Universal Access in the Information Society
https://doi.org/10.1007/s10209-024-01133-y

LONG PAPER

Impact of face swapping and data augmentation on sign language recognition

Marina Perea-Trigo[1] · Enrique J. López-Ortiz[2] · Luis M. Soria-Morillo[1] · Juan A. Álvarez-García[1] · J. J. Vegas-Olmos[3]

Accepted: 9 July 2024

Abstract

This study addresses the challenge of improving communication between the deaf and hearing community by exploring different sign language recognition (SLR) techniques. Due to privacy issues and the need for validation by interpreters, creating large-scale sign language (SL) datasets can be difficult. The authors address this by presenting a new Spanish isolated sign language recognition dataset, CALSE-1000, consisting of 5000 videos representing 1000 glosses, with various signers and scenarios. The study also proposes using different computer vision techniques, such as face swapping and affine transformations, to augment the SL dataset and improve the accuracy of the model I3D trained using them. The results show that the inclusion of these augmentations during training leads to an improvement in accuracy in top-1 metrics by up to 11.7 points, top-5 by up to 8.8 points and top-10 by up to 9 points. This has great potential to improve the state of the art in other datasets and other models. Furthermore, the analysis confirms the importance of facial expressions in the model by testing with a facial omission dataset and shows how face swapping can be used to include new anonymous signers without the costly and time-consuming process of recording.

Keywords Sign language · Sign language recognition · Face swapping · Data augmentation · Deep learning · Computer vision

1 Introduction

According to the World Health Organization [1], around 430 million people have hearing loss of moderate or higher severity, with the total number of people experiencing some degree of hearing loss increasing to over 1.5 billion, globally. As a result, almost 5% of the world population can be considered deaf, with that number increasing to almost 20% if we consider the hard-of-hearing people too.

Sign Languages (SL) are the main medium of communication for the deaf community, with about 300 different Sign Languages worldwide [2], which only 1% of the population (almost all deaf people themselves and their families) understand. As a result, a group faces challenges in communicating with individuals who can hear daily. This makes it even harder for them to access education, healthcare, employment, entertainment, and engage in social interactions where effective communication is crucial [3].

Therefore, providing a system capable of translating spoken languages into SL and vice versa, to facilitate the exchange of information between deaf and hearing people, who do not know Sign Language, remains a developing challenge, and technology is present to provide that help. SL are visual languages and they include non-manual features (facial and body expressions) beyond the manual gesture itself to provide additional information. Sign Languages have their own grammatical rules and are developed

✉ Marina Perea-Trigo
mptrigo@us.es

Enrique J. López-Ortiz
elortiz@us.es

Luis M. Soria-Morillo
lsoria@us.es

Juan A. Álvarez-García
jaalvarez@us.es

J. J. Vegas-Olmos
juanj@nvidia.com

[1] Department of Languages and Computer Systems, Universidad de Sevilla, 41012 Sevilla, Spain

[2] Department of Computer Science and Artificial Intelligence, Universidad de Sevilla, 41012 Sevilla, Spain

[3] NVIDIA Corporation, Ltd., Hermon Buildin, 20692 Yokneam, Israel

Published online: 24 July 2024

independently of spoken languages [4], which is why they have such a low comprehension rate for those who do not know the language, with no gloss-to-word correspondence between Sign Languages and their related spoken languages.

To address the challenge of improving communication between the deaf and hearing community, there are different techniques already explored in the literature, such as Sign Language Recognition (SLR) [5], which is the process of identifying signs that include manual and non-manual gestures and translating them into one or more glosses (representation of a sign), Sign Language Translation (SLT) [6, 7], whereby a spoken/written language sentence is extracted from a video in which signs are performed continuously, and Sign Language Production (SLP) [8, 9], whose goal is the generation of videos or a sequence of static images from a text or spoken language. Also, SLR includes two main categories: Isolated Sign Language Recognition (ISLR, sometimes referred in literature as word-level), which aims to recognize isolated glosses, which we will focus in this work, and Continuous Sign Language Recognition (CSLR), where the main goal is to recognize each gloss that comprises an SL sentence.

The study of these techniques requires large datasets with sufficient vocabulary and variability, which is challenging to produce because of the need for professional interpreters to validate them. In Spanish Sign Language (LSE) in particular, there are few datasets with a limited number of signs as can be seen in Table 5. In addition, privacy concerns about the visual nature of [10] data need to be addressed. Another problem that can arise is the variation of the same LS between different regions, where completely different signs are used to indicate the same gloss without there being a specific dictionary to clarify the different signs used.

Given these considerations, with a specific focus on Isolated Sign Language Recognition (ISLR), the main contributions of this paper are:

- We provide a new dataset for Spanish Isolated Sign Language Recognition: CALSE-1000 with the largest number of LSE videos for ISLR to date;
- We propose a new technique based on face swapping and affine transformations to increase the size of ISLR datasets without increasing the recording time and ensuring anonymity;
- We improve the accuracy of the recognition model I3D [11] using our proposal in top-1 metrics by up to 11.7 points, top-5 by up to 8.8 points and top-10 by up to 9 points.

The rest of this paper is organized as follows: Sect. 2 provides a context on the scarcity of Sign Language datasets and methods used to increase their size, Sect. 3 describes our method with the created dataset and the techniques applied to it. Section 4 details the experimental procedure performed and presents the results obtained and the dataset creation process, showing the experiments performed and the results obtained in Sect. 5. Finally, we conclude the paper in Sect. 6 by discussing our findings and outlining some possible future work.

2 Related work

Sign Language Recognition is an arduous and complex task due to the scarcity of adequate and consistent datasets [10]. Creating an SLR dataset in a controlled environment under optimal conditions is a time-consuming process and requires validation and supervision by professional interpreters.

2.1 Data scarcity

Sign Language Recognition is a computer vision task that has been explored for years. Although several techniques such as Hidden Markov Models (HMM) [12], Neural Network based methods [13] or Deep Learning methods [14] have been used to study it, having a proper amount of available, coherent datasets is the most fundamental prerequisite for working with this problem. Providing a public and large dataset for Sign Language Recognition is often a difficult task; due to the fact that these datasets consist of videos, it is important to create them in a controlled environment under optimal conditions (no occlusions, same illumination conditions, different viewpoints, etc.), in order to provide realistic samples.

