The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-2/W9-2025
ISPRS Intl. Workshop "Photogrammetric and computer vision techniques for environmental and infraStructure monitoring, Biometrics and Biomedicine"
PSBB25 , 9–11 June 2025, Moscow, Russia

G-MAE: Gesture-aware Masked Autoencoder for Human-Machine Interaction

Elena Ryumina, Dmitry Ryumin *, Denis Ivanko

St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russian Federation –
(ryumina.e, ryumin.d, ivanko.d)@iias.spb.su

Keywords: Masked Autoencoder, Multi-Scale Transformer, Multi-Head Self-Attention, Gesture Recognition, Human-Machine Interaction.

Abstract

Gesture recognition remains a critical challenge in human-computer interaction due to issues such as lighting variations, background noise, and limited annotated datasets, particularly for underrepresented sign languages. To address these limitations, we propose G-MAE (Gesture-aware Masked Autoencoder), a self-supervised framework leveraging a Gesture-aware Multi-Scale Transformer (GMST) backbone that integrates multi-scale dilated convolutions (MSDC), multi-head self-attention (MHSA), and a multi-scale contextual feedforward network (MSC-FFN) to capture both local and long-range spatiotemporal dependencies. Pre-trained on the Slovo corpus with 50–70% masking and fine-tuned on TheRusLan, G-MAE achieves 94.48% accuracy, with ablation studies confirming the contributions of each component. Removing MSDC, MSC-FFN, or MHSA reduces accuracy to 92.67%, 91.95%, and 90.54%, respectively. The optimal masking ratio (50–70%) balances information retention and learning efficiency, demonstrating robust performance even with limited labeled data, thus advancing gesture recognition in resource-constrained scenarios.

1. Introduction

Gesture recognition is a critical technology in modern human-machine interaction, enabling intuitive and touchless control mechanisms in various applications (Ryumin et al., 2023a, Qi et al., 2024). However, neural network-based models often struggle with challenges such as lighting variations, background noise and gesture variability, which affect the accurate recognition of both static and dynamic gestures (Hashi et al., 2024). In addition, gesture recognition corpora are often relatively small, partly due to the challenges of annotation and data collection, except for large corpora such as AUTSL (Sincan and Keles, 2020), Slovo (Kapitanov et al., 2023) and HaGRID (Kapitanov et al., 2024). Gesture annotation requires high-quality labeling which can be time-consuming and effort intensive, especially for video data where precise timing, type, and context of gestures need to be specified. In addition, the collection of such corpora can be difficult, as it involves capturing a variety of gestures under different conditions, which affects the size of the corpus. Many sign languages remain poorly represented, and their digital recognition requires the creation of new specialized corpora. This challenge is further exacerbated by the diversity of sign languages across countries and cultures, each with its own unique characteristics requiring specific approaches to data annotation and collection.

With recent advances in computer vision and machine learning, automatic gesture recognition has received increasing research attention (Ni et al., 2024). End-to-end trained deep neural networks allow for autonomous extraction of salient features from raw input data (Ikne et al., 2024a, Vostrikov et al., 2024), eliminating the need for hand-crafted feature engineering and classifier design. This method improves model learnability, robustness and prediction accuracy, making it more suitable for real-world applications. While end-to-end deep neural networks have improved feature extraction and reduced reliance on hand-crafted methods (Ikne et al., 2024a, Vostrikov et al., 2024), real-world deployment still faces limitations in terms of efficiency, robustness, and generalization - especially with limited labeled data. Transfer learning, meta-learning, and self-supervised learning aim to reduce annotation costs and improve model adaptability, but current solutions still often depend on pre-training on large external corpora.

To overcome these challenges, we propose G-MAE (Gesture-aware Masked Autoencoder), a self-supervised learning framework for gesture recognition. G-MAE leverages a novel Gesture-aware Multi-Scale Transformer (GMST) backbone that integrates convolutional and attention mechanisms to capture both local and long-range spatio-temporal dependencies. Unlike conventional MAE frameworks based on vision transformers (He et al., 2022), which require large-scale pre-training, G-MAE is designed for effective training on small corpora, reducing the dependency on external data and pre-trained models.

The rest of the paper is organized as follows. In Section 2, we analyze the state-of-the-art methods for gesture recognition. Section 3 provides a detailed description of the method proposed. Experimental results are presented in Section 4. Finally, Section 5 presents the conclusions and future work.

