundaries are not known. BSL-1K [12] is also a very large dataset, proposed for British IDSLR and comprising the largest number of samples (280K, 1060h in total). BOBSL [13] was proposed for British CSLR and comprises videos of 1467h in total making it the longest dataset in terms of duration. Phoenix [60] and Phoenix T [34] remain the most suitable datasets for German sign language, covering a large sign vocabulary. Nevertheless, the small number of samples and the low diversity - due to the limited number of signers and the constrained environment - make the collection of a new, larger and more diverse dataset for DGS. For Chinese SLR, NMFs-CSL [95] is significantly smaller than CSLD [229], CSL [102] and NCSL [228] but includes much more signs (1064 instead of less than 500). CSL-Daily [257] remains the only one for Chinese CSLR, covering a large sign vocabulary of 2000 signs in total.

10. Evaluation metrics

In this section, we present the key evaluation metrics adopted in the literature for measuring the performance of HGR methods. Broadly, metrics fall into two main categories: *classification metrics* (used in static and isolated dynamic HGR) and *sequence-to-sequence metrics* (used in continuous dynamic HGR). Additionally, other indicators such as computational resource usage and overfitting-related metrics are also considered.

10.1. Classification metrics

In multi-class HGR problems—especially when dealing with class imbalance—researchers often report both **micro-averaged** (per-instance) and **macro-averaged** (per-class) versions of common classification metrics. Micro-averaged metrics aggregate contributions from all instances across classes, thus being sensitive to class imbalance. Macro-averaged metrics compute the metric independently for each class and then average, treating all classes equally regardless of their frequency. The formulae for each metric can be found in Table 13.

Table 13
Commonly utilized metrics for the evaluation of hand gesture classification. Here, TP_i, FP_i, TN_i, and FN_i denote the number of true positives, false positives, true negatives, and false negatives for the i-th class, respectively, and N is the total number of classes.

Metrics	Micro-averaged formulae	Macro-averaged formulae
Accuracy	$\frac{\sum_{i=1}^{N}(TP_i+TN_i)}{\sum_{i=1}^{N}(TP_i+TN_i+FP_i+FN_i)}$	$\frac{1}{N}\sum_{i=1}^{N}\frac{TP_i+TN_i}{TP_i+FP_i+TN_i+FN_i}$
Precision	$\frac{\sum_{i=1}^{N}TP_i}{\sum_{i=1}^{N}(TP_i+FP_i)}$	$\frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FP_i}$
Recall	$\frac{\sum_{i=1}^{N}TP_i}{\sum_{i=1}^{N}(TP_i+FN_i)}$	$\frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FN_i}$
Jaccard-Index	$\frac{\sum_{i=1}^{N}TP_i}{\sum_{i=1}^{N}(TP_i+FP_i+FN_i)}$	$\frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FP_i+FN_i}$
F1-Score	$2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$	

10.2. Sequence-to-sequence metrics

For continuous HGR (CDHGR), particularly in Continuous Sign Language Recognition (CSLR), sequence-level metrics are used. **Word Error Rate (WER)** [2,157] is the standard metric, measuring the minimum number of substitutions (S), deletions (D), and insertions (I) needed to transform the predicted sequence into the ground truth sequence of length N:

$$WER = \frac{S + D + I}{N} \tag{1}$$

Edit Distance Accuracy [61,230] is a complementary metric representing the ratio of correctly predicted elements in the sequence:

$$\text{Edit-Accuracy} = \frac{N - S - D - I}{N} \tag{2}$$

These metrics provide insight not only into classification accuracy but also into temporal and sequential consistency.

10.3. Computational resource metrics

In addition to recognition performance, computational efficiency is often evaluated, especially for real-time or embedded HGR applications. Common metrics include:

- **Inference Time:** Time (in ms or seconds) taken to process one input sample.
- **Frames Per Second (FPS):** Number of frames processed per second; higher values indicate faster processing.
- **Model Size in Megabytes (MBs):** The total memory footprint of the model.
- **Number of Parameters:** The number of trainable parameters.
- **Floating Point Operations (FLOPs):** Measures the number of numerical operations required to run a model (either during training or inference time). Being independent of the hardware utilized, it can be considered—along with the number of parameters—more reliable.

These metrics are crucial when deploying HGR systems on edge devices or in latency-sensitive environments.

11. Current challenges and future research directions

As mentioned in the Methods sections (Sections 4–6), researchers have developed sophisticated techniques and large-scale datasets to achieve satisfactory performance. However, as shown in the relevant tables (Tables 3, 5 and 7), recognition accuracy remains low in many cases—particularly in more complex tasks such as CSLR. In this section,

we discuss the current challenges and limitations that hinder efficient VHGR, posing ongoing concerns for the research community and shaping future research directions. Additionally, we map the main bottlenecks to existing mitigation strategies and emerging research opportunities, aiming to provide a more forward-looking research route for the field. These observations address research question Q4 (see Section 1).

11.1. Gesture-irrelevant factors

11.1.1. Current challenges

The presence of gesture-irrelevant factors (or redundant information) hinders accurate recognition in diverse real-world scenarios. Some of these factors include the following [10,21,104,181,216,254]:

- **Environmental factors:** Cluttered backgrounds, varying illumination conditions, and self-occlusions make it difficult or even impossible to automatically spot the hand region, which is an important step for efficient HGR.
- **Subject characteristics:** Subject independence—i.e., the ability to correctly identify a gesture regardless of the signer's characteristics such as skin color, hand shape variations, or clothing—is a key prerequisite for achieving sufficient generalization and enabling real-world deployment.
- **Variations in perception device position:** The distance from the camera and the viewing angle may reduce the recognition performance.
- **Input data quality:** Low resolution and blurriness may corrupt the discriminative characteristics of the hand gesture region.

11.1.2. Mitigation strategies & future directions

Aiming to address these factors, researchers could apply various methods to eliminate their interference:

- **Hand cropping:** Authors often deploy palm segmentation to isolate the hand region from the background [46,196,256].
- **Spatial normalization:** Object detectors are applied to strengthen scale invariance, especially in long range recognition scenarios [10, 21,24].
- **Super resolution:** Super resolution methods can enhance image quality [21] to reduce the impact of low resolution and motion blur.
- **Attention mechanisms:** They force the visual encoders to focus on the most informative regions [254].
- **Other:** Information-theoretic losses [128] and feature selection techniques [218] have been proposed to reduce information redundancy.

However, many of these approaches rely on computationally expensive off-the-shelf models (e.g., segmentation models and object detectors), which increase the complexity of the overall HGR pipelines and set an upper bound on their accuracy. Further, the efficacy of custom attention modules relies on the available data. Therefore, mitigating these challenges still remains a promising field for future work; the following could be considered as potential solutions:

- **Explainable Artificial Intelligence (XAI):** XAI [191] could be adopted to better understand the behavior of the networks. For instance, Grad-CAM could be further utilized to determine the image region that most affects the predicted gesture class (as has already been done by a few researchers: [16,99]). Such results may indicate the effectiveness of attention mechanisms or other related modules embedded in the network in emphasizing the most important parts of the input image. t-SNE visualizations could also be leveraged to evaluate the quality of the extracted features, and in particular, how much they are influenced by the presence of redundant information. Thus, guided by XAI tools or similar frameworks, the interference of gesture-irrelevant factors could be mitigated through the careful design of the future architectures.
- **Specialized data augmentation and collection:** Researchers could consider applying data augmentation methods specifically to mitigate these issues. Further, the data collection should be driven by the edge cases; it is essential to include many environments, distances, viewing angles and image qualities, to make the models invariant and robust to these conditions. In other words, the data should be captured in the wild. For example, authors could explore the potential of using internet data.

11.2. Computational cost

11.2.1. Current challenges

Even lightweight methods for SHGR and IDHGR achieve high accuracy; however, for the more complex task of CDHGR, the cost is a critical factor and can generally be decomposed into two parts:

- **Floating Point Operations (FLOPs):** The RGB-based CSLR methods listed in Table 7, which generally outperform pose-only methods on the reported benchmarks, require more than 366.3 GFLOPs ([241]) and reach up to 2950.5 GFLOPs ([98]). This issue prevents not only the deployment of such methods but also the efficient training, which takes longer. Pose-based methods [107] are significantly lighter, but the pose estimation cost is not taken into account.
- **Number of parameters:** The number of parameters does not change dramatically between IDHGR and CDHGR methods, while many SHGR methods require more parameters (e.g., Gesture-CNN [200] has 67.25M parameters). Yet, recent approaches use large-scale generative models (e.g., Uni-Sign [130] leverages an LLM and comprises 592.1M parameters in total).

The need for large computational resources prevents not only the real-time and commercial deployment of VHGR systems, but also poses barriers to the evolution of the field, as the conducted experiments take many hours or even days to complete.

11.2.2. Mitigation strategies and future directions

To reduce the computational complexity of the models, researchers can consider the following:

- **Non-parametric layers:** Zero-parameter layers can reduce the computational cost without affecting the capacity, as shown in [67] and [229], where attention was replaced with pooling layers and a non-parametric temporal modeling layer, respectively.
- **Smaller architectures:** Minimizing the network size while maintaining accuracy has been studied by some researchers [245].
- **Model compression:** DL model compression [147] has not yet been applied to VHGR methods despite rapid evolution. Therefore, relevant techniques such as pruning (i.e., removing less important weights) and quantization (i.e., reducing the arithmetic precision of the parameters), may constitute future research.

11.3. Data scarcity

11.3.1. Current challenges

The inadequacy of most of the existing datasets can be broken down as:

- **Limited diversity:** Even though there are datasets that contain a sufficient number of samples, they lack diversity. In particular, among SHGR, only the versions of HaGRID [111,167] have been collected from more than one hundred participants, although some others also contain a plausible number of samples [21,89]. A similar situation is observed in IDHGR datasets, where only the Jester dataset [149] includes data from over one hundred subjects.
- **Laboratory samples:** Datasets are still often captured in constrained environments [2,117,172,192,265], potentially degrading in-the-wild recognition performance.

- **Data imbalance:** The number of samples for specific modalities, such as IR data, is limited compared with the size of RGB-based datasets [227]. Furthermore, most of the datasets for SLU consider more widespread sign languages, so there are only a few datasets for rare ones.

The lack of large-scale, diverse datasets leads to overfitting and limited generalization, which is closely related to the challenges discussed earlier. This issue is further exacerbated in tasks that involve costly annotation processes and require data-intensive deep learning models with a large number of parameters, such as CSLR.

11.3.2. Mitigation strategies and future directions

Several strategies have been proposed to mitigate data scarcity, while other promising directions remain underexplored:

- **Multilingual Data Aggregation:** Combining datasets (e.g., CSL and DGS) to exploit cross-lingual visual similarities increases the overall amount of training data [234].
- **Self-Supervised Learning:** In the context of the self-supervised learning setting [119], gloss-to-gloss translation tasks have been used for parameter initialization [253], providing linguistic priors to sign language models. Pretraining with pose reconstruction has also been investigated to incorporate prior knowledge of hand structure [92,93,103].
- **Advanced Data Augmentation:** Adversarial learning has been employed to optimize augmentation hyperparameters [164], making the models robust to spatial deformations.
- **Few-/Zero-shot Learning**: Approaches using only a small number of samples to correctly recognize gestures—often based on CLIP—have also been proposed [28,29,186,187,264].
- **Generative AI:** Diffusion models (DMs) for image augmentation [14] could help address data scarcity in HGR; however they have not yet been examined. DMs can increase dataset diversity by generating additional samples with varied scenes, subjects, and lighting, or support specialized domains such as rare sign languages and human-robot interaction in firefighting or military scenarios. Thus, leveraging DMs represents a promising direction for future work.

