Dynamic Cross-Feature Fusion for American Sign Language Translation

Tejaswini Ananthanarayana[1], Nikunj Kotecha[1], Priyanshu Srivastava[1], Lipisha Chaudhary[1], Nicholas Wilkins[2], Ifeoma Nwogu[1]
[1] Rochester Institute of Technology, Rochester, NY, USA
[2] Sign-Speak, NY, USA

2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) | 978-1-6654-3176-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/FG52635.2021.9667027

***Abstract*— While a significant amount of work has been done on the commonly used, tightly-constrained weather-based, German sign language (GSL) dataset, little has been done for continuous sign language translation (SLT) in more realistic settings, including American sign language (ASL) translation. Also, while CNN-based features have been consistently shown to work well on the GSL dataset, it is not clear whether such features will work as well in more realistic settings when there are more heterogeneous signers in non-uniform backgrounds. To this end, in this work, we introduce a new, realistic phrase-level ASL dataset (ASLing), and explore the role of different types of visual features (CNN embeddings, human body keypoints, and optical flow vectors) in translating it to spoken American English. We propose a novel Transformer-based, visual feature learning method for ASL translation. We demonstrate the explainability efficacy of our proposed learning methods by visualizing activation weights under various input conditions and discover that the body keypoints are consistently the most reliable set of input features. Using our model, we successfully transfer-learn from the larger GSL dataset to ASLing, resulting in significant BLEU score improvements. In summary, this work goes a long way in bringing together the AI resources required for automated ASL translation in unconstrained environments.**

I. Introduction

As many as 5% of Americans are currently Deaf and Hard-of-Hearing (DHH) [26] and as such, American Sign Language (ASL) is their primary mode of communication. But in the hearing-centric world we live in, the majority of the general population do not understand ASL and inadvertently require DHH individuals to communicate in ways other than their natural mode of communication. But in spite of its popularity, ASL has received very little computational research attention, probably due to limited data and the complexity associated with being a visual language in a mostly hearing-centric environment.

But the growing successes of translating between spoken languages have inspired machine translations of visual languages such as sign language into spoken/written ones and we discuss several of these methods in this paper. But the absence of annotated large-scale, parallel phrase-level ASL datasets and applicable NLP tools potentially qualifies it as a low-resource language. Hence, in this work, we contribute to the growing body of work in continuous sign language analysis through the following:

This material is based upon work partially supported by the National Science Foundation under Grant No. 1846076.

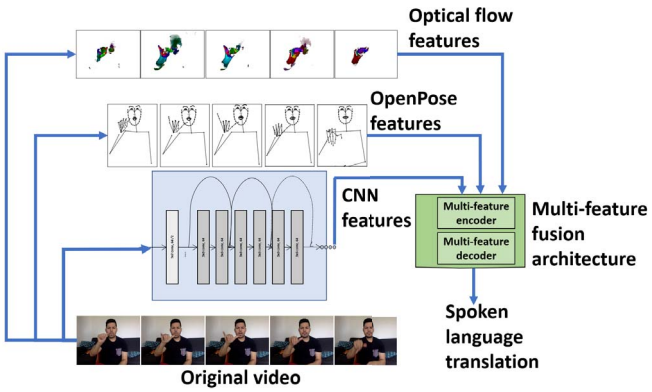


Fig. 1: Multi-feature fusion for sign language translation.

- We collect an ASL dataset by recording expert signers as they sign written English phrases provided. Although relatively small, to the best of our knowledge it is the largest phrase-level ASL dataset available. This ASL linguistic dataset (which we refer to as ASLing) will be made publicly available, to add to the growing body of sign language resources. Details are provided in Section IV.
- We develop a novel multi-feature fusion architecture, cross-feature dual fusion (CFDF), based on transformer network. This network was designed to translate signing videos of varying quality, focusing purely on translation without gloss, the intermediary written symbolic representation of a sign language.
- We perform experiments using the multi-feature architecture on ASLing and compare the results with experiments from the baseline transformer using individual features on the same dataset.
- We explore the explainability efficacy of the models and determine how each of the input features contributes to the final translation. We accomplish this by visualizing the attention weights over time, highlighting the influence of each feature over the sequence.
- We improve the performance of the unconstrained and uncontrolled ASLing dataset by transfer learning from the larger well-tested German sign language (GSL) dataset.

In a real-life scenario, the input sign language video cannot always be captured in a controlled, and well-balanced environment. Hence, in this work, we perform automated

SLT on an unconstrained and uncontrolled sign language dataset (ASLing) using a multi-feature fusion architecture shown in Fig. 1. The model is adaptive to the variations in the input frames and attends to different features based on the level of information rendered.

