Automatic dense annotation of large-vocabulary sign language videos

Liliane Momeni[1⋆], Hannah Bull[2⋆], K R Prajwal[1⋆], Samuel Albanie[3], Gül Varol[4], and Andrew Zisserman[1]

[1] Visual Geometry Group, University of Oxford, UK
[2] LISN, Univ Paris-Saclay, CNRS, France
[3] Department of Engineering, University of Cambridge, UK
[4] LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France
{liliane,prajwal,albanie,gul,az}@robots.ox.ac.uk;
hannah.bull@lisn.upsaclay.fr
https://www.robots.ox.ac.uk/~vgg/research/bsldensify/

Abstract. Recently, sign language researchers have turned to sign language interpreted TV broadcasts, comprising (i) a video of continuous signing and (ii) subtitles corresponding to the audio content, as a readily available and large-scale source of training data. One key challenge in the usability of such data is the lack of sign annotations. Previous work exploiting such weakly-aligned data only found *sparse* correspondences between keywords in the subtitle and individual signs. In this work, we propose a simple, scalable framework to *vastly* increase the *density* of automatic annotations. Our contributions are the following: (1) we significantly improve previous annotation methods by making use of synonyms and subtitle-signing alignment; (2) we show the value of pseudo-labelling from a sign recognition model as a way of sign spotting; (3) we propose a novel approach for increasing our annotations of *known* and *unknown* classes based on *in-domain exemplars*; (4) on the BOBSL BSL sign language corpus, we increase the number of confident automatic annotations from 670K to 5M. We make these annotations publicly available to support the sign language research community.

Keywords: Sign Language Recognition, Automatic Dataset Construction, Novel Class Discovery.

1 Introduction

Sign languages are visual-spatial languages that have evolved among deaf communities. They possess rich grammar structures and lexicons that differ considerably from those found among spoken languages [58]. An important factor impeding progress in automatic sign language recognition – in contrast to automatic speech recognition – has been the lack of large-scale training data. To address this issue, researchers have recently made use of sign language interpreted TV broadcasts, comprising (i) a video of continuous signing, and (ii) subtitles

⋆ Equal contribution

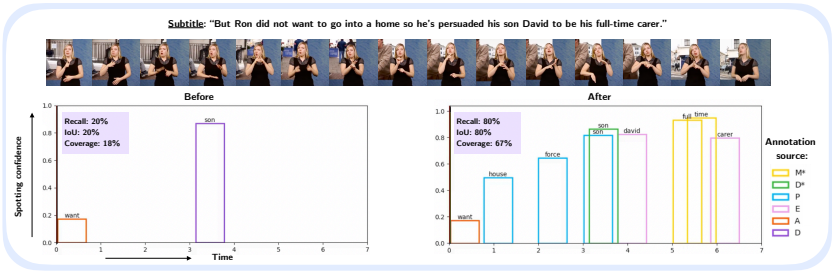

Fig. 1: **Densification**: For continuous sign language, we show automatic sign annotation timelines, along with their confidence and annotation source, *before* and *after* our framework is applied. M, D, A refer to automatic annotations from previous methods from mouthings [2], dictionaries [44] and the Transformer attention [61]. M*, D*, P, E, N refer to new and improved automatic annotations collected in this work. Annotation methods are compared in the appendix.

corresponding to the audio content, to build datasets such as Content4All [9] (190 hours) and BOBSL [1] (1460 hours).

Although such datasets are orders of magnitude larger than the long-standing RWTH-PHOENIX [10] benchmark (9 hours) and cover a much wider domain of discourse (not restricted to only weather news), the supervision they provide on the signed content is limited in that it is *weak* and *noisy*. It is weak because the subtitles are temporally aligned with the audio content and not necessarily with the signing. The supervision is also noisy because the presence of a word in the subtitle does not necessarily imply that the word is signed; and subtitles can be signed in different ways. Recent works have shown that training automatic sign language translation models on such *weak* and *noisy* supervision leads to low performance [1, 9, 61].

In an attempt to increase the value of such interpreted datasets, multiple works [2, 44, 61] have leveraged the subtitles to perform lexical *sign spotting* in an approximately aligned continuous signing segment – where the aim is to determine *whether* and *when* a subtitle word is signed. Methods include using visual keyword spotting to identify signer mouthings [2], learning a joint embedding with sign language dictionary video clips [44], and exploiting the attention mechanism of a transformer translation model trained on weak, noisy subtitle-signing pairs [61]. These works leverage the approximate subtitle timings and subtitle content to significantly reduce the correspondence search space between temporal windows of signs and spoken language words. Although such methods are effective at automatically annotating signs, they only find *sparse* correspondences between keywords in the subtitle and individual signs.

Our goal in this work is to produce *dense* sign annotations, as shown in Fig. 1. We define densification in two ways: (i) reducing gaps in the timeline so that we have a densely spotted signing sequence; and also (ii) increasing the number of words we recall in the corresponding subtitle. This process can be seen as automatic annotation of lexical signs. Automatic dense annotation of large-vocabulary sign language videos has a large range of applications including: (i) *substantially* improving recall for retrieval or intelligent fast forwards of online sign language videos; (ii) enabling *large-scale* linguistic analysis between spoken

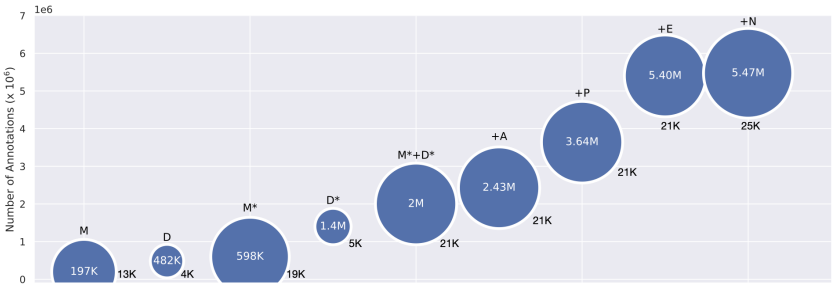

Fig. 2: **Yield of automatic annotations and vocabulary size:** We highlight the increase in the number of automatic annotations and vocabulary size at each stage in our proposed framework. M, D, A refer to annotations from previous methods. M*, D*, P, E, N refer to new and improved annotations collected in this work. The number of annotations is shown within each circle. The vocabulary size is reported below each circle and also represented by the circle diameter.

and signed languages; (iii) providing *supervision* and *improved alignment* for continuous sign language recognition and translation systems.

In this paper, we ask the following questions: (1) Can we improve current methods to improve the yield of automatic sign annotations whilst maintaining precision? (2) Can we increase the vocabulary of annotated signs over previous methods? (3) Can we 'fill in the gaps' that current spotting methods miss? The answer is yes, to all three questions, and we demonstrate this on the recently released BOBSL dataset of British Sign Language (BSL) signer interpreted video.