This makes the creation of an SL dataset to perform recognition tasks a time-consuming task, and as it is an unfamiliar language to many people, it requires the validation and supervision of professional interpreters to represent the information correctly, so not many datasets are ready for SLR. In addition, a dataset must include a sufficient variety of interpreters to allow a degree of variability so that trained models can generalise when the model is put into production. It is also influenced by the number of sign languages in existence, around 300 [2], which means that in the research field, the largest datasets are for American Sign Language. On the one hand, they may be incomplete in many cases, providing only gloss information or showing only one possible view [15], and on the other hand, the annotation format of the information tends to be inconsistent among the available sets, as there is no specific convention or protocol for organising the content. Consequently, the type and format of the information included do not always match among the different corpora and, in some cases, may not even be compatible [16]. However, some datasets are designed to be complete and, at the same time, challenging, provide also additional information, with different views and depth data.

Table 1 Overview of Isolated SLR (ISLR) datasets with their main characteristics. In the Signers section, the total number of signers participating in the dataset is listed first, followed by the average per video

	Characteristics				
Datasets	Gloss	Signers	Videos	Language	Year
Purdue RVL-SLLL [19]	104	14/–	2567	American	2002
Boston ASLLVD [20]	3,300+	–/~3.6	9794	American	2008
ASL-LEX [21]	1,000	–	1000	American	2016
MS-ASL [22]	1,000	222/–	25,513	American	2018
WLASL-2000 [18]	2,000	119/~10	20,863	American	2019
ASL-LEX 2.0 [23]	2,723	–	2723	American	2021
ASLLRP [24]	–	33	23,452	American	2022
DGS Kinect 40 [25]	40	15/–	3000	German	2012
SMILE [26]	100	30/–	–	Swiss-German	2018
DEVISIGN-L [27]	2,000	8/3	24,000	Chinese	2015
CSL [28]	500	50/–	125,000	Chinese	2016
MSL [29]	30	4/4	3000	Mexican	2022
LSA64 [30]	64	10/10	3200	Argentinian	2016
BosphorusSign22k [31]	744	6/–	22,542	Turkish	2020
AUTSL [17]	226	43/–	38,336	Turkish	2020
LSE-sign [32]	2400	2/1	2400	Spanish	2016
LSE_UVIGO (LSE_Lex40) [33]	40	32/ 32	1368	Spanish	2020
CALSE-100 (ours)	100	15/6	600	Spanish	2023
CALSE-1000 (ours)	1000	15/5	5000	Spanish	2023

Another fact to consider is the quality and the realism of the samples, as exposed by Sincan et al. in [17], where they presented the AUTSL dataset, an alternative for the less realistic datasets, as PHOENIX-2014-T [6] or WLASL [18], in which signers have similar body shapes, clothes and even backgrounds. In the AUTSL dataset, different environments, positions, and body types are considered as responses to the mentioned problem. This idea of increasing the dataset complexity has been an inspiration for the work carried out in Sect. 4. Table 1 presents a summary of the available published datasets including isolated signs.

As can be seen, two new datasets are included, one with 1000 glosses and 6 signers per gloss and a smaller one that can be used for faster testing of 100 glosses.

2.2 Data augmentation

Data augmentation comprises several methods that improve the quality and size of the training dataset, allowing to reduce overfitting and thus helping the model used to extract more information from the original dataset. The main objective is to add "new" data from modifications of the original data.

Many data augmentation techniques are used for different tasks [34, 35]. For image-based tasks (classification, object detection, etc.), one can apply augmentations by data deformation such as colour and geometric transformations (reflecting, cropping, rotation, flipping, etc.), random erasure, kernel filters, or adversarial training. Subsampling enhancements are also used, creating synthetic instances that are added to the training set such as image blending, feature space augmentation, or Generative Adversarial Networks (GAN), which have proven to be really efficient in augmenting datasets [36]. For video-based tasks (action recognition, object detection and segmentation among others), the temporal dimension is used in addition to the image augmentation techniques described above. In both video and image cases, using masks to modify the background or the distribution of the main objects through the frame/s has also proven to be quite effective as a transformation, as seen in [37].

Data augmentation has previously been used in other Sign Language Recognition studies, mostly to prevent overfitting [38] and to increase dataset size [39, 40], as well as for Sign Language Translation (SLT) tasks in the semantic part of the problem, attempting to improve gloss-to-text translation through synonym replacement [41] or by using syntactic rules and word order modifications to create synthetic gloss data [42].

2.3 Face swapping

Deepfakes algorithms are those that combine techniques to manipulate and create fake images and videos by transferring important features from the source image (or video) to the target image (or video) such that humans cannot distinguish them from real ones [43, 44]. Deepfakes can be created by combining traditional visual effects or computer graphics approaches, although the technique most recently

