Learnt Contrastive Concept Embeddings for Sign Recognition

Ryan Wong[1], Necati Cihan Camgoz[2], Richard Bowden[1]
[1]University of Surrey, United Kingdom
[2]Meta Reality Labs, Switzerland
{r.wong, r.bowden}@surrey.ac.uk, neccam@meta.com

Abstract

In natural language processing (NLP) of spoken languages, word embeddings have been shown to be a useful method to encode the meaning of words. Sign languages are visual languages, which require sign embeddings to capture the visual and linguistic semantics of sign.

Unlike many common approaches to Sign Recognition, we focus on explicitly creating sign embeddings that bridge the gap between sign language and spoken language. We propose a learning framework to derive LCC (Learnt Contrastive Concept) embeddings for sign language, a weakly supervised contrastive approach to learning sign embeddings. We train a vocabulary of embeddings that are based on the linguistic labels for sign video. Additionally, we develop a conceptual similarity loss which is able to utilise word embeddings from NLP methods to create sign embeddings that have better sign language to spoken language correspondence. These learnt representations allow the model to automatically localise the sign in time.

Our approach achieves state-of-the-art keypoint-based sign recognition performance on the WLASL and BOBSL datasets.

1. Introduction

Sign languages are visual languages used by Deaf communities around the world. Each country will typically have its own sign language which differs in their vocabulary. While the vocabulary of lexicons will vary, all sign languages share common attributes in terms of their use of hand shape, motion, 3D space, body posture, facial expression and mouthings in order to communicate. We can think of these as different channels of information.

The visual complexities of sign language make automatic Sign Language Translation (SLT) a challenging task for both natural language processing (NLP) and computer vision. Many methods have been proposed to break the SLT problem down into simpler tasks. These tasks include Isolated Sign Recognition, where the goal is to identify a single sign in a given temporal window or Sign Spotting, which is the task of identifying and temporally locating signs within a continuous sequence. All these tasks are used to help solve the underlying SLT problem of translating a sign video to the spoken language equivalent.

Many approaches tackle sign recognition as an extension to gesture recognition by making use of gesture recognition models [1, 2, 18, 25]. While these approaches have been shown to be successful at classification, for the datasets they are trained on, they do not explicitly create a sign representation associated for the sign label. We, therefore, propose creating learnable sign embeddings which are associated with signs from the target vocabulary, similar to word embeddings used in spoken language NLP. Signs with similar meaning or context are usually visually similar signs [36]. Motivated by this, we leverage word embeddings from NLP to learn better sign embeddings. Word embeddings, such as Word2Vec [28], fastText [4] and GloVe [32], are useful in NLP as the vector representations capture semantic similarities between words, where words with similar meanings are closer in the embedding space. In this paper, we guide the learning of sign embeddings by bringing together signs with similar meanings in the embedding space.

We summarise our contributions as follows: (1) We introduce a Learnt Contrastive Concept (LCC) embedding framework, which is a weakly supervised contrastive learning approach to explicitly learn sign embeddings with capabilities for automatic sign localisation. (2) We propose a novel method to integrate spoken language word embeddings with sign labels to produce a representation that has better sign-to-spoken language correspondence. (3) Our loss function is able to improve skeletal and RGB-based model performance compared to cross entropy loss used in past approaches. (4) Our approach is able to outperform previous skeleton-based state-of-the-art models on the WLASL and BOBSL recognition tasks.

2. Related Work

The advancements in computational power and deep learning have shifted the focus of sign recognition from

hand-crafted features to data-driven methods. Various approaches to solving sign recognition have been explored over the years, such as breaking the problem into subproblems by creating hand and mouthing shapes to be used for sign related tasks [11, 24]. These hand and mouthing shapes have been shown to be useful for both sign recognition and SLT [6, 7], but they require labels to be manually annotated at the frame level or learnt with specialised sign classification models [23]. One possible source of phonetic representation comes from the linguistic annotation in the form of HamNoSys or SignWriting [14, 37]. Such annotation systems provide detailed transcription for sign languages but require expert annotators to transcribe the video which is an expensive and time consuming task.

The development of large scale sign recognition datasets has allowed the exploration of deep learning based approaches. Continuous Sign Recognition datasets were initially created such as RWTH-PHOENIX-Weather-2014 [13] and RWTH-PHOENIX-Weather-2014-T [5] for the task of predicting all signs in a given sign video. Many approaches tackle this task with models trained with Connectionist Temporal Classification (CTC) loss [7, 15, 29].

