Leveraging Unannotated Sign Language Data via a Robust Data Augmentation Method for Contrastive Representation Learning

Ariel Basso Madjoukeng, Pierre Poitier, Edith Belise Kenmogne
Adélaïde Couplet, Margaux Leleu, Benoît Frénay
NADI/HuMaLearn at University of Namur
Rue Grangagnage 21, Namur, Belgium
arielbassodev@gmail.com

Abstract

Contrastive learning is a deep learning paradigm that allows learning of useful representations without annotations. In many fields, including sign language recognition (SLR), contrastive approaches have proven to be very effective for developing pretrained models. To learn representations, they generate augmented variants of an instance through augmentation techniques and then maximize their similarities. The quality of the learned representations is strongly correlated with the augmentations used during training. In several fields, specialized augmentations have been developed and adopted. However, in SLR, we observed two trends: contrastive-based SLR approaches often rely on augmentations that are not realistic for the application (e.g., vertical flip, excessive rotations); specialized augmentation methods lack of robustness. Hence, when used as a starting point for contrastive algorithms, the learned representations are often irrelevant and sometimes sensitive. These issues considerably affect the accuracy of SLR models on downstream tasks. In response, this paper proposes a robust augmentation method specially designed for contrastive approaches applied to SLR. The results show an improvement in accuracy during linear evaluation and semi-supervised learning with only 30% of annotations.

Keywords: Sign language recognition, Contrastive representation learning, Data augmentation

1. Introduction

Worldwide, there exist more than 300 sign languages, each with its own properties. In this field, the annotation process is a laborious and time-consuming task that requires linguistic expertise. Studies report around 100 hours of work to annotate one hour of video (Renz et al., 2021). Among the existing sign languages, many are underrepresented. Even when they are represented, they are often partially annotated (Fink et al., 2021; Albanie et al., 2021). For this reason, sign language suffers from a scarcity of annotated data. Faced with this limitation, recent studies increasingly leverage unannotated data to build pretrained models (Wong et al., 2025; Jiang et al., 2024; Gueuwou et al., 2025). These models are designed to learn meaningful representations from unannotated data, then be fine-tuned with fewer annotations and achieve good results. To build pretrained models, several paradigms exist; contrastive representation learning is one of the most well-known (He et al., 2020). It is simple, flexible, and requires fewer computational resources compared to language-model-based approaches (Chen et al., 2020).

Contrastive approaches learn by maximizing the similarity between augmented views of an instance (positive pairs) while minimizing their similarity with other instances (negative pairs). They have the particularity of learning representations that are invariant to the augmentations used during the pretraining. To be useful, augmentations used during training should reflect the target application (Mansfield et al., 2023; Madjoukeng et al., 2025b). Hence, to enable the learning of relevant representations, specialized augmentation methods have been proposed in several fields (time series forecasting, plant anomaly detection, etc.). These specialized augmentations generate positive pairs by degrading the parts of the sample that are not relevant for the application (e.g., background in the case of images (Madjoukeng et al., 2025b,a)).

In SLR, Madjoukeng et al. (2026) observed that due to coarticulation (Poitier et al., 2024) and repositioning movements, not all parts of sign videos are relevant for their identification. Based on this observation, they generate positive pairs by degrading the non-relevant parts of the videos while keeping the relevant parts unchanged. This strategy enables contrastive approaches to learn representations that are invariant to these non-informative segments, thereby improving their overall quality. This method has led to a significant success. However, in real-world SLR scenarios, there exist variabilities between signers (hand shapes, positions, speed, etc.) and due to factors such as occlusions or others, noise may affect the parts that are relevant parts of signs. For a robust and accurate representation learning, augmentation methods should not focus solely on non-relevant parts but also consider variations in the relevant ones. Existing augmentation, by focusing solely on relevant parts completely ignores these factors. This results in representations that lack robustness, fail

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 10–16
16 May 2026.

to account for real-world scenarios, and poor accuracy in downstream tasks. To the best of our knowledge, these issues have not been addressed in SLR. In response, this paper presents a robust augmentation method designed to satisfy two objectives: (i) generating positive pairs while degrading non-relevant parts, and (ii) simulating subtle and realistic variations on the relevant parts. When applied to a contrastive approach, this method allows learning useful and robust representations from unannotated data. By allowing better leveraging of unannotated data, the proposed method is highly valuable for low-resource sign languages. The results show an improvement in accuracy during the linear evaluation and fine-tuning across several sign language datasets.

