Geometric Knowledge Distillation via Procrustes Analysis for Efficient Motion Sequence Classification

Bikram De*, Kostas Blekos†, Vasilis Pikoulis†, Dimitrios Kosmopoulos†, and Vangelis Metsis*

*Computer Science, Texas State University, San Marcos TX 78666, USA
Email: bikramkumarde@txstate.edu, vmetsis@txstate.edu
†Computer Engineering and Informatics, University of Patras, GR 26504, Greece
Email: mplekos@upatras.gr, pikoulis@ceid.upatras.gr, dkosmo@upatras.gr

***Abstract*—Motion sequence classification methods that rely on geometric approaches—such as Procrustes analysis and Dynamic Time Warping (DTW)—offer high accuracy but are often unsuitable for real-time applications due to their computational cost. In this paper, we present a novel geometric knowledge distillation framework that bridges the gap between accuracy and efficiency by transferring rich geometric insights from a Procrustes-DTW-based distance metric into a transformer-based neural network. By generating soft probability distributions from pre-computed Procrustes-DTW distances, we effectively guide the student model's training to preserve essential geometric properties like shape similarity, temporal alignment, and spatial transformation invariance. Our method enables fast and scalable motion sequence classification while retaining the benefits of geometric interpretability. We evaluate our framework on two benchmark tasks: sign language recognition using the SIGNUM dataset and human action recognition on UTD-MHAD. Results show that our distillation approach significantly improves classification accuracy over standard supervised learning and achieves dramatically lower inference time compared to traditional geometric methods—making it ideal for real-time motion understanding in wearable, robotic, and interactive systems.**

***Keywords*—Knowledge Distillation, Procrustes Analysis, Sign Language Recognition, Action Recognition, Geometric Distance Learning, Efficient Inference, Skeletal Sequences.**

I. Introduction

Motion sequence classification from 3D tracking data has emerged as a crucial technology with applications spanning human-computer interaction, healthcare monitoring, and assistive technologies. While deep learning approaches have shown impressive results in this domain, they often struggle to preserve the complex geometric relationships inherent in human motion and require large amounts of annotated data—a challenge especially pronounced in domains like action and sign language recognition.

Procrustes analysis and its generalizations have emerged as powerful techniques for shape comparison and alignment [1], with successful applications in gait analysis [2], stroke patient evaluation [3], and hand-grasping tasks [4]. Unlike raw coordinate-based comparisons, Procrustes distance accounts for transformations such as translation, scaling, and rotation, allowing for more meaningful similarity measurements between shapes. One of the key advantages of Procrustes analysis is its *invariance to transformations*. When analyzing human poses, facial features, or skeletal structures, raw coordinates may vary significantly due to differences in positioning, camera angles, or individual body proportions. By aligning shapes to a common reference frame, Procrustes analysis ensures that comparisons focus on structural similarities rather than extraneous variations. When combined with Dynamic Time Warping (DTW), as demonstrated in [5], this approach can effectively compare motion sequences while preserving their geometric properties. However, the computational complexity of these methods at inference time creates significant barriers to real-world deployment.

In this paper, we propose a novel approach that bridges this gap through geometric knowledge distillation. Knowledge distillation, first introduced by Hinton et al. [6], offers a promising solution to this challenge. This approach transfers knowledge from a complex but accurate "teacher" model to a simpler, faster "student" model. Recent work has explored various distillation approaches for motion-related tasks. Wang et al. [7] propose a multi-modal distillation framework for action recognition using RGB and infrared data. For hand gesture recognition, Li et al. [8] employ multi-task learning with self-distillation. Bian et al. [9] introduce structural knowledge distillation for skeleton-based action recognition, while Gao et al. [10] apply cross-modal distillation to continuous sign language recognition.

However, these existing approaches focus primarily on transferring feature representations or classification logits, without explicitly preserving the geometric relationships crucial for motion understanding. Our work addresses this limitation through a novel distillation framework that leverages Procrustes-DTW distance as the foundation for knowledge transfer. By using pre-computed Procrustes-DTW distances to generate soft probability distributions as our teacher signal, we ensure that the geometric understanding of motion sequences—including shape similarities, temporal alignments, and invariance to spatial transformations—is effectively transferred to the student network.