In this paper we focus on the sign recognition task. The AUTSL [34], WLASL [25] and MSASL [19] datasets have been developed for sign recognition in an isolated setting, as well as developments of sign recognition from coarticulated videos such as BOBSL [2]. To tackle the sign recognition tasks on these datasets, action recognition models are used as inspiration. The inflated 3D ConvNet (I3D) [9] and ResNet2+1D [38] used in human action recognition has proven useful for transfer learning from large action recognition datasets to smaller isolated sign recognition datasets [1, 2, 17]. The pretrained sign models are used as feature extractors for the sign language translation task, reducing the RGB frames to a lower dimensional vector [2, 6].

Such methods that use raw video as input assume the model will learn the person independent features. An alternative approach is to use skeleton-based models to reduce the impact of background noise and the person's appearance. Most skeleton-based approaches make use of pose estimators like OpenPose or MediaPipe to detect and extract human body, hand and facial keypoints [8, 27, 40]. Skeletal inputs have been shown to be useful in Action Recognition tasks using Spatial-Temporal Graph Convolutional Networks to automatically learn the spatial and temporal patterns from the data [39]. Jiang *et al*. proposed the Sign Language Graph Convolution Network (SL-GCN) for sign recognition [18]. One of the drawbacks to skeletal-based models is the requirement for accurate keypoint detection in the presence of motion, making sign recognition highly reliant on the pose estimator's accuracy [31].

Learning sign language specific representations is currently an under explored topic but learning representations via contrastive learning is popular in computer vision [10, 21, 33]. These approaches attempt to learn embeddings where visually similar samples are close and dissimilar ones are far away in the embedding space. Sign language is a fine grain classification task which requires a subtle understanding of the small differences between pose, hand shape, motion and mouthings/face gestures. The multiple channels convey complementary information and machine learning models need to learn how these channels interact to understand the sign content. This is especially challenging since models need to learn features that are signer independent across datasets that may have a fairly low number of signers. Bilge *et al*. look into creating sign representations using zero-shot sign recognition approaches but require large quantities of annotations of textual descriptions of the motion and body pose on these datasets [3]. Albanie *et al*. have demonstrated how I3D feature representations are useful for automatic dense annotations for large-vocabulary sign language videos using cosine similarity to create a dataset of dense sign spottings [30].

While the above approaches have focused on visual similarities in sign, Dafnis *et al*. highlight the importance of linguistics priors [12]. They modify the labels of WLASL sign recognition dataset to provide a 1-1 correspondence between the signs and glosses by comparing the labels of videos from the WLASL dataset to videos from a sign bank. They additionally use an external sign dataset to explicitly learn to detect the start and end frames of the signs as a preprocessing step before sign recognition. In this work, we demonstrate an approach that eliminates the need for preprocessing and automatically localises the signs while classifying them.

Since sign languages are visual languages, we take a different approach and look towards NLP to tackle sign recognition. Word embeddings have been found to be useful to represent spoken words as vectors for text analysis, where words with similar embeddings have similar meaning. Word2Vec [28] and fastText [4] are examples of such methods to train meaningful embedding vectors. Motivated by this, we aim to learn embeddings for the sign recognition task which brings signs that are visually and linguistically similar closer together in the embedding space.

3. Method

As shown in Figure 1, our proposed framework has three core components: an *LCC Embedding*, an *Embedding Similarity Network* and a *Sign Recognition Head* for the sign recognition task. The LCC embedding captures the sign information which has similar functionality to word embeddings in NLP, where similar signs will have similar representations. The embedding similarity network and sign recognition head are used to automatically localise and predict the sign within the sequence. Unlike previous ap-

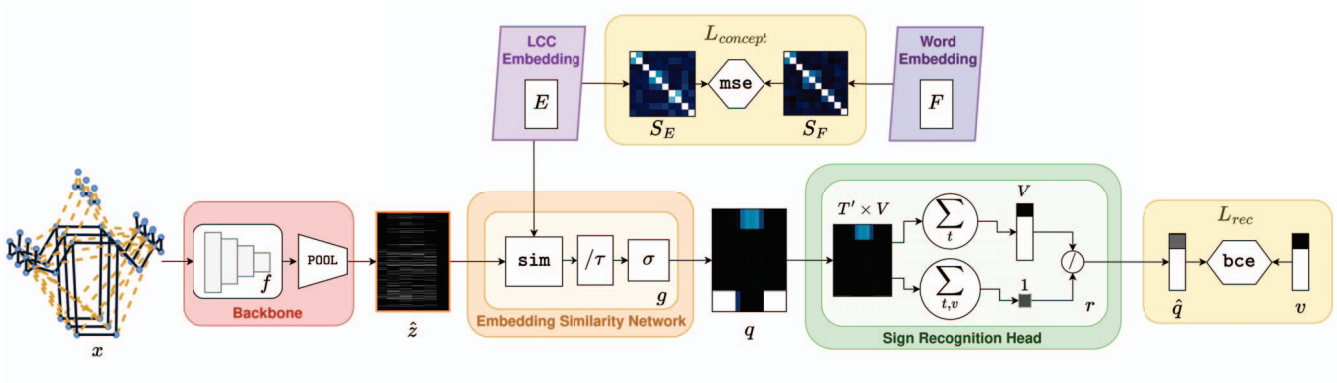

Figure 1. Overview of our proposed Learnt Contrastive Concept (LCC) embedding framework.

proaches, which make use of only the sign labels to indirectly learn sign representations from visual features, we explicitly learn the sign representations and incorporate linguistic knowledge to guide the model to learn visual-linguistic representations.