A. Sign language transcription versus translation

Completely transcribing sign language gestures (or signs) into written language, or attempting to use only written language to fully represent signs is an extremely challenging task. Many signs incorporate movement and space (within the sign) to modify the meaning. It is therefore challenging to encapsulate such spatially oriented information into words.

Also, much of sign language grammar is conveyed specifically by facial and body movements, not present in written texts, thereby rendering them even more challenging to encapsulate. For these reasons, sign languages are often transcribed first into an intermediary written representation called *gloss*, which captures both the sign-for-sign word ordering and the different notations needed to account for spatial-temporal, facial, and body grammar.

Sign language *recognition* or *transcription* involves the process of converting signed visual phrases into gloss, whereas sign language *translation* involves going directly from signs to the spoken version of the language. Unlike many recent works in sign language analysis [10], [6], where recognition is used as an added step to boost translation, in this work, we focus strictly on the challenging task of translation only, especially when gloss is not available as is the case for many low-resource sign languages.

II. Related Works

Natural language processing (NLP) has laid a strong foundation for sign language translation. In the following subsections, we will discuss how NLP tasks have evolved to perform sign language translation as well as some of the recent multi-modality and multi-feature techniques that inspired our work in this paper.

A. Evolution of sign language translation:

Neural machine translation (NMT) in NLP involves translating text from one language to another. A probabilistic continuous translation model was introduced by *Kalchbrenner and Blunsom* [18] and in the deep learning era *Sutskever et al.* [40] introduced multilayered LSTM for text-to-text translation. Further improvements on the sequence modeling were achieved by using attention mechanisms. Some of the SOTA work with attention using long short-term memory (LSTMs) for the text-to-text tasks followed in [2], [46], [25]. To account for longer dependencies *Vaswani et al.* [43] introduced the transformer model with a multi-head self-attention mechanism without using any recurrent neural networks (RNN) for text-to-text translation. This work further expanded towards other text-to-text tasks such as text classification, question-answering (Q&A), summarization, etc, in the Generative Pre-training (GPT) models (GPT 1 -3) [34], [35], [4]. Unsupervised mechanisms were introduced in bidirectional encoder representations from transformers (BERT) [13] for Q&A tasks.

The text translation tasks further evolved towards video captioning. One of the most popular SOTA methods was introduced by *Venugopalan, et al.* [44] where the sequence of video frames was passed as input to a two-layer LSTM, and one word at a time describing the video was predicted as output. Other methods for video-to-text with LSTM models improved upon the base models by using attention mechanism variations [50], [15], [30], [3], [29]. The transformer model was also expanded further for video captioning, visual Q&A tasks using the transformer and BERT models and by performing co-attention [52], [8], [38], [39], [24], [11].