We make the following four contributions: (1) we significantly improve previous methods by making use of synonyms and subtitle-signing alignment; (2) we show the value of pseudo-labelling from a sign recognition model as a way of sign spotting; (3) we propose a novel approach for increasing our annotations of *known* and *unknown* sign classes based on in-domain exemplars; (4) we will make all 5 million automatic annotations publicly available to support the sign language research community. Our increased yield and vocabulary size is shown in Fig. 2. Our final vocabulary of 24.8K represents the vocabulary of English words (including named entities) from the subtitles which have been automatically associated to a sign instance; different words may have the same sign.

We note that this work is focused on *interpreted* data, which can differ from *conversational* signing in terms of style, vocabulary and speed [6]. Although our long-term aim is to move to conversational signing, learning good representations of signs from interpreted data can be a 'stepping stone' in this direction. Moreover, non-lexical signs, such as a pointing sign and spatially located signs, are essential elements of sign language, but our method is limited to the annotation of lexical signs associated to words in the text.

2 Related Work

Our work relates to several themes which we give a brief overview of below.

Sign Spotting. One line of research has focused on the task of *sign spotting*, which seeks to detect signs from a given vocabulary in a target video. Early

efforts for sign spotting employed lower-level features (colour histograms and geometric cues) in combination with Conditional Random Fields [67], Hidden Markov Models (HMMs) [62] and Sequential Interval Patterns [48] for temporal modelling. A related body of work has sought to localise signs while leveraging weak supervision from audio-aligned subtitles. These include the use of external dictionaries [26, 42, 44] and other localisation cues such as mouthing [2] and Transformer attention [61]. The performance of these approaches depends on the quality of the visual features, keywords, and the search window. In this work, we show improved yield of existing sign spotting techniques by employing automatic subtitle alignment techniques to adjust the time window and incorporating synonyms when forming the keywords. Going further beyond the spotting task explored in prior work, we use the automatic spottings to initiate additional algorithms for sign discovery based on *in-domain* exemplar matching (Sec. 3.1). This is similar to dictionary-based sign spotting techniques [26, 44] except we do not source the exemplars from external dictionaries, avoiding the domain gap issue. Besides in-domain *sign* exemplars as in [26], we explore the weak *subtitle* exemplars with unknown sign locations.

A recent progress in mouthing-based keyword spotting was presented by *Transpotter* [51]. This architecture comprises a transformer joint encoder of visual features and phoneme features that is trained to regress both the presence and location of the target keyword in a sequence from mouthing patterns. Preliminary small-scale experimental results reported by Prajwal et al. [51] demonstrated that Transpotter can perform visual keyword spotting in signing footage. Here, we showcase its suitability for the large-scale annotation regime, and further train it on sign language data to obtain a greater density of sign annotations.

In this work, we demonstrate the additional value of *pseudo-labelling* [38, 68] with a sign classifier as an effective mechanism for sign spotting. While pseudo-labelling has been explored previously for category-agnostic sign segmentation [52] and temporal alignment of glosses [14, 37] to the best of our knowledge, this is the first use of pseudo-labelling for sign spotting by directly leveraging the predictions of a sign classifier in combination with a pseudo-label filter constructed from the subtitles themselves.

Sign Language Recognition. Efforts to develop visual systems for sign recognition stretch back to work in 1988 from Tamaura and Kawasaki [59], who sought to classify signs from hand location and motion features. There were later efforts to design hand-crafted features for sign recognition [13, 47, 57, 64, 65]. Deep convolutional neural networks then came to dominate sign representation [36], particularly via 3D convolutional architectures [2, 31, 39, 42] with extensions to focus model capacity around human skeletons [25] and non-manual features [24].

In the domain of continuous sign language recognition, in which the objective is to infer a sequence of sign glosses, prior work has explored HMMs [3, 35] in combination with Dynamic Time Warping (DTW) [70], RNNs [17] and architectures capable of learning effectively from CTC losses [14, 72]. Recently, sign representation learning methods inspired by BERT [19] have shown the potential to learn effective representations for both isolated [23] and continuous [73] recognition. Koller [34] provides an extensive survey of the sign recognition literature, highlighting the extremely limited supply of datasets with large-scale vocabularies suitable for continuous sign language recognition. In our work, we aim to take a step towards addressing this gap by developing "densification" techniques for constructing such datasets automatically.

Sign Language Translation. The task of translating sign language video to spoken language sentences was first tackled with neural machine translation by Camgöz et al. [10], who also introduced the PHOENIX-Weather-2014T dataset to facilitate research on this topic. Several frameworks have been proposed to employ transformers for this task [12, 69], with extensions to improve temporal modelling [41], multi-channel cues [11] and signer independence [27]. Related work has also sought to contribute to progress on this task by exploiting monolingual data [71] and gloss sequence synthesis [40, 45]. To date, various works have shown promise on the PHOENIX-Weather 2014T [10] and CSL Daily [71] benchmarks. However, sign language translation has not yet been demonstrated for a large vocabulary across multiple domains of discourse. Differently from the works above, this paper focuses on developing methods that are applicable to large/open vocabulary regimes.

Weakly-supervised Object Discovery and Localisation. Our approach is also related to the rich body of literature on object cosegmentation [29, 33, 53, 54], weakly supervised object localisation [18, 22, 46, 55, 66], object colocalisation [30, 60] and unsupervised object discovery and localisation [15, 63]. Here, we propose an algorithm for discovering and localising novel signs (i.e. for which we have no labelled examples), but instead have weak supervision in the form of subtitles containing keywords of interest. Moving beyond initial work that sought to learn from subtitles in an aligned setting [20], classical approaches for sign discovery using subtitles have included Multiple Instance Learning where the subtitles are considered as positive and negative bags for a particular keyword [7, 32, 50] and a priori mining [16]. Differently from these works, we first bootstrap our sign discovery process with sign spotting to both obtain initial candidates and learn robust sign representations, then propagate these examples across video data by leveraging the similarities between the resulting representations together with noisy constraints imposed by the subtitle content.

3 Densification

Our goal is to leverage several ways of sign spotting to achieve dense annotation on continuous signing data. To this end, we introduce both new sources of automatic annotations, and also improve the existing sign spotting techniques. We start by presenting two new spotting methods using in-domain exemplars: to mine more sign instances with individual *exemplar signs* (Sec. 3.1) and to discover novel signs with weak *exemplar subtitles* (Sec. 3.2). We also show the value of pseudo-labelling from a sign recognition model for sign spotting (Sec. 3.3). We then describe key improvements to previous work which substantially increase the yield of automatic annotations (Sec. 3.4). Finally, we present our evaluation framework to measure the quality of our sign spottings in a large-vocabulary setting (Sec. 3.5). The contributions of each source of annotation are assessed in the experimental results.