In the following section, we describe the architectural changes to support our framework. Then we elaborate on the incorporation of our visual-linguistic loss functions namely, a contrastive recognition loss and a conceptual similarity loss.

3.1. Model Architecture

Backbone $f(\cdot)$: Our backbone uses human pose keypoints as input (See Section 3.4 for further details about the input structure). Given a sign sequence $x \in \mathbb{R}^{T \times D \times N}$ of T length with N nodes and feature channels D, we want to maximise the accuracy of our model to predict the target label from a vocabulary V. The most common approach to train such a recognition model is:

$$\hat{y} = \texttt{FC}(\texttt{POOL}_{\texttt{global}}(f(x))) \tag{1}$$

where $\hat{y} \in \mathbb{R}^V$ is the logit class prediction. FC is a fully connected layer and $\texttt{POOL}_{\texttt{global}}$ is the spatio-temporal global average pooling layer which takes the output representation from the model $f(x) = z \in \mathbb{R}^{\frac{T}{\sigma_t} \times C \times \frac{N}{\sigma_n}}$ to a vector $\hat{z} \in \mathbb{R}^C$. σ_t and σ_n are the dimension reduction factors from the backbone network for T and N respectively.

Instead of spatio-temporal global average pooling, we use spatial global average pooling such that $\hat{z} \in \mathbb{R}^{\frac{T}{\sigma_t} \times C}$, which allows our model to make more fine-grained predictions. For simplicity we will refer to $\frac{T}{\sigma_t}$ as T'.

LCC Embedding E: To disentangle the embedding representation from the model representation ($\hat{z}$), we introduce LCC embeddings $E \in \mathbb{R}^{C \times V' \times M}$ as the sign embedding representation, where V' is the selected vocabulary size and M is the number of variations associated with each item in the embedding vocabulary. We select V' where $V < V'$. The motivation behind this choice is that not all representations within the sign sequence should be associated with an embedding in the given target vocabulary (from index 1 to V) as they may be *background*, such as signers in their resting pose, *transitions*, or *out of vocabulary signs*.

Embedding Similarity Network $g(\cdot)$: We introduce an embedding similarity network to allow the model to learn good representations with a strong correlation to the relevant LCC embeddings. $\hat{z}$ and E are compared using the cosine similarity function:

$$c_i = \texttt{sim}_i(\hat{z}_i, E) = \frac{\hat{z}_i \cdot E}{\|\hat{z}_i\| \|E\|} \tag{2}$$

where i is the temporal index of $\hat{z}$ to produce the similarity score $c_i \in [-1, 1]^{V' \times M}$. The resulting score is calculated by taking the average cosine similarity plus maximum cosine similarity across variations M where $\hat{c}_i \in \mathbb{R}^{V'}$. This is computed for each index in T' such that $\hat{c} \in \mathbb{R}^{T' \times V'}$. Using $\hat{c}$ we apply a temperature scalar τ followed by a softmax function as:

$$q = g(z, E) = \texttt{softmax}(\hat{c}/\tau) \tag{3}$$

where $q \in \mathbb{R}^{T' \times V'}$. Our motivation is for the model to explicitly learn representations for the sign vocabulary for each time segment T'.

Sign Recognition Head $r(\cdot)$: The network g requires the label associated to each time segment T', which are not available in sign recognition datasets. Sign recognition datasets only provide information that a sign exists within the given video clip. Isolated sign recognition datasets typically have signers in the resting pose which are irrelevant for the sign label. We, therefore, create a sign recognition head r with the assumption that given a sufficiently large temporal window T for a given sequence x, if we know that the sign v_j from vocabulary V exist somewhere in the sequence x, then we can set the label v_j to our target value.

For our model to learn the existence of a sign we need a function $r(\cdot)$ that takes $q \in \mathbb{R}^{T' \times V'}$ and outputs $\hat{q} \in [0,1]^V$.

More formally, our objective for r is to output $\hat{q}_j = 1$ when the target sign corresponds to the j^{th} element in the vocabulary, and if not output $\hat{q}_j = 0$. To achieve this, we formulate r as:

$$\hat{q} = r(q) = \left[\frac{\sum_{t=0}^{T'} q_{jt}}{\sum_{t=0}^{T'} \sum_{k=1}^{V} q_{kt}} \right]^{j:(1..V)} \tag{4}$$

The output is then a V length vector with values between 0 and 1, which indicate the existence of a sign from our given target vocabulary.