The rest of this paper is organized as follows: Section 2 presents contrastive approaches; Section 3 presents SLR; Section 4 presents existing data augmentations for SLR; Section 5 presents the proposed augmentation and Section 6 presents the results and discussion.

2. Contrastive Representation Learning

In contrastive learning, several methods exist; SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) are two popular approaches. They learn through four common steps: positive pairs generation, encoding, projection and loss computation. For each instance, augmentation methods are applied to generate positive pairs. They are then passed through an encoder (e.g., ResNet, Transformer) followed by a projection head that computes numerical representations. On these representations, a contrastive loss is computed to maximize the similarity between positive pairs while minimizing their similarity with other instances. The main difference between these approaches is that MoCo stores instances from previous iterations in a queue, increasing the quality of the learned representations. Furthermore, several contrastive methods that rely only on positive pairs have emerged. Among them, BYOL (Grill et al., 2020), SimSiam (Chen and He, 2021), and SL-FPN (Madjoukeng et al., 2026) are three well-known approaches. BYOL uses two encoders: the online and the target encoder. During training, the online encoder is trained to predict the output of the target encoder. SimSiam uses a single encoder for the positive pairs and, following a Siamese architecture (Koch et al., 2015), aligns the representations of the different views. SL-FPN leverages positive pairs and original instances, aligning their representations. These approaches have been used to build pretrained SLR models and have achieved great success. The next section presents previous works on SLR.

3. Sign Language Recognition

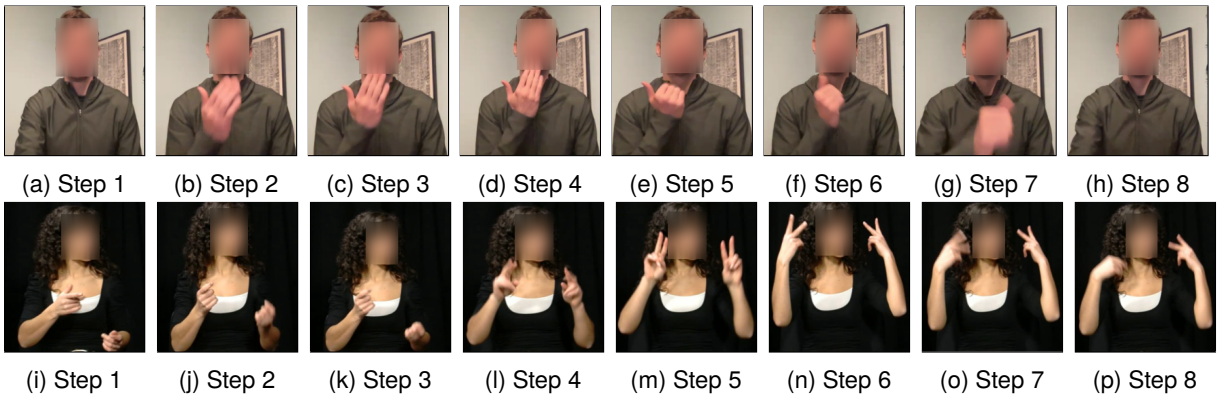
For several sign languages, depending on the amount of data available, SLR models have been proposed. For the French Belgian Sign Language (LSFB), Fink et al. (2021) conducted pioneering work, introducing a dataset consisting of more than 4,567 distinct signs. Using this dataset, they leveraged a Vision Transformer (Dosovitskiy et al., 2020) and achieved 54.4% accuracy on a subset of 700 different signs. For the Argentinian sign language (LSA), Masood et al. (2018) proposed a dataset consisting of multiple instances of 64 different signs. Building upon this dataset, Alyami et al. (2024) benchmarked several architectures and reported 98.25% accuracy with a Transformer-based model. For the Greek sign language, Adaloglou et al. (2021) presented a comprehensive dataset consisting of 310 different signs. Using this dataset, they combined an I3D (Carreira and Zisserman, 2017) with a BiLSTM (Huang et al., 2015) and achieved 89.74% accuracy. Papadimitriou et al. (2024) proposed a multimodal approach with both appearance and skeleton information and achieved 96.21% accuracy.