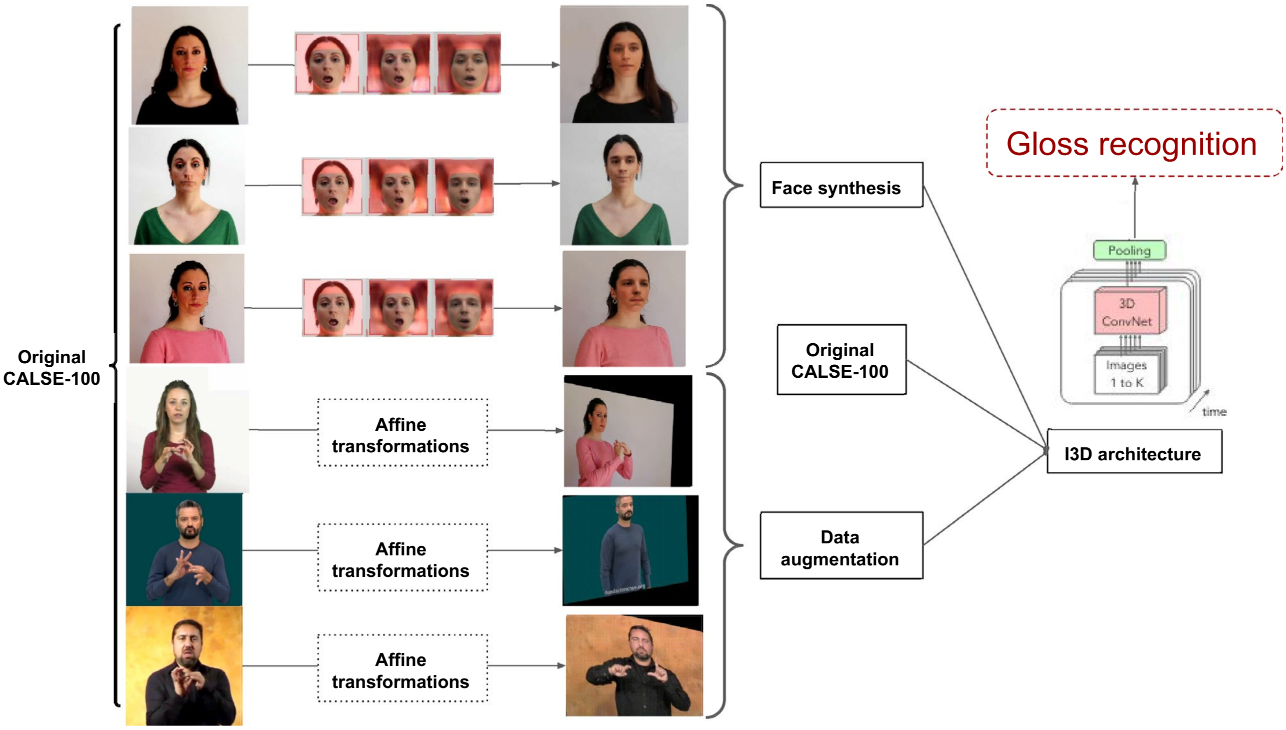

Fig. 1 Pipeline explaining the methodology employed

applied is Deep Learning models, such as GANs and autoencoders, which are widely used in computer vision [45, 46].

Mirsky et al. [47] categorize the media content generated by deepfakes into four types: reenactment [48], where the source is used to drive the expression, mouth, gaze, pose, or body of the desired target; replacement [49], among which is face swapping, replacing the target's content with that of the input, thus preserving the input's identity; editing [50] involves the removal, addition, or alteration of attributes (changes in clothing, ethnicity, age, etc.) of the target; and finally synthesis [51], which is when the deepfake is created without any target as a base.

Due to privacy concerns regarding the visual nature of Sign Language datasets, anonymization is a task that has been undertaken to increase participation readiness in the creation of SL datasets [10]. To this end, tasks such as pixelation [52], blackening [53] and greyscale filtering [10] are not applicable in SLR tasks because, during sign execution, facial and body information not only play an important role in the correct meaning of the sign, but are strictly necessary in order to provide real meaning, thus avoiding that their absence results in having only gestures without meaning. So the generation of realistic face swapping synthetic data is an alternative to the privacy problem, since all the desired characteristics of the original dataset are contained, but without sensitive content, making it impossible to identify individuals [54, 55], as well as a solution in applications such as increasing the number of unbalanced or insufficient datasets [56].

3 Methodology

Our approach starts with the collection of the dataset, which will be explained in Sect. 4, and then tests the influence of data augmentation and fae swapping through the combination with different variants of the original dataset. Once these variants have been obtained, the I3D model [11] is applied for ISLR on each of the generated datasets, in order to observe the improvement with respect to the application of the model on the original data. Figure 1 shows a pipeline explaining the methodology employed. For the experiments performed, the cross-dataset technique [57] has been applied, separating one of the sets as a test and leaving the rest of the data for training.

In addition, the influence of facial expression on the model will also be tested by applying facial omission on the test set. The results obtained are specified in Sect. 5.

All experiments were performed with an NVIDIA Geforce RTX 3090 and an NVIDIA A100.

To increase the dataset and avoid possible problems in maintaining identity through anonymization of the data, the FaceSwap [58] tool has been used, which employs Deep

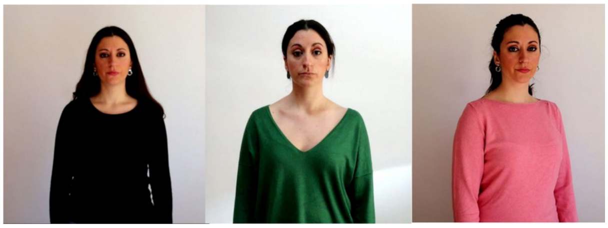

Fig. 2 Scenarios A, B and C, respectively in which the collaborating interpreter performs each sign

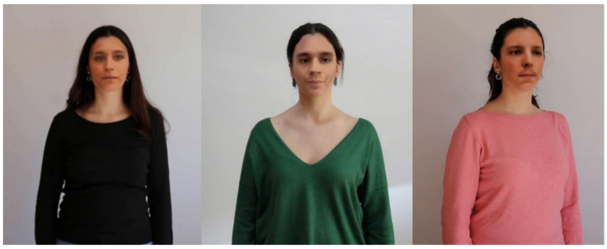

Fig. 3 Results of applying face swapping with 3 different models

Learning techniques to recognize and swap faces in each of the signers that make up the CALSE-100 dataset.

The face swapping to generate deepfakes is composed of 3 stages:

1. *Extraction* In this first stage, the extraction of faces from the target video for the later training takes place, in which face landmarks are recognized and the images are cropped, saving the faces to be used for training. In this first step, it is important to have a large set of images containing the face of the subject to be trained, as well as to consider the data quality and the variety of angles and expressions.
2. *Training* In this step, the training of the 'Phaze-a' model [59] (the latest model for Faceswap) is executed for 35,000–40,000 iterations (it depends on the models used for the face swapping), with a batch size of 10.
3. *Conversion* In this last stage, the face extraction is performed again (in this case, on the source video) and then the face swapping is performed to obtain the final set of videos with face swapping applied.

The face swapping technique is applied to the original data set shown in Fig. 2, obtaining the results presented in Fig. 3. It is important to highlight that this technique has been used only on the dataset from our interpreter. This is because the tool used to extract, train and convert the face swapping must be performed independently for each person, which means that obtaining DeepFakes for the entire public set of DILSE, SACU and Spread the Sign, which contains a wide range of signers, is a very costly process.