3.2. Learning Objective

Contrastive Recognition Loss L_{rec}: Unlike previous approaches which apply global average pooling from Eq. (1), our approach instead learns to localise the sign within the sequence using our weakly supervised contrastive recognition loss.

From Eq. (4), r allows the model to learn the existence of the sign in the given sequence since we only take the sum across the given target vocabulary $[1, ..., V]$ and not the extended vocabulary $[(V+1), ..., V']$. We apply a binary cross entropy loss for the contrastive recognition loss L_{rec}, where the target is the one-hot encoded vector of the label associated with the sign sequence. This allows our model to bring matching signs closer together within the embedding space while pushing dissimilar signs further apart.

Conceptual Similarity Loss $L_{concept}$: Since sign language is a visual language, there are many signs that look very similar or the same, with small discriminating factors. For example, stomach and abdomen are signed in almost identical ways in the WLASL dataset. Therefore the representations should be similar. Generally speaking, when this is the case, the semantic similarity of the spoken word or gloss should also be similar.

We introduce a new method to integrate sign language with spoken languages. We use fastText embeddings [4] of the glosses and learn visual embeddings which learn the correlation between sign embeddings and spoken language embeddings from the target vocabulary. We propose a conceptual similarity loss $L_{concept}$, where we take the cosine similarity between the fastText embeddings F for the target vocabulary to create a similarity matrix $S_F \in [-1,1]^{V \times V}$. The matrix is used to provide the model with a measure of the similarity of words. We then repeat the process with the LCC embeddings $E_{1..V}$ to create the visual embedding cosine similarity $S_E \in [-1,1]^{V \times V}$. A Mean Squared Error loss minimises the distance between S_F and S_E. This allows our sign embeddings to learn similar linguistic embedding distributions as the word embeddings.

Combined Loss: We apply our learning objectives simultaneously during training. This allows the model to learn visual-linguistic representations and automatically localise signs within the given sequence. We apply a weighted loss L_{rec} and $L_{concept}$ to create our LCC loss $\mathcal{L}$:

$$\mathcal{L} = \alpha * L_{concept} + \beta * L_{rec} \tag{5}$$

3.3. Drop Feature Mask

In the standard classification task, cross entropy loss is used with dropout before the final fully connected layer to improve the model's performance [35]. Since we measure cosine similarity between features, we introduce drop feature masking to our representations $\hat{z}$ and E where we zero out a portion of the same index channels in dimension C for both $\hat{z}$ and E. We also apply this at the temporal level which randomly zeros out the features at certain time segments. This technique allows our model to create a better embedding representation by utilising more of the feature channels.

3.4. Multi-channel Learning

We use a skeleton-based GCN as our backbone model. Skeleton representations, or keypoints, naturally provide a high degree of person independence by discarding irrelevant information, such as background, clothing or appearance. Jiang *et al*. highlights the issues of using a large number of keypoints and employs a graph reduction technique to significantly improve performance on sign recognition [17]. We use an alternative approach where different sign channels (body, hands and face) are input into separate GCN models. This allows our model's graph structure to remain small and gives it the ability to learn from each sign channel separately before fusing the sign features together. We use Mediapipe Holistic [27] to extract the keypoints. We extract 4 different channels which include two hands, of 21 keypoints each, 40 keypoints representing the mouth and 17 keypoints for the body pose. For the links between keypoints we use the default human skeletal connections with additional links on the body for a better hand-to-body interaction as shown in Figure 2.

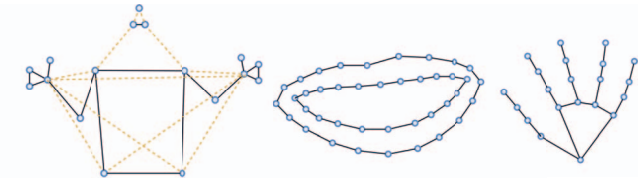

Figure 2. Keypoints extracted from Mediapipe used as the input into the different channels. Additional keypoint links shown by the dashed orange lines were added to allow the model to learn better hand-to-body interactions.

We create three separate GCN models and apply our learning objective from Eq. (5) on the output of the hands

(where each hand has a shared backbone GCN and average the output hand features for $\hat{z}$), mouthing and pose as $\mathcal{L}_h, \mathcal{L}_m$ and $\mathcal{L}_p$. We also concatenate the $\hat{z}$ feature outputs across all GCNs outputs to create a global representation and use a fully connected layer to reduce the dimensions back to the original dimensionality and apply the same learning objective ($\mathcal{L}_{global}$) with a separate global LCC network head (with the LCC embedding, embedding similarity network and sign recognition head). The final learning objective is:

$$\mathcal{L}_{overall} = \mathcal{L}_{global} + \mathcal{L}_h + \mathcal{L}_m + \mathcal{L}_p \tag{6}$$

At inference, the predicted label is selected based on the index of the largest value of $\hat{q}$ from the global LCC network head.