3.1 Mining more Spottings through In-domain Exemplars (E)

The key idea is: given a continuous signing video clip and a set of exemplar clips of a particular sign, we can use the exemplars to search for that sign within the video clip. In our case, the exemplars are obtained from other *automatic* spotting methods (M*, D*, A, P), described in Sec. 3.3 and Sec. 3.4, and come from the

same *domain* of sign language interpreted data, i.e. the same training set. We hypothesise that signs from the same domain are more likely to be signed in a similar way and in turn help recognition; in contrast, for example, to signs from a different domain such as dictionaries.

Formally, suppose we have a reference video V_0 in which we wish to localise a particular sign w, whose corresponding word occurs in the subtitle. We also have N video exemplars $V_1, \ldots, V_N$ of the sign w. For each video, V_i, let $\mathcal{C}_i$ denote the set of possible temporal locations of the sign w and let $c = (f, p) \in \mathcal{C}_i$ denote a candidate with features f at temporal location p. We compute a score map between our reference video V_0 and each exemplar $V_1, \ldots, V_N$ by computing the cosine similarity between each feature at each position in $c_0 \in \mathcal{C}_0$ and $(c_1, c_2, \ldots c_n) \in \mathcal{C}_1 \times \cdots \times \mathcal{C}_N$. This results in N score maps of dimension $|\mathcal{C}_0| \times |\mathcal{C}_i|$ for $i = 1 \ldots N$. We then apply a max operation over the temporal dimension of the exemplars, giving us N vectors of length $|\mathcal{C}_0|$, which we call $M_1, \ldots, M_N$.

We subsequently apply a voting scheme to find the location of the common sign w in V_0. Specifically, we let $L = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{(M_i > h)}$ for a threshold h, where the vector $\mathbb{1}_{(M_i > h)}$ takes the value 1 for entries of M_i which are greater than h and 0 otherwise. The candidate location of w in V_0 is then $c = (f, p) \in \mathcal{C}_0$ where p corresponds to the position of the maximum non-zero entry in the vector L (see Fig. 3 for a visual illustration). If there are multiple maxima, we assign p to be the midpoint of the largest connected component. If all entries in L are zero, we conclude w is not present. We perform two variants of this approach using mean and max pooling of the score maps (instead of voting); these are described in the appendix. We note that for a given signing sequence, we only focus on finding signs for words in the subtitle that have *not* been annotated by other methods.

3.2 Discovering Novel Sign Classes (N)

One limitation of our proposed method in Sec. 3.1 is that we are only able to collect more sign instances from a *closed* vocabulary, determined by sign exemplars obtained from other methods (described in Sec. 3.3 and Sec. 3.4). Here, we extend our approach to localise *novel* signs, for which we have no exemplar signs but whose corresponding word appears in the subtitle text. We follow our approach described in Sec. 3.1, computing score maps between our reference video and exemplar subtitles (instead of exemplar signs, see Fig. 3). We note that by 'exemplar subtitle', we are referring to the video frames corresponding to the subtitle timestamps. Non-lexical signs, such as pointing signs or pause gestures, are very common in sign language. To avoid annotating such non-lexical signs as the common sign across V_0 and $V_1, \ldots, V_N$, we also choose N^- negative subtitle exemplars $U_1 \ldots U_{N^-}$ presumed to not contain w (due to the absence of w in the subtitle). We compute L^+ and L^- using the score maps from positive exemplars $V_1, \ldots, V_N$ and negative exemplars $U_1, \ldots, U_{N^-}$ respectively. We then let $L = L^+ - L^-$. Implementation details on the number of positive and negative exemplars used can be found in the appendix.

3.3 Pseudo-labelling as a Form of Sign Spotting (P)

We propose to re-purpose a pretrained large-vocabulary sign classification model (see vocabulary expansion in Sec. 3.5) for the task of sign spotting. Specifically,

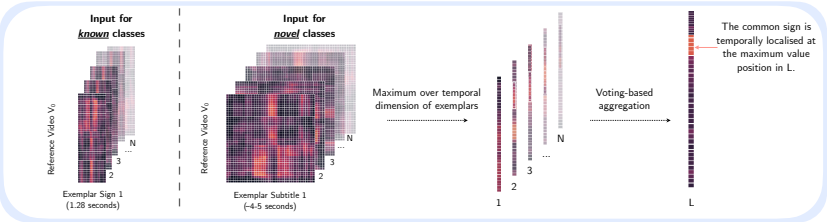

Fig. 3: **Sign spotting through exemplars to find instances of known classes (E) and novel classes (N)**: By comparing a reference video V_0 to a set of exemplars (either sign exemplars for known sign class instances or weak subtitle exemplars for novel sign class instances), we can find the common lexical sign in the collection. We (1) form a set of score maps by calculating the cosine similarities between reference and exemplar representations; (2) we perform a maximum operation over the temporal dimension of exemplars; (3) we apply a voting-based aggregation to find the temporal location of the commmon sign in V_0. The duration of exemplar signs is fixed.

we predict a sign class from a fixed vocabulary for each time step in a continuous signing video clip. We subsequently filter the predicted signs to words which occur in the corresponding English subtitle. Similarly to [61], here the task is to recognise the sign from scratch, without a query keyword. The subtitle is only used as a post-processing step to filter out signs which are less likely performed (due to absence in the subtitle).

3.4 Improving the Old (M*, D*)

Here, we briefly describe our improvements over the existing sign spotting techniques, additional details are provided in the appendix.

Better Mouthings with an Upgraded KWS from Transpotter [51]. In previous work [1], an improved BiLSTM-based visual-only keyword spotting model of Stafylakis et al. [56] from [43] (named "P2G [56] baseline") is used to automatically annotate signs via mouthings. In this work, we make use of the recently proposed transformer-based *Transpotter* architecture [51], provided by the authors, that achieves state-of-the-art results in visual keyword spotting on lipreading datasets. We follow the procedure described in [1, 2] to query words in the subtitle in continuous signing video clips.

Finetuning KWS on Sign Language Data through Bootstrapping. The visual keyword spotting Transpotter architecture in [51] is trained on silent speech segments, which differ considerably from signer mouthings. In fact, signers do not mouth continuously and sometimes only partly mouth words [5]. In order to reduce this severe domain gap, we propose a dual-stage finetuning strategy. First, we extract high-confidence mouthing annotations using the pre-trained Transpotter from [51] on the BOBSL training data. We query for the words in the subtitle and obtain the temporal localization of the word in the video. We finetune on this pseudo-labeled data using the same training pipeline of [51], where the spotted mouthings (word-video pairs) act as positive samples. For the negative samples, we pair a given word with a randomly sampled video segment from the dataset. As we observe the Transpotter to predict a large number of

false positives, we remedy this by sampling a larger number of negative pairs in each batch. We also do a second round of fine-tuning by training on the pseudo-labels from the finetuned model of the first stage. We did not achieve significant improvements with further iterations.