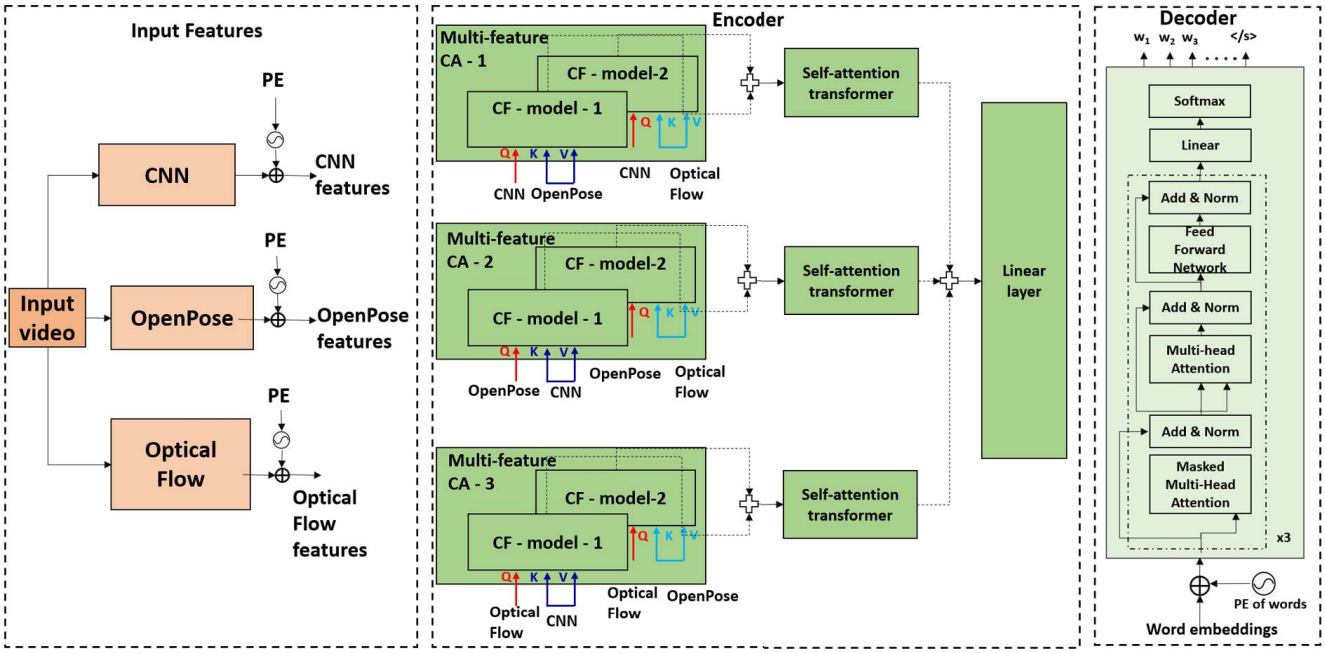
Video captioning methods can be extended to sign language translation and recognition tasks. Word/gloss level sign language recognition became popular from the image captioning/ gesture recognition tasks for predicting different hand, mouth shapes, and signs [27], [20], [21], [23], [47], [32]. Continuous sign language translation is a complex task when compared to word-level sign language translation which is more of a classification task. Variants of RNNs for continuous sign language were seen in DeepASL [14] with bidirectional deep RNN, hybrid model using temporal convolutions and bidirectional gated recurrent unit (BGRU) in [45], a multi-layered attention-based LSTM network in [10], and a two-layered LSTM network with different hand, body, face features for Chinese sign language (CSL) [51] for SLT, to name a few. Connectionist temporal classification (CTC) loss based processing was used by [16], [33], [9] to improve upon the SLT task. Transformer based models are also gaining popularity for SLT tasks. *Camgoz et al.*[6] used the transformer model with CTC loss and performed sign language recognition and translation with features obtained from their CNN-LSTM-HMM model trained on gloss information [19]. Other variations of the transformer model considering gloss were seen in [48], [49].

B. Gloss-aligned visual features

In this section, we briefly review one of the more successful input features that has been used for sign language understanding. The model used by several researchers [5], [6], yielding good metrics on benchmark datasets are based on a CNN-LSTM-HMM model [19]. The CNN-LSTM (Convolution Neural Network, Long Short-Term Memory) part is initially trained as a classifier in a weakly supervised manner to identify the gloss based classes, hand, and mouth shapes; then the probabilities from the CNN-LSTM model are fed to a hidden Markov model (HMM), further used for alignment. This CNN model is then initialized with the pretrained weights and used to extract features for the sign language video under consideration. Hence gloss is used also in the feature extraction process.

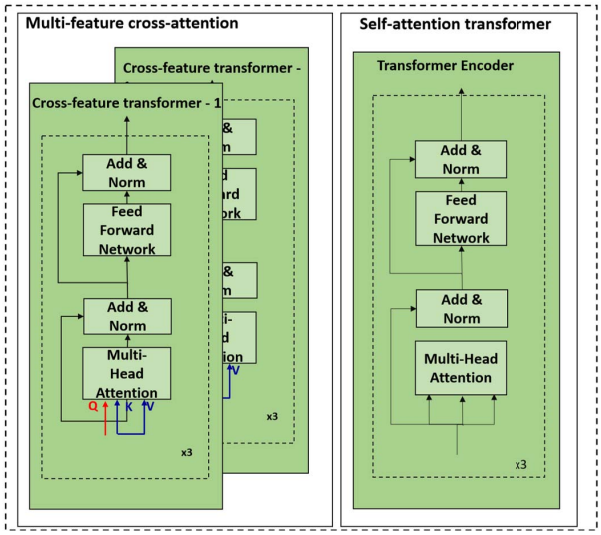
III. Methodology

We investigate the efficacy of a novel multi-feature architecture (inspired by [43], [42], [1]), the CFDF transformer



(a)

Fig. 2: Cross-feature fusion based transformer model for sign language translation.



(b)

Fig. 3: Multi-feature cross-attention and Self-attention for sign language translation.

model and compare it with a single-feature transformer-based architecture.

A. *The single-feature transformer:*

For each of the feature embeddings, the positional encoding is added to take into consideration, the frame order as shown in (1).

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \quad (1)$$

This input is then fed to a 1D convolutional layer and the embeddings are fed to the transformer encoder which learns multi-head self-attention. The query (Q), key (K), and values V for an individual transformer encoder is calculated as (2).

$$K_f = linear\_key(input_f)$$
$$Q_f = linear\_query(input_f)$$
$$V_f = linear\_value(input_f) \quad (2)$$

where, *linear_key*, *linear_query*, and *linear_value* are three individual linear layers learned for the input feature ($input_f$) and $f \in \{\text{CNN, OP, OF}\}$. Hence, we compute the key, query and value parameters for CNN, OpenPose (OP), optical flow (OF), extracted from the sign language video. These highlight the different modalities being evaluated.