2. Related Work

Deep learning has significantly advanced gesture recognition by enabling the modeling of complex spatio-temporal motion dependencies (Ryumin et al., 2023b). State-of-the-art methods integrate convolutional (Alonazi et al., 2023), transformer-based (Hampiholi et al., 2023, Garg et al., 2024), recurrent-based (LSTM (Axyonov et al., 2021a) and Mamba (Altaher et al., 2025)) and graph-based (Ikne et al., 2024b) neural architectures to optimize feature extraction and classification. Some studies propose the use of deformable 3D convolutions and modified graph neural networks to better capture gesture variability (Papadimitriou and Potamianos, 2023), while others exploit spatio-temporal features to improve recognition accuracy (Ryumin et al., 2023b). These methods highlight the importance

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-241-2025 | © Author(s) 2025. CC BY 4.0 License.

of architectural optimizations and multimodal analysis in the development of more robust and efficient gesture recognition systems.

Lightweight neural architectures have been explored to address the computational constraints of real-time gesture recognition. A sparsity-aware 3D convolutional model (Kim et al., 2024) uses inter-frame differential information and region-of-interest based computation to optimize feature extraction while significantly reducing computational costs. By incorporating activation and weight sparsity, this method achieves a significant parameter reduction while maintaining a high recognition accuracy. Similarly, the multimodal method (Christidis et al., 2024), which fuses 2D skeleton sequences with localized image patches, improves recognition performance without imposing an excessive computational cost. Despite these advances, challenges remain, particularly in balancing model efficiency and generalization when training on limited labeled data. Reducing annotation costs while ensuring model robustness is a key challenge for the deployment of deep learning-based gesture recognition systems in practical applications.

Transfer learning has emerged as a critical technique for improving the generalization of deep learning models in gesture recognition, facilitating knowledge transfer across domains to account for inter-user variability and environmental variations (Ojeda-Castelo et al., 2022). For example, TL-MKCNN (Zou and Cheng, 2021) improves adaptability between users and sessions by using distribution normalization and alignment modules, significantly improving classification accuracy in cross-user and cross-day scenarios. Similarly, in mIV3Net (Karsh et al., 2024), a modified Inception V3 network fine-tunes convolutional layers to focus on salient gesture features, mitigating challenges associated with complex backgrounds and similarities between classes. In addition to traditional transfer learning, meta-learning techniques such as cross-lingual few-shot learning (Bilge et al., 2024) enable models to recognize unseen gestures with minimal labelled data, proving effective for low-resource domains. In addition, self-supervised learning has been integrated into transfer learning frameworks to leverage large unlabeled gesture corpora and construct auxiliary tasks that exploit the spatio-temporal structures of gesture movements to reduce reliance on manual annotation. A notable example is DFCNet+ (Feng et al., 2024), which incorporates dynamic motion features and gloss-level alignment to improve continuous sign recognition, using contrastive learning to capture fine-grained temporal dependencies. Together, these investigations highlight the potential of transfer learning in developing robust, adaptive and efficient gesture recognition systems, bridging the gap between research advances and real-world applications.

Self-supervised learning using masked autoencoders (MAEs) (He et al., 2022) has received considerable attention for its ability to effectively extract semantic features from images without relying on traditional data augmentation methods. However, the application of MAEs to gesture recognition remains limited. In particular, MAEs that use vision transformers as their backbone require pre-training on large corpora, making them difficult to apply to smaller corpora. Unlike vision transformers, swin transformers (Liu et al., 2021) incorporate inductive biases similar to those in convolutional neural networks, making them easier to train on limited data. For example, the research (Xu et al., 2023) introduces a swin-MAE method that effectively trains on small medical images without the need for pre-trained models. In addition, the research (Liu et al., 2023) presents the mix-MAE method, which accelerates pre-training and increases the efficiency of models across different hierarchical vision transformers. In addition, methods such as the anatomically guided spatio-temporal MAE (Ikne et al., 2025) integrate anatomical constraints into the self-supervised training of spatio-temporal MAEs, thereby enhancing 3D keypoint learning for real-time hand gesture recognition. Another notable method based on MAEs (Zhao et al., 2024) introduces a motion-aware strategy and a semantic alignment module for sign language recognition, explicitly exploring dynamic motion cues while aligning global semantic features, leading to state-of-the-art results on benchmark corpora. These methods demonstrate that integrating MAEs with advanced neural architectures such as swin transformers, and incorporating anatomical and motion-aware guidance, can lead to more efficient and adaptive gesture recognition systems capable of operating with limited labeled data.