11.4. Optimization difficulties

11.4.1. Current challenges

Training the visual encoder effectively remains difficult, especially in deep architectures for CSLR:

- **Gradient Attenuation:** As shown in CTCA [74] (Fig. 11), increasing the depth of the temporal module (e.g., 1D-TCN followed by BiLSTM) improves its generalization but hampers the encoder's learning due to gradient attenuation during backpropagation.
- **CTC Spike Phenomenon:** The model often over-focuses on a few keyframes causing rapid overfitting and limiting encoder training [81,157].

11.4.2. Mitigation strategies and future directions

To address these issues, several strategies have been proposed and could be expanded in the future:

- **Auxiliary Supervision:** Adding early-stage classifiers to the encoder or short-term module has been proposed as a strategy to guide training by providing additional supervision [97,99,142,157,184,261].
- **Temporal Module Redesign:** Placing the 1D-TCN and BiLSTM in parallel instead of sequentially [74].
- **Linguistic Constraints and Gloss Segmentation:** Using gloss boundaries [81] or prior knowledge [75] to regularize learning.

The aforementioned approaches – particularly auxiliary supervision [157] – have served as the foundation for many existing methods.

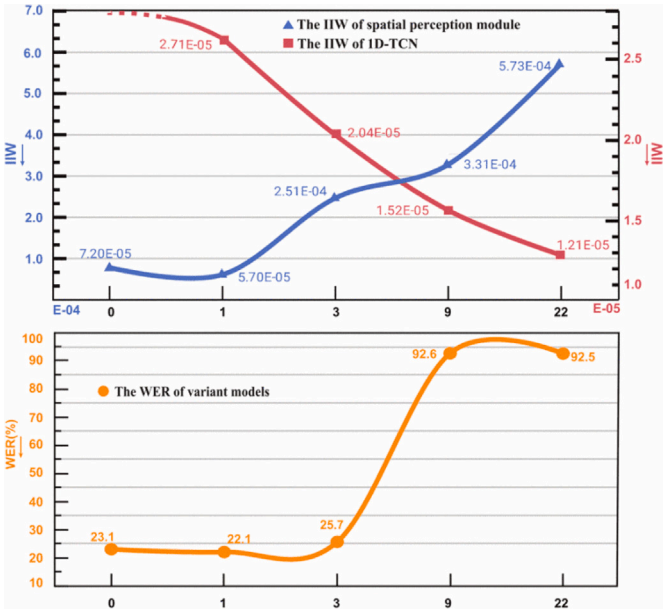


Fig. 11. Experimental results from Guo et al. [74] showing the effect of the depth of the temporal aggregation module—a 1D-TCN with a varying number of layers followed by a two-layer BiLSTM—on the training of the visual encoder (spatial perception module). In both plots, the x-axis indicates the number of 1D-TCN layers. The top plot shows the information stored in weights (IIW) [232], which reflects generalization ability, while the bottom plot reports the WER. Increasing the number of 1D-TCN layers improves short-term temporal generalization but degrades overall performance because the visual encoder becomes undertrained.

However, the increasing complexity of modern deep architectures often exceeds the effectiveness of these techniques. As a result, optimization remains a significant challenge – one that may be further exacerbated by the continued development of even deeper models.

11.5. Limitations of architectures

11.5.1. Current challenges

DL methods for HGR have increased the recognition performance. However, the existing architectures suffer from limitations, including the following:

- **Block-fixed models:** The existing architectures are based on handcrafted designs, consisting of blocks of layers with hyperparameters (e.g., number of channels, convolution kernel size) that are defined manually by the researchers. This approach is suboptimal [254], as several other possible configurations have not been explored.
- **Insufficient graph representations of GCNs:** The standard GCNs utilize fixed-topology graphs, often based on the shape of the hand, to better capture the interactions between the various skeleton joints. However, this configuration has been proven to be suboptimal [137], as it is unable to efficiently capture the temporal dynamics.
- **Ineffective multimodal fusion:** Among the existing multimodal methods, most of them perform late fusion to incorporate the information of multiple streams [92,93,105,106,132,250–252]. This scheme does not fully leverage the available modalities, as the features extracted by each stream are isolated from the others during backbone processing. Thus, the obtained vector representations do not benefit from the utilization of multiple modalities.

11.5.2. Mitigation strategies and future directions

To address these limitations, several mitigation strategies and future directions can be identified:

- **Neural Architecture Search (NAS):** NAS can be used to identify optimal model structures by searching over a large set of configurations [254].
- **Adaptive hand topology:** Learnable or time-dependent skeleton adjacency matrices that implicitly determine the optimal way to integrate the features of each joint [39,41,101,137].
- **Multi-level fusion:** Multi-level fusion for dual-stream approaches enhances the quality of the extracted features by effectively incorporating information from various sources [78]. To the best of our knowledge, multi-level and mid-level multimodality fusion have not been investigated for three or more streams. Thus, future work could consider the intermediate fusion in such cases, e.g., modifying existing fusion modules [78,222] and identifying the optimal way to connect the different streams.

11.6. Summary of challenges & research route

In summary, the previous analysis indicates that future VHGR research is shaped by a small set of recurring bottlenecks rather than isolated technical issues. The main research routes can be summarized as follows:

- **Eliminating gesture-irrelevant factors:** Tackling the interference of redundant information, such as cluttered backgrounds, lighting variations, occlusions, and viewpoint changes, remains essential for improving real-world accuracy [21,128,254].
- **Improving data coverage and diversity:** Larger, more diverse, and more realistic datasets are needed, especially for rare sign languages, in-the-wild settings, underrepresented modalities, and subject-independent evaluation. Multilingual aggregation, self-supervised learning, and generative augmentation are promising routes for mitigating data scarcity [14,92,234].
- **Reducing computational cost:** Minimizing computational overhead is crucial for real-time and resource-constrained deployment. Lightweight architectures, non-parametric modules, and model compression provide promising directions for improving deployability [67,147,245].
- **Addressing optimization difficulties:** The CSLR 2DCNN + 1DCNN + BiLSTM pipeline still suffers from inadequate training of the lower parts of the network, necessitating more effective learning strategies, auxiliary supervision, and temporal modeling schemes [74,75,81].
- **Advancing multi-stream fusion:** For architectures involving more than two streams, late fusion remains dominant [105,106]. More sophisticated techniques, such as multi-level and intermediate fusion, could improve cross-modal feature extraction and reduce the redundancy of independent streams.
- **Incorporating foundation models efficiently:** Foundation models can support stronger visual encoding, linguistic priors, and open-vocabulary recognition, but their benefits have not yet been fully exploited in VHGR and currently come with significant computational cost [15,113,130].

12. Conclusions

The undeniable importance of Vision-based Hand Gesture Recognition (VHGR) stems from its wide range of applications—spanning from human-robot interaction to sign language understanding. This broad applicability has made it a vibrant and active research field. Recent advances in deep learning (DL) have played a pivotal role in driving progress, enabling the development of increasingly sophisticated architectures and training strategies aimed at tackling more complex and fine-grained tasks, such as continuous sign language recognition (CSLR).

The current study aims to address four key research questions, as defined in Section 1. Specifically, this survey determines the principal aspects of VHGR by providing a thorough taxonomy in Section 3 (answering Q1) and identifies the current SOTA and key developments for each field following a task-oriented presentation (Sections 4–7), as required by Q2. Additionally, it performs a comparative analysis of the selected publications in the aforementioned sections to address Q3 and highlights challenges and future research directions in Section 11, as requested by Q4. The summaries of datasets (Section 9) and evaluation metrics (Section 10) are considered essential for facilitating the previous analysis.

This study focuses on DL-based VHGR methods. These methods were highlighted not only for their superior performance compared to traditional hand-crafted approaches, but also for their greater potential for deployment in real-world scenarios, unlike more cumbersome sensor-based techniques. From the surveyed literature, it becomes clear that even relatively simple and shallow DL models can achieve high accuracy in static HGR tasks. In such cases, the primary challenges are typical of general computer vision problems and are not inherently tied to HGR itself. For isolated dynamic and continuous VHGR, more advanced approaches have been proposed, aiming to capture more complex context. The complexity of the architectures (especially for CSLR) causes serious optimization difficulties, leading researchers to develop sophisticated training setups.

Nevertheless, the most important challenges regarding VHGR remain unsolved, leaving substantial research gaps; data scarcity—especially for more specialized tasks, such as the recognition of rare sign languages—and high computational cost remain essentially unresolved. Furthermore, the existing methods are limited to a predefined list of gestures, which may not be optimal for real-world interaction with humanoid robots or interpretation of body language. Thus, VHGR could be extended to open-vocabulary settings, enabling deployment in broader applications. The potential of leveraging the extensive knowledge of foundation models [113] in the context of VHGR enabling zero-shot gesture recognition, as well as the utilization of model compression techniques [147] to reduce the computational complexity is also rarely explored, leaving space for future work.

CRediT authorship contribution statement

Konstantinos Foteinos: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation. **Manousos Linardakis:** Methodology, Investigation. **Panagiotis Radoglou-Grammatikis:** Writing – review & editing. **Vasileios Argyriou:** Writing – review & editing. **Panagiotis Sarigiannidis:** Writing – review & editing. **Iraklis Varlamis:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Georgios Th. Papadopoulos:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization.

Funding

The research leading to these results received funding from the European Commission under Grant Agreement No. 101168042 (TRIFFID) and No. 101189557 (TORNADO).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] S.B. Abdullahi, K. Chamnongthai, V. Bolon-Canedo, B. Cancela, Spatial–temporal feature-based end-to-end Fourier network for 3D sign language recognition, Expert Syst. Appl. 248 (2024) 123258, https://doi.org/10.1016/j.eswa.2024.123258, https://www.sciencedirect.com/science/article/pii/S0957417424001234.

[2] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G.T. Papadopoulos, V. Zacharopoulou, G.J. Xydopoulos, K. Atzakas, D. Papazachariou, P. Daras, A