Specifically, our contributions are:

- A knowledge distillation framework[1] that transfers geometric understanding from Procrustes-DTW to a neural network, enabling fast inference while maintaining high accuracy.

This work was co-funded by the European Union / Greek State Scholarships Foundation under Erasmus+ grant number 2024-1-EL01-KA220-HED-000257847, Visual Interactive System for Teaching and Assessment of Sign Languages VISTA-SL.

[1]Code: https://github.com/imics-lab/geometric-knowledge-distillation

- An efficient training procedure using pre-computed Procrustes-DTW distances, particularly suitable for limited-data scenarios.
- Comprehensive evaluation on sign language recognition (SIGNUM dataset) and human action recognition (UTD-MHAD dataset), demonstrating 3% and 7% improvement in accuracy, respectively, without significant sacrifice in inference speed compared to a direct deep learning model.

II. Procrustes Distance calculation

We present a knowledge distillation framework that leverages the Procrustes-DTW distance [5] to train an efficient neural network for time series classification. Our approach transfers the geometric understanding captured by Procrustes analysis to a deep neural network through distillation, enabling fast inference while maintaining high accuracy.

A. Standard Procrustes algorithm

The standard Procrustes distance is a measure of similarity between two shapes. It quantifies how much one shape needs to be transformed (translated, scaled, and rotated) to best align with another shape. After aligning the shapes, the Procrustes distance is the square root of the sum of squared differences between corresponding points in the two shapes.

$$d_{proc}(\boldsymbol{X}, \boldsymbol{Y}) = \min_{\boldsymbol{R},s,\boldsymbol{t}} \|\boldsymbol{X} - (s\boldsymbol{Y}\boldsymbol{R} + \boldsymbol{t})\|_F \tag{1}$$

Where:

- $\boldsymbol{X}$: The point configuration matrix of the first shape.
- $\boldsymbol{Y}$: The point configuration matrix of the second shape.
- $\boldsymbol{R} \in \mathbb{R}^{d\times d}$: The rotation matrix that aligns $\boldsymbol{Y}$ to $\boldsymbol{X}$.
- $s \in \mathbb{R}$: The scaling factor that normalizes the size of $\boldsymbol{Y}$.
- $\boldsymbol{t} \in \mathbb{R}^{1\times d}$: The translation vector that shifts $\boldsymbol{Y}$ to align with $\boldsymbol{X}$.
- $\|\cdot\|_F$: The Frobenius norm, $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$.
- $d(\boldsymbol{X}, \boldsymbol{Y})$: The Procrustes distance, representing the minimum residual error between the aligned shapes.

The translation $\boldsymbol{t}$ becomes zero by changing the coordinate system to a common base (e.g., the wrist point in the case of a hand). The covariance matrix among the two point sets (after scaling) is computed as: $\boldsymbol{H} = (s\boldsymbol{Y})^\top \boldsymbol{X}$.

We perform the singular value decomposition of $\boldsymbol{H}$: $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, and $\boldsymbol{\Sigma}$ is a diagonal matrix of singular values. Then, the optimal rotation matrix is given by: $\boldsymbol{R} = \boldsymbol{V}\boldsymbol{U}^\top$.

B. Modified Procrustes Distance with Scale Normalization

For the problem of gesture recognition, we propose a modified Procrustes distance that integrates penalties for rotation and translation changes when comparing two hands. This way, we penalize how well two hand shapes match each other and their position in the 3D space after the complete body skeletons of the two users have been normalized and aligned. We also normalize each error term automatically using empirical data. This approach avoids manual hyperparameter tuning by scaling each component based on intra-class variability. In particular, we define a translation threshold δ so that only translations larger than what is typical for the same sign incur a penalty.

Note: Rotation and translation penalties are not necessary when comparing the entire skeleton of two users. In that case, the standard Procrustes algorithm can be used.

1) Mathematical Formulation: Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote the point configuration matrices (e.g., handshape landmarks) of two samples. The modified Procrustes distance is defined as

$$d_{\text{proc}}(\boldsymbol{X}, \boldsymbol{Y}) = \min_{\boldsymbol{R},s,\boldsymbol{t}} \left\{ \frac{\|\boldsymbol{X} - (s\,\boldsymbol{Y}\,\boldsymbol{R} + \boldsymbol{t})\|_F}{\sigma_1} + \frac{\|\log(\boldsymbol{R})\|_F^2}{\sigma_2} + \frac{\max(0, \|\boldsymbol{t}\| - \delta)^2}{\sigma_3} \right\} \tag{2}$$

subject to $\boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}$, where:

- σ_1, σ_2, and σ_3 are the empirical standard deviations of the alignment error, rotation penalty, and translation magnitude (used for penalty) respectively, and
- δ is a threshold on the translation magnitude, determined empirically (see Section II-B2).