For the stand-alone feature test, each of the three different input features (CNN, OP, and OF) are fed to the standalone transformer encoders one at a time and the performance results are recorded. The governing equation is:

$$Attention(K_f, Q_f, V_f) = softmax(\frac{Q_f K_f^T}{\sqrt{d_k}})V_f \quad (3)$$

B. Cross-feature dual fusion transformer:

The cross-feature architecture is shown in Fig. 2. Three different input features are simultaneously fed to the multi-feature cross-attention (CA) block after adding the positional encoding (PE) information. These inputs are passed through a 1D convolutional network before passing to the next stage.

To understand the CFDF architecture, let us consider one cross-attention block (Multi-feature cross-attention-1) from Fig 2, details expanded and shown in Fig. 3.

Equation (4) describes how the attention for the two cross-feature transformer models attending on the CNN features as the base feature is calculated:

$$\begin{aligned} Cross\_attention_1 &= softmax(\frac{Q_{CNN}K_{OP}^T}{\sqrt{d_k}})V_{OP} \\ Cross\_attention_2 &= softmax(\frac{Q_{CNN}K_{OF}^T}{\sqrt{d_k}})V_{OF} \end{aligned} \tag{4}$$

where, Q_{CNN} (CNN feature modality), K_{OP} (OpenPose feature modality), V_{OP} (OpenPose feature modality) acts as the query, key, and value inputs, respectively, for the first cross-feature transformer and Q_{CNN} (CNN feature modality), K_{OF} (optical flow modality), V_{OF} (optical flow modality), respectively, for the second. Similarly, cross-attentions are calculated for the other multi-feature cross-attention block 2 and 3 with their respective base features.

The fused output from the cross-feature transformer block is passed over to the self-attention block (expanded and shown in Fig. 3) where the model learns a higher degree of attention between the cross-feature attention blocks. Similar representations are obtained from the other cross-multi-feature attention blocks and self-attention blocks. The output from each of the self-attention blocks is further fused before passing through a final linear layer to learn the projections.

The output of the encoder which represents the relation between different features of the input video is passed as input to the multi-head attention block of the decoder thus learning the cross-attentions between the different sign features and the attentions from the words. The decoder takes in as input the word embeddings and performs masked-multi head attention by masking the future words. CFDF encoder output is fed to the multi-head attention block in the decoder where it learns the encoder-decoder attention and predicts the words after passing through a feed-forward network, linear, and softmax layers.

C. Training Information

The CFDF transformer model is trained on 1027 ASLing samples and tested on 257 held out samples. Adam optimization is used with a learning rate of $1e^{-03}$ and a weight decay rate of $1e^{-03}$. The maximum length of frames in each batch is chosen as the input sequence length for the encoder and the decoder is fixed at a maximum caption length of 30 based on the average length of the captions. The model has **128.17** million trainable parameters.

CFDF models were trained for 70 - 150 epochs. Other optimal model settings used are encoder-decoder embedding size of 512, along with 3 encoder and decoder layers, and 8 multi-head attention blocks.

IV. Datasets

A. Sign Language Datasets

We evaluate our cross-feature fusion method on the low-resource American sign language dataset (ASLing) and the well annotated German sign language dataset (GSL).

Fig. 4: Dataset samples: the top row shows the GSL signers in a controlled and constrained setting and the bottom row shows the ASLing signers in a real-life, more naturalistic setting. Compare the clothing of the signers as well as the backgrounds in the two datasets. Also, observe the lighting conditions in both datasets.

B. American Sign Language (ASL)

The American Sign Language (ASLing) dataset consists of 1027 training and 257 testing samples. We interchangeably use ASLing and ASL, both refer to the same dataset in our work. These videos were collected at 10 frames per second and were annotated by 7 signers. Each frame is of 450×600 size. The ASL dataset consists of a wide variety of topics unlike GSL that is more constrained on weather-related topics.

Although ASLing is currently the largest phrase-based ASL corpus, the data is very noisy (collected in real-life settings) and is thus challenging to analyze. The presence of poor illumination during collection results in low quality images, making the dataset even more challenging to analyze. These issues are in contrast with the GSL dataset which was collected under significantly more controlled conditions.

The word cloud in Fig. 5 (a) shows the organization of different words in the dataset. Note the variation in ASL word topics compared to GSL. The word and sentence repetition frequency for the ASLing dataset is shown in Fig. 7. Please contact the authors for the dataset.