4. Experiments

4.1. Datasets

We evaluate our approach on two different sign recognition tasks: isolated sign recognition and the recognition of a co-articulated sign in continuous footage.

WLASL2000 is a large-scale video dataset for American Sign Language (ASL) recognition. It is a challenging isolated sign recognition dataset as it was collected from unconstrained records with a large vocabulary of 2000 unique signs. Additional challenges are the signer-independent setting with over 100 signers and limited examples for each sign.

BOBSL is a large-scale co-articulated sign dataset for British Sign Language (BSL) obtained from broadcast videos. The recognition task on the BOBSL dataset poses new challenges for sign recognition as continuous co-articulated sign is typically signed faster than in an isolated setting. The dataset contains 2281 sign classes obtained by using automatic spotting tools from mouthings and dictionary sources, which has the added difficulty of noisy labels and large class imbalances.

4.2. Implementation Details

We use the MS-G3N as our backbone GCN model due to its high performance on skeletal action recognition [26]. For simplicity, we keep all model hyperparameters the same across each GCN. We keep the original hyperparameters from [26] but set the number of scales to 5 and use a window size of 5 with a dilation of 1 for the MS-G3D pathways.

We use a sequence length of 64 and 16 for WLASL2000 and BOBSL, respectively. During training, rotation, scaling and shifting are used as keypoint data augmentation. We set the size of V' to be $V + 10$ with α of 5.0 and β of 10.0 for the scaling factors of $L_{concept}$ and L_{rec}, respectively. We train the model with a batch size of 64 with a learning rate of 0.0012. On the WLASL dataset, we schedule our learning rate with a warm-up of 10 epochs and decay with a cosine policy to match the training strategy of [18] on the WLASL dataset, but train our model for 100 epochs instead of 200. We use a multi-step learning rate scheduler for the BOBSL dataset by reducing the learning rate by 0.1 at 10 and 20 epochs and train our model for 25 epochs in total.

The model is optimized using the Adam optimizer [22] with a weight decay of 0.0001, where we select the model which has the best accuracy on the validation set for evaluation on the test set.

4.3. Evaluation Protocol

We evaluate our models on the top-1 and top-5 per-instance classification accuracy as well as the top-1 and top-5 per-class accuracy. Our proposed training strategy is compared to our model trained with cross entropy loss approach, using global average pooling and a classification layer, to directly analyse the impact of our approach. Then, we compare our approach to state-of-the-art keypoints results on both WLASL2000 and BOBSL datasets. We also demonstrate the effectiveness of our framework to RGB-based sign recognition models.

4.4. Results on Isolated Sign Recognition

Models	Instance Acc.		Class Acc.	
	top-1	top-5	top-1	top-5
SignBERT (H+P) [16]	47.46	83.32	45.17	82.32
BEST (Keypoint) [41]	46.25	79.33	43.52	77.65
SL-GCN [18]	51.50	84.94	48.87	84.02
Ours (CE)	51.95	82.52	48.89	81.15
Ours (LCC)	**59.38**	**89.82**	**56.57**	**88.90**

Table 1. Per-instance and per-class accuracy on the WLASL2000 test set for the keypoint modality models. CE and LCC correspond to our models trained with cross entropy and LCC loss, respectively.

Comparisons to baseline: We compare our proposed loss to cross entropy loss used in previous methods on the WLASL2000 dataset. In Table 1, we find that our model trained with cross entropy loss achieves similar performance to the SL-GCN model. The addition of our loss significantly improves the test accuracy by 7.43% top-1 instance accuracy. Furthermore, the proposed approach has the additional benefit of providing localisation of the target sign as shown in Section 4.7.

Comparison to state-of-the-art: In Table 1, we find that our multi-stream keypoint model is able to significantly outperform other keypoint modality models. Our multi-stream

keypoint model is able to outperform the previous multi-stream keypoint model SL-GCN by improving the top-1 instance accuracy by 7.88%. While we are aware of Bidirectional Skeleton-Based Graph Convolutional Networks [12], we are unable to directly compare against it as the approach was evaluated on an unreleased modified WLASL dataset with 1449 lexical signs and contains additional preprocessing to detect the sign's start and stopping frames. Our model has the additional benefit of automatically detecting the sign's start and stopping frames.

We find that our keypoint modality results also achieve competitive results with a multi-modal ensemble such as SAM-SLR v1 and v2 [18, 17] which make use of keypoints, features, RGB frames and RGB flow as input modalities to the multiple models to achieve 59.39% and 91.48% top-1 and top-5 per-instance accuracy respectively.

