PseudoDepth-SLR: Generating Depth Data for Sign Language Recognition

Noha Sarhan, Jan M. Willruth, and Simone Fritnrop

Universität Hamburg, Vogt-Kölln-Str. 30, 22527 Hamburg, Germany
{noha.sarhan, simone.frintrop}@uni-hamburg.de
jan.willruth@studium.uni-hamburg.de

Abstract. In this paper, we investigate the significance of depth data in Sign Language Recognition (SLR) and propose a novel approach for generating pseudo depth information from RGB data to boost performance and enable generalizability in scenarios where depth data is not available. For the depth generation, we rely on an approach that utilizes vision transformers as a backbone for depth prediction. We examine the effect of pseudo depth data on the performance of automatic SLR systems and conduct a comparative analysis between the generated pseudo depth data and actual depth data to evaluate their effectiveness and demonstrate the value of depth data in accurately recognizing sign language gestures. Our experiments show that our proposed generative depth architecture outperforms an RGB-only counterpart.

Keywords: Sign Language Recognition · Deep Learning · Depth Data · 3D Convolutional Neural Networks.

1 Introduction

Sign languages are rich and complex visual languages used by the deaf community for communication. With their own grammar, syntax, and vocabulary, sign languages serve as vital means of expression and facilitate communication among individuals who are deaf or hard of hearing. However, the comprehension and interpretation of sign languages remain a significant challenge for the wider hearing population. Automatic Sign Language Recognition (SLR) has emerged as a promising solution to bridge the communication gap between sign language users and non-signers, aiming to develop systems that can automatically recognize and interpret sign language gestures [12, 6, 20].

SLR can be viewed as a very specific case of human action recognition, a rather challenging one. This is attributed to the unique nature of sign language, which incorporates both overall body motion and intricate arm/hand gestures to convey its meaning. Facial expressions also play a role in conveying emotions [13]. In addition, different signers may perform gestures differently, e.g. in terms of speed, left- or right-handed, etc. Consequently, SLR becomes even more challenging due to the need for diverse data samples from numerous signers, however, sign language data is hard to acquire, owing to several challenges such as privacy and the need for experts to perform and annotate datasets.

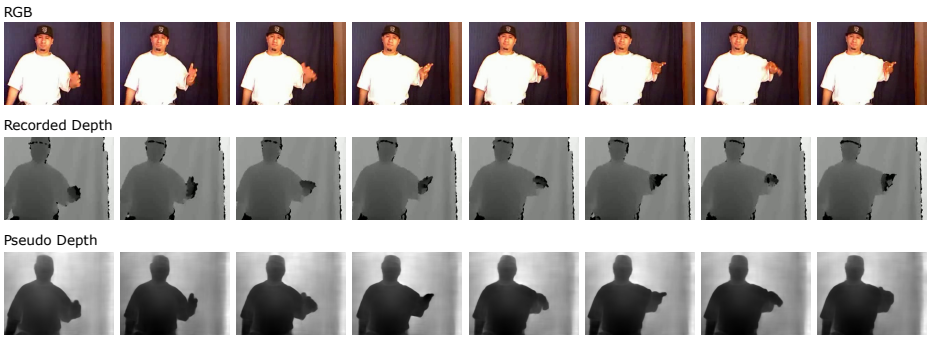

Fig. 1. Example of RGB (top) and recorded depth (middle) and generated pseudo-Depth for the gesture DivingSignal/SomethingWrong.

In order to capture the full dynamics of the gesture, SLR methods rely on the use of several input modalities [3, 10, 14]. Besides RGB data, one of the key modalities commonly utilized in SLR is depth data, offering rich spatial and depth information that enhance the discrimination of signs that would otherwise seem similar. In addition, the use of depth in conjunction with RGB data can be helpful when distinguishing the signer from a cluttered background, and hands from body, such ensembles lead to improved recognition accuracy and robustness. Consequently, the majority of state-of-the-art SLR systems heavily rely on depth data [10, 15, 8, 19, 16]. While depth data has some benefits such as robustness to lighting conditions, some existing sign language dataset lack depth information (e.g. from news broadcasts [12]). In addition, relying on depth data hinders the generalizability of existing models to SLR and vice versa.