C. German Sign Language (GSL)

The German Sign Language (GSL) dataset is obtained from the weather forecast airings from the RWTH-PHOENIX-Weather dataset publicly available [28]. The dataset consists of 7096/519/642 train/val/test samples. Each frame is of 210×260 size and the videos were recorded at 25 frames per second. The dataset is annotated by 9 signers. The word cloud in Fig. 5 (b) depicts that the dataset is limited to a particular subject area. The sentence repetition and word repetition are shown in Fig. 7.

Fig. 5: Word cloud for the GSL and ASLing datasets (best viewed in color).

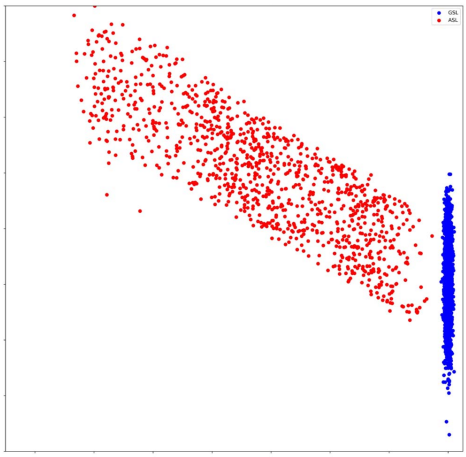


Fig. 6: Embedding space for the GSL and ASLing datasets using PCA (best viewed in color) (Red - ASLing, Blue - GSL).

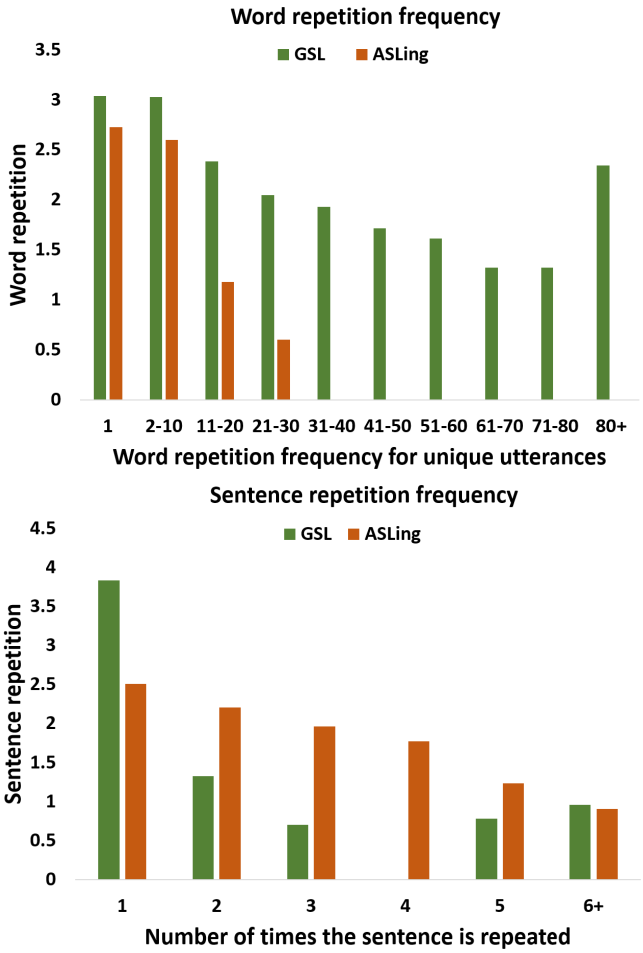


Fig. 7: Top: Word repetition frequency for unique utterances. Bottom: Sentence repetition frequency. Y-axis represents the log10 scale.

To understand the distribution of the two datasets, we converted the German words to English and projected the word2vec embeddings using PCA for both the GSL and ASL dataset as shown in Fig. 6. The ASL (shown in red) dataset is widely spread out whereas the GSL (shown in blue) dataset covers a very confined subject.

D. Feature Selection

To take full advantage of the multi-feature fusion models, we consider three different features from the input video that highlight different modalities. We extract 2048 dimension CNN ResNet50 [17] features for each frame, pretrained on ImageNet [12]; these provide information on the visual RGB frames. We obtain 25 points for the body, 21 points for each hand, and 70 face points from OpenPose [7]. These points are (x, y) locations of the body, hands, and face. We convert these raw locations into a canonical, smoothed, and normalized form which helps the model in better learning and training without exploding gradients. The canonical form is obtained by scaling and centering the points. To perform smoothing we use Savitzky-Golay (SavGol) [36] and perform frame-to-frame smoothing and finally normalize the points to fall between $0-1$. We also extract 2048 dimension vector for each frame from optical flow [37], [41] which provide the flow-based information between consecutive frames.