3. Proposed Method

The aim of the current work is to adapt MAE to small gesture corpora. We propose G-MAE: Gesture-aware Masked Autoencoder (see Figure 1) for human-machine interaction, which uses the Gesture-aware Multi-Scale Transformer ($GMST$) as its backbone. $GMST$ combines convolutions and self-attention to extract multi-scale local features, using multi-scale dilated convolutions ($MSDC$) and multi-head self-attention ($MHSA$) to effectively capture inter-scale and long-range correlations in gestures. In addition, we introduce a multi-scale contextual feedforward network ($MSC\text{-}FFN$) module that enhances feature representation at multiple scales to improve recognition accuracy. $GMST$ uses a pyramid structure to process gesture data at different scales, capturing different levels of detail for improved performance. In G-MAE, similar to MAE, the decoder reconstructs the complete gesture from the encoder output.

3.1 Multi-Scale Dilated Convolution

To effectively capture the multi-scale spatial and temporal patterns inherent in human gestures, we incorporate an $MSDC$ module into the $GMST$ encoder. The primary focus of this module is to expand the receptive field of the convolutional operations without increasing the number of parameters or reducing the resolution of the feature maps, which is essential for preserving fine details such as finger positions while modeling broader contexts such as hand pose and body orientation.

The $MSDC$ module processes the input feature map through multiple parallel convolutional branches. Specifically, a 1×1 convolution operation is applied to preserve fine-grained local information. In parallel, three 3×3 dilated convolution operations with dilation rates $d = \{1, 2, 3\}$ are applied to capture local, mid-range and long-range dependencies within the gesture data. The output of each branch is normalized by batch normalization (BN) and passed through a GELU activation function to enable stable, non-linear feature transformations. Given an input feature map $F \in \mathbb{R}^{H \times W \times C}$, the computation of the parallel branches is calculated as 1:

$$\begin{cases} F_0 = \text{GELU}\left(\text{BN}\left(\text{Conv}_{1\times1}(F)\right)\right), \\ F_i = \text{GELU}\left(\text{BN}\left(\text{DConv}_{d=i-1}(F)\right)\right), \quad i \in \{1, 2, 3\} \end{cases} \tag{1}$$

The outputs from all four branches are concatenated along the channel dimension to produce a fused multi-scale feature map, calculated as 2:

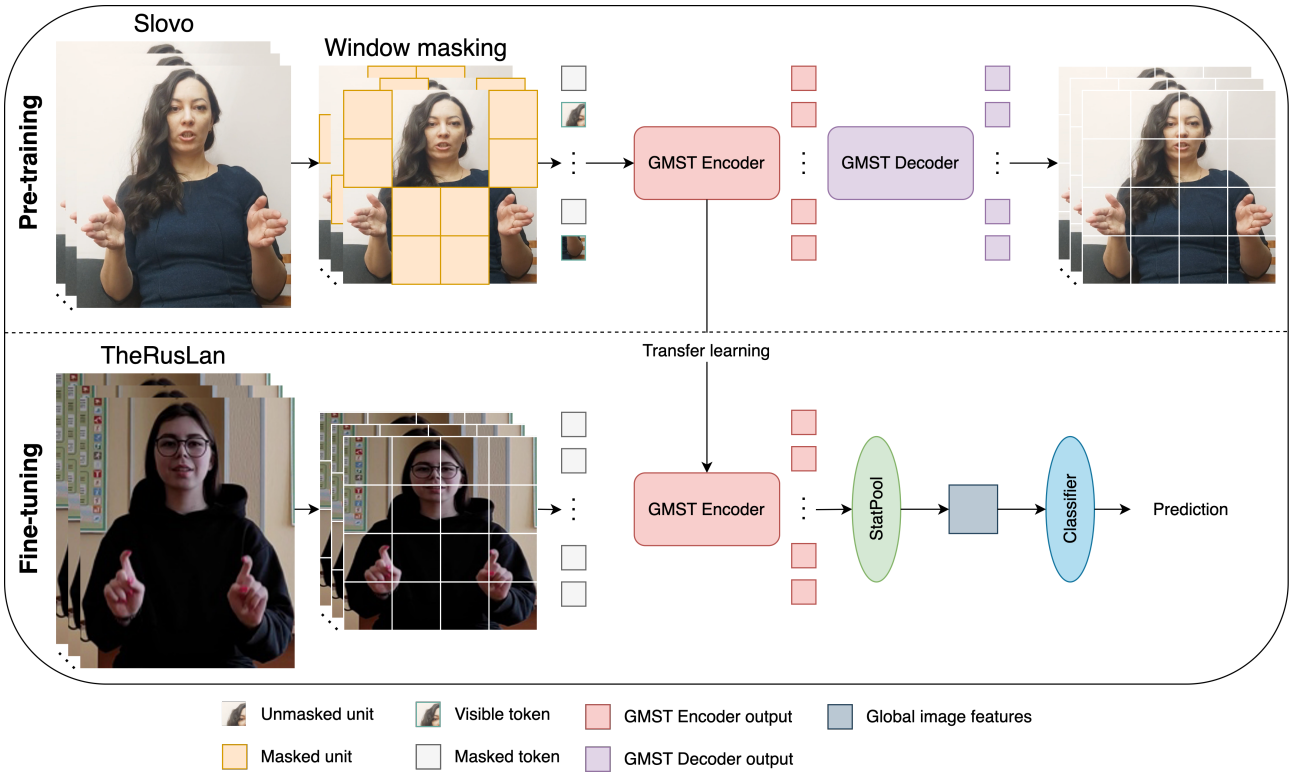

Figure 1. Gesture-aware Masked Autoencoder (G-MAE) pipeline

$$F_{\text{concat}} = \text{Concat}(F_0, F_1, F_2, F_3) \tag{2}$$

To fuse these multi-scale features while controlling dimensionality, we apply a 1×1 convolution to the concatenated tensor, producing the output Multi-Scale Feature Fusion ($MSFF$), as shown in 3:

$$MSFF = \text{GELU}\left(\text{BN}\left(\text{Conv}_{1\times1}(F_{\text{concat}})\right)\right) \tag{3}$$

This module effectively achieves receptive fields equivalent to 3×3, 5×5 and 7×7 by controlling the dilation rates in the parallel branches, as shown in Figure 2. In the feature fusion stage, these $MSDC$ complement each other, allowing the model to capture both fine-grained local structures and broad contextual dependencies. This significantly improves the completeness of the feature representation across different spatial scales.

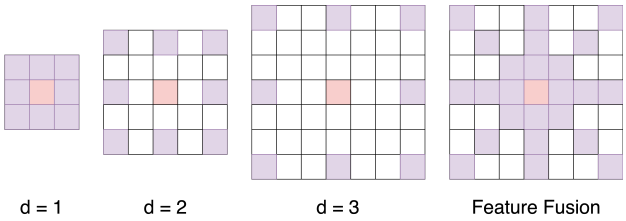

Figure 2. Receptive fields of $MSDC$ dilated convolutions.

To further improve feature selectivity, the $MSDC$ module incorporates a spatial attention (SA) mechanism that highlights important spatial regions by applying both average pooling and max pooling operations along the channel axis of $MSFF$. The resulting two spatial maps are concatenated and processed using a convolutional layer with a 7×7 kernel, followed by sigmoid activation, to generate the spatial attention mask, as calculated in 4:

$$\begin{aligned} F_{\text{concat}} &= \text{Concat}\left(\text{AvgPool}(F), \text{MaxPool}(F)\right), \\ \text{SA} &= \sigma\left(\text{Conv}_{7\times7}(F_{\text{concat}})\right) \times F \end{aligned} \tag{4}$$

In parallel, a channel attention mechanism based on the squeeze-and-excitation (SE) block is applied to adaptively recalibrate channel-wise feature responses. Global average pooling is first performed on $MSFF$ to obtain a channel descriptor, as shown in 5:

$$\text{SE} = \text{GlobalAvgPool}(MSFF) \tag{5}$$

This descriptor is then passed through two fully connected (FC) layers with a $ReLU$ activation in between and a sigmoid activation at the output, as shown in 6:

$$F_{\text{excited}} = \sigma\left(\text{FC}_2\left(\text{ReLU}\left(\text{FC}_1 \times \text{SE}\right)\right)\right) \tag{6}$$

The resulting excitation vector is multiplied by the original feature map by channel-wise multiplication, resulting in the recalibrated feature map CA, the channel-attended feature map, as shown in 7:

$$\text{CA} = F_{\text{excited}} \times MSFF \tag{7}$$

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-241-2025

Finally, the output of the $MSDC$ module is calculated by applying a residual connection between the recalibrated feature map CA and the original input feature map F, as shown in 8:

$$MSDC = \text{CA} + F \tag{8}$$

The residual connection preserves both original and multi-scale enhanced features, ensuring stable training and improved gradient flow. The $MSDC$ module captures fine-grained gesture details and global contextual information, improving recognition of gestures with subtle variations in different conditions.