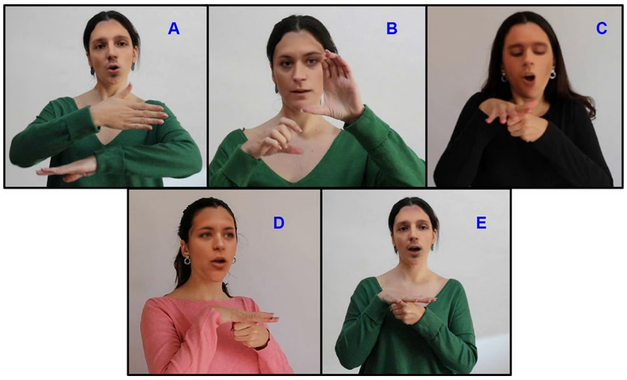

Fig. 4 Different frames obtained by applying face swapping

Moreover, Fig. 4 contains different frames of the results after applying face swapping in more detail. Among them we can compare how the results can be quite accurate when there are no hand occlusions or much expressiveness in the face (frames A and E), while in those where there is more effusiveness (frames C and D) or occlusions are produced when passing the hand in front of the face (frame B), details of the original image are lost. We assume that, although there may be frames in the videos that are not maximally accurate for these reasons, the results obtained are good enough to be taken into account during training.

In addition, a new, publicly available library[1] has been used to increase datasets size that will launch augmentations on the videos.

Although several functions are implemented to apply transformations, we will only consider affine transformations. An affine transformation is the result of a combination of linear transformations. In the linear transformation, lines are converted while retaining points, lines and planes, thus maintaining their parallelism, but not necessarily Euclidean distances and angles, therefore, it includes the classical transformations, i.e. translations, reflections, scalings and rotations. In addition, the function applies a random affine transformation, which may therefore result in the application of one of the above transformations. However, the library adds the possibility to apply these transformations individually, in order to provide a complete tool.

Previous works have utilized anonymization techniques such as pixelation or blackening to preserve privacy. However, since facial expression plays a crucial role in interpreting signs correctly, the test set underwent face omission. This helped evaluate the importance of facial expressions

[1] *https://github.com/RodGal-2020/video_augmentationRodGal-2020/*.

for the model to recognize signs accurately. The results of this evaluation will determine whether incorporating face swapping techniques into the data is worthwhile.

4 Experimental setting

4.1 Dataset

In this section, we detail how we have collected the dataset used to achieve our objectives. The dataset, defined as CALSE ("Conjunto Aislado de Lengua de Signos Española"), has been formed by obtaining videos from 3 different, publicly available data sources, which are the Dictionary of Spanish Sign Language (DILSE) [60] and Spread the Sign (STS) [61] dictionaries, as well as the dataset from the University Community Assistance Service (SACU) of the University of Seville [62], which is a new tool to meet the needs associated with hearing impaired students who use SL as their means of communication. Examples of these 3 sources that compose the data set can be seen in Fig. 5.

The CALSE-1000 set is composed of 1000 glosses, with at least two video samples of each gloss extracted from the sources described above (because not all vocabulary is present in the 3 public data sources). In addition, this set has also been signed by a professional Spanish Sign Language interpreter, thus being able to add more examples for each word in the set. Thanks to this collaboration, 3 more items of each word have been incorporated, thus obtaining a total of 5 examples of each word.

It is important to ensure the model has the ability to adapt for situations and scenarios that are not covered by the training input data, so that once the model is trained, it can be equally accurate to new input data that is different from the training data. To ensure this, it is crucial to incorporate variations in the appearance and style of the signs during the training process, while always faithfully maintaining the meaning of the sign. For this purpose, we have designed several scenarios for video recording of the signs performed by the interpreter. These scenarios cover changes in perspective (such as front or side view), intensity when performing the signs, and variations in clothing, among other aspects.

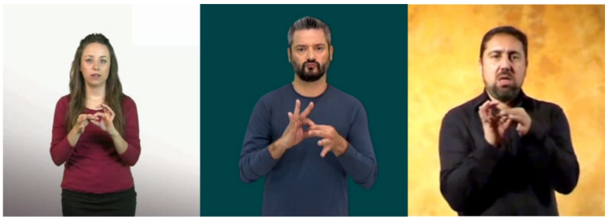

Fig. 5 Capture of the different public data sources: SACU, DILSE and STS from left to right

This approach helps to introduce diversity among the different elements of the same signed words, ultimately resulting in a complete, enriched dataset and a more robust model in different situations.

Figure 2 shows the 3 different scenarios in which the collaborating interpreter performed every sign. For each of the scenarios the clothing used is different. In the case of scenario A, it has been defined to perform the recording with the hair up emphasizing each of the signs facing forward; for scenario B, each sign has been performed from the front, with a normal focus of the signs and the hair down; finally, scenario C shows a profile perspective of the signs without emphasis.

This subset can be accessed and downloaded through the OneDrive folder.[2]

Due to the limited availability of training data for signs, it has been decided to omit the validation set. With only 600 videos available, distributed over 6 videos per sign, any additional data separation for a validation set would significantly reduce the size of the training set. In this context, the priority is to maximize the amount of training data to allow our models to effectively learn the distinctive features of each sign. By not using a validation set, we can take full advantage of our limited resources and train more robust and generalized models that better fit the available data. While we understand the importance of evaluating model performance on unseen data, we believe that in this particular case, the quality and quantity of training data are crucial to the success of the research. Therefore, we split the samples into training and test in a 5:1 ratio for the different experiments. The training process ends when 60 epochs are reached, since this is the point at which the loss metric stabilizes and stops decreasing.

This dataset is available for use and download through our project's GitHub repository[3].

4.2 Implementation details

For the isolated recognition we have used the I3D [11] network architecture implemented in PyTorch, being the same as the one used in [18]. All experiments utilized identical configurations, employing the Adam optimizer [63], a batch size of 10, a learning rate of $10^{-}3$ and a total of 60 epochs. Due to the limited number of samples available for each sign, the dataset was divided into training and test sets only, applying the cross-dataset approach [57], so that one of the datasets is separated as a test set (equivalent to one example per sign) and the rest as a training set.

To assess the performance of the models, we calculate the average scores for top-K classification accuracy with

[2] *https://sharepoint.com/Deepknowledge-US/CALSE-100*.

[3] *https://github.com/Deepknowledge-US/TAL-IA*.

K = 1, 5, 10. This evaluation is conducted across all sign instances.

5 Results

5.1 Experimental set

A first experiment was performed with the complete CALSE-100 set without data augmentation during training, setting aside a total of 100 original videos for the test set.