2) Empirical Estimation of σ_k and δ: To automatically normalize the different error components and to set δ, we use intra-class pairs of samples (i.e., pairs that belong to the same sign) as follows:

1) **Intra-Class Pair Collection:** For each sign class c, select a set of pairs $\{(\boldsymbol{X}_i, \boldsymbol{X}_j)\}$ where both $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ represent the same sign.
2) **Computation of Transformation Errors:** For each pair $(\boldsymbol{X}_i, \boldsymbol{X}_j)$, compute the optimal transformation parameters $(\boldsymbol{R}_{ij}, s_{ij}, \boldsymbol{t}_{ij})$ by performing standard Procrustes alignment (without penalization). Then, record:

$$f_1^{(ij)} = \|\boldsymbol{X}_i - (s_{ij}\,\boldsymbol{X}_j\,\boldsymbol{R}_{ij} + \boldsymbol{t}_{ij})\|_F, \tag{3}$$

$$f_2^{(ij)} = \|\log(\boldsymbol{R}_{ij})\|_F^2, \tag{4}$$

$$f_3^{(ij)} = \|\boldsymbol{t}_{ij}\|. \tag{5}$$

3) **Empirical Standard Deviations:** Compute the standard deviations for each error term over all intra-class pairs: $\sigma_1 = \text{std}\{f_1^{(ij)}\}, \quad \sigma_2 = \text{std}\{f_2^{(ij)}\}, \quad \sigma_3 = \text{std}\{f_3^{(ij)}\}$.
4) **Determination of δ:** To set the translation threshold δ, we examine the empirical distribution of the translation magnitudes $\{f_3^{(ij)}\}$. We then define δ as the α-th percentile of this distribution: $\delta = \text{Percentile}_\alpha\left(\{f_3^{(ij)}\}\right)$, where a typical choice is $\alpha = 0.90$. This means that for 90% of intra-class pairs, the translation magnitude is below δ, and only larger-than-usual translations incur a penalty.

III. Procrustes - DTW as a Learning Problem

Consider a training dataset $\mathcal{D} = \{(\boldsymbol{S}_i, y_i)\}_{i=1}^N$ where each $\boldsymbol{S}_i \in \mathbb{R}^{T\times D}$ represents a time series sequence of length T with D channels (features), and $y_i \in \{1, ..., C\}$ denotes the corresponding class label. Our goal is to train a neural network that can efficiently classify these sequences while preserving the geometric relationships captured by Procrustes-DTW distance.

1) Procrustes-DTW Teacher: The teacher component of our framework is based on the Procrustes-DTW distance to measure similarity between sequences. For any two sequences $\boldsymbol{S}_i$ and $\boldsymbol{S}_j$, their Procrustes-DTW distance is defined as:

$$d_P(\boldsymbol{S}_i, \boldsymbol{S}_j) = \min_{\boldsymbol{W}} \sum_{k=1}^{K} d_{proc}(\boldsymbol{S}_i[w_k^x], \boldsymbol{S}_j[w_k^y]) \tag{6}$$

where $\boldsymbol{W} = (w_1, ..., w_K)$ is a warping path, K represents the length of the warping path, and d_{proc} is the Procrustes distance between individual frames, as defined in equation 1 or 2, depending on the application.

During the pre-computation phase, we compute all pairwise Procrustes-DTW distances between training sequences. The teacher then uses these pre-computed distances to generate soft probability distributions over classes. For an input sequence $\boldsymbol{S}$, the teacher's prediction for class c is:

$$p_t(c|\boldsymbol{S}) = \frac{\sum_{j \in \mathcal{N}_c} \exp(-d_P(\boldsymbol{S}, \boldsymbol{S}_j)/\tau_t)}{\sum_{k=1}^{N} \exp(-d_P(\boldsymbol{S}, \boldsymbol{S}_k)/\tau_t)} \tag{7}$$

where $\mathcal{N}_c$ is the set of training examples from class c, τ_t is the temperature parameter, and all d_P values are retrieved from the pre-computed distance matrix.