Recent research has explored alternative approaches that aim to eliminate the requirement for depth information, proposing the use of solely RGB data [20, 21]. These approaches challenge the assumption that depth data is indispensable, suggesting that comparable performance can be achieved without it.

In this paper, we investigate the value of depth data in SLR systems and its impact on overall performance in comparison to RGB-based systems. Additionally, we propose an alternative approach to address the limitations posed by datasets lacking depth information, aiming to bridge the depth gap in SLR. Specifically, we explore the generation of pseudo depth data, which allows for the creation of depth-like information from RGB data. For depth generation, we utilize an architecture, namely DPT (Dense Prediction Transformer) [17] that uses vision transformers [7] as a backbone for the dense depth prediction. Figure 1 shows an example of RGB and depth input modalities and the corresponding generated depth data for a sign language gesture. By comparing the generated depth data with the actual depth data, we validate the efficacy of our method and its potential for enhancing SLR in scenarios where depth data is scarce, or extend it to gesture recognition applications where depth data might be non-existent, e.g. automotive, sports training, etc.

Our main contributions can be summarized as follows:

- We evaluate and analyze the significance of depth data for sign language recognition.
- We propose an alternative approach in case of absence of depth data by the generation of pseudo depth data. We examine its implications on the performance of SLR systems and compare it to the use of actual depth data.
- Our proposed alternative method outperforms methods that disregard depth data completely, while still relying only on RGB modality.

2 Related Work

Significant progress has been made in SLR since the advancements in depth sensing technologies, such as time-of-flight cameras and structured light sensors, have enabled more accurate and detailed depth measurements. Having depth data together with RGB data has helped capture the complex spatial and temporal dynamics of sign language gestures. Since then, depth data has been recognized as a valuable input modality in SLR, and its incorporation has been evident across various methodologies employed over the years. This section provides an overview of related works, highlighting the persistent usage of depth data in SLR, from early hand-crafted feature-based approaches [24, 19, 1, 5] to recent state-of-the-art methods that rely on the advancements in machine learning and computer vision techniques [10, 8, 20].

Early approaches in SLR primarily relied on handcrafted features, such as shape, motion and appearance descriptors, combined with traditional machine learning algorithms such as support vector machines and hidden Markov Models [24, 19, 1, 16]. This resulted in systems that have very limited generalization capabilities, unable to go beyond applications that lack these extra modalities. However, the introduction of deep learning architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), revolutionized SLR by leveraging their ability to automatically learn discriminating features from data [21, 20, 15].

Automatic SLR(ASLR) has long relied on depth data as a fundamental component for accurate and robust recognition. To date, most deep learning based methods still rely on the depth modality. Wang *et al.* [23] relied on both RGB and depth information. They utilized full frames to represent the fully body, along with hand crops of both modalities. They fused together a 4-stream ConvNet and 3D ConvLSTMs-based classification to get an average-score fusion. Miao *et al.* [15] extracted information from RGB and depth input data using and concatenated RGB, flow and depth features using a SVM classifier.

Furthermore, some research rely on extracting even more information from the input depth maps, e.g. depth saliency, depth flow, etc. Jiang *et al.* [10] propose an ensemble of five 3D CNN streams, two of which rely on depth data. For the first stream, they extract flow information from the input depth maps to feed to the network, for the second stream they extract HHA (Horizontal disparity, Height above the ground, and Angle normal) features from the depth