For the American sign language (ASL), Desai et al. (2023) presented the ASL-Citizen dataset, which consists of 2,731 signs. On this dataset, they trained an ST-GCN and achieved 59.52% accuracy. Furthermore, Jiang et al. (2024) proposed SignCLIP, an architecture that leverages two modalities (text and video), with one encoder for each modality and aligns their representations using a contrastive loss. On this dataset, they achieved 46% accuracy during fine-tuning. The main challenge with this approach is the fact that it requires a vast amount of data from diverse sources to be trained.

In many cases, sign language datasets are not fully annotated (Albanie et al., 2021; Fink et al., 2021). Often, practitioners prefer a model that can be trained on unannotated data from their specific sign language and fine-tuned on data of the same type. Contrastive approaches have been shown to be very effective for this task (Chen et al., 2020; He et al., 2020; Grill et al., 2020). However, the relevance of the representations learned by these methods is directly linked to the augmentations used during pretraining. For SLR, several augmentations are commonly used. The next section presents them.

4. SLR Data Augmentation Methods

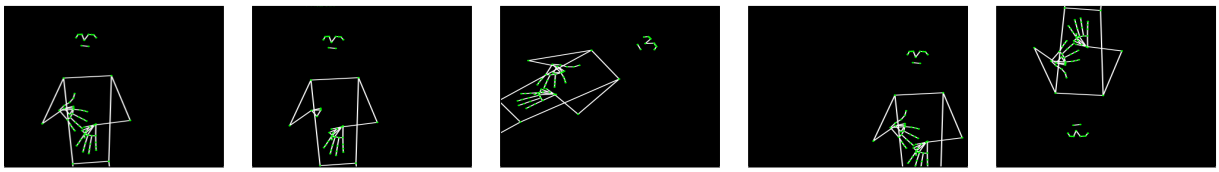
In SLR, data augmentation methods are generally divided into two categories: spatial and temporal augmentations. Spatial augmentations are those that modify the spatial distribution of frames. They include Gaussian noise, translation, flipping, rota-



(a) Step 1 (b) Step 2 (c) Step 3 (d) Step 4 (e) Step 5 (f) Step 6 (g) Step 7 (h) Step 8

(i) Step 1 (j) Step 2 (k) Step 3 (l) Step 4 (m) Step 5 (n) Step 6 (o) Step 7 (p) Step 8

Figure 1: Repositioning and coarticulation mouvement on the ASL (Desai et al., 2023) (first row), and LSFB (Fink et al., 2021) (second row). For privacy reasons, the faces of the signers have been blurred.



(a) Original sign (b) Masking (c) Rotation (d) Translation (e) Vertical flip

Figure 2: A frame and its augmented variants extracted from a video of the LSFB dataset using MediaPipe (Lugaresi et al., 2019).

tion (Lingg et al., 2022) and many others. Figure 2 shows an example of spatial augmentations applied to a skeleton. This skeleton corresponds to one of the frames extracted from a video in the LSFB dataset and processed using MediaPipe (Lugaresi et al., 2019).