A. Student Network

The student network $f_\theta : \mathbb{R}^{T \times D} \to \mathbb{R}^C$ is implemented as a deep neural network that maps input sequences directly to class probabilities. For an input sequence $\boldsymbol{S}$, the student's prediction is:

$$p_s(c|\boldsymbol{S}) = \text{softmax}(f_\theta(\boldsymbol{S})/\tau_s)_c \tag{8}$$

where τ_s is the student's temperature parameter.

B. Knowledge Distillation Framework

Our training objective combines standard supervised learning with knowledge distillation:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{CE} + \beta \mathcal{L}_{KL} \tag{9}$$

The cross-entropy loss $\mathcal{L}_{CE}$ measures the student's performance against ground truth labels:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} \sum_{c=1}^{C} y_{ic} \log(p_s(c|\boldsymbol{S}_i)) \tag{10}$$

The Kullback-Leibler divergence loss $\mathcal{L}_{KL}$ encourages the student to mimic the teacher's soft predictions:

$$\mathcal{L}_{KL} = \sum_{i=1}^{N} \sum_{c=1}^{C} p_t(c|\boldsymbol{S}_i) \log\left(\frac{p_t(c|\boldsymbol{S}_i)}{p_s(c|\boldsymbol{S}_i)}\right) \tag{11}$$

The weights α and β control the contribution of each loss term. To ensure effective knowledge distillation, we use the same temperature parameter $\tau = \tau_s = \tau_t$ for both teacher and student during training. This shared temperature controls the softness of the probability distributions and thus the amount of information transferred from teacher to student. During inference, we use $\tau_s = 1$ for the student network's predictions, as is standard practice in knowledge distillation.

C. Training Procedure

The training process consists of two phases:

1) Pre-computation Phase: To ensure efficient training, we pre-compute Procrustes-DTW distances between all pairs of training sequences. These distances are stored in a memory-efficient format for quick lookup during training.

2) Training Phase: For each mini-batch $\mathcal{B}$, we:

1) Retrieve pre-computed teacher predictions for the batch sequences
2) Forward pass the sequences through student network
3) Compute combined loss $\mathcal{L}_{total}$ using temperature-scaled predictions
4) Update student parameters via gradient descent:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{total} \tag{12}$$

where η is the learning rate.

The optimization problem can be formally stated as:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{S},y)\sim\mathcal{D}}[\alpha \mathcal{L}_{CE} + \beta \mathcal{L}_{KL}] \tag{13}$$

where $\mathcal{D}$ is our training dataset.

D. Inference

During inference, only the student network is used, enabling efficient prediction without the need for computing Procrustes-DTW distances:

$$\hat{y} = \arg, \max_c p_s(c|\boldsymbol{S}) \tag{14}$$

This approach provides significant speedup compared to the distance-based classifiers, such as the nearest-neighbor classifier (KNN), while maintaining the geometric understanding learned through distillation when training a deep learning model.

IV. Experiments

We evaluate our proposed knowledge distillation framework on two challenging public datasets: the SIGNUM dataset for sign language recognition [11] and the UTD-MHAD dataset for human action recognition [12]. Our experiments assess both the classification performance and computational efficiency of three approaches: (1) k-nearest neighbor classification using Procrustes-DTW distance, (2) direct supervised learning with a transformer model, and (3) our proposed knowledge distillation approach.

A. Datasets and Implementation Details

1) SIGNUM Dataset: The SIGNUM dataset [11] contains 450 basic signs (*classes*) from German Sign Language (DGS), performed by 25 different signers. We extract 3D hand landmarks (21 points per hand) using MediaPipe Holistic [13], resulting in a 126-dimensional feature vector per frame (21 landmarks × 3 coordinates × 2 hands). The sequence length is 80 timesteps. Following standard practice for signer-independent evaluation, we split the dataset by signer ID:

- Training set: 14 signers (IDs 1-14), 6,300 sequences
- Validation set: 4 signers (IDs 15-18), 1,800 sequences
- Test set: 7 signers (IDs 19-25), 1,901 sequences

TABLE I. RESULTS ON SIGNUM DATASET (TEST SET)