3.2 Multi-Head Self-Attention

$MHSA$ is used to capture global dependencies in gesture data, facilitating the interaction between local and global features. In the G-MAE model, the $MHSA$ module follows the $MSDC$ module to enhance the network's ability to capture complex interactions between spatial and temporal patterns in gestures. The number of attention heads is set to 4, 8, 16 and 32 across the model stages, allowing the model to efficiently process multi-scale representations of gestures.

Given the input features F, the module first applies a FC layer to generate the query (Q), key (K) and value (V). The attention map (A) is then computed by performing a dot product between Q and K, followed by a softmax operation for normalization, as shown in 9:

$$A = \text{Softmax}(K \times Q/\alpha) \tag{9}$$

where α is an adaptive scaling parameter for softmax control.

The attention-weighted features are obtained by multiplying A by V, and the final output feature map $\hat{F}$ is computed by passing the result through another FC layer, as shown in 10:

$$\hat{F} = \text{FC}(A(Q, K, V)) \tag{10}$$

The attention heads are split along the channel dimension, allowing the model to learn different attention patterns for different parts of the gesture simultaneously.

3.3 Multi-Scale Contextual Feed-forward Network

The $MSC\text{-}FFN$ is designed to capture multi-scale contextual information, which is essential for recognizing gestures with both fine details (e.g., finger positions) and broader contexts (e.g., hand pose and body orientation). The network refines the features extracted from the previous $MSDC$ and $MHSA$ modules to provide a more complete and contextually aware representation of the input gesture.

The $MSC\text{-}FFN$ processes features from multiple scales in parallel. The network first applies dilated convolutions at different dilation rates to capture different spatial and temporal contexts. The outputs of these convolutions are concatenated along the channel dimension to form a multi-scale feature map.

The structure of $MSC\text{-}FFN$ consists of two main stages: feature extraction using dilated convolutions and feature refinement through linear layers. The first stage applies 1×1 convolutions to preserve fine-grained local features, while the second stage applies dilated convolutions at different rates d to capture local, mid-range, and long-range dependencies. The outputs are concatenated along the channel dimension to form a multi-scale feature map, as shown in 11:

$$\text{MS}(X) = \text{Concat}\left(\text{DConv}_d(X), \quad d \in \{1, 2, 3\}\right) \tag{11}$$

Next, a 1×1 convolution is applied to fuse these multi-scale features while maintaining computational efficiency, as shown in 12:

$$\text{MSF}(X) = \text{MS}(X) + \text{Conv}_{1\times1}(X) \tag{12}$$

The resulting multi-scale feature map is fed through a two-layer feed-forward network (FFN), where the first layer increases the dimensionality by a factor of 4 and the second layer reduces it by the same factor. This ensures a balanced flow of information. The $MSC\text{-}FFN$ is formulated as shown in 13:

$$MSC\text{-}FFN(X) = \text{Conv}_{1\times1}\left(\text{GELU}(\text{FC}_1(\text{MSF}))\right) \times \text{FC}_2 \tag{13}$$

The final output is a refined multi-scale feature representation. This improves the accuracy of gesture recognition.

3.4 Pre-training and Fine Tuning

The training of G-MAE follows a two-stage paradigm: unsupervised pre-training and supervised fine-tuning. This method allows the model to first learn informative spatio-temporal representations from large-scale gesture corpus, and then adapt these representations to the specific task of gesture classification.

In the pre-training stage, the G-MAE model is trained as a masked autoencoder, where a significant portion of the input gesture sequence is randomly masked. The encoder processes only the visible (unmasked) parts of the sequence, generating a latent representation. The decoder then attempts to reconstruct the missing frames based on this latent representation and the available context. The objective function used in this stage is the mean squared error (MSE) (Sara et al., 2019) between the reconstructed and original gesture sequences, as shown in 14:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(x_{\text{true},i} - x_{\text{pred},i})^2 \tag{14}$$

This pre-training strategy allows the encoder to capture robust and transferable representations of gesture sequences by exploiting temporal and spatial continuity, even in the absence of explicit labels.

In the fine-tuning stage, the pre-trained coder is retained, and the decoder is discarded. A classification head is attached to the encoder for gesture classification. During the initial fine-tuning epochs, the encoder weights are frozen to allow the classification head to stabilize. After this warm-up period, the entire model is fine-tuned together in an end-to-end manner.