comprehensive study on deep learning-based methods for sign language recognition, IEEE Trans. Multimed. 24 (2022) 1750–1762, https://doi.org/10.1109/TMM.2021.3070438
[3] I.A. Adeyanju, O.O. Bello, M.A. Adegboye, Machine learning methods for sign language recognition: a critical review and analysis, Intell. Syst. Appl. 12 (2021) 200056.
[4] D. Ahlström, K. Hasan, P. Irani, Are you comfortable doing that? Acceptance studies of around-device gestures in and for public settings, in: Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '14), ACM, 2014, pp. 193–202, https://doi.org/10.1145/2628363.2628381
[5] J. Ahn, Y. Jang, J.S. Chung, Slowfast network for continuous sign language recognition, in: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 3920–3924, https://doi.org/10.1109/ICASSP48485.2024.10445841
[6] S. Aich, J. Ruiz-Santaquiteria, Z. Lu, P. Garg, K.J. Joseph, A.F. Garcia, V.N. Balasubramanian, K. Kin, C. Wan, N.C. Camgoz, S. Ma, F. De La Torre, Data-free class-incremental hand gesture recognition, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20901–20910, https://doi.org/10.1109/ICCV51070.2023.01916
[7] B. Al Abdullah, G. Amoudi, H. Alghamdi, Advancements in sign language recognition: a comprehensive review and future prospects, IEEE Access 12 (2024).
[8] M. Al-Qurishi, T. Khalid, R. Souissi, Deep learning for sign language recognition: current techniques, benchmarks, and open issues, IEEE Access 9 (2021) 126917–126951.
[9] M. Alaftekin, I. Pacal, K. Cicek, Real-time sign language recognition based on YOLO algorithm, Neural Comput. Appl. 36 (2024) 7609–7624.
[10] M.M. Alam, M.T. Islam, S.M. Rahman, Unified learning approach for egocentric hand gesture recognition and fingertip detection, Pattern Recognit. 121 (2022) 108200, https://doi.org/10.1016/j.patcog.2021.108200, https://www.sciencedirect.com/science/article/pii/S0031320321003824.
[11] F.S. Alamri, S. Bala Abdullahi, A.R. Khan, T. Saba, Enhanced weak spatial modeling through CNN-based deep sign language skeletal feature transformation, IEEE Access 12 (2024) 77019–77040, https://doi.org/10.1109/ACCESS.2024.3405341
[12] S. Albanie, G. Varol, L. Momeni, T. Afouras, J.S. Chung, N. Fox, A. Zisserman, BSL-1K: scaling up co-articulated sign language recognition using mouthing cues, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, Springer, 2020, pp. 35–53.
[13] S. Albanie, G. Varol, L. Momeni, H. Bull, T. Afouras, H. Chowdhury, N. Fox, B. Woll, R. Cooper, A. McParland, A. Zisserman, BBC-Oxford British sign language dataset, arXiv:2111.03635, 2021.
[14] P. Alimisis, I. Mademlis, P. Radoglou-Grammatikis, P. Sarigiannidis, G.T. Papadopoulos, Advances in diffusion models for image data augmentation: a review of methods, models, evaluation metrics and future research directions, Artif. Intell. Rev. 58 (2025) 1–55.
[15] S. Alyami, H. Luqman, CLIP-SLA: parameter-efficient CLIP adaptation for continuous sign language recognition, arXiv preprint arXiv:2504.01666, 2025.
[16] S. Alyami, H. Luqman, Swin-MSTP: swin transformer with multi-scale temporal perception for continuous sign language recognition, Neurocomputing 617 (2025) 129015.
[17] G. Amprimo, A. Ancilotto, A. Savino, F. Quazzolo, C. Ferraris, G. Olmo, E. Farella, S. Di Carlo, EHWGesture–A dataset for multimodal understanding of clinical gestures, arXiv preprint arXiv:2509.07525, 2025.
[18] F. Argelaguet, C. Andújar, A survey of 3D object selection techniques for virtual environments, Comput. Graph. 37 (2013) 121–136, https://doi.org/10.1016/j.cag.2012.12.003
[19] D. Avola, M. Bernardi, L. Cinque, G.L. Foresti, C. Massaroni, Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures, IEEE Trans. Multimed. 21 (2019) 234–245, https://doi.org/10.1109/TMM.2018.2856094
[20] G. Bailly, J. Müller, M. Rohs, D. Wigdor, S. Kratz, ShoeSense: a new perspective on gestural interaction and wearable applications, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12), ACM, 2012, pp. 1239–1248, https://doi.org/10.1145/2207676.2208576
[21] E. Bamani, E. Nissinman, I. Meir, L. Koenigsberg, A. Sintov, Ultra-range gesture recognition using a web-camera in human–robot interaction, Eng. Appl. Artif. Intell. 132 (2024) 108443, https://doi.org/10.1016/j.engappai.2024.108443, https://www.sciencedirect.com/science/article/pii/S0952197624006018.
[22] S. Bambach, S. Lee, D.J. Crandall, C. Yu, Lending a hand: detecting hands and recognizing activities in complex egocentric interactions, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2015, pp. 1949–1957, https://doi.org/10.1109/ICCV.2015.226
[23] O. Bau, W.E. Mackay, OctoPocus: a dynamic guide for learning gesture-based command sets, in: Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology (UIST '08), ACM, 2008, pp. 37–46, https://doi.org/10.1145/1449715.1449724
[24] E.B. Beeri, E. Nissinman, A. Sintov, Robust dynamic gesture recognition at ultra-long distances, arXiv preprint arXiv:2411.18413, 2024.
[25] G. Benitez-Garcia, J. Olivares-Mercado, G. Sanchez-Perez, K. Yanai, IPN hand: a video dataset and benchmark for real-time continuous hand gesture recognition, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 4340–4347, https://doi.org/10.1109/ICPR48806.2021.9412317
[26] S. Bharti, A. Balmik, A. Nandy, Novel error correction-based key frame extraction technique for dynamic hand gesture recognition, Neural Comput. Appl. 35 (2023) 21165–21180.
[27] G. Bhaumik, M.C. Govil, SpAtNet: a spatial feature attention network for hand gesture recognition, Multimed. Tools Appl. 83 (2024) 41805–41822.
[28] Y.C. Bilge, R.G. Cinbis, N. Ikizler-Cinbis, Towards zero-shot sign language recognition, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 1217–1232, https://doi.org/10.1109/TPAMI.2022.3143074
[29] Y.C. Bilge, N. Ikizler-Cinbis, R.G. Cinbis, Cross-lingual few-shot sign language recognition, Pattern Recognit. 151 (2024) 110374, https://doi.org/10.1016/j.patcog.2024.110374, https://www.sciencedirect.com/science/article/pii/S0031320324001250.
[30] M. Bohacek, M. Hrúz, Learning from what is already out there: few-shot sign language recognition with online dictionaries, in: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), 2023, pp. 1–6, https://doi.org/10.1109/FG57933.2023.10042544
[31] D.A. Bowman, L.F. Hodges, An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments, in: Proceedings of the 1997 Symposium on Interactive 3D Graphics (I3D '97), ACM, 1997, pp. 35–38, https://doi.org/10.1145/253284.253301
[32] V. Buchmann, S. Violich, M. Billinghurst, A. Cockburn, FingARtips: gesture based direct manipulation in augmented reality, in: Proceedings of the 2nd International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia (GRAPHITE '04), ACM, 2004, pp. 212–221, https://doi.org/10.1145/988834.988871
[33] G. Caggianese, N. Capece, U. Erra, L. Gallo, M. Rinaldi, Freehand-steering locomotion techniques for immersive virtual environments: a comparative evaluation, Int. J. Hum.–comput. Interact. 36 (2020) 1734–1755.
[34] N.C. Camgoz, S. Hadfield, O. Koller, H. Ney, R. Bowden, Neural sign language translation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7784–7793, https://doi.org/10.1109/CVPR.2018.00812
[35] X. Cao, P. Virupaksha, W. Jia, B. Lai, F. Ryan, S. Lee, J.M. Rehg, SocialGesture: delving into multi-person gesture understanding, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 19509–19519.
[36] Z. Cao, I. Radosavovic, A. Kanazawa, J. Malik, Reconstructing hand-object interactions in the wild, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2021, pp. 12417–12426, https://doi.org/10.1109/ICCV48922.2021.01219
[37] E. Chan, T. Seyed, W. Stuerzlinger, X.D. Yang, F. Maurer, User elicitation on single-hand microgestures, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16), ACM, 2016, pp. 3403–3414, https://doi.org/10.1145/2858036.2858589
[38] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. LIU, B. Mak, Two-stream network for sign language recognition and translation, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2022, pp. 17043–17056, https://proceedings.neurips.cc/paper_files/paper/2022/file/6cd3ac24cdb789beeaa9f7145670fcae-Paper-Conference.pdf.
[39] J. Cheng, D. Shi, C. Li, Y. Li, H. Ni, L. Jin, X. Zhang, Skeleton-based gesture recognition with learnable paths and signature features, IEEE Trans. Multimed. 26 (2024) 3951–3961, https://doi.org/10.1109/TMM.2023.3318242
[40] S. Constantin, F.I. Eyiokur, D. Yaman, L. Bärmann, A. Waibel, Interactive multimodal robot dialog using pointing gesture recognition, in: L. Karlinsky, T. Michaeli, K. Nishino (Eds.), Computer Vision – ECCV 2022 Workshops, Springer Nature Switzerland, Cham, 2023, pp. 640–657.
[41] H. Cui, R. Huang, R. Zhang, T. Hayama, Dstsa-gcn: advancing skeleton-based gesture recognition with semantic-aware spatio-temporal topology modeling, Neurocomputing 637 (2025) 130066.
[42] R. Cui, H. Liu, C. Zhang, A deep neural framework for continuous sign language recognition by iterative training, IEEE Trans. Multimed. 21 (2019) 1880–1891, https://doi.org/10.1109/TMM.2018.2889563
[43] F. Cunico, F. Girella, A. Avogaro, M. Emporio, A. Giachetti, M. Cristani, OOdMVMT: a deep multi-view multi-task classification framework for real-time 3D hand gesture classification and segmentation, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2023, pp. 2745–2754, https://doi.org/10.1109/CVPRW59228.2023.00275
[44] L.D. Cutler, B. Fröhlich, P. Hanrahan, Two-handed direct manipulation on the responsive workbench, in: Proceedings of the 1997 Symposium on Interactive 3D Graphics (I3D '97), ACM, 1997, pp. 107–114, https://doi.org/10.1145/253284.253315
[45] T.L. Dang, H.T. Nguyen, D.M. Dao, H.V. Nguyen, D.L. Luong, B.T. Nguyen, S. Kim, N. Monet, SHAPE: a dataset for hand gesture recognition, Neural Comput. Appl. 34 (2022) 21849–21862.
[46] T.L. Dang, T.H. Pham, Q.M. Dang, N. Monet, A lightweight architecture for hand gesture recognition, Multimed. Tools Appl. 82 (2023) 28569–28587.
[47] T.L. Dang, T.H. Pham, D.M. Dao, H.V. Nguyen, Q.M. Dang, B.T. Nguyen, N. Monet, DATE: a video dataset and benchmark for dynamic hand gesture recognition, Neural Comput. Appl. 36 (2024) 17311–17325.
[48] M. De Coster, M. Van Herreweghe, J. Dambre, Isolated sign recognition from RGB video using pose flow and self-attention, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021, pp. 3436–3445, https://doi.org/10.1109/CVPRW53098.2021.00383
[49] Q. De Smedt, H. Wannous, J.P. Vandeborre, Skeleton-based dynamic hand gesture recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2016, pp. 1206–1214, https://doi.org/10.1109/CVPRW.2016.153
[50] Z. Deng, Y. Leng, J. Chen, X. Yu, Y. Zhang, Q. Gao, TMS-Net: a multi-feature multi-stream multi-level information sharing network for skeleton-based sign language recognition, Neurocomputing 572 (2024) 127194, https://doi.org/