Method	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)	Infer. Time (ms/sample)
Procrustes-DTW (k-NN)	63.9	68.2	64.4	63.1	3.6×10^6
Transformer (Direct)	86.9	89.5	87.1	86.6	0.22
Ours (Distillation)	90.2	91.7	90.2	89.8	0.35

TABLE II. RESULTS ON UTD-MHAD DATASET (TEST SET)

Method	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)	Infer. Time (ms/sample)
Procrustes-DTW (k-NN)	31.9	38.9	32.1	28.4	1.89×10^5
Transformer (Direct)	57.5	60.9	57.6	55.6	0.21
Ours (Distillation)	64.9	67.2	64.9	63.9	0.83

2) UTD-MHAD Dataset: The UTD-MHAD dataset [12] comprises 27 actions performed by 8 subjects, with each action repeated 4 times. We use the 3D skeleton data captured by a Kinect sensor, which provides 20 joint positions in 3D space. The sequence length varies from 45 to 125 timesteps. The dataset contains 861 sequences after removing corrupted samples. For subject-independent evaluation, we use:

- Training set: 5 subjects (IDs 1-5), 539 sequences
- Test set: 3 subjects (IDs 6-8), 322 sequences

Due to the small number of subjects in this dataset, we do not use a validation set, as extensive hyperparameter tuning is not the objective of this study.

Experiments were run on Azure Standard_NC4as_T4_v3 nodes, each with 4 AMD EPYC 7V12 vCPUs (2.45 GHz), 28 GiB RAM, 176 GiB local SSD, and an NVIDIA Tesla T4 GPU (16 GiB). We implemented our models using PyTorch. The transformer architecture consists of 4 encoder layers with 9 attention heads, a hidden dimension of 256, and dropout rate of 0.1. For training, we used batch size: 16; learning rate: 1e-4; number of epochs: 400. For knowledge distillation, we set temperature (τ): 3.0; cross-entropy loss weight (α): 0.5; KL divergence loss weight (β): 0.5.

B. Results and Discussion

Tables I and II present the classification performance and computational efficiency of each approach on the SIGNUM and UTD-MHAD datasets, respectively.

1) Classification Performance: As a baseline, for the 450-class sign language recognition problem, the One-Nearest Neighbor (1-NN) classifier using the Procrustes-DTW as the distance metric achieves an accuracy of 63.9%. For reference, in addition to standard KNN classification accuracy, we evaluate Top-10 Nearest Neighbor Accuracy (Recall@10), which measures the proportion of test samples where the true class appears among the 10 nearest training samples ranked by the Procrustes-DTW distance. The *Recall@10* accuracy is 92.8%. The pure Transformer-based neural network model achieves an accuracy of 86.9%, while our distillation approach uses the exact same neural network architecture and training hyperparameters and achieves a 90.2% accuracy. This $\sim$3% improvement in accuracy is attributed to the geometric knowledge distillation. Figure 1 shows the training loss and accuracy curves for the training and validation sets on the SIGNUM dataset. We observe a similar behavior for both direct transformer and distillation models.

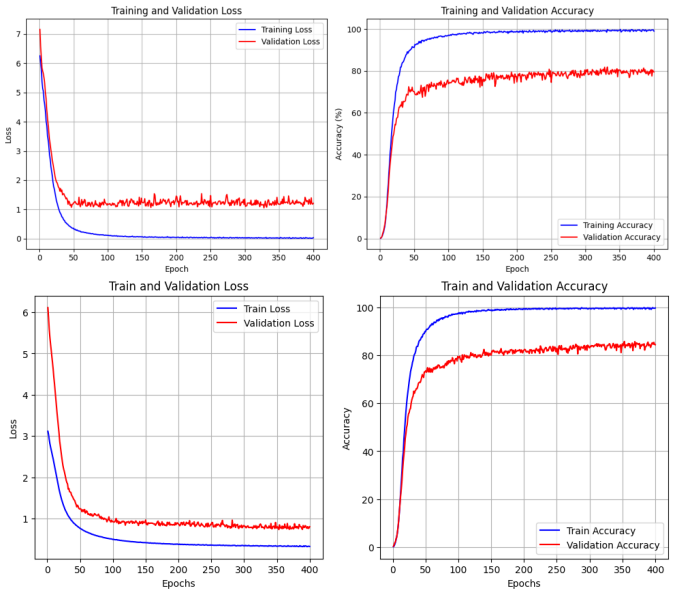

Fig. 1. Training curves comparing loss and accuracy for direct transformer training (top) vs. distillation (bottom) on the SIGNUM dataset.