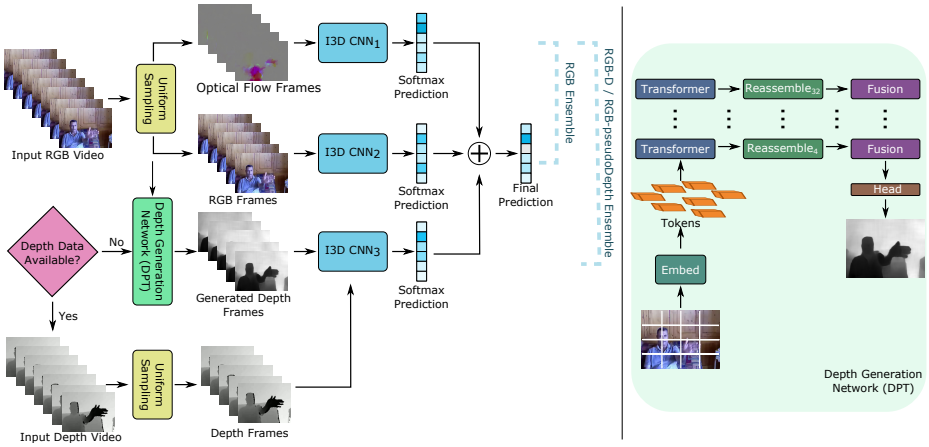

Fig. 2. *Left:* Architecture overview. The RGB ensemble comprises 2 streams of Inflated 3D CNNs (I3D), one is fed RGB frames, and the other optical flow frames. For the inclusion of depth, a third stream is used. It is fed depth frames if depth data is available, otherwise, they are first generated from the RGB frames. For each of the 3 streams, the softmax predictions are averaged together to yield a final label. *Right:* Depth generation network architecture, adapted from [17]. An RGB frame is first transformed into tokens by extracting non-overlapping patches. The tokens are then passed through multiple transformer stages, and reassembled from different stages into an image-like representation at multiple resolutions. Finally, fusion modules progressively fuse and upsample the representations to generate the depth prediction. Details of the Reassemble and Fusion units can be found in [17].

stream, encoding depth information into a 3-channel RGB-like output. Duan *et al.* [8] proposed a two-stream consensus voting network, extracting spatial information from RGB and depth input, and temporal information from RGB and depth flow data. They also aggregated a 3D depth-saliency ConvNet stream in parallel to identify subtle motion characteristics. Late score fusion was adopted for the final recognition.

In attempt to refrain from using depth data, Sarhan and Frintrop [20] proposed a mathod that relies only on RGB data. However, depth data brings about additional information and robustness (e.g. against illumination changes, noise, and background clutter) that RGB alone does not deliver. Therefore, in this work we attempt to bridge the depth gap and propose the generation of pseudo depth data from RGB data for sign language recognition.

3 Methodology

In this section, we first introduce our proposed model, which utilizes depth data along with RGB data. Afterwards, we will explain our method for the generation of pseudo depth data.

3.1 Proposed Architecture

Figure 2 shows our proposed architecture which is made up of 2 ensembles comprising 3 streams in total. The most widely used 3D CNN architectures are Inflated 3D CNNs (I3D) [4], ResNeXt3D-101 [9], and separable 3D CNNs (S3D) [25]. Therefore, for the RGB ensemble, we opted for the I3D CNN-based architecture proposed by Sarhan and Frintrop [20] as I3D CNNs have shown outstanding performance in isolated SLR [20, 21, 10]. Each stream has one I3D CNN, where the 2D $k \times k$ filters and pooling kernels are inflated into 3D kernels by adding a new dimension t and become cubic through this transformation $t \times k \times k$, spanning t frames. The first stream is fed full-frame RGB sequence as input, while the second stream is fed optical flow data, which are generated from the RGB stream using Dual TV-L1 algorithm [26].

In order to capture features from depth data, we introduced a new third stream, which is fed the recorded depth sequence. Together with the two RGB streams, we present this as the RGB-D ensemble. All input videos are first uniformly sampled to extract a fixed number of frames before being fed to the I3D CNNs.

As a final step, we opt for a late fusion scheme, where the softmax predictions of all three streams are averaged together to yield a final prediction for the signed gesture.

3.2 Pseudo Depth Data Generation

In order to refrain from using recorded depth data while still including depth information, we propose an alternative which is to generate pseudo depth data from the RGB images. In that case, the RGB frames are first used to generate the pseudo depth images, and then fed to the I3D CNN as shown in Figure 2 (left). Together with the 2 RGB streams, we refer to this as the RGB-pseudoDepth ensemble.