Better Search Window with Subtitle Alignment with SAT [8]. One challenge in using sign language interpreted TV broadcasts is that the original subtitles are not aligned to the signing, but to the audio track. In [1], a signing query window is defined as the audio-aligned subtitle timings together with padding on both sides to account for the misalignment. We automatically align spoken language text subtitles to the signing video by using the SAT model introduced in [8], trained on manually aligned and pseudo-labelled subtitles as described in [1]. By using subtitles which are better aligned to the signing, we reduce the probability of missing spottings.

Better Keywords with Synonyms and Similar Words. To determine whether a keyword belongs to a subtitle, previous works [1] check whether the raw form, the lemmatised form, or the text normalised form (e.g. *two* instead of *2*) appears in the subtitle text. We notice that this is sub-optimal as multiple words may correspond to the same sign, often due to (i) English synonyms, (ii) identical signs for similar words, or (iii) ambiguities in spoken language. For example, *dad* and *father* or *today* and *now* can be the same signs in BSL. In this work, we investigate whether the automatic annotation yield could be improved by querying words beyond the subtitle, by querying synonyms and similar words to the words in the subtitle. We collect the additional words to query through (i) English synonyms from WordNet [21], (ii) the metadata present in online sign language dictionaries such as SignBSL[5] [44] and BSL Sign-Bank[6] which provide a set of 'related words' for each sign video entry; (iii) words with GloVe [49] cosine similarity above 0.9 to account for ambiguities in spoken language.

3.5 Evaluation Framework

Our framework consists of three stages: (a) a costly end-to-end classification training to learn sign category aware video features given an initial set of sign-clip annotation pairs; (b) a lightweight classification training given pre-extracted video features for a large number of annotations; (c) a sliding window evaluation of the trained lightweight model by comparing dense sign predictions against the subtitles (see Sec. 4.1). These stages are illustrated in Fig. 4. Note that the *annotations* we refer to are always *automatically* localised sign spottings from continuous videos using subtitle information. The motivation for the video backbone and lightweight classifier is purely related to computational costs. Unlike traditional video recognition datasets, we work with untrimmed video data of 1400 hours, where the set of sign-clip pairs is not fixed. Instead, our goal is to increase the number of sign-clip pairs within the continuous stream, and assess the quality of the expanded annotation yield on the proxy task of continuous sign language recognition. Next, we describe the training stages for the video backbone and the lightweight classifier.

Improving the I3D Feature Extractor through Vocabulary Expansion. Following previous works [1, 2, 31, 39], we use the I3D spatio-temporal convolutional architecture to train an end-to-end sign recognition model. We input 16

[5] www.signbsl.com

[6] bslsignbank.ucl.ac.uk

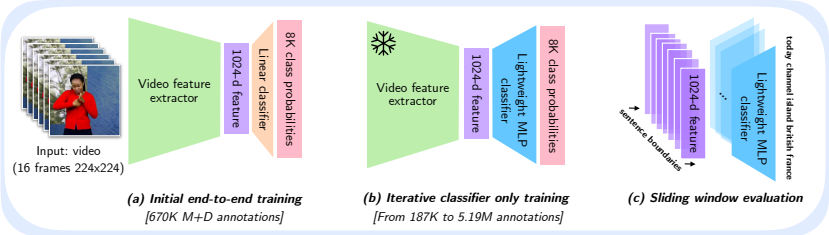

Fig. 4: **Evaluation framework**: (a) Video features are obtained by training an I3D architecture end-to-end given M + D annotations from [1]. The I3D ingests 16-frames of video and has a linear classifier for 8K sign categories. The end-to-end training is a costly procedure which is not affordable to repeat for each set of our new sign spottings that are on the order of several million training samples. (b) As new sets of spottings are generated, a light weight MLP classifier is trained on the pre-extracted I3D features. This relatively inexpensive training procedure means that we benefit from new annotations without the expense of end-to-end training. (c) The MLP is applied in a sliding window fashion to the signing sequence to generate sign predictions.

consecutive RGB frames and output class probabilities. The details about optimisation are provided in the appendix. As explained above, this model forms the basis of sign video representation which corresponds to the spatio-temporally pooled latent embedding before the classification layer. The prior work of [1] trains this classifier on the BOBSL dataset (see Sec. 4.1) with 2K categories obtained through the vocabulary of mouthing spottings. As a first step, we perform a vocabulary expansion and construct a significantly increased vocabulary of 8K categories. This is achieved by including each sign that has at least 5 training spottings above 0.7 confidence from both mouthing (M) and dictionary (D) annotations. The confidence for the mouthing annotation corresponds to the probability that a text keyword (corresponding to the sign) is mouthed at a certain time frame, as computed in [2]. The confidence for the dictionary annotation corresponds to the cosine similarity (normalised between 0-1) between the representations of a dictionary clip of the sign and the continuous signing at each time frame, as in [44]. The resulting M+D training set comprises 670K annotations, with a long-tailed distribution. Furthermore, we note that the categories are noisy where multiple categories may correspond to the same sign, and vice versa. Despite this noise, we empirically show that this model provides better performance than its 2K-vocabulary counterpart. We use our improved I3D model for two purposes: as the frozen feature extractor and as the source of pseudo-labelling for sign spotting (see Sec. 3.3).

Lightweight Sign Recognition Model. Following [61], we opt for a 4-layer MLP module (with one residual connection) to assess the quality of different sets of annotations. Given pre-extracted features, this model is trained for sign recognition into 8K categories. We note that we do not train on a larger vocabulary to avoid the presence of many singletons in the training set. The efficiency of the MLP allows faster experimentation to analyse the value of each of our sign spotting sets. The input is one randomly sampled feature around the sign spotting location (the receptive field of one feature 16 frames). The MLP weights are

randomly initialised. Additional training and implementation details are given in the appendix Sec. B.5 and B.6.

4 Experiments

We start by describing our dataset and evaluation metrics (Sec. 4.1). We then present experimental results on the contribution of each source of annotation and show qualitative examples (Sec. 4.2).

4.1 Data and Evaluation Protocol

BOBSL [1] is a public dataset consisting of British Sign Language interpreted BBC broadcast footage, along with English subtitles corresponding to the audio content. The data contains 1,962 episodes, which have a total duration of 1,467 hours spanning 426 different TV shows. BOBSL has a total 1,193K subtitles covering a total vocabulary of 78K words. We note that in this work we use the word *subtitle* to refer to the processed BOBSL sentences from [1] as opposed to the raw subtitles. There are a total of 39 signers in the dataset. Further dataset statistics can be found in [1]. For a subset of 36 episodes in BOBSL, referred to as SENT-TEST in [1], the English subtitles have been manually aligned *temporally* to the continuous signing video. We make use of this test set to evaluate the quality of our predicted automatic annotations. SENT-TEST covers a total duration of 31 hours and contains 20,870 English subtitles. The total vocabulary of English words is 13,641, of which 5,604 are singletons. The 3 signers in SENT-TEST are different to the signers in the training set, this enables signer-independent BSL recognition to be evaluated.