For the human action dataset, we observe an improvement of $\sim$7% in accuracy, although the overall accuracy for all methods is lower due to the dataset's difficulty and smaller size.

2) Computational Efficiency: The primary advantage of our approach becomes apparent when considering inference time. The Procrustes-DTW k-NN classifier requires several orders of magnitude higher computation time (3.6×10^6 ms/sample for SIGNUM, 1.9×10^5 ms/sample for UTD-MHAD) due to the need to compute distances to all training samples. Even though several efficiency improvements can be made to distance-based classification methods, inference time remains prohibitively high for real-time applications. In contrast, our distilled model, similar to the direct transformer, achieves sub-millisecond inference time per example (0.35 ms/sample for SIGNUM, 0.83 ms/sample for UTD-MHAD), representing a speedup of over 10^6x.

V. CONCLUSION

We presented a knowledge distillation framework transferring geometric understanding from Procrustes-DTW to efficient neural networks. Our approach achieves significant accuracy improvements over standard transformer training while maintaining sub-millisecond inference times—combining the geometric understanding of Procrustes analysis with the computational efficiency of deep learning. This makes our method particularly valuable for real-time applications requiring precise geometric understanding, such as sign language translation and motion-based interfaces. Future work could extend this approach to continuous recognition tasks and other domains where shape and motion properties are essential.

REFERENCES

[1] F. L. Bookstein, *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge: Cambridge University Press, 1992. [Online]. Available: https://doi.org/10.1017/CBO9780511573064

[2] A. R. Anwary, H. Yu, and M. Vassallo, "Gait evaluation using procrustes and euclidean distance matrix analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, pp. 2021–2029, 2019. [Online]. Available: https://doi.org/10.1109/JBHI.2018.2875812

[3] A. L. Wong, S. A. Jax, L. L. Smith, L. J. Buxbaum, and J. W. Krakauer, "Movement imitation via an abstract trajectory representation in dorsal premotor cortex," *Journal of Neuroscience*, vol. 39, pp. 3320–3331, 2019. [Online]. Available: https://doi.org/10.1523/JNEUROSCI.2597-18.2019

[4] J. Manuello, C. Maronati, M. Rocca, R. Guidotti, T. Costa, and A. Cavallo, "Motor styles in action: Developing a computational framework for operationalization of motor distances," *Behavior Research Methods*, vol. 57, no. 1, p. 13, 2024. [Online]. Available: https://doi.org/10.3758/s13428-024-02530-0

[5] N. Arvanitis, E. Sartinas, and D. Kosmopoulos, "Procrustes-dtw: Dynamic time warping variant for the recognition of sign language utterances," in *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2023, pp. 1–5.

[6] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[7] Z. Quan, Q. Chen, K. Zhao, Z. Liu, and Y. Li, "Knowledge distillation for action recognition based on rgb and infrared videos," in *International Forum on Digital TV and Wireless Multimedia Communications*. Springer, 2021, pp. 18–29.

[8] J.-Y. Li, H. Prawiro, C.-C. Chiang, H.-Y. Chang, T.-Y. Pan, C.-T. Huang, and M.-C. Hu, "Efficient hand gesture recognition using multi-task multi-modal learning and self-distillation," in *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, 2023, pp. 1–7.

[9] C. Bian, W. Feng, L. Wan, and S. Wang, "Structural knowledge distillation for efficient skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 2963–2976, 2021.

[10] L. Gao, P. Shi, L. Hu, J. Feng, L. Zhu, L. Wan, and W. Feng, "Cross-modal knowledge distillation for continuous sign language recognition," *Neural Networks*, vol. 179, p. 106587, 2024.

[11] U. von Agris and K.-F. Kraiss, "Signum database: Video corpus for signer-independent continuous sign language recognition," in *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, 2010, pp. 243–246.

[12] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE International conference on image processing (ICIP)*. IEEE, 2015, pp. 168–172.

[13] Google AI Edge, "Mediapipe solutions," 2023, primary documentation available at: https://developers.google.com/mediapipe, Code repository: https://github.com/google-ai-edge/mediapipe, Accessed: 2025-03-03.