V. Results

A. CFDF performance

We test our cross feature model on the low-resource ASLing dataset. We initially trained our model using only a single attention block belonging to individual features which mainly learns self-attention to understand how much each of the individual features contributes. Further, we train our CFDF multi-feature fusion model using all three feature inputs. Our CFDF multi-feature fusion model performs better than the individual features by obtaining a $3-4$ points increase across BLEU 1-BLEU 4 scores. BLEU-BiLingual Evaluation Understudy [31] is a popular metric used to test the efficacy of predicted sentences with respect to the ground truth. n-grams (number of words) from predicted caption are compared with n-grams from the ground truth. Comparing individual words (1-gram) is referred to as BLEU 1, two words (2-gram) as BLEU 2, and so on.

Further, to test the efficacy of our architectures we train our CFDF model on the well-constrained, and well-

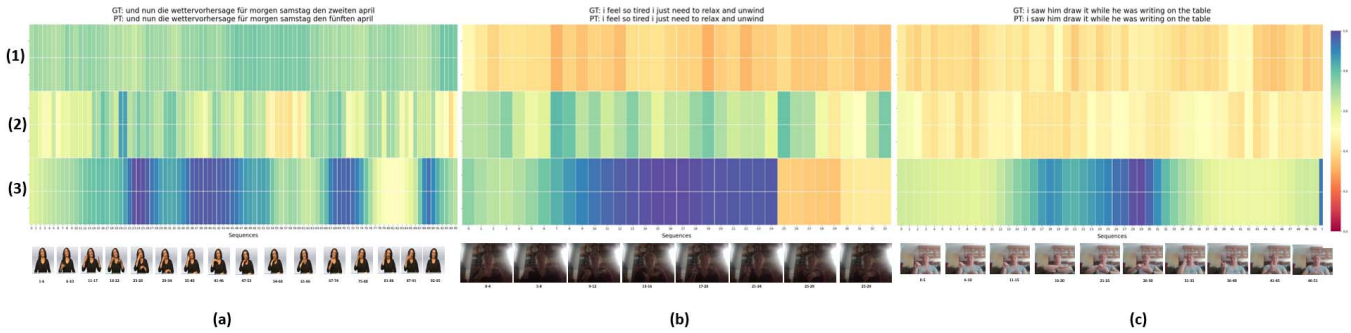


Fig. 8: A0tention visualization (best viewed in color). (a) best sample from the GSL dataset, (b), (c) best samples from the ASL dataset. (1) ResNet50 based fused features, (2) optical flow based fused features, (3) OpenPose based fused features. The samples chosen for visualization were one of the few where the BLEU 1 - BLEU 4 scores were close to 100%.

TABLE I: Ablation study comparing the performance of the CFDF multi-feature architecture with the single-feature architecture. The ablation study was performed on the low-resource ASLing dataset. The results show the BLEU 1 to BLEU 4 (B1, B2, etc) values; TL is the result of transfer learning from the GSL dataset, R50 - ResNet50, OP - OpenPose, OF - Optical Flow

gray!50Architecture	B1	B2	B3	B4
ResNet50 features only (R50)	10.70	4.67	3.21	2.66
OpenPose features only (OP)	10.23	6.79	5.8	5.36
Optical Flow features only (OF)	8.95	4.70	3.39	2.86
CFDF	13.98	7.64	5.60	4.59
CFDF (TL)	**22.39**	**15.96**	**13.56**	**12.25**

controlled[1] GSL dataset and fine-tune the ASLing model by transfer learning from GSL to ASLing. Transfer learning shows significant improvement in the BLEU scores where the BLEU 1 and BLEU 4 scores improved by approximately 8 points. We are continually expanding the current ASLing dataset so that increased data size along with transfer learning will continue to boost the BLEU scores.

The high BLEU scores achieved in some of the state-of-the-art SLT work [6], [10], [5] is due to the presence of gloss or gloss-based features. With respect to an architecture that performs translation without gloss [22] like ours, we perform very similarly on the GSL dataset by achieving a 12.09 BLEU 4 scores against their 13.48 with different features.

Achieving a decent BLEU score on a well-balanced dataset is a challenge in itself. Our work takes it one step further by focusing on a low-resource, unconstrained dataset ASLing where we achieve acceptable BLEU scores as seen in Table I.