Evaluation protocol. Given an English subtitle and the *temporally* aligned continuous signing video clip, we evaluate our predicted signs for the clip using (i) *intersection over union* (IoU); (ii) *recall* between signs and the English word sequence; and (iii) *temporal coverage*: this is defined as the proportion of frames in the clip assigned to signs that occur in the word sequence, where a sign is given a fixed duration of 16 frames (for 25Hz video). Note that none of these metrics depend on the word order of the English subtitle, only the words it contains. All metrics are rescaled from the range 0-1 to 0-100 percentage for readability.

For this evaluation, stop words are filtered out since often they are not signed. This reduces the number of test subtitles from 20,870 to 20,547: subtitles such as “is it?”, “Oh!”, “but no” are removed. The sign and word sequences are also lemmatised. We also remove repetitions from the predicted sign sequence and allow the prediction of synonyms of words in the English subtitle. This processing is highlighted in Fig. 5, where the IoU and recall are computed for a pair of predicted signs and English text. While this evaluation is suboptimal due to the simplified word-sign correspondence assumption, it tests the capacity of the sign recognition model in a large-vocabulary scenario, necessary for open-vocabulary sign language technologies.

Note, the predicted signs for a clip can be produced in two ways. In the first way, the signs are obtained from the automatic annotations using knowledge of the content of the English subtitle – we refer to these as *Spottings*. In the second, signs are predicted directly from the clip using the MLP sign predictions, without access to the corresponding English subtitle. These are referred to as

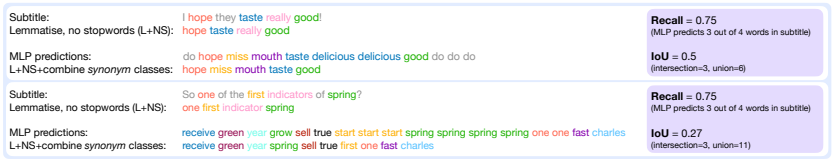

Fig. 5: **Evaluation illustration on sample prediction:** We illustrate the processing applied to the predicted sign sequence from the MLP predictions and corresponding English subtitle for calculating our metrics. As the MLP model predicts one sign per time-step, some predictions are repeated and irrelevant words appear at transition periods between signs, decreasing the IoU. Some signs are not predicted as they are not signed, showing the limitations of using the subtitle to measure performance.

Table 1: **Comparison of I3D video features**: We highlight the improved performance of I3D on the test set (SENT-TEST) when trained on a larger vocabulary (8K instead of 2K) with more samples (670K instead of 426K).

Annot. source	Num. I3D train annot.	Vocab. size	I3D predictions (subtitle independent)		
			Recall	IoU	Coverage
M [2]+D [44]	426K	2K	25.5	6.4	15.5
M [2]+D [44]	670K	8K	**26.3**	**7.9**	**16.3**

MLP predictions. Spottings are evaluated using all the words; this metric is important to monitor how dense we can automatically annotate the data. The MLP evaluation is limited to the fixed classification vocabulary (of size 8K in our experiments). We note that when different annotations are combined, the sign spotting methods are applied independently.

4.2 Results

Comparison of Video Features. By finetuning our Kinetics pretrained I3D model on BOBSL M+D annotations from [1] using an 8K vocabulary instead of a 2K vocabulary, we improve predictions on the test set, as shown in Tab. 1. We increase the recall from 25.5 to 26.3 and the coverage from 15.5 to 16.3. We therefore use the 8K M+D model for the rest our experiments as the frozen feature extractor. We note that we restrict the M+D annotations to the high-confidence ones (over 0.8 threshold) used for the I3D baseline in [1], as these present an appropriate signal-to-noise ratio. We use the same threshold for subsequent automatic annotations unless stated otherwise.

Oracle. As the MLPs are trained on a restricted 8K vocabulary, it is not possible to predict the full vocabulary of 13,641 words present in the test set subtitles. Furthermore, not all words in the subtitle are signed and vice versa. This means a recall, IoU and coverage of 100% is not achievable between predicted signs and English subtitle words. However, we propose an oracle in Tab. 2 whereby we measure the recall and IoU assuming each word in the subtitle, which either falls within the 8K vocabulary or corresponds to a synonym of a word in the 8K vocabulary, is signed and correctly predicted. The oracle achieves a recall of 86.7 and IoU of 86.3. For the coverage metric, we assume each correctly predicted sign has a duration of 16 frames and no signs overlap. The resulting oracle coverage

Table 2: **Improved mouthing and dictionary spottings:** We evaluate different sets of spottings and their respective MLP predictions. M [51] shows our finetuned version for all the rows in the last block. We quantify the effects of subtitle alignment and querying synonyms. We also show the oracle performance and a translation baseline.

			Training set			Spottings [full] (subtitle dependent)			MLAP predictions [8K] (subtitle independent)		
Annotation source	Subtitle alignment	Synonyms	full vocab	#ann. [full]	#ann. [8K]	Recall	IoU	Coverage	Recall	IoU	Coverage
Oracle			-	-	-	-	-	-	86.7	86.3	55.2
Translation baseline [1]			-	-	-	-	-	-	11.7	8.3	7.6
M [2]			13.6K	197K	187K	2.5	2.2	1.3	15.1	3.2	8.7
M [51] (no finetuning)			21.5K	725K	661K	9.4	8.3	4.9	20.4	4.8	11.9
M [51]			18.6K	445K	412K	7.1	6.5	3.9	23.6	4.8	13.8
M [51] (M*)	✓		19.6K	598K	552K	8.9	8.2	4.9	27.4	6.3	16.7
M [51]	✓	✓	19.6K	1.38M	1.25M	11.8	10.4	6.1	25.3	6.2	16.3
D [44]			4.4K	482K	482K	6.5	6.3	3.7	24.0	7.2	15.1
D [44]	✓		4.5K	535K	535K	7.0	6.9	4.0	24.2	7.3	15.3
D [44] (D*)	✓	✓	5.0K	1.40M	1.39M	12.5	11.6	7.0	26.0	7.3	16.9
M*+ D*	✓	✓(D-only)	20.9K	2.00M	1.94M	19.0	17.6	10.5	29.0	7.9	18.4
M*+ D*+ A [61]	✓	✓(D-only)	20.9K	2.43M	2.37M	**21.9**	**20.1**	**11.8**	**29.6**	**9.1**	**19.0**

is 55.2. This low coverage is partly due to the signer pausing within subtitles and also due to the presence of non-lexical signs. In fact, the percentage of fully lexical signs in three other sign language corpora (Auslan [28], ASL [28] and LSF [4]) is estimated to be only 70-85% of total signing.

Translation Baseline. Although the goal in this work is not translation, but achieving dense annotations, we can nevertheless compare our MLP predictions to the translation baseline in [1]. Using the test set translation predictions from this model, we perform the same processing as highlighted in Fig. 5 to calculate our metrics. As shown in Tab. 2, all our simple MLP models clearly outperform the transformer-based translation model used in [1], demonstrating that we are able to recognise more signs in the English subtitle.