Further analysis of the individual streams of joint, bone, joint motion and bone motion in Table 2 shows significant individual stream performance improvements compared to SL-GCN. Our joint-based model is also able to outperform all previous single model results.

Stream	Models	Instance Acc.	
		top-1	top-5
Joint	SL-GCN	45.61	77.79
	Ours (LCC)	55.52	86.76
Bone	SL-GCN	43.27	75.58
	Ours (LCC)	54.37	85.78
Joint Motion	SL-GCN	27.23	56.73
	Ours (LCC)	46.56	78.76
Bone Motion	SL-GCN	31.26	60.35
	Ours (LCC)	46.66	76.37
Multi-stream	SL-GCN	51.50	84.94
	Ours (LCC)	**59.38**	**89.82**

Table 2. Comparison between our LCC approach and SL-GCN on individual keypoint streams on the WLASL test set.

4.5. Results on Sign Recognition in continuous footage

Models	Instance Acc.		Class Acc.	
	top-1	top-5	top-1	top-5
2D Pose → Sign [2]	61.8	82.1	30.6	56.6
Ours (CE)	67.8	87.0	34.9	60.9
Ours (LCC)	**71.7**	**89.3**	**37.3**	**64.5**

Table 3. Comparison of accuracy on the BOBSL test set. CE: Our model trained with Cross Entropy loss, LCC: Our model trained with LCC loss.

Comparison to baseline: In Table 3, we show that our model trained with our loss is able to improve multi-stream model performance by almost 4% on the top-1 instance accuracy compared to cross entropy loss on the BOBSL recognition task.

Further analysis of the keypoint modality on the individual streams of joint, bone, joint motion and bone motion in Table 4 shows similar improvements compared to cross entropy loss on individual streams. Our joint based model is able to produce the best individual stream results.

Stream	Loss	Instance Acc.	
		top-1	top-5
Joint	CE	66.28	86.02
	LCC	70.89	88.83
Bone	CE	66.02	85.71
	LCC	70.54	88.61
Joint Motion	CE	61.05	81.60
	LCC	67.49	86.18
Bone Motion	CE	61.02	81.80
	LCC	66.93	85.68
Multi-stream	CE	67.75	86.99
	LCC	**71.72**	**89.32**

Table 4. Comparison between our model trained with cross entropy loss (CE) versus our LCC loss on the BOBSL test set.

Comparison to state-of-the-art: In Table 3, we show that our approach is able to outperform the previous keypoint based model (2D Pose→Sign) by almost 10% top-1 accuracy. All of our individual stream results are able to outperform 2D Pose→Sign by more than 5% top-1 accuracy.

4.6. Ablation Study

We perform our ablation study on the WLASL validation dataset using the joint-based stream for our model.

Impact of temperature: The temperature τ plays an important part in the model's performance. We find that when the temperature is too high, it has a negative impact on the model's performance. In Table 5, $\tau = 0.1$ gives the strongest performance improvements.

Impact of $L_{concept}$: Due to the recognition loss L_{rec} having a contrastive objective to match representations from the model to the target vocabulary embedding, the model tends to push other representations further apart. This is detrimental in cases where signs are visually similar. Our Conceptual Similarity Loss $L_{concept}$ attempts to alleviate these