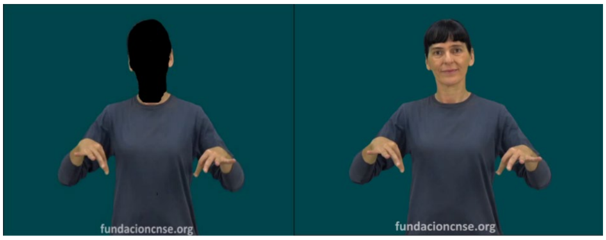

Fig. 6 Result of applying face omission on a DILSE set signer

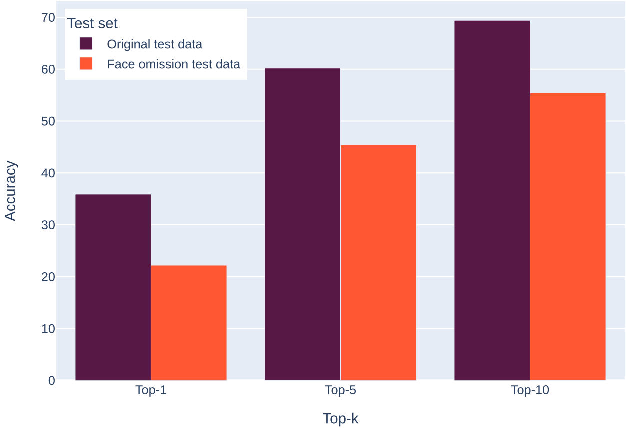

Fig. 7 Training results with original data applied to the original DILSE test set and with face omission

Similarly, the result of this training was tested with the same 100 videos by face-parsing them over the entire face of the signer.

In Fig. 6 we can see the output obtained from applying the facial omission on the DILSE test set. This process has been performed on all the videos in the set through face analysis and segmentation. Source code has been obtained from the Face-parsing [64] repository, which provides PyTorch implementations of common models and algorithms for this kind of tasks.

Figure 7 illustrates the training results obtained from testing with original and face omission videos. The omission of facial expression information leads to a significant decrease in results. This observation underscores the critical role of facial expressions in the model's performance. Additionally, it emphasizes the rationale behind utilizing face swapping instead of alternative techniques such as pixelation or blackening, which may anonymize the data but also remove valuable information in the process.

To assess the effect of data augmentation on CALSE-100 training, a test battery was created that merged original data from the training set with data generated through face swapping and other augmentations. For face swapping, three models (two male and one female) were utilized, with 200 new videos generated for each model. Not all of the face swapping videos created during training were used, thus denoting that FS1 experiment is formed by an increment of 200 videos in which the applied corresponds to model 1. Additionally, affine transformations were randomly applied to each video in the dataset, with AF1 and AF2 differing in the dataset to which they were applied. Specifically, AF1 corresponds to the affine transformation applied to our interpreter videos, while AF2 applies to videos from the public datasets.

To evaluate the influence of data augmentation techniques on model performance, the test battery described in Table 2 was applied to different subsets using the cross-dataset approach, where a public repository was selected as the test set, and the remaining videos in the dataset

Table 2 Preliminary experiments. Column AF1 denotes the affine transformation applied to our interpreter videos, while AF2 corresponds to videos from the public datasets

Experiment	Face swapping			Affine transformations		Train size
	Model1	Model2	Model3	AF1	AF2	
Baseline	×	×	×	×	×	500
FS1	✓	×	×	×	×	700
AF1	×	×	×	✓	×	700
AF2	×	×	×	×	✓	700
FS1-AF1	✓	×	×	✓	×	900
FS1-AF2	✓	×	×	×	✓	900

Table 3 Experiment results executed using STS as a test set

Experiment	top-1	top-5	top-10
Baseline	34.6	60.4	71.9
FS1	39.6 (+5)	63.1 (+2.7)	71.6 (−0.3)
AF1	37.6 (+3)	62.9 (+2.5)	72.4 (+0.5)
AF2	34.4 (−0.2)	60.3 (−0.1)	71.2 (−0.7)
FS1-AF1	32.5 (−2.1)	63.4 (+3)	**76.2 (+4.3)**
FS1-AF2	**40.1 (+5.5)**	**64.2 (+3.8)**	74.7 (+2.8)

Best results for each top-k are marked in bold

Table 4 Experiment results executed using DILSE as a test set. Face omission row denotes results when testing with the DILSE set after applying the face omission

Experiment	top-1	top-5	top-10
Baseline	35.9	60.2	69.4
Face omission	22.2	45.4	55.4
FS1	**47.6 (+11.7)**	**69 (+8.8)**	75.8 (+6.4)
AF1	46.3 (+10.4)	67.5 (+7.3)	74.2 (+4.8)
AF2	40.6 (+4.7)	65.6 (+5.4)	**76.8 (+7.4)**
FS1-AF1	37.8 (+1.9)	62.6 (+2.4)	73.8 (+4.4)
FS1-AF2	37.2 (+1.3)	64.7 (+4.5)	73.4 (+4)

Best results for each top-k are marked in bold

Table 5 Experiment results executed using SACU as a test set

Experiment	top-1	top-5	top-10
Baseline	32.2	52.1	62.9
FS1	32.3 (+0.1)	54.1 (+2)	67.1 (+4.2)
AF1	**37.1 (+4.9)**	**59.8 (+7.7)**	**71.9 (+9)**
AF2	32.2 (0)	54.6 (+2.5)	63.7 (+0.8)
FS1-AF1	31.4 (−0.8)	54.7 (+2.6)	68.2 (+5.3)
FS1-AF2	**37.1 (+4.9)**	57.1 (+5)	67.6 (+4.7)

Best results for each top-k are marked in bold

were used for training. Therefore, the experimental series was repeated three times: once with the SACU as test set, once with STS, and finally leaving DILSE out of the training set. By comparing the results obtained with these different subsets, we can determine the effectiveness of the data augmentation techniques in improving model performance.

The first experiment performed could be identified as a baseline experiment because no data were added through face swapping or affine transformations during training. Since there are no previous investigations on the newly created dataset, it is not possible to compare the results obtained with those of other studies. The results of subset 100 of the WLASL dataset, also trained for the I3D model, correspond to **65**.**89**, **84**.**11** and **89**.**92** for top-1, top-5 and top-10 accuracy respectively. It is worth noting that the WLASL100 subset is significantly larger than our dataset, which consists of 600 videos. This subset includes 2038 videos, with over 20 video samples per sign, resulting in a dataset almost four times larger than ours.