Temporal augmentations alter the temporal dynamics of video sequences, including frame permutation, random frame dropping, frame speed-up, and frame slow-down. These augmentations are generic and, when used in contrastive learning for SLR to generate positive pairs, do not allow the learning of targeted representations. Faced with this issue, Madjoukeng et al. (2026) proposed a data augmentation method specifically designed for contrastive learning applied to SLR. They observe that in sign language videos, due to certain movements (repositioning, coarticulation (Poitier et al., 2024), etc.), not all parts are relevant for classification. Based on this observation, they propose a data augmentation method specially designed for contrastive approaches in SLR. Their method consists of generating positive pairs by altering the order of the frames in the non-relevant parts while keeping the relevant frames unchanged. This augmentation has proved to be effective with contrastive algorithms in SLR. However, the fact that any transformation is applied on the relevant parts of the videos represent a major limitation. Indeed, contrastive algorithms are therefore not constrained to be robust to small perturbations or to variations in the signer's position on these crucial segments. Due to the highly visual and dynamic nature of the movements in the SLR context, this limitation is particularly problematic. To address this limitation, the next section introduces a dedicated robust augmentation method.

5. Proposed Augmentation Method

To learn meaningful representations, contrastive approaches require data augmentations capable of generating positive pairs while taking into account the particularities of the application domain. In SLR, a video contains both relevant and non-relevant segments. To learn meaningful representations, a contrastive approach should be able to ignore non-relevant segments while being efficient (robust to subtle perturbations, variations, etc.) on the relevant ones. To enable contrastive approaches to achieve these objectives, this section introduces a novel data augmentation specifically designed for contrastive approaches in SLR. It first explains why not all parts of a video are relevant for a classification model. Then, it describes an existing method for identifying the relevant and non-relevant segments of a video. Finally, presents the proposed augmentation.

To illustrate the fact that in sign language, not all parts of a video are relevant for sign recognition, let us consider Figure 1. The first row is a video from the ASL- Citizen and the second from

the LSFB dataset. The different videos were split into 8 frames (steps 1 to 8). From this figure, we note that the signers begin the motion in the second frame (step 2) and continues signing until the fifth frame (for the ASL video) and the sixth frame (for the LSFB video). After these frames, the signers tend to reposition their hands. Such movements are not the essence of the sign and a reliable model should not focus on these frames to classify the signs.

At this step, a natural question arises: how to determine the relevant and non-relevant parts of the signs. For this task, Madjoukeng et al. (2026) proposed an empirical strategy based on the variation of the accuracy according to the degradation of the frames. Their approach consists of degrading progressively (through permutations) the frames from the first to the last and from the last to the first. After each permutation, they computed the accuracy. They assume that, for a sign to make sense, the order of the frames is important (due to the continuous nature of the movement). Therefore, if the relevant parts for identifying a sign are altered, the accuracy during the evaluation will be affected. Hence, they determine from which frame the movement starts and ends to be informative for a model.

In their approach, they generated positive pairs by focusing only on the non-relevant segments without helping the contrastive model to learn robust representations on the relevant parts. For contrastive approaches in SLR, an effective augmentation should satisfy at least these two properties: (i) generating positive pairs by degrading non-relevant parts, and (ii) introducing subtle and realistic perturbations on the relevant ones. Degrading non-relevant parts allows contrastive approaches to abstract away non-relevant movements, while subtle perturbations to relevant parts enable contrastive approaches to become robust to factors such as positional variations and joint-level errors that may occur in the data.

The proposed method is specifically designed to meet these criteria. Algorithm 1 summarizes the proposed approach. For a sign x, the relevant (x^r) and non-relevant (x^{nr}) parts are determined using a function called `relevant&non_relevant_parts` following Madjoukeng et al. (2026). The non-relevant parts are split into two parts: the first non-relevant part (x^f) and last non-relevant part (x^l). They are typically corresponding to the coarticulation and repositioning movements. After this decomposition, temporal permutations (π_1, π_2) and (π'_1, π'_2) are applied to the non-relevant parts (x^f, x^l) while a composition of translation and noise ($t_1 \circ g_1$ and $t_2 \circ g_2$) is applied on the relevant parts (x^r). Since the meaning of a sign is linked to the order of the frames in the video, temporal permutation induces the learning of representations that are invariant to the order of the frames in these parts. This encourages the model to ignore these parts. The composition of translation and noise simulates subtle perturbations and enhances the robustness of the models. Indeed, noise simulates inaccuracies arising from pose estimation, sensor noise, while translation models small global shifts in hand and body position that commonly occur across different signers and recording setups. By applying these perturbations exclusively to relevant frames, the model is encouraged to become robust to natural variability in informative segments. Note that this method can be used by any contrastive approach. The next section presents an evaluation on several contrastive approaches.