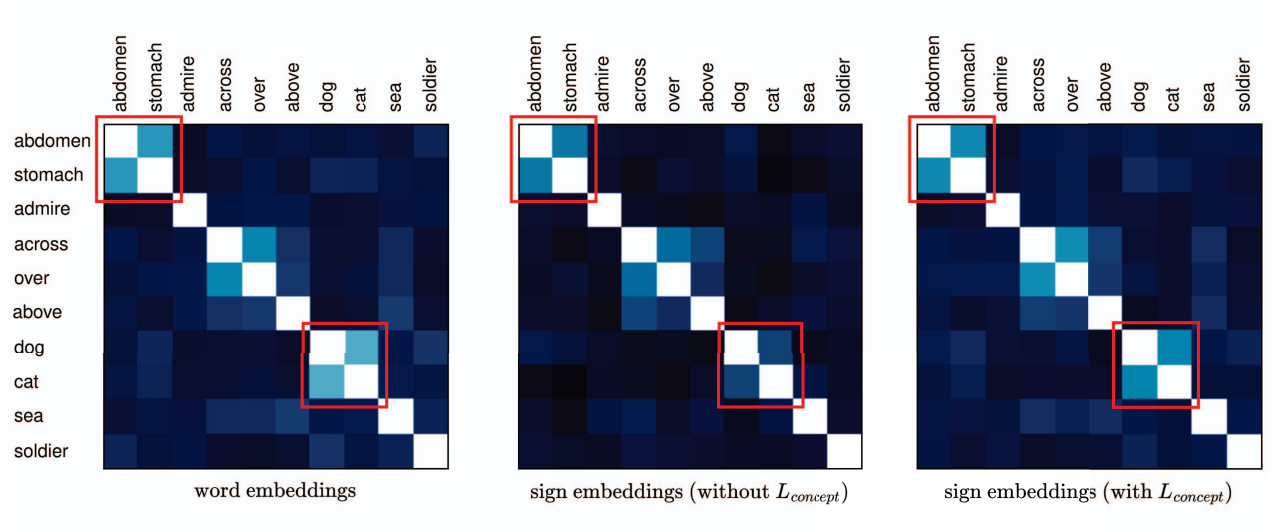

Figure 3. Comparison of similarity matrices for the word embeddings from fastText (left), sign embeddings from a model trained without conceptual similarity loss (middle) and sign embeddings trained with conceptual similarity loss (right). Red regions highlight that the similarity matrix trained with the conceptual similarity has similar linguistic embedding distributions to the word embeddings.

τ	Instance Acc.	
	top-1	top-5
0.5	0.1938	0.4816
0.2	0.3897	0.7181
0.1	**0.5181**	**0.8307**
0.05	0.5023	0.8159

Table 5. Table demonstrating the impact of the temperature value has on the WLASL validation accuracy

issues by creating visual embeddings which match the distances between words from word embeddings. In Table 6, we show that the inclusion of $L_{concept}$ greatly improves the accuracy.

L_{rec}	$L_{concept}$	Instance Acc.	
β	α	top-1	top-5
10.0	0.0	52.76	82.79
10.0	1.0	53.32	84.53
10.0	5.0	**54.29**	**84.60**
10.0	10.0	54.11	83.96

Table 6. Table demonstrating the importance and impact that the cosine similarity loss has on the WLASL validation accuracy

In Figure 3, we show the impact the conceptual similarity loss has on the embeddings by computing the cosine similarity between our sign embeddings. As shown, without the similarity loss the model tends to learn a similar structure to the word embedding. The conceptual similarity loss is able to regularise features in our embedding space to match the distribution of the associated word embedding space. For example, the signs for stomach and abdomen in Figure 3 have higher similarity scores (brighter) with a model trained with $L_{concept}$ (right) than without (middle).

Impact of drop feature masking: We introduce drop feature masking on the embedding features for both the model outputs and vocabulary embedding to create representations that reduce overfitting and over-reliance on certain feature channels. In Table 7, we show that the inclusion of drop masking improves performance.

dfm	Instance Acc.	
	top-1	top-5
–	54.52	85.32
✓	**56.66**	**87.16**

Table 7. Table demonstrating the impact of drop feature masking on WLASL validation results.

Impact of backbone: As an additional study, we evaluate our approach using an RGB backbone model to demonstrate that our framework is input-agnostic. A limitation of multi-channel RGB inputs is the heavy computational requirements. Separate video crops for each sign channel are more memory intensive compared to the keypoint based approach, we therefore use the I3D model with full frames as input. Previous methods have shown that transfer learning of models on other sign language datasets improves sign recognition performance [1, 17]. We, therefore, use Inception I3D pretrained on Kinetic dataset [20] to directly

evaluate the impact of our approach compared to cross entropy loss without any pretraining on external sign language datasets. We apply our learning objective from Eq. (5). For data augmentation during training, we first apply frame resizing to 256×256 then random cropping of 224×224 with random horizontal flipping and colour jitter.

Models	Instance Acc.		Class Acc.	
	top-1	top-5	top-1	top-5
I3D (CE)	41.38	74.98	38.93	73.94
I3D (LCC)	**43.92**	**77.80**	**41.10**	**75.97**

Table 8. Table demonstrating the difference between our LCC loss with a RGB backbone (LCC) versus cross entropy loss (CE) on the WLASL test set.

In Table 8, we show that our proposed loss is able to improve performance compared to cross entropy loss. This experiment demonstrates that our model is capable of generalising to RGB models.

4.7. Sign Localisation

Our model is able to temporally locate the target sign in a given sign sequence using the output value q. In Figure 4, we show the visualisation of the localisation of the signs. We find that in WLASL2000 there are many sequences with signers in their resting pose, which are irrelevant to the target vocabulary. Our model is able to identify those background segments. Similarly, for BOBSL, our model is able to identify when a sign is out of vocabulary and localise the target vocabulary signs.