To generate high quality dense depth maps, we tested two different methods for depth prediction: DPT by Ranftl *et al.* [17], and DenseDepth by Bhat *et al.* [2]. According to our ablation study in Section 6.2, We opted for the encoder-decoder based method by Ranftl *et al.* [17], namely DPT, since according to our ablation study in Section 6.2 it turned out to be better. Their architecture is depicted in Figure 2 (right). They leverage vision transformers [7] as the backbone for dense prediction, where tokens from various stages of the transformer are assembled into image-like representations at various resolutions and progressively combine them into full-resolution predictions using a convolutional decoder. The transformer backbone processes representations at a constant and relatively high resolution and has a global receptive field at every stage.

4 Experimental Details

In this section, we present the evaluation of our proposed approach on the ChaLearn249 IsoGD dataset [22]. We start by a brief summary about the dataset

and how we evaluate our approach on it. Then, we explain how we evaluate the quality of the generated pseudo depth data. Afterwards, we provide more implementation details, that would aid in making this work reproducible.

4.1 Dataset and Evaluation

Evaluation of proposed Architecture. We evaluate our proposed method on the ChaLearn249 IsoGD dataset [22], a large dataset for isolated SLR. The dataset comprises 47,933 videos, captured by a Microsoft Kinect camera, hence providing RGB and recorded depth images. The dataset is signer-independent, and is one of the mostly used dataset for isolated sign language gestures [15, 8, 23, 14].

For evaluation, we follow the same protocol provided by the dataset. It is split into 35,878 videos for training, 5,784 videos for validation, and 6,271 videos for testing. For all our experiments, we report and compare the accuracy of both the validation and test sets.

Depth Generation Evaluation. To calculate the error between the ground truth recorded depth images and their generated pseudo-depth counterparts, root-mean-square error (RMSE) was calculated as in Equation 1, where p is an individual pixel, n is the number of pixels in each image, y_p is the ground truth, and $\hat{y}_p$ is the estimated value for pixel y.

$$RMSE = \sqrt{\frac{1}{n}\sum_{p=1}^{n}(y_p - \hat{y}_p)^2} \tag{1}$$

In addition, we also calculate the structural similarity index (SSIM) to determine how structurally similar two images are, in this case how similar the generated depth image is to the recorded equivalent. The SSIM of two images x and y is defined in Equation 2.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{2}$$

where μ_x and μ_y are the average of x and y, respectively, σ_x^2 and σ_y^2 are the variance of x and y, respectively, and σ_{xy} is the covariance of x and y. $C_1 = (K_1L)^2$ and $C_2 = (K_2L)^2$ are variables with K_1, $K_2 \ll 1$ and $\mathrm{L} = 2^{\text{bit depth}} - 1$. The value of the SSIM lies between -1 and 1, where a score of -1 means the images are complementary, and a score of 1 means that they are identical.

4.2 Implementation Details

Preprocessing. The video sequences are uniformly sampled into a fixed number of frames. The frames are cropped around the center to a spatial size of 224×224. Optical flow frames have been generated from the RGB videos using the Dual-TVL[1] algorithm [26].

Table 1. Accuracy results on the ChaLearn249 dataset using the RGB ensemble, RGB-D ensemble (RGB + recorded depth), and RGB-pseudoDepth (RGB + generated Depth).

Modality	Validation	Test
RGB [20]	61.76 %	64.97 %
RGB-D	64.54 %	70.63 %
RGB-pseudoDepth	62.5 %	66.02 %

Training. For all streams, we adopt the training scheme used by [20]. The I3D CNN was originally trained on ImageNet [18] before inflation into 3D, and then pretrained on Kinetics dataset [4]. The top, randomly initialized layers are first trained for 3 epochs while freezing the pretrained layers, at a learning of 1×10^{-3}. Afterwards, all layers are fine-tuned and the learning rate is lowered to 1×10^{-4}. Here, early-stopping was adopted to halt training once the validation loss has not improved for 3 consecutive epochs. Adam [11] was used as an optimizer in conjunction with a mini-batch size of 4, and categorical cross-entropy as the loss function.

Data Augmentation. Implementing data augmentation in SLR poses challenges despite its significance for small and medium-sized datasets. Common data augmentation techniques such as image flipping or rotation can directly impact the conveyed sign itself. To address this concern, our approaches focuses on data augmentation by shifting images along the x- and y- axes and adjusting brightness levels.