Furthermore, we show that CNN features obtained are not always the best as it mainly depends on the quality of the input video. With a less noisy dataset like GSL, the CNN features extracted from 3D convolution networks, ResNet50, work well. However, they can fail in settings where the image quality is poor. We, therefore, use a fusion of features (CNN, location-based, flow-based) and the models proposed adapt to these features based on the quality of frames and information rendered from the frames. We confirm this further by visualizing the attention weights for select frames from GSL and ASLing datasets, described in the next subsection.

[1]Well-constrained and controlled refers to videos collected with a uniform background, the homogeneous appearances of the signers, their similar distances from the camera.

B. Attention visualization

To understand the contribution of these three sets of features we looked at one of the test samples that gave the best BLEU scores for the GSL and the ASL dataset. The attention visualization is shown in Fig. 8.

The x-axis on the visualization are the different frames of the video under consideration and the y-axis is the combination of features. To visualize we store the attention weights from the last layer of each CFDF encoder and plot the heatmap against the sequence of frames.

Fig. 8 (a) shows the attention for one of the best test samples of the GSL dataset. From the frames, we can see that all the frames are pretty uniform in terms of the person annotating and the background. Thus, the model seems to be attending to ResNet50 based fused features almost equally in all the frames. The model attends to the optical flow based fused features where there is motion between the frames, for example, frames 13 - 32, 63 - 70, see a lot of changes between the consecutive frames leading to higher attention in these areas whereas, frames $47-61$ seem to have less motion between the signs leading to less attention. The model attends to OpenPose based fused features strongly for some sections of the video. Similarly, we picked a couple of samples from the ASL dataset which performed the best. From Fig. 8 (b) we can see that the frames under consideration are very dark in terms of the RGB visualization. The quality of the frames has a direct impact on the ResNet50 based fused features hence the model does not attend as well when the image quality is poor. We see a similar pattern in both the datasets where the model attends fairly well to optical flow based features. We see a similar trend with OpenPose based fused features as well. We see similar trends in the other ASL sample as shown in Fig. 8 (c) where the image quality is low so the model does not attend as much to ResNet50 and optical flow based fused features.

Overall, we can see that the model often attends to OpenPose based fused features the most as OpenPose di-

rectly targets the hand, body, face keypoints which are the most essential for sign language. The model benefits from ResNet50 whenever there is a good quality video and benefits from optical flow whenever the flow movements between the frames are prominent. In a real-life scenario, there is no surety of having a good quality sign video, controlled image, prominent hand, and body movements, and/or face expressions. In situations like these, if only one of the features is used as input to the model, the output may not be accurate as the model could not attend to details based on the input feature. However, with fusion using CFDF, the model learns to attend to information from each of the features whenever and wherever applicable making the model robust to the changes in a real-life scenario.

VI. Conclusion

In this work, we introduced the unconstrained ASLing dataset collected in real-world settings, where the participants could dress in their regular everyday clothes (with multiple textures and colors) andthe signing videos were collected in arbitrary, unconstrained settings including the dorm rooms of DHH students. We focused on developing models to help us understand how best to perform high-quality ASL translation with multi-feature models, and in the absence of any gloss information.

To this end, we introduced the cross-feature dual-fusion (CFDF) architecture, and provided it with multiple features (ResNet50 visual embeddings, OpenPose - body, hands, and face keypoints, and optical flow - frame-to-frame motion vectors) as inputs. We observed that this model dynamically attended to the ResNet50 features when the visual quality of the input frame was good, but the model attended more to the keypoints (body, hands, and face) for most of the frames. The model also attended to optical flow whenever there was a lot of movement and good flow-based information in the inputs.

We discovered that we could successfully fine tune from a larger dataset (GSL) to boost the performance of the multi-feature fusion models on the ASLing dataset. This transfer learning paradigm significantly increased the resulting BLEU 1-4 scores on ASLing. To understand the inner workings of the models, we visualized the attention weights based on the three fused features and found that the model dynamically learned and attended to each of these features based on the input frame type.

In summary, in our research, we focus on improving the AI resources for ASL, which is still a low-resource language, in spite of all the resources readily available for its spoken/written counterpart.

References

[1] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the ACL*, pages 2236–2246. Association for Computational Linguistics, July 2018.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] L. Baraldi, C. Grana, and R. Cucchiara. Hierarchical boundary-aware neural encoder for video captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. 2020.

[5] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden. Multi-channel transformers for multi-articulatory sign language translation. *arXiv preprint arXiv:2009.00299*, 2020.

[6] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. *arXiv preprint arXiv:2003.13830*, 2020.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[8] M. Chen, Y. Li, Z. Zhang, and S. Huang. Tvt: Two-view transformer network for video captioning. In J. Zhu and I. Takeuchi, editors, *ACML*, volume 95, pages 847–862. PMLR, 14–16 Nov 2018.