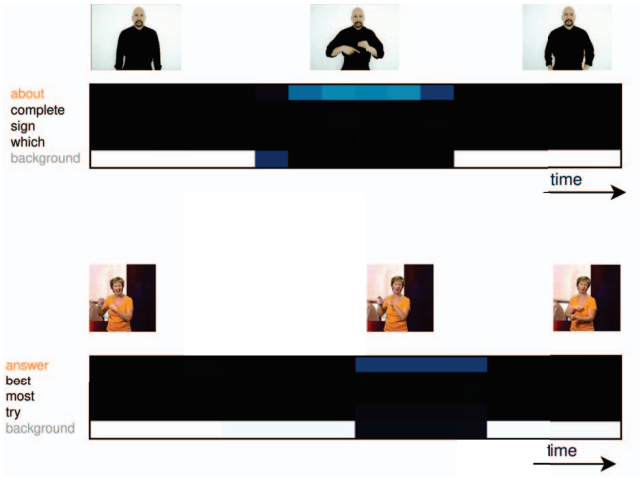

Figure 4. Localisation of identified signs, target word highlighted in orange, where temporal activations are shown in brighter colours. The background class is the sum of the probabilities across extended vocabulary.

4.8. Limitations

One of the limitations of the GCN model is its reliance on the accuracy of keypoint inputs. Past research has highlighted that existing pretrained pose estimators may fail in hand-to-hand or hand-to-face interactions [31].

Secondly, our framework assumes that the conceptual similarity of the words often implies visual similarity in sign language. While we show experiments on the inclusion of the conceptual similarity loss that supports our hypothesis, it may not always be the case for all signs, for example on the signs of names and places.

5. Conclusion

In this work, we introduce Learnt Contrastive Concept embeddings framework, a novel training strategy to learn sign embeddings by employing a weakly supervised contrastive training pipeline that is able to learn sign embeddings from the sign recognition task. We demonstrate the effectiveness of our approach for the localisation of signs and how our framework can be used to improve results on sign recognition compared to cross entropy loss. Our model is able to significantly outperform previous keypoint recognition results on both WLASL and BOBSL datasets. Our approach is able to utilise word embeddings to create sign embeddings that incorporate visual-linguistic features that will hopefully be useful for future work in sign language translation. For future work, exploration of improving the framework for continuous sign recognition may be useful to solve as the next step for automatic sign language translation research.

Acknowledgements

This work was supported by the EPSRC project ExTOL (EP/R03298X/1), SNSF project 'SMILE II' (CRSII5 193686), European Union's Horizon2020 programme ('EASIER' grant agreement 101016982) and the Innosuisse IICT Flagship (PFFS-21-47). This work reflects only the authors view and the Commission is not responsible for any use that may be made of the information it contains

References

[1] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *European Conference on Computer Vision*, pages 35–53. Springer, 2020.

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

[3] Yunus Can Bilge, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Towards zero-shot sign language recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[5] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793, 2018.

[6] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Multi-channel transformers for multi-articulatory sign language translation. In *European Conference on Computer Vision*, pages 301–319. Springer, 2020.

[7] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033, 2020.

[8] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

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

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[11] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3056–3065, 2017.

[12] Konstantinos M Dafnis, Evgenia Chroni, Carol Neidle, and Dimitri Metaxas. Bidirectional skeleton-based isolated sign recognition using graph convolutional networks. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7328–7338, 2022.

[13] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *LREC*, pages 1911–1916, 2014.

[14] Thomas Hanke. Hamnosys-representing sign language data in language resources and language processing contexts. In *LREC*, volume 4, pages 1–6, 2004.

[15] Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11303–11312, 2021.

[16] Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. Signbert: pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11087–11096, 2021.

[17] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Sign language recognition via skeleton-aware multi-model ensemble. *arXiv preprint arXiv:2110.06161*, 2021.

[18] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3413–3423, 2021.

[19] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*, 2018.

[20] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2306–2320, 2019.

[24] Oscar Koller, Hermann Ney, and Richard Bowden. Deep learning of mouth shapes for sign language. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 85–91, 2015.

[25] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 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*, pages 1459–1469, 2020.

[26] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020.

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

[28] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[29] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11542–11551, 2021.

[30] Liliane Momeni, Hannah Bull, KR Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman. Automatic dense annotation of large-vocabulary sign language videos. In *ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 671–690. Springer, 2022.

[31] Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgoz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Muller, and Sarah 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*, pages 3434–3440, 2021.

[32] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[33] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[34] Ozge Mercanoglu Sincan and Hacer Yalim Keles. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355, 2020.

[35] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[36] William C Stokoe. Sign language structure. *Annual review of anthropology*, pages 365–390, 1980.

[37] Valerie Sutton. Signwriting. *Sl: sn*, page 9, 2009.

[38] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[39] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[40] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020.

[41] Weichao Zhao, Hezhen Hu, Wengang Zhou, Jiaxin Shi, and Houqiang Li. Best: Bert pre-training for sign language recognition with coupling tokenization. *arXiv preprint arXiv:2302.05075*, 2023.