The fine-tuning process is monitored and optimized using a multi-class cross-entropy loss function, as show in 15:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{c=1}^{C} y_{i,c}\log \hat{y}_{i,c}, \tag{15}$$

where M is the number of training samples, C is the number of gesture classes, $y_{i,c}$ is a binary indicator (1 if sample i belongs to class c, 0 otherwise), and $\hat{y}_{i,c}$ is the predicted probability for class c for sample i.

This two-stage training scheme enables G-MAE to achieve state-of-the-art (SOTA) performance in gesture recognition tasks, especially in scenarios with limited labelled data, by exploiting self-supervised representation learning during pre-training.

3.5 Masking Strategy and Decoder

A core component of the G-MAE method is the design of the masking strategy and the decoder, which together enable effective self-supervised learning through masked gesture reconstruction.

During pre-training, a large portion of the input gesture sequence is randomly masked. Specifically, a masking ratio (ranging from 50% to 70%) is applied to the input frame along the spatial dimension, such that only a few pixels of the frame remain visible to the encoder. This masking is performed uniformly and randomly, without regard to gesture boundaries or motion intensity, which enhances the ability of the model to generalize to different missing data scenarios.

After masking, the entire sequence of frames - each with its own spatial masking pattern - is flattened into a sequence of spatio-temporal tokens. Positional embeddings are added to each token to preserve both spatial and temporal information. The resulting token sequence is then passed through the encoder, which extracts a latent spatio-temporal representation from the visible parts of the sequence.

The decoder, which is only used during the pre-training stage, receives two types of tokens:

- The latent tokens output by the encoder, corresponding to unmasked patches.
- Learnable mask tokens representing the positions of the masked patches.

Positional embeddings are also added to both types of tokens to maintain the correct spatial and temporal order.

The decoder reconstructs the original gesture sequence by predicting the pixel values (or patch embeddings) at the positions of the masked patches. The reconstruction objective is to minimize the MSE between the original gesture frames G^{orig} and the reconstructed frames G^{recon} at the masked positions. The reconstruction loss is shows in 16:

$$\mathcal{L}_{\text{recon}} = \frac{1}{N}\sum_{i=1}^{N}\left\|G_i^{\text{orig}} - G_i^{\text{recon}}\right\|_2^2 \tag{16}$$

where N is the total number of masked patches across all frames in the sequence. This loss encourages the model to infer plausible content for the missing parts based on the available spatio-temporal context.

After pre-training, the decoder is discarded, and the pre-trained encoder is now capable of modelling complex gesture dynamics from incomplete sequences.

4. Experimental Results

4.1 Research Corpora

The proposed G-MAE method is pre-trained on the Slovo gesture corpus (Kapitanov et al., 2023), where 50% to 70% of the data is randomly masked and fed into the encoder. The decoder reconstructs the gesture data from the encoder's latent representations. Subsequently, G-MAE is fine-tuned on the TheRusLan corpus (Kagirov et al., 2020), where only the encoder processes the unmasked data. The output features are globally averaged before being passed to a classifier for gesture recognition.

4.2 Implementation Details

All experiments were implemented in the PyTorch framework version 2.4. The AdamW optimizer was used with training hyperparameters and a cosine annealing learning rate schedule configured analogously to the research (Axyonov et al., 2024). The batch size was set to 64 for all training stages. During the pre-training phase on the Slovo corpus (Kapitanov et al., 2023), a total of 500 epochs were run with a masking ratio of 0.75. In the fine-tuning phase, the model was trained for 250 epochs on the TheRusLan corpus (Kagirov et al., 2020). All experiments were performed on a high-performance computing infrastructure based on an NVIDIA A100 GPU with 80 GB of video memory. The software environment was based on Python 3.12 with PyTorch 2.4 and CUDA support. The operating system was CentOS 7 with Linux kernel, which provided a stable and reproducible environment for all experimental procedures. Docker was used to manage the dependencies and to ensure fixed initial random seeds for consistent experimental replication.

4.3 Recognition Performance

The recognition results demonstrate that G-MAE method outperforms previous SOTA methods. Specifically, on the TheRusLan corpus (Kagirov et al., 2020), G-MAE achieved a recognition accuracy of 94.48%, outperforming competing methods, as shown in Table 1.

4.4 Ablation Studies

An ablation study was performed by selectively removing key modules and varying the masking ratio during pre-training, in order to assess the impact of different components of the proposed architecture, as shown in Table 2.