Improving Mouthing and Dictionary Spottings. As shown in Tab. 2, by using the Transpotter [51] for spotting mouthings M, our yield of total annotations triples from 197K to 725K. The quality of these new annotations is reflected in the increased performance of the MLP: the recall increases from 15.1 to 20.4 and the coverage from 8.7 to 11.9. Finetuning the keyword spotter on sign language data through pseudo-labelling also helps considerably despite the drop in the number of training annotations since there are less false positives; recall increases from 20.4 to 23.6 and coverage from 11.9 to 13.8. Subtitle alignment improves the yield of both mouthing and dictionary annotations, as shown in Tab. 2. This translates to a significant boost for mouthings on the MLP performance; the recall increases from 23.6 to 27.4 and the coverage from 13.8 to 16.7. For dictionary annotations, the improvement by using aligned subtitles is less striking. By querying synonyms when searching for mouthings, the yield more than doubles. However, these additional annotations seem to be quite noisy as they decrease the performance of our MLP. Due to the nature of sign language interpretation, it is possible that signers are far more likely to mouth a word which is actually in the written subtitle than a synonym of that word. We therefore do not query synonyms for mouthing spottings. For dictionary spottings, we observe the opposite effect. By incorporating synonyms, the yield of dictionary spottings more than doubles and the recall of the MLP predictions also increases from 24.2 to 26.0. We denote our best performing mouthing and dictionary spot-

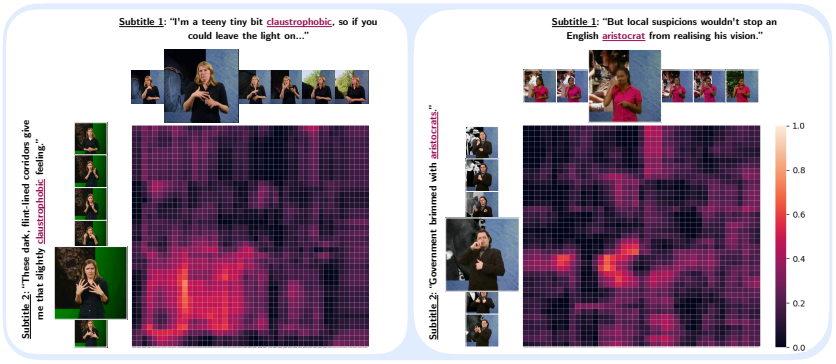

Fig. 6: **Discovering novel sign classes (N)**: For two pairs of continuous signing sentences, we plot the score maps (as described in Sec. 3.1) between their feature sequences. We highlight the ability of our approach to spot novel sign classes.

Table 3: **Ablation on mining exemplar-based spottings for known signs E:** We perform different ablations for mining known signs which have been unannotated by previous methods (M*, D*, A, P). We experiment with the source of exemplar data (same episode, same signer, all data), the confidence of exemplar signs (0,0.5,0.8), the number of samples of exemplar data (5,10,20) and the pooling mechanism (average, max, vote). We evaluate on the test set (SENT-TEST).

					Training set			**Spottings [full] (subtitle dependent)**			**MLP predictions [8K] (subtitle independent)**		
Ann. src.	ex. data	ex. thres	ex. #	ex. pooling	full vocab	#ann. [full]	#ann. [8K]	Recall	IoU	Coverage	Recall	IoU	Coverage
E	same ep.	0	var	avg	11.6K	869K	833K	10.4	9.6	5.8	25.1	6.9	15.3
E	same signer	0	20	avg	15.9K	505K	421K	7.8	7.5	4.4	23.1	5.6	14.2
E	all	0	20	avg	16.7K	351K	252K	5.7	5.7	3.3	21.5	5.1	13.4
E	all	0.5	20	avg	16.6K	370K	261K	5.9	5.8	3.4	21.9	5.2	13.5
E	all	0.8	20	avg	16.6K	458K	358K	7.4	7.3	4.3	25.2	6.2	15.7
E	all	0.8	20	max	15.4K	1.48M	1.38M	20.2	18.6	10.8	27.6	8.4	17.7
E	all	0.8	10	max	15.4K	1.07M	982K	15.2	14.0	8.3	27.9	8.0	17.7
E	all	0.8	5	max	15.3K	740K	664K	10.7	10.0	6.0	27.6	7.6	17.4
E	all	0.8	20	vote	15.9K	1.76M	1.63M	**25.8**	**23.3**	**13.5**	**28.4**	**8.5**	**18.1**
E	all	0.8	10	vote	15.8K	1.32M	1.21M	20.0	18.1	10.7	28.4	8.3	18.1

tings with M* and D*, respectively. Adding attention spottings from [1] (with a threshold of 0) adds around 400K additional annotations and boosts the MLP performance; increasing recall from 29.0 to 29.6 and coverage from 18.4 to 19.0, compared to the oracle recall of 86.7 and coverage of 55.2.

Sign Recognition as a Form of Pseudo-labelling. Pseudo-labels P are a source of over 1M new annotations (when using a threshold of 0.5) on top of our best M*, D*, A spottings. As shown in Tab. 4, they greatly increase the spottings recall from 21.9 to 25.4 and coverage from 11.8 to 13.9, while only marginally increasing the recall and coverage for MLP predictions. As the pseudo-labels come from our 8K I3D model in Tab. 1 whose frozen features are also used for training the MLP, P may not be providing additional information for our downstream evaluation. Nevertheless, they provide a great source of additional spottings (not found by previous methods) for our goal of dense annotation.

Mining more Examples of Known and Novel Sign Classes with In-domain Exemplars. By explicitly querying words in the subtitle text which

Table 4: **Pseudo-label spottings P & Exemplar-based sign spottings for known E and novel classes N:** We highlight the boost in annotations by adding our pseudo-label annotations (P) as well as exemplar-based spottings of known (E) and novel (N) classes. We evaluate Spottings and MLP predictions on the test set (SENT-TEST). For the novel classes, we only show the evaluation of spottings since these are beyond the 8K training vocabulary of the MLP.

	Training set			**Spottings [full] (subtitle dependent)**			**MLP predictions [8K] (subtitle independent)**		
Annotation source	full vocab	#ann. [full]	#ann. [8K]	Recall	IoU	Coverage	Recall	IoU	Coverage
M* + D* + A [61] + P	20.9K	3.64M	3.56M	25.4	23.5	13.9	29.8	8.9	19.2
M* + D* + A [61] + P + E	20.9K	5.40M	5.19M	45.3	40.7	23.3	**30.7**	**9.5**	**19.8**
M* + D* + A [61] + P + E + N	24.8K	5.47M	-	**45.6**	**40.9**	**23.4**	-	-	-

are not present in our annotations, we can obtain significantly more annotations. Tab. 3 shows multiple methods to use exemplar signs to find additional annotations for these signs. The best performing method takes spotting exemplars from across the whole training set, irrespective of signer or episode, and uses the voting scheme described in Sec. 3.1 to localise signs. By using 20 spotting exemplars, we acquire 1.63M additional annotations. An MLP model trained *only* on these additional annotations achieves a recall of 28.4 and coverage of 18.1. Tab. 4 illustrates the impact of combining these additional annotations from spotting exemplars to M*, D*, A and P annotations. With the additional exemplar-based annotations E, recall increases from 29.8 to 30.7 and coverage increases from 19.2 to 19.8, where the oracle recall and coverage are 86.7 and 55.2. Furthermore, by mining instances of novel sign classes N (see Fig. 6), we increase our total vocabulary to 24.8K and total number of annotations to 5.47M.