5 Results and Analysis

In this section, we show the results of using depth and generated depth data on the ChaLearn dataset, along with the per-class accuracies. In addition, we compare their performance to state-of-the-art results on this dataset.

5.1 How significant is depth data?

Results on ChaLearn dataset. In this section, we verify the importance of depth information for SLR. Table 1 shows the validation and test accuracy when using the RGB, RGB-D, and RGB-pseudoDepth ensembles. Recorded depth data (RGB-D) shows the best performance results, 64.54 % on the validation set, and 70.63 % on the test set. While the use of generated Depth data (RGB-pseudoDepth) achieves lower accuracy in comparison, 62.5 % and 66.02 % for the validation and test sets respectively, it still outperforms using only RGB data. This shows that generated depth data is indeed valuable, and that the RGB-pseudoDepth ensemble still captures more features, while relying only on RGB input.

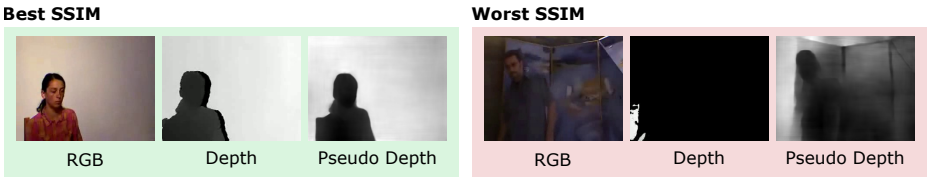

Fig. 3. Example of pseudo depth images with best and worst SSIM measure and their RGB and recorded depth counterparts.

Evaluating the depth streams separately. We also evaluate the accuracy of the recorded and generated depth streams on their own, without the RGB ensemble. The recorded depth stream achieved an accuracy of 50.71 % and 60.56 % on the validation and test sets, while the generated depth stream achieved 38.04 % and 44.97 % on the validation and test sets. While the generated depth stream does not score a very high recognition rate by itself, they still add valuable information to the RGB input, evident by their higher accuracy in the RGB-pseudoDepth ensemble than the RGB ensemble. To evaluate the quality of the generated depth data, the RMSE was 79.42, while the SSIM was 0.67. In Figure 3 we show two examples of generated pseudo depth frames with highest and lowest SSIM. It is clear from these images how the quality of RGB image affects the generated depth data.

Per class accuracy. In addition to the recognition accuracies, to verify the potential from depth and generated depth information, we compared the RGB-D and RGB-pseudoDepth ensembles by calculating the per class accuracy change with respect to the RGB ensemble. In Figure 4 we plot these differences. The blue line represents RGB-D ensemble, while the red line represents the RBG-pseudoDepth ensemble. A positive/negative difference means that the inclusion of depth/generated depth brought about a positive/negative effect over the RGB ensemble. A zero means no improvement over the RGB accuracy. As shown in Figure 4, for the class range c122-c137, using the RGB stream only works well, however in several other classes, such as the range c058-c065 and c192-c197 depth information has resulted in significant improvement. Generally, the higher the positive value of the accuracy difference, the more number of samples there were originally predicted wrong by RGB only, are now predicted correctly. Overall, the average difference in the case of RGB-D ensemble is +3.10, and +0.67 for RGB-pseudoDepth.

5.2 Comparison With State-of-the-Art Results

In Table 2, we compare with top competitors on the leader board and state-of-the-art results on the ChaLearn249 dataset. Our proposed RGB-D architecture that relies on RGB, optical flow and recorded depth data has outperformed the other methods by more than 2.5% on the validation set, and more than 3% on the test set. The use of generated depth was the second best performing method

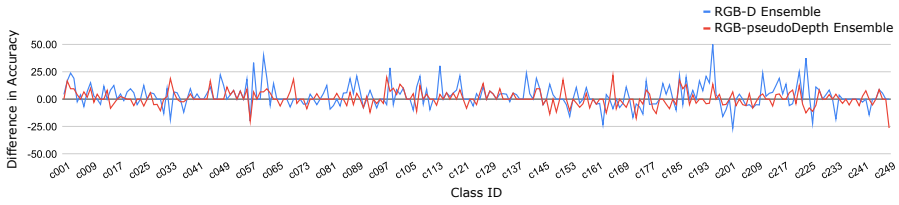

Fig. 4. Differences in class accuracies for RGB-D (blue) and RGB-pseudoDepth (red) ensembles with respect to the RGB ensemble. A positive/negative value means a higher/lower accuracy for that class with respect to the RGB ensemble, a zero means no improvement over the RGB ensemble.