10.1016/j.neucom.2023.127194, https://www.sciencedirect.com/science/article/pii/S0925231223013176.
[51] N.S. Dinh, T.D. Nguyen, D.T. Tran, N.D.H. Pham, T.H. Tran, N.A. Tong, Q.H. Hoang, P. Le Nguyen, Sign language recognition: a large-scale multi-view dataset and comprehensive evaluation, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, pp. 7887–7897.
[52] Z. Doždor, Z. Kalafatić, Z. Ban, T. Hrkać, TY-Net: transforming YOLO for hand gesture recognition, IEEE Access 11 (2023) 140382–140394, https://doi.org/10.1109/ACCESS.2023.3341702
[53] A.D. Dragan, K.C.T. Lee, S.S. Srinivasa, Legibility and predictability of robot motion, in: Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI '13), IEEE Press, 2013, pp. 301–308, https://doi.org/10.1109/HRI.2013.6483603
[54] Y. Du, P. Xie, M. Wang, X. Hu, Z. Zhao, J. Liu, Full transformer network with masking future for word-level sign language recognition, Neurocomputing 500 (2022) 115–123, https://doi.org/10.1016/j.neucom.2022.05.051, https://www.sciencedirect.com/science/article/pii/S0925231222006178.
[55] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, X. Giro-I Nieto, How2Sign: a large-scale multimodal dataset for continuous American sign language, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2734–2743, https://doi.org/10.1109/CVPR46437.2021.00276
[56] S. Ebling, N.C. Camgöz, P.B. Braem, K. Tissi, S. Sidler-Miserez, S. Stoll, S. Hadfield, T. Haug, R. Bowden, S. Tornay, et al., SMILE Swiss German sign language dataset, in: Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC) 2018, The European Language Resources Association (ELRA), 2018.
[57] E.S.M. El-Alfy, H. Luqman, A comprehensive survey and taxonomy of sign language research, Eng. Appl. Artif. Intell. 114 (2022) 105198.
[58] M. Emporio, A. Ghasemaghaei, J.J. Laviola Jr, A. Giachetti, Continuous hand gesture recognition: benchmarks and methods, Comput. Vis. Image Underst. (2025) 104435.
[59] F. Fang, Z. Liao, Z. Kan, G. Wang, W. Yang, MDSI: pluggable multi-strategy decoupling with semantic integration for RGB-D gesture recognition, Pattern Recognit. 166 (2025) 111653.
[60] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, H. Ney, Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-Weather, 2014, p. 1911–1916. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85020237430&partnerID=40&md5=e17a51f131e02531fa45248c5bd212bc. Cited by: 130.
[61] Z. Fu, J. Chen, K. Jiang, S. Wang, J. Wen, M. Yang, D. Yang, Traffic police 3D gesture recognition based on spatial–temporal fully adaptive graph convolutional network, IEEE Trans. Intell. Transp. Syst. 24 (2023) 9518–9531, https://doi.org/10.1109/TITS.2023.3276345
[62] M. Gan, J. Liu, Y. He, A. Chen, Q. Ma, Keyframe selection via deep reinforcement learning for skeleton-based gesture recognition, IEEE Robot. Autom. Lett. 8 (2023) 7807–7814, https://doi.org/10.1109/LRA.2023.3322645
[63] S. Gan, Y. Yin, Z. Jiang, H. Wen, L. Xie, S. Lu, SignGraph: a sign sequence is worth graphs of nodes, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13470–13479, https://doi.org/10.1109/CVPR52733.2024.01279
[64] L. Gao, L. Zhu, L. Hu, P. Shi, L. Wan, W. Feng, A structure-based disentangled network with contrastive regularization for sign language recognition, Expert Syst. Appl. 271 (2025) 126623.
[65] Q. Gao, Y. Chen, Z. Ju, Y. Liang, Dynamic hand gesture recognition based on 3D hand pose estimation for human–robot interaction, IEEE Sens. J. 22 (2022) 17421–17430, https://doi.org/10.1109/JSEN.2021.3059685
[66] G. Garcia-Hernando, S. Yuan, S. Baek, T.K. Kim, First-person hand action benchmark with RGB-D videos and 3D hand pose annotations, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 409–419, https://doi.org/10.1109/CVPR.2018.00050
[67] M. Garg, D. Ghosh, P.M. Pradhan, GestFormer: multiscale wavelet pooling transformer network for dynamic hand gesture recognition, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2024, pp. 2473–2483, https://doi.org/10.1109/CVPRW63382.2024.00254
[68] M. Garg, D. Ghosh, P.M. Pradhan, MVTN: a multiscale video transformer network for hand gesture recognition, arXiv:2409.03890, 2024.
[69] M. Garg, D. Ghosh, P.M. Pradhan, ConvMixFormer- a resource-efficient convolution mixer for transformer-based dynamic hand gesture recognition, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2025, pp. 6156–6166, https://doi.org/10.1109/WACV61041.2025.00600
[70] X. Geng, Y. Li, Z. Ma, Z. Lu, Q. Miao, No blind alignment but generation: a different view of continuous sign language recognition based on diffusion, Pattern Recognit. 169 (2026) 111960.
[71] A. Ghorai, U. Nandi, C. Changdar, T. Si, M.M. Singh, J.K. Mondal, Indian sign language recognition system using network deconvolution and spatial transformer network, Neural Comput. Appl. 35 (2023) 20889–20907.
[72] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 369–376.
[73] M. Guan, Y. Wang, G. Ma, J. Liu, M. Sun, MSKA: multi-stream keypoint attention network for sign language recognition and translation, 2025, https://doi.org/10.1016/j.patcog.2025.111602. https://www.sciencedirect.com/science/article/pii/S0031320325002626.
[74] L. Guo, W. Xue, Q. Guo, B. Liu, K. Zhang, T. Yuan, S. Chen, Distilling cross-temporal contexts for continuous sign language recognition, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10771–10780, https://doi.org/10.1109/CVPR52729.2023.01037
[75] L. Guo, W. Xue, B. Liu, K. Zhang, T. Yuan, D. Metaxas, Gloss prior guided visual feature learning for continuous sign language recognition, IEEE Trans. Image Process. 33 (2024) 3486–3495, https://doi.org/10.1109/TIP.2024.3404869
[76] X. Guo, Q. Zhu, Y. Wang, Y. Mo, MG-GCT: a motion-guided graph convolutional transformer for traffic gesture recognition, IEEE Trans. Intell. Transp. Syst. 25 (2024) 14031–14039, https://doi.org/10.1109/TITS.2024.3394911
[77] S. Gustafson, D. Bierwirth, P. Baudisch, Imaginary interfaces: spatial interaction with empty hands and without visual feedback, in: Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology (UIST '10), ACM, 2010, pp. 3–12, https://doi.org/10.1145/1866029.1866033
[78] B. Hampiholi, C. Jarvers, W. Mader, H. Neumann, Convolutional transformer fusion blocks for multi-modal gesture recognition, IEEE Access 11 (2023) 34094–34103, https://doi.org/10.1109/ACCESS.2023.3263812
[79] W. Han, M. Hao, Y. Yuan, P. Liu, Fusion enhancement of YOLOv5 and copula Bayesian classifier for hand gesture recognition in smart sports venues, IEEE Access 12 (2024) 67005–67012, https://doi.org/10.1109/ACCESS.2024.3398142
[80] Y. Han, Y. Han, Q. Jiang, A study on the STGCN-LSTM sign language recognition model based on phonological features of sign language, IEEE Access 13 (2025) 74811–74820, https://doi.org/10.1109/ACCESS.2025.3560779
[81] A. Hao, Y. Min, X. Chen, Self-mutual distillation learning for continuous sign language recognition, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11283–11292, https://doi.org/10.1109/ICCV48922.2021.01111
[82] S. Hao, M. Fu, X. Liu, B. Zheng, Dynamic gesture recognition based on two-scale 3-D-ConvNeXt, IEEE Sens. J. 23 (2023) 29227–29234, https://doi.org/10.1109/JSEN.2023.3324479
[83] C. Harrison, D. Tan, D. Morris, Skinput: appropriating the body as an input surface, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10), ACM, 2010, pp. 453–462, https://doi.org/10.1145/1753326.1753394
[84] A.O. Hashi, S.Z.M. Hashim, A.B. Asamah, A systematic review of hand gesture recognition: an update from 2018 to 2024, IEEE Access 12 (2024).
[85] D.R.T. Hax, P. Penava, S. Krodel, L. Razova, R. Buettner, A novel hybrid deep learning architecture for dynamic hand gesture recognition, IEEE Access 12 (2024) 28761–28774, https://doi.org/10.1109/ACCESS.2024.3365274
[86] J. He, C. Zhang, X. He, R. Dong, Visual recognition of traffic police gestures with convolutional pose machine and handcrafted features, Neurocomputing 390 (2020) 248–259, https://doi.org/10.1016/j.neucom.2019.07.103, https://www.sciencedirect.com/science/article/pii/S0925231219314420.
[87] J. Hertel, S. Karaosmanoğlu, S. Schmidt, J. Bräker, M. Semmann, F. Steinicke, A taxonomy of interaction techniques for immersive augmented reality based on an iterative literature review, in: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, 2021, pp. 431–440, https://doi.org/10.1109/ISMAR52148.2021.00060
[88] J.D. Hincapié-Ramos, X. Guo, P. Moghadasian, P. Irani, Consumed endurance: a metric to quantify arm fatigue of mid-air interactions, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14), ACM, 2014, pp. 1063–1072, https://doi.org/10.1145/2556288.2557130
[89] Y.S. Hsiao, J. Sanchez-Riera, T. Lim, K.L. Hua, W.H. Cheng, LaRED: a large RGB-D extensible hand gesture dataset, in: Proceedings of the 5th ACM Multimedia Systems Conference, Association for Computing Machinery, New York, NY, USA, 2014, pp. 53–58, https://doi.org/10.1145/2557642.2563669
[90] H. Hu, J. Pu, W. Zhou, H. Fang, H. Li, Prior-aware cross modality augmentation learning for continuous sign language recognition, IEEE Trans. Multimed. 26 (2024) 593–606, https://doi.org/10.1109/TMM.2023.3268368
[91] H. Hu, J. Pu, W. Zhou, H. Li, Collaborative multilingual continuous sign language recognition: a unified framework, IEEE Trans. Multimed. 25 (2023) 7559–7570, https://doi.org/10.1109/TMM.2022.3223260
[92] H. Hu, W. Zhao, W. Zhou, H. Li, SignBERT+: hand-model-aware self-supervised pre-training for sign language understanding, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 11221–11239, https://doi.org/10.1109/TPAMI.2023.3269220
[93] H. Hu, W. Zhao, W. Zhou, Y. Wang, H. Li, SignBERT: pre-training of hand-model-aware representation for sign language recognition, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11067–11076, https://doi.org/10.1109/ICCV48922.2021.01090
[94] H. Hu, W. Zhou, H. Li, Hand-model-aware sign language recognition, Proc. AAAI Conf. Artif. Intell. 35 (2021) 1558–1566, https://doi.org/10.1609/aaai.v35i2.16247, https://ojs.aaai.org/index.php/AAAI/article/view/16247.
[95] H. Hu, W. Zhou, J. Pu, H. Li, Global-local enhancement network for NMF-aware sign language recognition, ACM Trans. Multimed. Comput. Commun. Appl. 17 (2021), https://doi.org/10.1145/3436754
[96] J. Hu, S. Liu, M. Liu, T. Zhou, J. Lu, X. Zuo, ST-CGNet: a spatiotemporal gesture recognition network with triplet attention and dual feature fusion, Pattern Recognit. (2025) 111767.
[97] L. Hu, W. Feng, L. Gao, Z. Liu, L. Wan, CorrNet+: sign language recognition and translation via spatial-temporal correlation, arXiv:2404.11111, 2024.
[98] L. Hu, L. Gao, Z. Liu, W. Feng, Temporal lift pooling for continuous sign language recognition, in: European Conference on Computer Vision, Springer, 2022, pp. 511–527.
[99] L. Hu, L. Gao, Z. Liu, W. Feng, Continuous sign language recognition with correlation network, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2529–2539, https://doi.org/10.1109/CVPR52729.2023.00249
[100] L. Hu, L. Gao, Z. Liu, W. Feng, Self-emphasizing network for continuous sign language recognition, Proc. AAAI Conf. Artif. Intell. 37 (2023) 854–862, https://