5.2 Results analysis

Tables 3, 4 and 5 indicates that, with some exceptions, using data augmentation during training improves the results. For training by isolating the DILSE set for test, as we can see in Table 4 an improvement always occurs regardless of the type of augmentation used, achieving an increase of up to 11.7 points in the top-1 accuracy with respect to the baseline experiment without augmentations, even though it is not the largest training set used.

On the other hand, Table 5 reflects that unanimously in the top-1, top-5 and top-10, for that training set the best performing augmentation is the use of affine transformations over the set of our collaborating interpreter.

Table 6 shows a summary of the results obtained in all the experiments performed by dataset. As can be seen, it is generally the DILSE set as a test that obtains the best results in top-1, top-5 and top-10 accuracy, with respect to SACU these results being up to 15.3 points higher.

Figure 8 displays the percentage improvement in the top-1 accuracy metric achieved by including various

Table 6 Results obtained in all the experiments performed

Experiment	Top-1			Top-5			Top-10		
	STS	DILSE	SACU	STS	DILSE	SACU	STS	DILSE	SACU
Baseline	34.6	**35.9**	32.2	**60.4**	60.2	52.1	**71.9**	69.4	62.9
FS1	39.6	**47.6**	32.3	63.1	**69**	54.1	71.6	**75.8**	67.1
AF1	37.6	**46.3**	37.1	62.9	**67.5**	59.8	72.4	**74.2**	71.9
AF2	34.4	**40.6**	32.2	60.3	**65.6**	54.6	71.2	**76.8**	63.7
FS1-AF1	32.5	**37.8**	31.4	**63.4**	62.6	54.7	**76.2**	73.8	68.2
FS1-AF2	**40.1**	37.2	37.1	64.2	**64.7**	57.1	**74.7**	73.4	67.6

Best results for each top-k are marked in bold

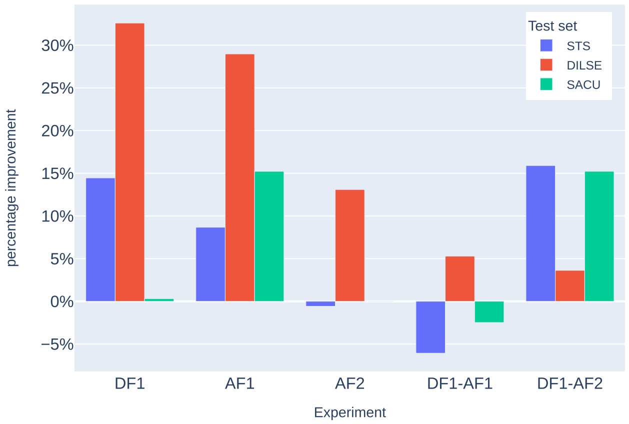

Fig. 8 Top-1 percentage improvement per experiment in each test subset

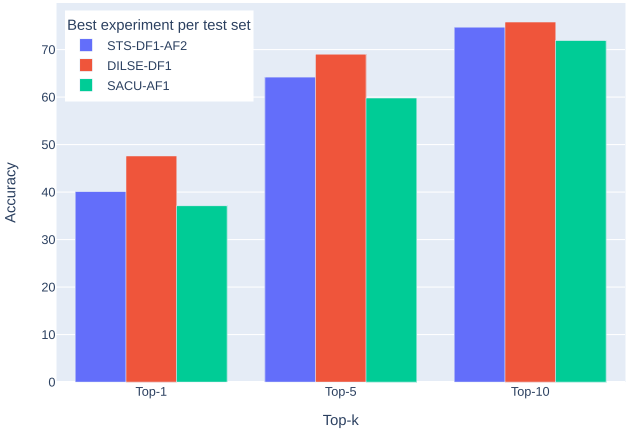

Fig. 9 Best result for each isolated set

augmentations during training, relative to the baseline experiment, for each test set evaluated separately.

As can be seen, the use of data augmentation during training can produce an improvement of up to 32.59% in top-1 accuracy with respect to its baseline.

Furthermore, it is evident that while the combination of face swapping and affine transformations in the same training process may improve results, it is when they are used separately that the highest performance is achieved. This isolated usage of face swapping and affine transformations is also beneficial in terms of reduced training set size, leading to shorter execution times.

Worse results are observed when face swapping videos are added to our signer dataset using three different models, as well as when face swapping videos are combined with affine transformations on our collaborator interpreter. This degradation in results may be attributed to the fact that during training, an excessive number of videos with augmentations applied to our interpreter are included, which can lead to a loss of dataset generalization.

Finally, Fig. 9 presents the best results achieved in each subset, where different augmentations were applied during training. As already shown in Table 6, it is the DILSE set as a test that achieves the highest accuracy in top-1, top-5, and top-10 metrics, with almost 50% accuracy in the first most probable prediction.

6 Conclusions and future work

This study introduces a novel CALSE-100 dataset for Spanish Sign Language, comprising 100 words, on which the I3D architecture was employed to assess the accuracy of Sign Language Recognition. This dataset gathers 600 videos in different scenarios, using diverse perspectives during sign execution and consisting of more than 15 signers.

Since Deep Learning techniques require a large amount of data, various data augmentation approaches such as affine transformations and face swapping were also suggested to enhance accuracy. The cross-dataset approach was employed to conduct the same experiments on different training and test sets. Our proposal showed that augmentations during training generally improved the accuracy in the top-1, top-5, and top-10 metrics compared to the baseline experiment. The improvements ranged up to 32.59% in the top-1 metric.

The importance of facial expressions in the model was confirmed by testing with a facial omission set, so, the incorporation of face swapping videos during training not only improved accuracy, but also ensured user anonymity while preserving facial information during sign execution. On the other hand, it is necessary to consider when it is interesting to use face swapping in this type of study. Creating realistic face swapping videos is an open problem that also presents difficulties, such as the amount of data needed to train and enable the execution of a model, how laborious and impractical it can be to train multiple identities or the high processing times required by this type of technique.

In future work, our focus will be on exploring alternative architectures for Sign Language Recognition, that will allow us to improve the accuracy of the results.

Acknowledgements We would like to thank José Antonio Rodríguez-Gallego who worked in the library to augment videos and Macarena Vilches-Suárez who is our collaborating interpreter an has signed part of the CALSE dataset. We are also grateful to the University Community Assistance Service (SACU) of the University of Seville for allowing us to use their videos and to our colleague Miguel A. Martínez-del-Amor, to whom NVIDIA donated an A100 GPU that we used in the experiments.