5 Conclusion

Progress in sign language research has been accelerated in recent years due to the availability of large-scale datasets, in particular sourced from interpreted TV broadcasts. However, a major obstacle for the use of such data is the lack of available sign level annotations. Previous methods [2, 44, 61] only found *sparse* correspondences between keywords in the subtitle and individual signs. In our work, we propose a framework which scales the number of confident automatic annotations from 670K to 5.47M (which we make publicly available). Potential future directions for research include: (1) increasing our number of annotations by incorporating context from *surrounding* signing to resolve ambiguities; (2) investigating *linguistic* differences between spoken English and British Sign Language such as the different word/sign ordering; (3) leveraging our automatic annotations for sign language translation.

Acknowledgements. This work was supported by EPSRC grant ExTol, a Royal Society Research Professorship and the ANR project CorVis ANR-21-CE23-0003-01. LM would like to thank Sagar Vaze for helpful discussions. HB would like to thank Annelies Braffort and Michèle Gouiffès for the support.

Bibliography

[1] Albanie, S., Varol, G., Momeni, L., Afouras, T., Bull, H., Chowdhury, H., Fox, N., Cooper, R., McParland, A., Woll, B., Zisserman, A.: BOBSL: BBC-Oxford British Sign Language Dataset. arXiv preprint arXiv:2111.03635 (2021) 2, 7, 8, 9, 10, 11, 12, 13

[2] Albanie, S., Varol, G., Momeni, L., Afouras, T., Chung, J.S., Fox, N., Zisserman, A.: BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In: ECCV (2020) 2, 4, 7, 8, 9, 11, 12, 14

[3] Bauer, B., Hienz, H.: Relevant features for video-based continuous sign language recognition. In: Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580). pp. 440–445. IEEE (2000) 4

[4] Belissen, V., Braffort, A., Gouiffès, M.: Experimenting the automatic recognition of non-conventionalized units in sign language. Algorithms **13**(12), 310 (2020) 12

[5] Boyes Braem, P., Sutton-Spence, R.: The Hands Are The Head of The Mouth. The Mouth as Articulator in Sign Languages. Hamburg: Signum Press (2001) 7

[6] Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., Ringel Morris, M.: Sign language recognition, generation, and translation: An interdisciplinary perspective. In: ACM SIGACCESS (2019) 3

[7] Buehler, P., Zisserman, A., Everingham, M.: Learning sign language by watching tv (using weakly aligned subtitles). In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2961–2968. IEEE (2009) 5

[8] Bull, H., Afouras, T., Varol, G., Albanie, S., Momeni, L., Zisserman, A.: Aligning subtitles in sign language videos. In: ICCV (2021) 8

[9] Camgoz, N., Saunders, B., Rochette, G., Giovanelli, M., Inches, G., Nachtrab-Ribback, R., Bowden, R.: Content4all open research sign language translation datasets. ArXiv **abs/2105.02351** (05 2021) 2

[10] 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. pp. 7784–7793 (2018) 2, 5

[11] Camgöz, N.C., Koller, O., Hadfield, S., Bowden, R.: Multi-channel transformers for multi-articulatory sign language translation. 16th European Conference on Computer Vision (ECCV), ACVR Workshop (2020) 5

[12] Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 5

[13] Charayaphan, C., Marble, A.: Image processing system for interpreting motion in american sign language. Journal of Biomedical Engineering **14**(5), 419–425 (1992) 4

[14] Cheng, K.L., Yang, Z., Chen, Q., Tai, Y.W.: Fully convolutional networks for continuous sign language recognition. In: European Conference on Computer Vision. pp. 697–714. Springer (2020) 4

[15] Cho, M., Kwak, S., Schmid, C., Ponce, J.: Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region

proposals. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1201–1210 (2015) 5
[16] Cooper, H., Bowden, R.: Learning signs from subtitles: A weakly supervised approach to sign language recognition. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2568–2574. IEEE (2009) 5
[17] Cui, R., Liu, H., Zhang, C.: Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7361–7369 (2017) 4
[18] Deselaers, T., Alexe, B., Ferrari, V.: Localizing objects while learning their appearance. In: European conference on computer vision. pp. 452–466. Springer (2010) 5
[19] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 4
[20] Farhadi, A., Forsyth, D.: Aligning ASL for statistical translation using a discriminative word model. In: CVPR (2006) 5
[21] Feinerer, I., Hornik, K.: wordnet: WordNet Interface (2020), https://CRAN.R-project.org/package=wordnet, r package version 0.1-15 8
[22] Gokberk Cinbis, R., Verbeek, J., Schmid, C.: Multi-fold mil training for weakly supervised object localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2409–2416 (2014) 5
[23] Hu, H., Zhao, W., Zhou, W., Wang, Y., Li, H.: Signbert: Pre-training of hand-model-aware representation for sign language recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11087–11096 (2021) 4
[24] Hu, H., Zhou, W., Pu, J., Li, H.: Global-local enhancement network for nmf-aware sign language recognition. ACM transactions on multimedia computing, communications, and applications (TOMM) **17**(3), 1–19 (2021) 4
[25] Huang, J., Zhou, W., Li, H., Li, W.: Attention-based 3d-cnns for large-vocabulary sign language recognition. IEEE Transactions on Circuits and Systems for Video Technology **29**(9), 2822–2832 (2018) 4
[26] Jiang, T., Camgoz, N.C., Bowden, R.: Looking for the signs: Identifying isolated sign instances in continuous video footage. IEEE International Conferene on Automatic Face and Gesture Recognition (2021) 4
[27] Jin, T., Zhao, Z.: Contrastive disentangled meta-learning for signer-independent sign language translation. Proceedings of the 29th ACM International Conference on Multimedia (2021) 5
[28] Johnston, T.: Lexical frequency in sign languages. Journal of deaf studies and deaf education **17**(2), 163–193 (2012) 12
[29] Joulin, A., Bach, F., Ponce, J.: Discriminative clustering for image co-segmentation. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1943–1950. IEEE (2010) 5
[30] Joulin, A., Tang, K., Fei-Fei, L.: Efficient image and video co-localization with frank-wolfe algorithm. In: European Conference on Computer Vision. pp. 253–268. Springer (2014) 5
[31] 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) 4, 8
[32] Kelly, D., Mc Donald, J., Markham, C.: Weakly supervised training of a sign language recognition system using multiple instance learning density