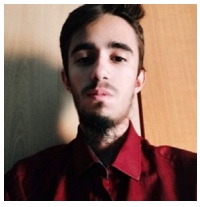
doi.org/10.1609/aaai.v37i1.25164, https://ojs.aaai.org/index.php/AAAI/article/view/25164.
[101] L. Hu, L. Gao, Z. Liu, W. Feng, Dynamic spatial-temporal aggregation for skeleton-aware sign language recognition, in: N. Calzolari, M.Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 5450–5460, https://aclanthology.org/2024.lrec-main.484/.
[102] J. Huang, W. Zhou, H. Li, W. Li, Attention-based 3D-CNNs for large-vocabulary sign language recognition, IEEE Trans. Circuits Syst. Video Technol. 29 (2019) 2822–2832, https://doi.org/10.1109/TCSVT.2018.2870740
[103] O. Ikne, B. Allaert, H. Wannous, Skeleton-based self-supervised feature extraction for improved dynamic hand gesture recognition, in: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), 2024, pp. 1–10, https://doi.org/10.1109/FG59268.2024.10581975
[104] F. Jafari, A. Basu, Two-dimensional parallel spatio-temporal pyramid pooling for hand gesture recognition, IEEE Access 11 (2023) 133755–133766, https://doi.org/10.1109/ACCESS.2023.3336591
[105] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, Y. Fu, Sign language recognition via skeleton-aware multi-model ensemble, arXiv preprint arXiv:2110.06161, 2021.
[106] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, Y. Fu, Skeleton aware multi-modal sign language recognition, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021, pp. 3408–3418, https://doi.org/10.1109/CVPRW53098.2021.00380
[107] P. Jiao, Y. Min, Y. Li, X. Wang, L. Lei, X. Chen, CoSign: exploring co-occurrence signals in skeleton-based continuous sign language recognition, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20619–20629, https://doi.org/10.1109/ICCV51070.2023.01890
[108] B. Joksimoski, E. Zdravevski, P. Lameski, I.M. Pires, F.J. Melero, T.P. Martinez, N.M. Garcia, M. Mihajlov, I. Chorbev, V. Trajkovik, Technological solutions for sign language recognition: a scoping review of research trends, challenges, and opportunities, IEEE Access 10 (2022) 40979–40998.
[109] H.R.V. Joze, O. Koller, MS-ASL: a large-scale data set and benchmark for understanding American sign language, arXiv:1812.01053, 2019.
[110] S. Kang, E.I. Libao, J. Lee, W. Woo, DirectionQ: continuous mid-air hand input for selecting multiple targets through directional visual cues, in: 2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), IEEE, 2022, pp. 757–762.
[111] A. Kapitanov, K. Kvanchiani, A. Nagaev, R. Kraynov, A. Makhliarchuk, HaGRID–HAnd gesture recognition image dataset, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 4572–4581.
[112] B. Karsh, R.H. Laskar, R.K. Karsh, mXception and dynamic image for hand gesture recognition, Neural Comput. Appl. 36 (2024) 8281–8300.
[113] S. Käs, A. Burenko, L. Markert, O.A. Culha, D. Mack, T. Linder, B. Leibe, How do foundation models compare to skeleton-based approaches for gesture recognition in human-robot interaction? arXiv preprint arXiv:2506.20795, 2025.
[114] M. Kawulok, J. Kawulok, J. Nalepa, Spatial-based skin detection using discriminative skin-presence features, Pattern Recognit. Lett. 41 (2014) 3–13.
[115] A. Khan, S. Jin, G.H. Lee, G.E. Arzu, T.N. Nguyen, L.M. Dang, W. Choi, H. Moon, Deep learning approaches for continuous sign language recognition: a comprehensive review, IEEE Access 13 (2025).
[116] P.V.V. Kishore, D.A. Kumar, R.C. Tanguturi, K. Srinivasarao, P.P. Kumar, D. Srihari, Joint motion affinity maps (JMAM) and their impact on deep learning models for 3D sign language recognition, IEEE Access 12 (2024) 11258–11275, https://doi.org/10.1109/ACCESS.2024.3354775
[117] S.K. Ko, C.J. Kim, H. Jung, C. Cho, Neural sign language translation based on human keypoint estimation, Appl. Sci. 9 (2019), https://doi.org/10.3390/app9132683, https://www.mdpi.com/2076-3417/9/13/2683.
[118] M. Koelle, A. El Ali, V. Cobus, W. Heuten, S.C.J. Boll, All about acceptability? Identifying factors for the adoption of data glasses, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17), ACM, New York, NY, USA, 2017, pp. 295–300, https://doi.org/10.1145/3025453.3025749
[119] S. Konstantakos, J. Cani, I. Mademlis, D.I. Chalkiadaki, Y.M. Asano, E. Gavves, G.T. Papadopoulos, Self-supervised visual learning in the low-data regime: a comparative evaluation, Neurocomputing 620 (2025) 129199.
[120] D. Kothadiya, C. Bhatt, K. Sapariya, K. Patel, A.B. Gil-González, J.M. Corchado, Deepsign: sign language detection and recognition using deep learning, Electronics 11 (2022), https://doi.org/10.3390/electronics11111780, https://www.mdpi.com/2079-9292/11/11/1780.
[121] D.R. Kothadiya, C.M. Bhatt, H. Kharwa, F. Albu, Hybrid InceptionNet based enhanced architecture for isolated sign language recognition, IEEE Access 12 (2024) 90889–90899, https://doi.org/10.1109/ACCESS.2024.3420776
[122] D.R. Kothadiya, C.M. Bhatt, T. Saba, A. Rehman, S.A. Bahaj, SIGNFORMER: DeepVision transformer for sign language recognition, IEEE Access 11 (2023) 4730–4739, https://doi.org/10.1109/ACCESS.2022.3231130
[123] P.P. Kumar, P. Vadakkepat, A.P. Loh, Hand posture and face recognition using a fuzzy-rough approach, Int. J. Humanoid Robot. 7 (2010) 331–356.
[124] B. Kwolek, Continuous hand gesture recognition for human-robot collaborative assembly, in: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2023, pp. 1992–1999, https://doi.org/10.1109/ICCVW60793.2023.00214
[125] D. Laines, M. Gonzalez-Mendoza, G. Ochoa-Ruiz, G. Bejarano, Isolated sign language recognition based on tree structure skeleton images, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2023, pp. 276–284, https://doi.org/10.1109/CVPRW59228.2023.00033
[126] T. Lee, Y. Oh, K.M. Lee, Human part-wise 3D motion context learning for sign language recognition, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20683–20693, https://doi.org/10.1109/ICCV51070.2023.01896
[127] D. Li, C.R. Opazo, X. Yu, H. Li, Word-level deep sign language recognition from video: a new large-scale dataset and methods comparison, in: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1448–1458, https://doi.org/10.1109/WACV45572.2020.9093512
[128] Y. Li, H. Chen, G. Feng, Q. Miao, Learning robust representations with information bottleneck and memory network for RGB-D-based gesture recognition, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20911–20921, https://doi.org/10.1109/ICCV51070.2023.01917
[129] Y. Li, G. Wei, C. Desrosiers, Y. Zhou, Decoupled and boosted learning for skeleton-based dynamic hand gesture recognition, Pattern Recognit. 153 (2024) 110536, https://doi.org/10.1016/j.patcog.2024.110536, https://www.sciencedirect.com/science/article/pii/S0031320324002875.
[130] Z. Li, W. Zhou, W. Zhao, K. Wu, H. Hu, H. Li, Uni-sign: toward unified sign language understanding at scale, arXiv:2501.15187, 2025.
[131] H. Liang, L. Fei, S. Zhao, J. Wen, S. Teng, Y. Xu, Mask-guided multiscale feature aggregation network for hand gesture recognition, Pattern Recognit. 145 (2024) 109901, https://doi.org/10.1016/j.patcog.2023.109901, https://www.sciencedirect.com/science/article/pii/S003132032300599X.
[132] K. Lin, X. Wang, L. Zhu, B. Zhang, Y. Yang, SKIM: skeleton-based isolated sign language recognition with part mixing, IEEE Trans. Multimed. 26 (2024) 4271–4280, https://doi.org/10.1109/TMM.2023.3321502
[133] M. Linardakis, I. Varlamis, G.T. Papadopoulos, Distributed maze exploration using multiple agents and optimal goal assignment, IEEE Access 12 (2024) 101407–101418, https://doi.org/10.1109/ACCESS.2024.3431909
[134] M. Linardakis, I. Varlamis, G.T. Papadopoulos, Survey on hand gesture recognition from visual input, arXiv preprint arXiv:2501.11992, 2025.
[135] D. Liu, L. Zhang, Y. Wu, LD-ConGR: a large RGB-D video dataset for long-distance continuous gesture recognition, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3294–3302, https://doi.org/10.1109/CVPR52688.2022.00330
[136] H. Liu, L. Wang, Gesture recognition for human-robot collaboration: a review, Int. J. Ind. Ergon. 68 (2018) 355–367, https://doi.org/10.1016/j.ergon.2017.02.004
[137] J. Liu, X. Wang, C. Wang, Y. Gao, M. Liu, Temporal decoupling graph convolutional network for skeleton-based gesture recognition, IEEE Trans. Multimed. 26 (2024) 811–823, https://doi.org/10.1109/TMM.2023.3271811
[138] J. Liu, Y. Wang, S. Xiang, C. Pan, HAN: an efficient hierarchical self-attention network for skeleton-based gesture recognition, Pattern Recognit. 162 (2025) 111343.
[139] L. Liu, L. Shao, Learning discriminative representations from RGB-D video data, in: IJCAI, Beijing, People's Republic of China, 2013, pp. 3.
[140] T. Liu, T. Tao, Y. Zhao, M. Li, J. Zhu, A signer-independent sign language recognition method for the single-frequency dataset, Neurocomputing 582 (2024) 127479, https://doi.org/10.1016/j.neucom.2024.127479, https://www.sciencedirect.com/science/article/pii/S0925231224002509.
[141] T. Liu, T. Tao, Y. Zhao, J. Zhu, A two-stream sign language recognition network based on keyframe extraction method, Expert Syst. Appl. 253 (2024) 124268, https://doi.org/10.1016/j.eswa.2024.124268, https://www.sciencedirect.com/science/article/pii/S0957417424011345.
[142] H. Lu, A.A. Salah, R. Poppe, TCNet: continuous sign language recognition from trajectories and correlated regions, Proc. AAAI Conf. Artif. Intell. 38 (2024) 3891–3899, https://doi.org/10.1609/aaai.v38i4.28181, https://ojs.aaai.org/index.php/AAAI/article/view/28181.
[143] Z. Lu, S. Qin, P. Lv, L. Sun, B. Tang, Real-time continuous detection and recognition of dynamic hand gestures in untrimmed sequences based on end-to-end architecture with 3D DenseNet and LSTM, Multimed. Tools Appl. 83 (2024) 16275–16312.
[144] H. Luqman, An efficient two-stream network for isolated sign language recognition using accumulative video motion, IEEE Access 10 (2022) 93785–93798, https://doi.org/10.1109/ACCESS.2022.3204110
[145] Y. Ma, B. Zhou, R. Wang, P. Wang, Multi-stage factorized spatio-temporal representation for RGB-D action and gesture recognition, arXiv:2308.12006, 2023.
[146] F. Manganaro, S. Pini, G. Borghi, R. Vezzani, R. Cucchiara, Hand gestures for the human-car interaction: the briareo dataset, in: E. Ricci, S. Rota Bulò, C. Snoek, O. Lanz, S. Messelodi, N. Sebe (Eds.), Image Analysis and Processing – ICIAP 2019, Springer International Publishing, Cham, 2019, pp. 560–571.
[147] G.C. Marinó, A. Petrini, D. Malchiodi, M. Frasca, Deep neural networks compression: a comparative survey and choice recommendations, Neurocomputing 520 (2023) 152–170.
[148] M. Maruyama, S. Singh, K. Inoue, P. Pratim Roy, M. Iwamura, M. Yoshioka, Word-level sign language recognition with multi-stream neural networks focusing on local regions and skeletal information, IEEE Access 12 (2024) 167333–167346, https://doi.org/10.1109/ACCESS.2024.3494878
[149] J. Materzynska, G. Berger, I. Bax, R. Memisevic, The jester dataset: a large-scale video dataset of human gestures, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 2874–2882, https://doi.org/10.1109/ICCVW.2019.00349
[150] M. Matilainen, P. Sangi, J. Holappa, O. Silvén, OUHANDS database for hand detection and pose recognition, in: 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), 2016, pp. 1–5, https://doi.org/10.1109/IPTA.2016.7821025
[151] K.R. May, T.M. Gable, B.N. Walker, Designing an in-vehicle air gesture set using elicitation methods, in: Proceedings of the 9th International Conference on

Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '17), 2017, pp. 74–83, https://doi.org/10.1145/3122986.3123015
[152] O. Mazhar, B. Navarro, S. Ramdani, R. Passama, A. Cherubini, A real-time human-robot interaction framework with robust background invariant hand gesture detection, Robot. Comput.-Integr. Manuf. 60 (2019) 34–48.
[153] D. Mendes, F.M. Caputo, A. Giachetti, A. Ferreira, J. Jorge, A survey on 3D virtual object manipulation: from the desktop to immersive virtual environments, Comput. Graph. Forum 38 (2019) 21–45, https://doi.org/10.1111/cgf.13390
[154] O. Mercanoglu Sincan, H.Y. Keles, Using motion history images with 3D convolutional networks in isolated sign language recognition, IEEE Access 10 (2022) 18608–18618, https://doi.org/10.1109/ACCESS.2022.3151362
[155] A.S.M. Miah, M.A.M. Hasan, S. Nishimura, J. Shin, Sign language recognition using graph and general deep neural network based on large scale dataset, IEEE Access 12 (2024) 34553–34569, https://doi.org/10.1109/ACCESS.2024.3372425
[156] A.S.M. Miah, M.A.M. Hasan, J. Shin, Dynamic hand gesture recognition using multi-branch attention based graph and general deep learning model, IEEE Access 11 (2023) 4703–4716, https://doi.org/10.1109/ACCESS.2023.3235368
[157] Y. Min, A. Hao, X. Chai, X. Chen, Visual alignment constraint for continuous sign language recognition, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11522–11531, https://doi.org/10.1109/ICCV48922.2021.01134
[158] M.R. Mine, F.P. Brooks Jr, C.H. Sequin, Moving objects in space: exploiting proprioception in virtual-environment interaction, in: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '97), ACM, 1997, pp. 19–26, https://doi.org/10.1145/258734.258747
[159] N. Mohamed, M.B. Mustafa, N. Jomhari, A review of the hand gesture recognition system: current progress and future directions, IEEE Access 9 (2021) 157422–157436, https://doi.org/10.1109/ACCESS.2021.3129650
[160] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, J. 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), 2016, pp. 4207–4215, https://doi.org/10.1109/CVPR.2016.456
[161] A. Moryossef, I. Tsochantaridis, J. Dinn, N.C. Camgöz, R. Bowden, T. Jiang, A. Rios, M. Müller, S. Ebling, Evaluating the immediate applicability of pose estimation for sign language recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2021, pp. 3429–3435.
[162] M. Müller, M. Alikhani, E. Avramidis, R. Bowden, A. Braffort, N. Cihan Camgöz, S. Ebling, C. España-Bonet, A. Göhring, R. Grundkiewicz, M. Inan, Z. Jiang, O. Koller, A. Moryossef, A. Rios, D. Shterionov, S. Sidler-Miserez, K. Tissi, D. Van Landuyt, Findings of the second WMT shared task on sign language translation (WMT-SLT23), in: P. Koehn, B. Haddow, T. Kocmi, C. Monz (Eds.), Proceedings of the Eighth Conference on Machine Translation, Association for Computational Linguistics, Singapore, 2023, pp. 68–94, https://doi.org/10.18653/v1/2023.wmt-1.4, https://aclanthology.org/2023.wmt-1.4/.
[163] S. Nakamura, Y. Kawanishi, S. Nobuhara, K. Nishino, DeePoint: visual pointing recognition and direction estimation, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20520–20530, https://doi.org/10.1109/ICCV51070.2023.01881
[164] Y. Nakamura, L. Jing, Skeleton-based data augmentation for sign language recognition using adversarial learning, IEEE Access 13 (2025) 15290–15300, https://doi.org/10.1109/ACCESS.2024.3481254
[165] S. Narayan, A.P. Mazumdar, S.K. Vipparthi, SBI-DHGR: skeleton-based intelligent dynamic hand gestures recognition, Expert Syst. Appl. 232 (2023) 120735, https://doi.org/10.1016/j.eswa.2023.120735, https://www.sciencedirect.com/science/article/pii/S095741742301237X.
[166] Z. Niu, R. Zuo, B. Mak, F. Wei, A Hong Kong sign language corpus collected from sign-interpreted TV news, arXiv:2405.00980, 2024.
[167] A. Nuzhdin, A. Nagaev, A. Sautin, A. Kapitanov, K. Kvanchiani, HaGRIDv2: 1M images for static and dynamic hand gesture recognition, arXiv:2412.01508, 2024.
[168] J.J. Ojeda-Castelo, M.D.L.M. Capobianco-Uriarte, J.A. Piedra-Fernandez, R. Ayala, A survey on intelligent gesture recognition techniques, IEEE Access 10 (2022) 87135–87156.
[169] O. Özdemir, .M. Baytaş, L. Akarun, Multi-cue temporal modeling for skeleton-based sign language recognition, Front. Neurosci. 17 (2023) 1148191.
[170] O. Ozdemir, I.M. Baytas, L. Akarun, Cross attentive multi-cue fusion for skeleton-based sign language recognition, IEEE Access 13 (2025).
[171] G.T. Papadopoulos, M. Antona, C. Stephanidis, Towards open and expandable cognitive AI architectures for large-scale multi-agent human-robot collaborative learning, IEEE Access 9 (2021) 73890–73909, https://doi.org/10.1109/ACCESS.2021.3080517
[172] S. Patra, A. Maitra, M. Tiwari, K. Kumaran, S. Prabhu, S. Punyeshwarananda, S. Samanta, Hierarchical windowed graph attention network and a large scale dataset for isolated Indian sign language recognition, arXiv:2407.14224, 2024.
[173] K. Pfeuffer, B. Mayer, D. Mardanbegi, H. Gellersen, Gaze + pinch interaction in virtual reality, in: Proceedings of the 5th Symposium on Spatial User Interaction (SUI '17), ACM, 2017, pp. 99–108, https://doi.org/10.1145/3131277.3132180
[174] C.A. Pickering, K.J. Burnham, M.J. Richardson, A research study of hand gesture recognition technologies and applications for human vehicle interaction, in: Proceedings of the 3rd Institution of Engineering and Technology Conference on Automotive Electronics, IET, Warwick, U.K., 2007, pp. 1–15.
[175] P.K. Pisharady, P. Vadakkepat, A.P. Loh, Attention based detection and recognition of hand postures against complex backgrounds, Int. J. Comput. Vis. 101 (2013) 403–419.
[176] T. Piumsomboon, A. Clark, M. Billinghurst, A. Cockburn, User-defined gestures for augmented reality, in: Human-Computer Interaction – INTERACT 2013, Springer, 2013, pp. 282–299, https://doi.org/10.1007/978-3-642-40480-1_18
[177] N. Pugeault, R. Bowden, Spelling it out: real-time ASL fingerspelling recognition, in: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Ieee, 2011, pp. 1114–1119.
[178] T. Qazi, M. Rupesh Kumar, P. Mukherjee, B. Lall, EgoFormer: ego-gesture classification in context of autonomous driving, IEEE Sens. J. 24 (2024) 18133–18140, https://doi.org/10.1109/JSEN.2024.3390794
[179] W. Qi, S.E. Ovur, Z. Li, A. Marzullo, R. Song, Multi-sensor guided hand gesture recognition for a teleoperated robot using a recurrent neural network, IEEE Robot. Autom. Lett. 6 (2021) 6039–6045, https://doi.org/10.1109/LRA.2021.3089999
[180] M.M. Rahman, A. Uzzaman, F. Khatun, M. Aktaruzzaman, N. Siddique, A comparative study of advanced technologies and methods in hand gesture analysis and recognition systems, Expert Syst. Appl. (2024) 125929.
[181] V.A. Raj, G. Malu, EnGesto: an ensemble learning approach for classification of hand gestures, IEEE Access 12 (2024).
[182] E. Rajalakshmi, R. Elakkiya, V. Subramaniyaswamy, L.P. Alexey, G. Mikhail, M. Bakaev, K. Kotecha, L.A. Gabralla, A. Abraham, Multi-semantic discriminative feature learning for sign gesture recognition using hybrid deep neural architecture, IEEE Access 11 (2023) 2226–2238, https://doi.org/10.1109/ACCESS.2022.3233671
[183] H. Ranjbar, A. Taheri, ISLR101: an Iranian word-level sign language recognition dataset, IEEE Access 13 (2025) 96147–96158, https://doi.org/10.1109/ACCESS.2025.3574074
[184] Q. Rao, K. Sun, X. Wang, Q. Wang, B. Zhang, Cross-sentence gloss consistency for continuous sign language recognition, Proc. AAAI Conf. Artif. Intell. 38 (2024) 4650–4658, https://doi.org/10.1609/aaai.v38i5.28265, https://ojs.aaai.org/index.php/AAAI/article/view/28265.
[185] R. Rastgoo, K. Kiani, S. Escalera, Hand sign language recognition using multi-view hand skeleton, Expert Syst. Appl. 150 (2020) 113336, https://doi.org/10.1016/j.eswa.2020.113336, https://www.sciencedirect.com/science/article/pii/S0957417420301615.
[186] R. Rastgoo, K. Kiani, S. Escalera, ZS-GR: zero-shot gesture recognition from RGB-D videos, Multimed. Tools Appl. 82 (2023) 43781–43796.
[187] R. Rastgoo, K. Kiani, S. Escalera, M. Sabokrou, Multi-modal zero-shot dynamic hand gesture recognition, Expert Syst. Appl. 247 (2024) 123349.
[188] S. Reifinger, F. Wallhoff, M. Ablassmeier, T. Poitschke, G. Rigoll, Static and dynamic hand-gesture recognition for augmented reality applications, in: Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments. HCII 2007, Springer, 2007, pp. 728–737, https://doi.org/10.1007/978-3-540-73110-8_79
[189] Y. Ren, H. Li, Y. Li, J. Pu, X. Pu, S. Jing, P. Jin, L. He, Multi-modal isolated sign language recognition based on self-paced learning, Expert Syst. Appl. 291 (2025) 128340.
[190] J. Rico, S. Brewster, Usable gestures for mobile interfaces: evaluating social acceptability, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10), ACM, New York, NY, USA, 2010, pp. 887–896, https://doi.org/10.1145/1753326.1753458
[191] N. Rodis, C. Sardianos, P. Radoglou-Grammatikis, P. Sarigiannidis, I. Varlamis, G.T. Papadopoulos, Multimodal explainable artificial intelligence: a comprehensive review of methodological advances and future research directions, IEEE Access 12 (2024) 159794–159820, https://doi.org/10.1109/ACCESS.2024.3467062
[192] F. Ronchetti, F.M. Quiroga, C. Estrebou, L. Lanzarini, A. Rosete, LSA64: an argentinian sign language dataset, arXiv:2310.17429, 2023.
[193] H.A. Rubaiyeat, H. Mahmud, A. Habib, M.K. Hasan, BdSLW60: a word-level bangla sign language dataset, Multimed. Tools Appl. (2025) 1–25.
[194] I. Saadi, A. Taleb-Ahmed, A. Hadid, Y. El Hillali, et al., Driver's facial expression recognition: a comprehensive survey, Expert Syst. Appl. 242 (2024) 122784.
[195] A. Sabater, I. Alonso, L. Montesano, A.C. Murillo, Domain and view-point agnostic hand action recognition, IEEE Robot. Autom. Lett. 6 (2021) 7823–7830, https://doi.org/10.1109/LRA.2021.3101822
[196] J.P. Sahoo, S.P. Sahoo, S. Ari, S.K. Patra, Hand gesture recognition using densely connected deep residual network and channel attention module for mobile robot control, IEEE Trans. Instrum. Meas. 72 (2023) 1–11, https://doi.org/10.1109/TIM.2023.3246488
[197] J. Salonen, A. Kronqvist, T. Jantunen, The corpus of Finnish sign language, in: Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives, 2020, pp. 197–202.
[198] N. Sarhan, S. Frintrop, Unraveling a decade: a comprehensive survey on isolated sign language recognition, in: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2023, pp. 3202–3211, https://doi.org/10.1109/ICCVW60793.2023.00345
[199] F. Shafizadegan, A.R. Naghsh-Nilchi, E. Shabaninia, Multimodal vision-based human action recognition using deep learning: a review, Artif. Intell. Rev. 57 (2024) 178.
[200] S. Sharma, S. Singh, Vision-based hand gesture recognition using deep learning for the interpretation of sign language, Expert Syst. Appl. 182 (2021) 115657, https://doi.org/10.1016/j.eswa.2021.115657, https://www.sciencedirect.com/science/article/pii/S0957417421010484.
[201] S. Sheikholeslami, A. Moon, E.A. Croft, Cooperative gestures for industry: exploring the efficacy of robot hand configurations in expression of instructional gestures for human-robot interaction, Int. J. Robot. Res. 36 (2017) 699–720, https://doi.org/10.1177/0278364917709941
[202] X. Shen, S. Yuan, H. Sheng, H. Du, X. Yu, Auslan-daily: Australian sign language translation for daily communication and news, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2023,