Author contributions All authors contributed to the study conception and design. Material preparation were performed by Marina Perea-Trigo, Enrique J. López-Ortiz, Juan A. Álvarez-García y Luis M.

Soria-Morillo. Data collection and analysis were performed by Marina Perea-Trigo. The first draft of the manuscript was written by Marina Perea-Trigo and all authors commented, reviewed and edited previous versions of the manuscript. All authors read and approved the final manuscript. Supervision of the entire research and manuscript elaboration process was carried out by Juan A. Álvarez-García, Luis M. Soria-Morillo and J.J. Vegas-Olmos.

Funding Funding for open access publishing: Universidad de Sevilla/CBUA. This research has been partially funded by the projects DISARM (PDC2021-121197-C21) and HORUS (PID2021-126359OB-I00) funded by MCIN / AEI / 10.13039 / 501100011033, by the "European Union NextGenerationEU/PRTR" and Project ID2PPAC funded through the ECSEL framework program (Grant agreement ID: 101007254).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no Conflict of interest.

References

1. Organization, W. H., et al.: World report on hearing, World Health Organization (2021)
2. Peery, M.L.: World federation of the deaf, Encyclopedia of Special Education: A Reference for the Education of Children, Adolescents, and Adults with Disabilities and Other Exceptional Individuals. Wiiley, Hoboken (2013)
3. del Estado, J.: Ley 27/2007, de 23 de octubre, por la que se reconocen las lenguas de signos españolas y se regulan los medios de apoyo a la comunicación oral de las personas sordas, con discapacidad auditiva y sordociegas. Boletín Oficial del Estado. **255**(24), 10 (2007)
4. Baker, A., van den Bogaerde, B., Pfau, R., Schermer, T.: The Llinguistics of Sign Languages: An Introduction. John Benjamins Publishing Company, Amsterdam (2016)
5. Rastgoo, R., Kiani, K., Escalera, S.: Sign language recognition: a deep survey. Expert Syst. Appl. **164**, 113794 (2020)
6. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 7784–7793 (2018)
7. Luqman, H., Mahmoud, S.A.: Automatic translation of Arabic text-to-Arabic sign language. Univ. Access Inf. Soc. **18**, 939–951 (2019)
8. Rastgoo, R., Kiani, K., Escalera, S., Sabokrou, M.: Sign language production: a review, arXiv preprint arXiv:2103.15910 (2021)
9. Kahlon, N.K., Singh, W.: Machine translation from text to sign language: a systematic review. Univ. Access Inf. Soc. **22**(1), 1–35 (2023)
10. Bragg, D., Koller, O., Caselli, N., Thies, W.: Exploring collection of sign language datasets: privacy, participation, and model performance, In: The 22nd International ACM SIGACCESS Conference on Computers and Accessibility, pp. 1–14 (2020)
11. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
12. Al-Rousan, M., Assaleh, K., Tala'a, A.: Video-based signer-independent Arabic sign language recognition using hidden Markov models. Appl. Soft Comput. **9**(3), 990–999 (2009)
13. Pigou, L., Dieleman, S., Kindermans, P.-J., Schrauwen, B.: Sign language recognition using convolutional neural networks, In: Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13, Springer, pp. 572–578 (2015)
14. Huang, J., Zhou, W., Li, H., Li, W., Sign language recognition using 3d convolutional neural networks, In: IEEE International Conference on Multimedia and Expo (ICME). IEEE 2015, pp. 1–6 (2015)
15. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., Giro-i Nieto, X.: How2sign: a large-scale multimodal dataset for continuous American sign language, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2735–2744 (2021)
16. De Sisto, M., Vandeghinste, V., Gómez, S.E., De Coster, M., Shterionov, D., Seggion, H.: Challenges with sign language datasets for sign language recognition and translation, In: LREC2022, the 13th International Conference on Language Resources and Evaluation, pp. 2478–2487 (2022)
17. Sincan, O.M., Keles, H.Y.: Autsl: A large scale multi-modal Turkish sign language dataset and baseline methods. IEEE Access **8**, 181340–181355 (2020)
18. Li, D., Rodriguez, C., Yu, X., Li, H.: Word-level deep sign language recognition from video: a new large-scale dataset and methods comparison. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 1459–1469 (2020)
19. Martínez, A.M., Wilbur, R.B., Shay, R., Kak, A.C.: Purdue rvl-slll asl database for automatic recognition of american sign language, In: Proceedings. Fourth IEEE International Conference on Multimodal Interfaces, IEEE, pp. 167–172 (2002)
20. Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Yuan, Q., Thangali, A.: The American sign language lexicon video dataset, In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, pp. 1–8 (2008)
21. Caselli, N.K., Sehyr, Z.S., Cohen-Goldberg, A.M., Emmorey, K.: Asl-lex: a lexical database of American sign language. Behav. Res. Methods **49**(2), 784–801 (2017)
22. Joze, H.R.V., Koller, O.: Ms-asl: A large-scale data set and benchmark for understanding American sign language, arXiv preprint arXiv:1812.01053 (2018)
23. Sehyr, Z.S., Caselli, N., Cohen-Goldberg, A.M., Emmorey, K.: The asl-lex 2.0 project: a database of lexical and phonological properties for 2723 signs in American sign language. J. Deaf Stud. Deaf Educ. **26**(2), 263–277 (2021)
24. Neidle, C., Opoku, A., Ballard, C., Dafnis, K. M., Chroni, E., Metaxas, D.: Resources for computer-based sign recognition from video, and the criticality of consistency of gloss labeling across multiple large asl video corpora, In: Proceedings of the LREC2022 10th Workshop on the Representation and Processing

of Sign Languages: Multilingual Sign Language Resources, pp. 165–172 (2022)