Algorithm 1: Proposed Augmentation

Data: $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$: set of signs
Result: $\mathcal{X}'$: augmented sequences
for $x \in \mathcal{X}$ **do**
  $x^r, x^{nr} \leftarrow$ `relevant&non_relevant_parts`(x)
  $x^f, x^l \leftarrow x^{nr}$
  $A_1 \leftarrow t_1 \circ g_1$
  $A_2 \leftarrow t_2 \circ g_2$
  $x_1 \leftarrow \pi_1(x^f) \parallel A_1(x^r) \parallel \pi_2(x^l)$
  $x_2 \leftarrow \pi'_1(x^f) \parallel A_2(x^r) \parallel \pi'_2(x^l)$
  Add (x_1, x_2) to $\mathcal{X}'$
return $\mathcal{X}'$

6. Experiments

This section evaluates the proposed augmentation. It presents the experiments, datasets, results and discussion.

6.1. Conducted Experiments

To evaluate the effectiveness of the proposed approaches, two types of evaluations were used: a linear evaluation protocol and semi-supervised learning. Linear evaluation consists of pretraining a backbone on unannotated sign language data, freezing the backbone, and training a simple classifier on top of it. Semi-supervised learning consists of training a pretrained backbone with only 30% of annotations. This evaluation is important to simulate data scarcity and partially annotated sign language datasets.

6.2. Datasets

For this study, four datasets with different sizes were used. First, the American sign language dataset (Desai et al., 2023) which consists of 83,399 videos covering 2,731 signs. Second, the Greek sign language dataset (Adaloglou et al., 2021) containing 40,785 videos corresponding to 310 unique signs. Third, the Argentinian sign language dataset (Masood et al., 2018) comprising 3,200 videos representing 64 different signs. Finally, the LSFB dataset (Fink et al., 2021) containing more than 47,551 videos spanning 4,567 distinct signs. These datasets were split according to their original papers (e.g., 70/30 for the LSFB, 80/20 for the LSA, etc.). During the experiments, we used all the classes of the LSA and GSL datasets, and a subset of 500 different classes from the LSFB and ASL datasets. Each model was trained five consecutive times and the results were reported with 95% confidence.

6.3. Training setup

For the experiments, Python 3.10 and PyTorch 2.5 were used. The Lightly library was used to implement the different contrastive approaches[1]. For training, a ViT Transformer (Dosovitskiy et al., 2020) was used as the backbone. Indeed, it has been widely used as a backbone in various studies (Fink et al., 2023; Madjoukeng et al., 2026). As in several studies, the videos were first transformed into skeletons using MediaPipe (Lugaresi et al., 2019). This reduces the computational cost and the influence of elements such as the background. The different contrastive approaches were trained for 200 epochs, using the parameters specified in their original papers (Chen et al., 2020; He et al., 2020; Grill et al., 2020). The batch size was set to 512. During the fine-tuning stage, the models were trained for 1000 epochs using an SGD optimizer.

6.4. Results and Discussion

The first evaluation consisted of comparing the proposed method with existing specialized SLR augmentation. For this evaluation, we used five contrastive methods and the different datasets. On each dataset, we evaluated the proposed augmentation against the one proposed by Madjoukeng et al. (2026). Table 1 presents the results obtained. The best results are highlighted in bold. We observe that, in linear evaluation, the proposed approach outperforms existing method in most cases. This shows that, to learn relevant representations, contrastive approaches require augmentation methods that are likely to reflect the phenomena of the specific application domain. In semi-supervised learning, we observe that with our approach, contrastive methods perform better than using existing augmentation. These results are promising and highlight the potential impact of contrastive approaches in addressing the problem of scarcity of annotated data in sign language.