pp. 80455–80469, https://proceedings.neurips.cc/paper_files/paper/2023/file/feb34ce77fc8b94c85d12e608b23ce67-Paper-Datasets_and_Benchmarks.pdf.
[203] B. Shi, D. Brentari, G. Shakhnarovich, K. Livescu, Open-domain sign language translation learned from online video, arXiv:2205.12870, 2022.
[204] B. Shi, A.M.D. Rio, J. Keane, D. Brentari, G. Shakhnarovich, K. Livescu, Fingerspelling recognition in the wild with iterative visual attention, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5399–5408, https://doi.org/10.1109/ICCV.2019.00550
[205] J. Shin, A.S.M. Miah, Y. Akiba, K. Hirooka, N. Hassan, Y.S. Hwang, Korean sign language alphabet recognition through the integration of handcrafted and deep learning-based two-stream feature extraction approach, IEEE Access 12 (2024) 68303–68318, https://doi.org/10.1109/ACCESS.2024.3399839
[206] J. Shin, A.S.M. Miah, M.H. Kabir, M.A. Rahim, A. Al Shiam, A methodological and structural review of hand gesture recognition across diverse data modalities, IEEE Access 12 (2024).
[207] A.A.I. Sidig, H. Luqman, S. Mahmoud, M. Mohandes, KArSL: Arabic sign language database, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 20 (2021), https://doi.org/10.1145/3423420
[208] O.M. Sincan, H.Y. Keles, AUTSL: a large scale multi-modal Turkish sign language dataset and baseline methods, IEEE Access 8 (2020) 181340–181355, https://doi.org/10.1109/ACCESS.2020.3028072
[209] R. Slama, W. Rabah, H. Wannous, STr-GCN: dual spatial graph convolutional network and transformer graph encoder for 3D hand gesture recognition, in: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), 2023, pp. 1–6, https://doi.org/10.1109/FG57933.2023.10042643
[210] Q.D. Smedt, H. Wannous, J.P. Vandeborre, J. Guerry, B.L. Saux, D. Filliat, 3D hand gesture recognition using a depth and skeletal dataset, in: I. Pratikakis, F. Dupont, M. Ovsjanikov (Eds.), Eurographics Workshop on 3D Object Retrieval, The Eurographics Association, 2017, https://doi.org/10.2312/3dor.20171049
[211] J.H. Song, K. Kong, S.J. Kang, Dynamic hand gesture recognition using improved spatio-temporal graph convolutional network, IEEE Trans. Circuits Syst. Video Technol. 32 (2022) 6227–6239, https://doi.org/10.1109/TCSVT.2022.3165069
[212] N. Song, Y. Xiang, SLGTformer: an attention-based approach to sign language recognition, arXiv:2212.10746, 2022.
[213] A. Sridhar, R.G. Ganesan, P. Kumar, M. Khapra, INCLUDE: a large scale dataset for Indian sign language recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1366–1375, https://doi.org/10.1145/3394171.3413528
[214] S. Sridhar, A. Markussen, A. Oulasvirta, C. Theobalt, S. Boring, WatchSense: on- and above-skin input sensing through a wearable depth sensor, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17), ACM, 2017, pp. 3891–3902, https://doi.org/10.1145/3025453.3026005
[215] K.W. Strabala, M.K. Lee, A.D. Dragan, J.L. Forlizzi, S. Srinivasa, M. Cakmak, V. Micelli, Towards seamless human–robot handovers, J. Hum.-Robot Interact. 2 (2013) 112–132, https://doi.org/10.5898/JHRI.2.1.Strabala
[216] Y.S. Tan, K.M. Lim, C.P. Lee, Hand gesture recognition via enhanced densely connected convolutional neural network, Expert Syst. Appl. 175 (2021) 114797, https://doi.org/10.1016/j.eswa.2021.114797, https://www.sciencedirect.com/science/article/pii/S0957417421002384.
[217] S. Tang, D. Guo, R. Hong, M. Wang, Graph-based multimodal sequential embedding for sign language translation, IEEE Trans. Multimed. 24 (2021).
[218] X. Tang, Z. Yan, J. Peng, B. Hao, H. Wang, J. Li, Selective spatiotemporal features learning for dynamic gesture recognition, Expert Syst. Appl. 169 (2021) 114499, https://doi.org/10.1016/j.eswa.2020.114499, https://www.sciencedirect.com/science/article/pii/S095741742031143X.
[219] T. Tao, Y. Zhao, T. Liu, J. Zhu, Sign language recognition: a comprehensive review of traditional and deep learning approaches, datasets, and challenges, IEEE Access 12 (2024).
[220] J. Triesch, C. Von Der Malsburg, A system for person-independent hand posture recognition against complex backgrounds, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 1449–1453.
[221] A. Tunga, S.V. Nuthalapati, J. Wachs, Pose-based sign language recognition using GCN and BERT, in: 2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW), 2021, pp. 31–40, https://doi.org/10.1109/WACVW52041.2021.00008
[222] H.R. Vaezi Joze, A. Shaban, M.L. Iuzzolino, K. Koishida, MMTM: multimodal transfer module for CNN fusion, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13286–13296, https://doi.org/10.1109/CVPR42600.2020.01330
[223] A. Venugopalan, R. Reghunadhan, Applying deep neural networks for the automatic recognition of sign language words: a communication aid to deaf agriculturists, Expert Syst. Appl. 185 (2021) 115601, https://doi.org/10.1016/j.eswa.2021.115601, https://www.sciencedirect.com/science/article/pii/S0957417421010009.
[224] R. Walter, G. Bailly, J. Müller, StrikeAPose: revealing mid-air gestures on public displays, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13), ACM, 2013, pp. 841–850, https://doi.org/10.1145/2470654.2470774
[225] J. Wan, S.Z. Li, Y. Zhao, S. Zhou, I. Guyon, S. 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), 2016, pp. 761–769, https://doi.org/10.1109/CVPRW.2016.100
[226] S. Wan, L. Yang, K. Ding, D. Qiu, Dynamic gesture recognition based on three-stream coordinate attention network and knowledge distillation, IEEE Access 11 (2023) 50547–50559, https://doi.org/10.1109/ACCESS.2023.3278100
[227] B. Wang, L. Yu, B. Zhang, AL-MobileNet: a novel model for 2D gesture recognition in intelligent cockpit based on multi-modal data, Artif. Intell. Rev. 57 (2024) 282.
[228] F. Wang, Y. Du, G. Wang, Z. Zeng, L. Zhao, (2+1) D-SLR: an efficient network for video sign language recognition, Neural Comput. Appl. 34 (2022) 2413–2423.
[229] F. Wang, L. Zhang, H. Yan, S. Han, TIM-SLR: a lightweight network for video isolated sign language recognition, Neural Comput. Appl. 35 (2023) 22265–22280.
[230] S. Wang, K. Jiang, J. Chen, M. Yang, Z. Fu, T. Wen, D. Yang, Skeleton-based traffic command recognition at road intersections for intelligent vehicles, Neurocomputing 501 (2022) 123–134, https://doi.org/10.1016/j.neucom.2022.05.107, https://www.sciencedirect.com/science/article/pii/S0925231222006944.
[231] S. Wang, K. Wang, T. Yang, Y. Li, D. Fan, Improved 3D-ResNet sign language recognition algorithm with enhanced hand features, Sci. Rep. 12 (2022) 17812.
[232] Z. Wang, S.L. Huang, E.E. Kuruoglu, J. Sun, X. Chen, Y. Zheng, PAC-Bayes information bottleneck, in: International Conference on Learning Representations, 2022, https://openreview.net/forum?id=iLHOIDsPv1P.
[233] Z. Wang, D. Li, R. Jiang, M. Okumura, Continuous sign language recognition with multi-scale spatial-temporal feature enhancement, IEEE Access 13 (2025).
[234] F. Wei, Y. Chen, Improving continuous sign language recognition with cross-lingual signs, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23555–23564, https://doi.org/10.1109/ICCV51070.2023.02158
[235] J. Wiederer, A. Bouazizi, U. Kressel, V. Belagiannis, Traffic control gesture recognition for autonomous vehicles, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020, pp. 10676–10683, https://doi.org/10.1109/IROS45743.2020.9341214
[236] R. Wong, N.C. Camgöz, R. Bowden, Hierarchical i3d for sign spotting, in: European Conference on Computer Vision, Springer, 2022, pp. 243–255.
[237] P. Xie, Z. Cui, Y. Du, M. Zhao, J. Cui, B. Wang, X. Hu, Multi-scale local-temporal similarity fusion for continuous sign language recognition, Pattern Recognit. 136 (2023) 109233, https://doi.org/10.1016/j.patcog.2022.109233, https://www.sciencedirect.com/science/article/pii/S0031320322007129.
[238] P. Xie, M. Zhao, X. Hu, PiSLTRc: position-informed sign language transformer with content-aware convolution, IEEE Trans. Multimed. 24 (2022) 3908–3919, https://doi.org/10.1109/TMM.2021.3109665
[239] C. Xu, X. Wu, M. Wang, F. Qiu, Y. Liu, J. Ren, Improving dynamic gesture recognition in untrimmed videos by an online lightweight framework and a new gesture dataset ZJUGesture, Neurocomputing 523 (2023) 58–68, https://doi.org/10.1016/j.neucom.2022.12.022, https://www.sciencedirect.com/science/article/pii/S0925231222015193.
[240] Z. Xu, P. Sun, Y. Lu, H. Ge, M. Li, Y. Qi, Intuitive UAV operation: a novel dataset and benchmark for multi-distance gesture recognition, in: 2024 International Joint Conference on Neural Networks (IJCNN), 2024, pp. 1–8, https://doi.org/10.1109/IJCNN60899.2024.10651436
[241] M. Yu, A. Gan, C. Xue, G. Yan, DSTEN-CSLR: dual spatial–temporal enhancement network for continuous sign language recognition, Neural Comput. Appl. (2025) 1–24.
[242] Z. Yu, B. Zhou, J. Wan, P. Wang, H. Chen, X. Liu, S.Z. Li, G. Zhao, Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition, IEEE Trans. Image Process. 30 (2021) 5626–5640, https://doi.org/10.1109/TIP.2021.3087348
[243] H. ZainEldin, S.A. Gamel, F.M. Talaat, M. Aljohani, N.A. Baghdadi, A. Malki, M. Badawy, M.A. Elhosseini, Silent no more: a comprehensive review of artificial intelligence, deep learning, and machine learning in facilitating deaf and mute communication, Artif. Intell. Rev. 57 (2024) 188.
[244] H. Zhang, Z. Guo, Y. Yang, X. Liu, D. Hu, C2ST: cross-modal contextualized sequence transduction for continuous sign language recognition, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20996–21005, https://doi.org/10.1109/ICCV51070.2023.01925
[245] M. Zhang, Z. Zhou, T. Wang, W. Zhou, A lightweight network deployed on ARM devices for hand gesture recognition, IEEE Access 11 (2023) 45493–45503, https://doi.org/10.1109/ACCESS.2023.3273713
[246] Y. Zhang, C. Cao, J. Cheng, H. Lu, EgoGesture: a new dataset and benchmark for egocentric hand gesture recognition, IEEE Trans. Multimed. 20 (2018) 1038–1050, https://doi.org/10.1109/TMM.2018.2808769
[247] Y. Zhang, W. Xue, Y. Zhou, T. Yuan, S. Chen, CORE: multi-link graph attention network with inter-regional collaboration for continuous sign language recognition, Pattern Recognit. (2025) 111716.
[248] Z. Zhang, Z. Tian, M. Zhou, Handsense: smart multimodal hand gesture recognition based on deep neural networks, J. Ambient Intell. Humaniz. Comput. (2024) 1–16.
[249] D. Zhao, H. Li, S. Yan, Spatial–temporal synchronous transformer for skeleton-based hand gesture recognition, IEEE Trans. Circuits Syst. Video Technol. 34 (2024) 1403–1412, https://doi.org/10.1109/TCSVT.2023.3295084
[250] W. Zhao, H. Hu, W. Zhou, Y. Mao, M. Wang, H. Li, MASA: motion-aware masked autoencoder with semantic alignment for sign language recognition, IEEE Trans. Circuits Syst. Video Technol. 34 (2024) 10793–10804, https://doi.org/10.1109/TCSVT.2024.3409728
[251] W. Zhao, H. Hu, W. Zhou, J. Shi, H. Li, BEST: BERT pre-training for sign language recognition with coupling tokenization, Proc. AAAI Conf. Artif. Intell. 37 (2023) 3597–3605, https://doi.org/10.1609/aaai.v37i3.25470, https://ojs.aaai.org/index.php/AAAI/article/view/25470.
[252] W. Zhao, W. Zhou, H. Hu, M. Wang, H. Li, Self-supervised representation learning with spatial-temporal consistency for sign language recognition, IEEE Trans. Image Process. 33 (2024) 4188–4201, https://doi.org/10.1109/TIP.2024.3416881
[253] J. Zheng, Y. Wang, C. Tan, S. Li, G. Wang, J. Xia, Y. Chen, S.Z. Li, CVT-SLR: contrastive visual-textual transformation for sign language recognition with variational alignment, in: 2023 IEEE/CVF Conference on Computer Vision and