Table 2. Comparison with state-of-the-art on ChaLearn249 and their modalities. The best results are shown in red and the second best in blue.

Method	Modalities			Accuracy	
	RGB	Depth	pseudoDepth	Validation	Test
XDETVP [27]	✓	✓		58 %	60.47 %
AMRL [23]	✓	✓		60.81 %	65.59 %
RGB-pseudoDepth (ours)	✓		✓	62.5 %	66.2 %
SYSU_ISEE [14]	✓	✓		59.7 %	67.02 %
2SCVN-3DDSN [8]	✓	✓		49.17 %	67.26 %
ASU [15][1]	✓	✓		57.88 %	-
RGB-D (ours)	✓	✓		64.54 %	70.63 %

in comparison to the other methods that also relied on recorded depth, but performed slightly lower on the test set. The results in Tables 1 and 2 demonstrate the effectiveness of generated depth data, they do not only outperform RGB-only methods, but are also comparable with methods that rely on recorded depth data.

6 Ablation Study

In this section, we investigate the use of depth flow data as a fourth stream to our architecture. Additionally, we explore an alternative method for pseudo depth map generation. By conducting this ablation study, we aim to gain insights into the specific contributions and significance of each component within the proposed model.

6.1 Depth Flow Data

Since including RGB and optical flow streams has been successful in several approaches for SLR [20, 21, 10, 15], we experimented with adding a fourth stream

[1] We compare with their averaging fusion scheme, similar to what is used in our method for fair comparison. Test set results for that fusion scheme were not reported.

Table 3. Accuracy results on the ChaLearn249 dataset when including depth flow data as a fourth stream to our proposed architecture.

Method	Validation	Test
RGB-D	64.54 %	70.63 %
RGB-D + Depth flow	61.07 %	69.22 %
RGB-pseudoDepth	62.5 %	66.02 %
RGB-pseudoDepth + Depth flow	64.54 %	64.84 %

to our architecture, where the input was optical flow information extracted from the depth data (depth flow). We performed this experiment for both the recorded depth data, and generated depth data, and report these results in Table 3. The use of depth flow data lowered the recognition accuracy in both cases. One possible reason is that depth data usually suffers form noise and uncertainty, affecting the quality of optical flow estimation. These errors can propagate to the optical flow estimation process due to its inherent recurrence.

6.2 Pseudo Depth Data Generation

As an alternative method for the generation of dense depth maps from a single RGB image, we opted for the deep learning-based method by Bhat *et al.* [2], namely DenseDepth, that utilizes fully convolutional networks. Their architecture is composed of two main components: an encoder-decoder block and an adaptive bin-width estimator block called AdaBins. The used model was pretrained on NYU Depth V2 dataset [2]. The dataset is composed of images and depth maps for different indoor scenes, and has 120K training samples and 654 testing samples. As a post processing step, all images have been normalized using Min-Max Normalization.

The results are shown in Table 4. The use of visual transforms clearly outperforms the fully convolutional network method. The use of DenseDepth is still outperformed by using RGB-only ensemble. As for the evaluation of the generated depth images, DenseDepth had an RMSE of 146.64 (vs. 92.67 for DPT), and an SSIM of 0.281 (vs. 0.55 for DPT), explaining the poor results achieved by DenseDepth.

Table 4. Comparison of different depth generation methods.