[1] https://docs.lightly.ai/self-supervised-learning/index.html

7. Conclusion

This paper presented a robust data augmentation method designed to enable contrastive approaches to learn relevant representations. The proposed method consists of generating positive pairs by degrading the non-relevant parts of the videos, while applying subtle and realistic perturbations on the relevant ones. The proposed approach has been evaluated with various contrastive approaches on different datasets. The results showed an improvement in terms of accuracy during linear evaluation and semi-supervised learning on various datasets. These results demonstrate the potential of contrastive approaches to leverage unannotated data in sign language. Given the scarcity of annotated data in this field, this advancement represents a major contribution. As further work, we plan to extend the proposed method to continuous sign language recognition.

8. Acknowledgments

The present research benefited from computational resources made available on Lucia, the Tier-1 supercomputer of the Walloon Region, infrastructure funded by the Walloon Region under the grant agreement n°1910247.

Table 1: Comparison against (Madjoukeng et al., 2026) on linear evaluation, fine-tuning with 30% of annotations. The best results are highlighted in bold.

Method	Dataset	SimCLR	MoCo	SimSiam	BYOL	SL-FPN
Linear Evaluation Results						
Madjoukeng et al. (2026)	LSFB	$14.16\% \pm 0.24$	$13.68\% \pm 0.48$	$15.26\% \pm 0.67$	$14.72\% \pm 0.65$	$23.73\% \pm 0.53$
	ASL	$14.13\% \pm 0.42$	$14.69\% \pm 0.35$	$15.91\% \pm 0.56$	$16.43\% \pm 0.96$	$20.46\% \pm 1.21$
	GSL	$34.19\% \pm 0.85$	$36.15\% \pm 0.69$	$32.01\% \pm 0.54$	$34.09\% \pm 0.93$	$47.76\% \pm 0.79$
	LSA	$34.02\% \pm 1.24$	$35.69\% \pm 1.06$	$30.06\% \pm 2.14$	$37.47\% \pm 1.51$	$41.74\% \pm 1.08$
Ours	LSFB	**16.89**$\% \pm 0.33$	**16.13%** ± 0.81	**16.77%** ± 0.39	**17.71%** ± 0.10	$23.90\% \pm 0.81$
	ASL	**16.28%** ± 0.39	**17.09%** ± 0.30	$16.17\% \pm 0.56$	**18.39%** ± 0.20	**22.13%** ± 0.63
	GSL	**36.23%** ± 0.51	**37.14%** ± 0.17	**35.57%** ± 0.59	**36.17%** ± 0.69	$48.57\% \pm 0.42$
	LSA	**37.08%** ± 1.13	$36.21\% \pm 0.89$	$32.22\% \pm 0.19$	$38.58\% \pm 1.11$	**43.98%** ± 0.88
Partial Fine-tuning						
Madjoukeng et al. (2026)	LSFB	$42.69\% \pm 2.50$	$42.23\% \pm 2.14$	$43.69\% \pm 2.69$	$41.40\% \pm 2.04$	$49.93\% \pm 2.98$
	ASL	$47.43\% \pm 0.77$	$47.49\% \pm 0.54$	$47.23\% \pm 0.63$	$47.02\% \pm 0.51$	$49.28\% \pm 0.79$
	GSL	$78.82\% \pm 2.96$	$77.42\% \pm 2.87$	$77.02\% \pm 2.95$	$78.04\% \pm 2.65$	$83.86\% \pm 2.01$
	LSA	$87.69\% \pm 1.48$	$88.04\% \pm 1.69$	$87.96\% \pm 0.06$	$88.64\% \pm 1.36$	$92.76\% \pm 1.63$
Ours	LSFB	**44.57**$\% \pm 1.16$	**44.09**$\% \pm 0.98$	**48.03**$\% \pm 1.07$	$44.14\% \pm 1.14$	**51.97**$\% \pm 0.51$
	ASL	$47.21\% \pm 0.19$	**48.05**$\% \pm 0.34$	**48.13**$\% \pm 0.31$	**48.14**$\% \pm 0.99$	**52.56**$\% \pm 0.98$
	GSL	$79.33\% \pm 1.14$	$78.41\% \pm 0.59$	**79.07**$\% \pm 1.19$	**80.05**$\% \pm 1.02$	$83.07\% \pm 0.60$
	LSA	**89.01**$\% \pm 0.17$	$88.96\% \pm 0.56$	**88.57**$\% \pm 0.36$	**89.17**$\% \pm 0.51$	$92.88\% \pm 0.11$

9. Bibliographical References

Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. 2021. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE transactions on multimedia*, 24:1750–1762.