matrices. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **41**(2), 526–541 (2010) 5
[33] Kim, G., Xing, E.P., Fei-Fei, L., Kanade, T.: Distributed cosegmentation via submodular optimization on anisotropic diffusion. In: 2011 international conference on computer vision. pp. 169–176. IEEE (2011) 5
[34] Koller, O.: Quantitative survey of the state of the art in sign language recognition. ArXiv **abs/2008.09918** (2020) 4
[35] Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding **141**, 108–125 (2015) 4
[36] Koller, O., Ney, H., Bowden, R.: Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3793–3802 (2016) 4
[37] Koller, O., Zargaran, S., Ney, H.: Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4297–4305 (2017) 4
[38] Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896 (2013) 4
[39] 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. pp. 1459–1469 (2020) 4, 8
[40] Li, D., Xu, C., Liu, L., Zhong, Y., Wang, R., Petersson, L., Li, H.: Transcribing natural languages for the deaf via neural editing programs. arXiv preprint arXiv:2112.09600 (2021) 5
[41] Li, D., Xu, C., Yu, X., Zhang, K., Swift, B., Suominen, H., Li, H.: Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. arXiv preprint arXiv:2010.05468 (2020) 5
[42] Li, D., Yu, X., Xu, C., Petersson, L., Li, H.: Transferring cross-domain knowledge for video sign language recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6205–6214 (2020) 4
[43] Momeni, L., Afouras, T., Stafylakis, T., Albanie, S., Zisserman, A.: Seeing wake words: Audio-visual keyword spotting. BMVC (2020) 7
[44] Momeni, L., Varol, G., Albanie, S., Afouras, T., Zisserman, A.: Watch, read and lookup: Learning to spot signs from multiple supervisors. In: ACCV (2020) 2, 4, 8, 9, 11, 12, 14
[45] Moryossef, A., Yin, K., Neubig, G., Goldberg, Y.: Data augmentation for sign language gloss translation. In: MTSUMMIT (2021) 5
[46] Nguyen, M.H., Torresani, L., De La Torre, F., Rother, C.: Weakly supervised discriminative localization and classification: a joint learning process. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 1925–1932. IEEE (2009) 5
[47] 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. pp. 2200–2207. IEEE (2012) 4
[48] Ong, E.J., Koller, O., Pugeault, N., Bowden, R.: Sign spotting using hierarchical sequential patterns with temporal intervals. In: Proceedings of the

IEEE conference on computer vision and pattern recognition. pp. 1923–1930 (2014) 4
[49] Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014) 8
[50] Pfister, T., Charles, J., Zisserman, A.: Large-scale learning of sign language by watching tv (using co-occurrences). In: BMVC (2013) 5
[51] Prajwal, K., Momeni, L., Afouras, T., Zisserman, A.: Visual keyword spotting with attention. In: BMVC (2021) 4, 7, 12
[52] Renz, K., Stache, N.C., Fox, N., Varol, G., Albanie, S.: Sign segmentation with changepoint-modulated pseudo-labelling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3403–3412 (2021) 4
[53] Rother, C., Minka, T., Blake, A., Kolmogorov, V.: Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 1, pp. 993–1000. IEEE (2006) 5
[54] Rubinstein, M., Joulin, A., Kopf, J., Liu, C.: Unsupervised joint object discovery and segmentation in internet images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1939–1946 (2013) 5
[55] Shi, Z., Hospedales, T.M., Xiang, T.: Bayesian joint topic modelling for weakly supervised object localisation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2984–2991 (2013) 5
[56] Stafylakis, T., Tzimiropoulos, G.: Zero-shot keyword spotting for visual speech recognition in-the-wild. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 513–529 (2018) 7
[57] Starner, T.E.: Visual recognition of american sign language using hidden markov models. Tech. rep., Massachusetts Inst Of Tech Cambridge Dept Of Brain And Cognitive Sciences (1995) 4
[58] Sutton-Spence, R., Woll, B.: The Linguistics of British Sign Language: An Introduction. Cambridge University Press (1999) 1
[59] Tamura, S., Kawasaki, S.: Recognition of sign language motion images. Pattern recognition **21**(4), 343–353 (1988) 4
[60] Tang, K., Joulin, A., Li, L.J., Fei-Fei, L.: Co-localization in real-world images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1464–1471 (2014) 5
[61] Varol, G., Momeni, L., Albanie, S., Afouras, T., Zisserman, A.: Read and attend: Temporal localisation in sign language videos. In: CVPR (2021) 2, 4, 7, 9, 12, 14
[62] Viitaniemi, V., Jantunen, T., Savolainen, L., Karppa, M., Laaksonen, J.: S-pot–a benchmark in spotting signs within continuous signing. In: Proceedings of the 9th international conference on Language Resources and Evaluation (LREC 2014), ISBN 978-2-9517408-8-4. European Language Resources Association (LREC) (2014) 4
[63] Vo, V.H., Sizikova, E., Schmid, C., Pérez, P., Ponce, J.: Large-scale unsupervised object discovery. Advances in Neural Information Processing Systems **34** (2021) 5
[64] Vogler, C., Metaxas, D.: Adapting hidden markov models for asl recognition by using three-dimensional computer vision methods. In: 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation. vol. 1, pp. 156–161. IEEE (1997) 4

[65] Vogler, C., Metaxas, D.: Asl recognition based on a coupling between hmms and 3d motion analysis. In: Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271). pp. 363–369. IEEE (1998) 4
[66] Wang, C., Ren, W., Huang, K., Tan, T.: Weakly supervised object localization with latent category learning. In: European Conference on Computer Vision. pp. 431–445. Springer (2014) 5
[67] Yang, H.D., Sclaroff, S., Lee, S.W.: Sign language spotting with a threshold model based on conditional random fields. IEEE transactions on pattern analysis and machine intelligence **31**(7), 1264–1277 (2008) 4
[68] Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: 33rd annual meeting of the association for computational linguistics. pp. 189–196 (1995) 4
[69] Yin, K., Read, J.: Better sign language translation with stmc-transformer. In: COLING (2020) 5
[70] Zhang, J., Zhou, W., Li, H.: A threshold-based hmm-dtw approach for continuous sign language recognition. In: Proceedings of International Conference on Internet Multimedia Computing and Service. pp. 237–240 (2014) 4
[71] Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H.: Improving sign language translation with monolingual data by sign back-translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1316–1325 (2021) 5
[72] Zhou, H., Zhou, W., Zhou, Y., Li, H.: Spatial-temporal multi-cue network for continuous sign language recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020) 4
[73] Zhou, Z., Tam, V.W., Lam, E.Y.: Signbert: A bert-based deep learning framework for continuous sign language recognition. IEEE Access **9**, 161669–161682 (2021) 4