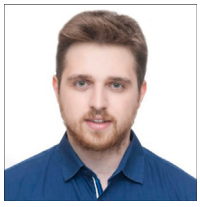
Pattern Recognition (CVPR), 2023, pp. 23141–23150, https://doi.org/10.1109/CVPR52729.2023.02216
[254] B. Zhou, Y. Li, J. Wan, Regional attention with architecture-rebuilt 3D network for RGB-D gesture recognition, Proc. AAAI Conf. Artif. Intell. 35 (2021) 3563–3571, https://doi.org/10.1609/aaai.v35i4.16471, https://ojs.aaai.org/index.php/AAAI/article/view/16471.
[255] B. Zhou, P. Wang, J. Wan, Y. Liang, F. Wang, D. Zhang, Z. Lei, H. Li, R. Jin, Decoupling and recoupling spatiotemporal representation for RGB-D-based motion recognition, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20122–20131, https://doi.org/10.1109/CVPR52688.2022.01952
[256] G. Zhou, Z. Cui, J. Qi, FGDSNet: a lightweight hand gesture recognition network for human robot interaction, IEEE Robot. Autom. Lett. 9 (2024) 3076–3083, https://doi.org/10.1109/LRA.2024.3362144
[257] H. Zhou, W. Zhou, W. Qi, J. Pu, H. Li, Improving sign language translation with monolingual data by sign back-translation, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1316–1325, https://doi.org/10.1109/CVPR46437.2021.00137
[258] H. Zhou, W. Zhou, Y. Zhou, H. Li, Spatial-temporal multi-cue network for sign language recognition and translation, IEEE Trans. Multimed. 24 (2022) 768–779, https://doi.org/10.1109/TMM.2021.3059098
[259] L. Zhou, C. Du, Z. Sun, T.L. Lam, Y. Xu, Long-range hand gesture recognition via attention-based SSD network, in: 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021, pp. 1832–1838, https://doi.org/10.1109/ICRA48506.2021.9561189
[260] Z. Zhou, V.W.L. Tam, E.Y. Lam, SignBERT: a BERT-based deep learning framework for continuous sign language recognition, IEEE Access 9 (2021) 161669–161682, https://doi.org/10.1109/ACCESS.2021.3132668
[261] R. Zuo, B. Mak, C2SLR: consistency-enhanced continuous sign language recognition, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5121–5130, https://doi.org/10.1109/CVPR52688.2022.00507
[262] R. Zuo, F. Wei, B. Mak, Natural language-assisted sign language recognition, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14890–14900, https://doi.org/10.1109/CVPR52729.2023.01430
[263] R. Zuo, F. Wei, B. Mak, Towards online continuous sign language recognition and translation, arXiv:2401.05336, 2024.
[264] G.S. Özcan, Y.C. Bilge, E. Sümer, Hand and pose-based feature selection for zero-shot sign language recognition, IEEE Access 12 (2024) 107757–107768, https://doi.org/10.1109/ACCESS.2024.3436671
[265] O. Özdemir, A.A. Kındıroğlu, N.C. Camgöz, L. Akarun, BosphorusSign22k sign language recognition dataset, arXiv:2004.01283, 2020.

Author biography



Konstantinos Foteinos is an undergraduate student at the Department of Informatics and Telematics, Harokopio University of Athens, where he also works as a research associate. His main research interests include computer vision, human-robot interaction, human gesture recognition, and robotics.

Manousos Linardakis is a B.Sc. graduate of the Department of Informatics and Telematics, Harokopio University of Athens, where he holds the all-time highest grade in the department's history as of this writing. He is currently pursuing an M.Sc. in Data Science and Machine Learning at the School of Electrical and Computer Engineering, National Technical University of Athens. He has conducted research in multi-robot exploration and computer vision, with practical experience in managing laboratory systems, implementing DevOps practices, and developing decision support software. He also participated in the Google Summer of Code as an open-source developer. His research interests include data science, machine learning and robotics with a focus on integrating theoretical advancements into practical, real-world applications.



Panagiotis Radoglou-Grammatikis received the Diploma and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Western Macedonia, Greece, in 2016 and 2023, respectively. He has published more than 50 research papers in international scientific journals, conferences and book chapters. He was included in Stanford University's list (shared by Elsevier) of the Top 2% of Scientists in the World for 2021 and 2022, respectively. He is currently working as a Research Director with K3Y Ltd., while he is also a Postdoc Researcher with the ITHACA Lab, University of Western Macedonia and a Co-Founder of MetaMind Innovations P.C. He is involved in several national and international projects. His main research interests focus on AI-driven cybersecurity, intrusion detection, and security games. He has received five best paper awards. Finally, he is a member of ACM and the Technical Chamber of Greece.

Vasileios Argyriou received the B.Sc. degree in computer science from the Aristotle University of Thessaloniki, Greece, in 2001, and the M.Sc. and Ph.D. degrees in electrical engineering working on registration from the University of Surrey, in 2003 and 2006, respectively. From 2001 to 2002, he held a research position with the Aristotle University of Thessaloniki, working on image and video watermarking. He joined the Communications and Signal Processing Department, Imperial College London, London, in 2007, where he was a Research Fellow working on 3D object reconstruction. He is currently a Professor with Kingston University London, working on computer vision and AI for crowd and human behavior analysis, computer games, entertainment, and medical applications. Additionally, research is conducted on educational games and on HCI for augmented and virtual reality (AR/VR) systems.

Panagiotis Sarigiannidis received the B.Sc. and Ph.D. degrees in computer science from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2001 and 2007, respectively. He is currently the Director of the ITHACA Laboratory, the Co-Founder of the 1st spin-off of the University of Western Macedonia: MetaMind Innovations P.C., and an Associate Professor with the Department of Electrical and Computer Engineering, University of Western Macedonia, Kozani, Greece. He has published more than 260 papers in international journals, conferences, and book chapters, including IEEE Communications Surveys and Tutorials, IEEE Transactions on Communications, IEEE Internet of Things Journal, IEEE Transactions on Broadcasting, IEEE Systems Journal, IEEE Wireless Communications Magazine, IEEE Open Journal of the Communications Society, IEEE/OSA Journal of Lightwave Technology, IEEE Transactions on Industrial Informatics, IEEE Access, and Computer Networks. He has been involved in several national, European, and international projects. He is currently the project coordinator of three H2020 Projects, namely H2020-DS-SC7-2017 (DS-07-2017), SPEAR: Secure and PrivatE smArt gRid; H2020-LC-SC3-EE-2020-1 (LC-SC3-EC4-2020), EVIDENT: bEhaVioral Insgihts anD Effective eNergy policy acTions; and H2020-ICT-2020-1 (ICT-56-2020), TERMINET: nexT gEneRation sMart INterconnectEd IoT, while he coordinates the Operational Program MARS: sMart fArming with dRoneS (Competitiveness, Entrepreneurship, and Innovation) and the Erasmus+ KA2 ARRANGE-ICT: SmartROOT: Smart faRming innOvatiOn Training. He also serves as a Principal Investigator for the H2020-SUDS-2018 (SU-DS04-2018), SDN-microSENSE: SDN-microgrid reSilient Electrical eNergy SystEm and in three Erasmus+ KA2: ARRANGE-ICT: pArtneRship foR AddressiNG mEgatrends in ICT, JAUNTY: Joint undergAduate coUrses for smart eNergy managemenT sYstems, and STRONG: advanced firST RespONders trainiG (Cooperation for Innovation and the Exchange of Good Practices). His research interests include telecommunication networks, the Internet of Things, and network security. He participates on the editorial boards of various journals.

Iraklis Varlamis received the M.Sc. degree in information systems engineering from the University of Manchester Institute of Science and Technology, Manchester, U.K., and the Ph.D. degree from Athens University of Economics and Business, Athens, Greece. He is currently a Professor of data management with the Department of Informatics and Telematics, Harokopio University of Athens (HUA), Kallithea, Greece. He has more than 250 papers published in international journals and conferences and more than 5000 citations for his work. He holds a patent from the Greek Patent Office for a system that thematically groups web documents using content and links. He is the Scientific Coordinator for HUA in several EU (H2020, ECSEL, REC) and Qatar (QNFR) projects, as well as in national projects. His research interests include data mining and social network analytics to recommender systems for social media and real-world applications.

Georgios Th. Papadopoulos received the Diploma and Ph.D. degrees in electrical and computer engineering from the Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece. He was a Postdoctoral Researcher with the Foundation for Research and Technology Hellas/Institute of Computer Science (FORTH/ICS) and the Centre for Research and Technology Hellas/Information Technologies Institute (CERTH/ITI). He is currently an Assistant Professor of computer graphics and computational vision with the Department of Informatics and Telematics, Harokopio University of Athens, Greece. He has published over 100 peer-reviewed research papers in international journals and conference proceedings. His research interests include computer vision, artificial intelligence, machine/deep learning, human action recognition, human-computer interaction, and explainable artificial intelligence. He is a member of the Technical Chamber of Greece.