[9] N. Cihan Camgoz, S. Hadfield, O. Koller, and R. Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *ICCV*, Oct 2017.

[10] N. Cihan Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden. Neural sign language translation. In *CVPR*, pages 7784–7793, 2018.

[11] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *CoRR*, abs/1901.02860, 2019.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 CVPR*, pages 248–255. Ieee, 2009.

[13] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[14] B. Fang, J. Co, and M. Zhang. Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, pages 1–13, 2017.

[15] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19(9):2045–2055, 2017.

[16] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd ICML*, pages 369–376, 2006.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] N. Kalchbrenner and P. Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, 2013.

[19] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *TPAMI*, 42(9):2306–2320, 2020.

[20] O. Koller, H. Ney, and R. Bowden. Deep learning of mouth shapes for sign language. In *ICCV*, pages 85–91, 2015.

[21] O. Koller, H. Ney, and R. Bowden. Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *CVPR*, pages 3793–3802, 2016.

[22] D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, and H. Li. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation, 2020.

[23] X. Li, C. Mao, S. Huang, and Z. Ye. Chinese sign language recognition based on shs descriptor and encoder-decoder lstm model. In *Chinese Conference on Biometric Recognition*, pages 719–728. Springer, 2017.

[24] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

[25] M. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025, 2015.

[26] R. E. Mitchell. How Many Deaf People Are There in the United States? Estimates From the Survey of Income and Program Participation. *Journal of Deaf Studies and Deaf Education*, 11(1):112–19, 2005.

[27] B. Mocialov, G. Turner, K. Lohan, and H. Hastie. Towards continuous sign language recognition with deep learning. In *Proc. of the Workshop on the Creating Meaning With Robot Assistants: The Gap Left by*

Smart Devices, 2017.
[28] O. K. H. N. R. B. Necati Cihan Camgöz, Simon Hadfield. Rwth-phoenix-weather 2014 t: Parallel corpus of sign language video, gloss and translation. ICPR 2018, 05 2018.
[29] S. Olivastri, G. Singh, and F. Cuzzolin. An end-to-end baseline for video captioning. *CoRR*, abs/1904.02628, 2019.
[30] Y. Pan, T. Yao, H. Li, and T. Mei. Video captioning with transferred semantic attributes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
[32] L. Pigou, M. Van Herreweghe, and J. Dambre. Gesture and sign language recognition with temporal residual networks. In *ICCVW*, pages 3086–3093. IEEE, 2017.
[33] J. Pu, W. Zhou, and H. Li. Dilated convolutional network with iterative optimization for continuous sign language recognition. In *IJCAI*, pages 885–891, 2018.
[34] A. Radford. Improving language understanding by generative pre-training. 2018.
[35] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.
[36] A. Savitzky and M. J. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.
[37] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, volume 27, pages 568–576, 2014.
[38] C. Sun, F. Baradel, K. Murphy, and C. Schmid. Contrastive bidirectional transformer for temporal representation learning. *CoRR*, abs/1906.05743, 2019.
[39] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. Videobert: A joint model for video and language representation learning. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
[40] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
[41] J. Sánchez Pérez, E. Meinhardt-Llopis, and G. Facciolo. TV-L1 Optical Flow Estimation. *Image Processing On Line*, 3:137–150, 2013.
[42] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, volume 2019, page 6558, 2019.
[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *NIPS*, 30:5998–6008, 2017.
[44] S. Venugopalan, M. Rohrbach, J. Donahue, R. J. Mooney, T. Darrell, and K. Saenko. Sequence to sequence - video to text. *CoRR*, abs/1505.00487, 2015.
[45] S. Wang, D. Guo, W.-g. Zhou, Z.-J. Zha, and M. Wang. Connectionist temporal fusion for sign language translation. In *Proceedings of the 26th ACM*, pages 1483–1491, 2018.
[46] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.
[47] Y. Ye, Y. Tian, M. Huenerfauth, J. Liu, N. Ruiz, E. Chong, J. M. Rehg, S. Palsson, E. Agustsson, R. Timofte, et al. Recognizing american sign language gestures from within continuous videos. In *CVPR*, pages 2064–2073, 2018.
[48] K. Yin. Sign language translation with transformers. *arXiv preprint arXiv:2004.00588*, 2020.
[49] K. Yin and J. Read. Attention is all you sign: Sign language translation with transformers. In *ECCV*, 2020.
[50] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, June 2016.
[51] T. Yuan, S. Sah, T. Ananthanarayana, C. Zhang, A. Bhat, S. Gandhi, and R. Ptucha. Large scale sign language interpretation. In *FG 2019*, pages 1–5, 2019.
[52] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong. End-to-end dense video captioning with masked transformer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.