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. 2021. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*.

Sarah Alyami, Hamzah Luqman, and Mohammad Hammoudeh. 2024. Isolated arabic sign language recognition using a transformer-based model and landmark keypoints. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(1):1–19.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. Proceedings of Machine Learning Research.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758.

Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. 2023. Asl citizen: a community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems*, 36:76893–76907.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Jérôme Fink, Benoît Frénay, Laurence Meurant, and Anthony Cleve. 2021. Lsfb-cont and lsfb-isol: Two new datasets for vision-based sign language recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*.

Jerome Fink, Pierre Poitier, Maxime André, Loup Meurice, Benoît Frénay, Anthony Cleve, Bruno Dumas, and Laurence Meurant. 2023. Sign language-to-text dictionary with lightweight transformer models. In *IJCAI*, pages 5968–5976.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, and Karen Livescu. 2025. Signmusketeers: An efficient multi-stream approach for sign language translation at scale. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22506–22521.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. Signclip: Connecting text and sign language by contrastive learning. *arXiv preprint arXiv:2407.01264*.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, pages 1–30. Lille.

Nico Lingg, Miguel Sarabia, Luca Zappella, and Barry-John Theobald. 2022. Contrastive self-supervised learning for skeleton representations. *arXiv preprint arXiv:2211.05304*.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Ariel Basso Madjoukeng, Jerome Fink, Pierre Poitier, Edith Belise Kenmogne, and Benoît Frénay. 2025a. Benchmarking data augmentation for contrastive learning in static sign language recognition. In *ESANN 2024: 32nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. i6doc. com.

Ariel Basso Madjoukeng, Jérôme Fink, Pierre Poitier, Edith Belise Kenmogne, and Benoit Frenay. 2026. SSL-SLR: Self-supervised representation learning for sign language recognition. *Transactions on Machine Learning Research*.

Ariel Basso Madjoukeng, Edith Bélise Kenmogne, Pierre Poitier, Benoît Frénay, and Jérôme Fink. 2025b. Local-global data augmentation for contrastive learning in static sign language recognition. In *International Symposium on Intelligent Data Analysis*, pages 54–66. Springer.

Philip Andrew Mansfield, Arash Afkanpour, Warren Richard Morningstar, and Karan Singhal. 2023. Random field augmentations for self-supervised representation learning. *CoRR*.

Sarfaraz Masood, Adhyan Srivastava, Harish Chandra Thuwal, and Musheer Ahmad. 2018. Real-time sign language gesture (word) recognition from video sequences using cnn and rnn. In *Intelligent Engineering Informatics: Proceedings of the 6th International Conference on FICTA*, pages 623–632. Springer.

Katerina Papadimitriou, Galini Sapountzaki, Kyriaki Vasilaki, Eleni Efthimiou, Stavroula-Evita Fotinea, and Gerasimos Potamianos. 2024. A large corpus for the recognition of greek sign language gestures. *Computer Vision and Image Understanding*, 249:104212.

Pierre Poitier, Jérôme Fink, and Benoît Frénay. 2024. Towards better transition modeling in recurrent neural networks: The case of sign language tokenization. *Neurocomputing*, 567:127018.

Katrin Renz, Nicolaj C Stache, Samuel Albanie, and Gül Varol. 2021. Sign language segmentation with temporal convolutional networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139. IEEE.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2025. Signrep: Enhancing self-supervised sign representations. *arXiv preprint arXiv:2503.08529*.