25. Ong, E.-J., Cooper, H., Pugeault, N., Bowden, R.: Sign language recognition using sequential pattern trees, In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 2200–2207 (2012)
26. Ebling, S., Camgöz, N.C., Braem, P.B., Tissi, K., S.-M. et al., Smile Swiss German sign language dataset. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
27. Chai, X., Wanga, H., Zhoub, M., Wub, G., Lic, H., Chena, X.: Devisign: dataset and evaluation for 3d sign language recognition, Technical report, Beijing. Techn, Rep (2015)
28. Pu, J., Zhou, W., Li, H.: Sign language recognition with multimodal features, In: Pacific Rim Conference on Multimedia, Springer, pp. 252–261 (2016)
29. Mejía-Peréz, K., Córdova-Esparza, D.-M., Terven, J., Herrera-Navarro, A.-M., García-Ramírez, T., Ramírez-Pedraza, A.: Automatic recognition of Mexican sign language using a depth camera and recurrent neural networks. Appl. Sci. **12**(11), 5523 (2022)
30. Ronchetti, F., Quiroga, F., Estrebou, C., Lanzarini, L., Rosete, A.: Lsa64: a dataset of argentinian sign language,In: XX II Congreso Argentino de Ciencias de la Computación (CACIC) (2016)
31. Özdemir, O., Kındıroğlu, A.A., Camgöz, N.C., Akarun, L.: Bosphorussign22k sign language recognition dataset, 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, pp. 181–188 (2020)
32. Gutierrez-Sigut, E., Costello, B., Baus, C., Carreiras, M.: Lse-sign: a lexical database for Spanish sign language. Behav. Res. Methods **48**(1), 123–137 (2016)
33. Docío-Fernández, L., Alba-Castro, J.L., Torres-Guijarro, S., Rodríguez-Banga, E., Rey-Area, M., Pérez-Pérez, A., Rico-Alonso, S., García-ateo, C.: 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, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 45–52. https:// aclanthology.org/2020.signlang-1.8. LSE_UVIGO: a multi-source database for Spanish Sign Language recognition
34. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. big data **6**(1), 1–48 (2019)
35. Perez, L., Wang, J.: The effectiveness of data augmentation in image classification using deep learning, arXiv preprint arXiv:1712.04621 (2017)
36. Marchesi, M.: Megapixel size image creation using generative adversarial networks, arXiv preprint arXiv:1706.00082 (2017)
37. Illarionova, S., Nesteruk, S., Shadrin, D., Ignatiev, V., Pukalchik, M., Oseledets, I.: Object-based augmentation for building semantic segmentation: Ventura and santa rosa case study. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1659–1668 (2021)
38. Boháček, M., Hrúz, M.: Sign pose-based transformer for word-level sign language recognition, In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 182–191, (2022)
39. Sharma, S., Singh, S.: Recognition of indian sign language (isl) using deep learning model. Wireless Pers. Commun. **123**(1), 671–692 (2022)
40. Dima, T.F., Ahmed, M.E.: Using yolov5 algorithm to detect and recognize American sign language, In: 2021 International Conference on Information Technology (ICIT), IEEE, pp. 603–607 (2021)
41. An, C., Han, E., Noh, D., Kwon, O., Lee, S., Han, H.: Building Korean sign language augmentation (Kosla) corpus with data augmentation technique, arXiv preprint arXiv:2207.05261 (2022)
42. Moryossef, A., Yin, K., Neubig, G., Goldberg, Y.: Data augmentation for sign language gloss translation, arXiv preprint arXiv:2105.07476 (2021)
43. Kietzmann, J., Lee, L.W., McCarthy, I.P., Kietzmann, T.C.: Deepfakes: Trick or treat? Bus. Horiz. **63**(2), 135–146 (2020)
44. Nguyen, T.T., Nguyen, C.M., Nguyen, D.T., Nguyen, D.T., Nahavandi, S.: Deep learning for deepfakes creation and detection, arXiv preprint arXiv:1909.11573 1 (2019) 2
45. Liu, M.-Y., Huang, X., Yu, J., Wang, T.-C., Mallya, A.: Generative adversarial networks for image and video synthesis: algorithms and applications. Proc. IEEE **109**(5), 839–862 (2021)
46. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders, arXiv preprint arXiv:1511.05644 (2015)
47. Mirsky, Y., Lee, W.: The creation and detection of deepfakes: a survey. ACM Comput. Surv. (CSUR) **54**(1), 1–41 (2021)
48. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: Advances in Neural Information Processing Systems, vol. 32, pp. 7137–7147 (2019)
49. Sun, Q., Tewari, A., Xu, W., Fritz, M., Theobalt, C., Schiele, B.: A hybrid model for identity obfuscation by face replacement. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 553–569 (2018)
50. Hao, K.: The biggest threat of deepfakes isn't the deepfakes themselves. MIT Technol. Rev. **21**, 2022 (2019)
51. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 8110–8119 (2020)
52. Rudge, L.A.: Analysing British sign language through the lens of systemic functional linguistics, Ph.D. thesis, University of the West of England (2018)
53. Bleicken, J., Hanke, T., Salden, U., Wagner, S.: Using a language technology infrastructure for German in order to anonymize German sign language corpus data, In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 3303–3306 (2016)
54. Xia, Z., Chen, Y., Zhangli, Q., Huenerfauth, M., Neidle, C., Metaxas, D.: Sign language video anonymization, In: sign-lang@ LREC 2022, European Language Resources Association (ELRA), pp. 202–211 (2022)
55. Saunders, B., Camgoz, N.C., Bowden, R.: Anonysign: Novel human appearance synthesis for sign language video anonymisation, In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), IEEE, pp. 1–8 (2021)
56. Waqas, N., Safie, S.I., Kadir, K.A., Khan, S., Khel, M.H.K.: Deepfake image synthesis for data augmentation. IEEE Access **10**, 80847–80857 (2022)
57. Ramis, S., Buades, J.M., Perales, F.J., Manresa-Yee, C.: A novel approach to cross dataset studies in facial expression recognition. Multimed. Tools Appl. **81**(27), 39507–39544 (2022)
58. Faceswap, https://github.com/deepfakes/faceswap, Accessed: 2022-11-15 (2023)
59. Phaze-a model, https://forum.faceswap.dev/viewtopic.php?f=27&t=1525, Accessed: 2022-11-15 (2023)
60. DILSE, Diccionario de la lengua de signos española, https://fundacioncnse-dilse.org/ (2023)
61. StS, Spread the sign, http://www.spreadthesign.com/es.es/search/ (2023)
62. de Asistencia S.: A la Comunidad Universitaria, Sacu. disability: Spanish sign language glossary, https://sacu.us.es/ne-prestaciones-discapacidad-glosario (2023)

63. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014)
64. Face-parsing, https://github.com/zllrunning/face-parsing.PyTorch/tree/master, Accessed: 2022-11-15 (2023)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.