The ablation results clearly show the contribution of each architectural component to the final recognition accuracy. Removing the $MSDC$ module resulted in a decrease in accuracy from 94.48% to 92.67%, confirming its importance in capturing multi-scale spatio-temporal dependencies in gesture sequences. Removing the $MSC\text{-}FFN$ resulted in a further drop

Method	Recognition rate, %
(Axyonov et al., 2021b)	53.07
	68.23
	69.74
	73.54
	74.28
	77.43
	79.98
	84.67
	87.38
	88.92
(Axyonov et al., 2022)	91.14
(Ryumin et al., 2023a)	93.33
Ours (G-MAE)	**94.48**

Table 1. Comparison of gesture recognition accuracy (%) on the TheRusLan corpus.

Configuration	Recognition rate, %
Masking 50–55%	93.41
Masking 55–60%	93.72
Masking 60–65%	93.88
Masking 65–70%	94.02
Masking 50–70%	**94.48**
Masking 75–80%	93.85
Masking 80–85%	92.97
Masking 85–90%	91.65
Without MSDC module	92.67
Without MSC-FFN module	91.95
Without MHSA module	90.54
Without MSDC & MSC-FFN	89.72

Table 2. Ablation study and masking ratio influence on recognition accuracy (%) on the TheRusLan corpus.

to 91.95%, highlighting its role in improving feature representation through multi-scale processing within the feed-forward layers. Removing the $MHSA$ module had an even more pronounced effect, reducing performance to 90.54%, highlighting the importance of attention-based mechanisms for modelling long-range dependencies in gesture data. The most significant degradation was observed when both the $MSDC$ and $MSC\text{-}FFN$ modules were excluded simultaneously, resulting in a recognition accuracy of 89.72%. This confirms the synergistic effect of these modules in preserving the rich and diverse spatio-temporal features necessary for accurate gesture recognition.

In addition, the analysis of the masking ratio shows that optimal performance is achieved when the masking ratio during the pre-training is in the range of 50% to 70%. The highest accuracy of 94.48% was recorded with a dynamic masking ratio sampled within this interval. Both lower (50–55%) and higher (75–90%) masking ratios resulted in decreased recognition accuracy, suggesting that excessive reduction or preservation of input information negatively affects the model's ability to learn robust latent representations. Notably, recognition accuracy declined above 70% masking, highlighting the need to balance information retention and learning pressure during pre-training.

5. Conclusion and Future Work

In this research, we introduced a novel method for gesture recognition based on G-MAE (Gesture-aware Masked Autoencoder). Through extensive experimentation, we have shown that our method improves recognition accuracy. On the TheRusLan corpus, it achieves 94.48%, outperforming all previous SOTA methods. The method uses self-supervised pre-training on the Slovo gesture corpus, followed by fine-tuning on the TheRusLan corpus. During this process, we investigated the effect of different masking ratios and found that a masking ratio in the range of 50% to 70% provided the best performance. Our ablation study also highlighted the critical importance of the key modules $MSDC$, $MSC\text{-}FFN$ and $MHSA$. Removal of these modules resulted in a decrease in performance, demonstrating their essential role in improving the model's ability to capture and process gesture-related features.

In addition, we observed that recognition accuracy decreased when the masking ratio exceeded 70%, confirming the need for a balanced method in the pre-training phase. This finding highlights the importance of carefully tuning the masking ratio to avoid excessive loss of information, which could otherwise hinder the model's ability to learn effectively. Overall, our method not only demonstrates the performance of SOTA, but also provides valuable insights into the design of self-supervised models for gesture recognition.

Future work will focus on expanding the corpus with a wider range of gestures to improve generalization and robustness. We plan to explore advanced self-supervised techniques such as contrastive learning for better feature extraction. Another direction is the integration of multimodal data to improve recognition in complex scenarios. In addition, we aim to optimize the method for smart devices and develop domain-specific fine-tuning strategies for different application contexts.

Acknowledgements

This research is financially supported by the Russian Science Foundation (`https://rscf.ru/en/project/24-71-00083/`, No. 24-71-00083).

References

Alonazi, M., Ansar, H., Al Mudawi, N., Alotaibi, S. S., Almujally, N. A., Alazeb, A., Jalal, A., Kim, J., Min, M., 2023. Smart Healthcare Hand Gesture Recognition using CNN-based Detector and Deep Belief Network. *IEEE Access*, 11, 84922–84933. doi.org/10.1109/ACCESS.2023.3289389.

Altaher, A. S., Bang, C., Alsharif, B., Altaher, A., Alanazi, M., Altaher, H., Zhuang, H., 2025. Mamba Vision Models: Automated American Sign Language Recognition. *Franklin Open*, 10, 100224. doi.org/10.1016/j.fraope.2025.100224.