Method	Validation	Test
RGB	61.76 %	64.97 %
RGB-D	64.54 %	70.63 %
RGB-pseudoDepth (DPT [17])	62.5 %	66.02 %
RGB-pseudoDepth (DenseDepth [2])	60.81 %	64.34 %

7 Conclusion

Depth data has long been recognized as a crucial input modality for SLR systems due to its ability to capture spatial and depth information. However, depth data is not always available, e.g. in news broadcasts, and acquiring it for sign language would be difficult and expensive. In this paper, we aimed to bridge the depth gap and proposed a novel approach for generating pseudo depth data from RGB inputs when recorded depth data is scarce. Our results and analysis further validates the effectiveness of our approach and its potential for improving recognition accuracy in depth-limited scenarios, and open up avenues for SLR research and applications, enabling depth-based insights even when depth data is lacking.

References

1. Badhe, P.C., Kulkarni, V.: Indian sign language translator using gesture recognition algorithm. In: 2015 IEEE international conference on computer graphics, vision and information security (CGVIS). pp. 195–200 (2015)
2. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4009–4018 (June 2021)
3. Boháček, M., Hrúz, M.: Sign pose-based transformer for word-level sign language recognition. In: WACV. pp. 182–191 (2022)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
5. Chai, X., Li, G., Lin, Y., Xu, Z., Tang, Y., Chen, X., Zhou, M.: Sign language recognition and translation with Kinect. In: IEEE conf. on AFGR. vol. 655, p. 4 (2013)
6. Cui, R., Liu, H., Zhang, C.: A deep neural framework for continuous sign language recognition by iterative training. IEEE Trans. Multimed. **21**(7), 1880–1891 (2019)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Duan, J., Wan, J., Zhou, S., Guo, X., Li, S.Z.: A unified framework for multi-modal isolated gesture recognition. TOMM **14**(1s), 1–16 (2018)
9. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2D CNNs and Imagenet? In: CVPR. pp. 6546–6555 (2018)
10. Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., Fu, Y.: Skeleton aware multi-modal sign language recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3413–3423 (2021)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. CVIU **141**, 108–125 (2015)
13. Koller, O., Ney, H., Bowden, R.: Deep learning of mouth shapes for sign language. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 85–91 (2015)

14. Li, B., Li, W., Tang, Y., Hu, J., Zheng, W.: GL-PAM RGB-D gesture recognition. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 3109–3113 (2018)
15. Miao, Q., Li, Y., Ouyang, W., Ma, Z., Xu, X., Shi, W., Cao, X.: Multimodal gesture recognition based on the RESC3D network. In: Proceedings of the IEEE International Cconference on Computer Vision Workshops. pp. 3047–3055 (2017)
16. Pigou, L., Dieleman, S., Kindermans, P., Schrauwen, B.: Sign language recognition using convolutional neural networks. In: European Conference on Computer Vision Workshops. pp. 572–578 (2014)
17. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12179–12188 (2021)
18. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision (IJCV) **115**, 211–252 (2015)
19. Sarhan, N., El-Sonbaty, Y., Youssef, S.: HMM-based Arabic sign language recognition using Kinect. In: Tenth International Conference on Digital Information Management (ICDIM). pp. 169–174 (2015)
20. Sarhan, N., Frintrop, S.: Transfer learning for videos: from action recognition to sign language recognition. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 1811–1815 (2020)
21. Sarhan, N., Frintrop, S.: Sign, Attend and Tell: Spatial attention for sign language recognition. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). pp. 1–8 (2021)
22. Wan, J., Zhao, Y., Zhou, S., Guyon, I., Escalera, S., Li, S.Z.: ChaLearn looking at people RGB-D isolated and continuous datasets for gesture recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition Workshops. pp. 56–64 (2016)
23. Wang, H., Wang, P., Song, Z., Li, W.: Large-scale multimodal gesture recognition using heterogeneous networks. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 3129–3137 (2017)
24. Xiaohan Nie, B., Xiong, C., Zhu, S.C.: Joint action recognition and pose estimation from video. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1293–1301 (2015)
25. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 305–321 (2018)
26. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime TV-L1 optical flow. In: Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29. pp. 214–223 (2007)
27. Zhang, L., Zhu, G., Shen, P., Song, J., Afaq Shah, S., Bennamoun, M.: Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition. In: Proceedings of the IEEE international Conference on Computer Vision Workshops. pp. 3120–3128 (2017)