Axyonov, A. A., Kagirov, I. A., Ryumin, D. A., 2022. A Method of Multimodal Machine Sign Language Translation for Natural Human-Computer Interaction. *Journal Scientific and Technical Of Information Technologies, Mechanics and Optics*, 139(3), 585. doi.org/10.17586/2226-1494-2022-22-3-585-593.

Axyonov, A., Ryumin, D., Ivanko, D., Kashevnik, A., Karpov, A., 2024. Audio-visual speech recognition in-the-wild: Multi-angle vehicle cabin corpus and attention-based method. *IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 8195–8199.

Axyonov, A., Ryumin, D., Kagirov, I., 2021a. Method of Multi-Modal Video Analysis of Hand Movements for Automatic Recognition of Isolated Signs of Russian Sign Language. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 44, 7–13. doi.org/10.5194/ISPRS-ARCHIVES-XLIV-2-W1-2021-7-2021.

Axyonov, A., Ryumin, D., Kagirov, I., 2021b. Method of Multi-Modal Video Analysis of Hand Movements for Automatic Recognition of Isolated Signs of Russian Sign Language. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIV-2/W1-2021, 7–13. doi.org/10.5194/ISPRS-ARCHIVES-XLIV-2-W1-2021-7-2021.

Bilge, Y. C., Ikizler-Cinbis, N., Cinbis, R. G., 2024. Cross-Lingual Few-Shot Sign Language Recognition. *Pattern Recognition*, 151, 110374. doi.org/10.1016/j.patcog.2024.110374.

Christidis, A., Papaioannidis, C., Mademlis, I., Pitas, I., 2024. Lightweight human gesture recognition using multimodal features. *European Signal Processing Conference (EUSIPCO)*, 977–981.

Feng, Y., Chen, N., Wu, Y., Jiang, C., Liu, S., Chen, S., 2024. DFCNet+: Cross-Modal Dynamic Feature Contrast et for Continuous Sign Language Recognition. *Image and Vision Computing*, 151, 105260. doi.org/10.1016/j.imavis.2024.105260.

Garg, M., Ghosh, D., Pradhan, P. M., 2024. Gestformer: Multiscale wavelet pooling transformer network for dynamic hand gesture recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2473–2483.

Hampiholi, B., Jarvers, C., Mader, W., Neumann, H., 2023. Convolutional Transformer Fusion Blocks for Multi-Modal Gesture Recognition. *IEEE Access*, 11, 34094–34103. doi.org/10.1109/ACCESS.2023.3263812.

Hashi, A. O., Hashim, S. Z. M., Asamah, A. B., 2024. A Systematic Review of Hand Gesture Recognition: An Update from 2018 to 2024. *IEEE Access*, 12, 143599-143626. doi.org/10.1109/ACCESS.2024.3421992.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

Ikne, O., Allaert, B., Wannous, H., 2024a. Skeleton-based self-supervised feature extraction for improved dynamic hand gesture recognition. *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, IEEE, 1–10.

Ikne, O., Allaert, B., Wannous, H., 2025. Ag-mae: Anatomically guided spatio-temporal masked auto-encoder for online hand gesture recognition. *International Conference on 3D Vision*, 1–11.

Ikne, O., Slama, R., Saoudi, H., Wannous, H., 2024b. Spatio-temporal sparse graph convolution network for hand gesture recognition. *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, IEEE, 1–5.

Kagirov, I., Ivanko, D., Ryumin, D., Axyonov, A., Karpov, A., 2020. TheRuSLan: Database of Russian Sign Language. *Language Resources and Evaluation Conference (LREC)*, 6079–6085.

Kapitanov, A., Karina, K., Nagaev, A., Elizaveta, P., 2023. Slovo: Russian sign language ddataset. *International Conference on Computer Vision Systems*, Springer, 63–73.

Kapitanov, A., Kvanchiani, K., Nagaev, A., Kraynov, R., Makhliarchuk, A., 2024. Hagrid - hand gesture recognition image dataset. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4572–4581.

Karsh, B., Laskar, R. H., Karsh, R. K., 2024. mIV3Net: Modified Inception V3 Network for Hand Gesture Recognition. *Multimedia Tools and Applications*, 83(4), 10587–10613. doi.org/10.1007/s11042-023-15865-1.

Kim, S., Jung, J., Lee, K. J., 2024. A Real-Time Sparsity-Aware 3D-CNN Processor for Mobile Hand Gesture Recognition. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 71(8), 3695–3707. doi.org/10.1109/TCSI.2024.3408072.

Liu, J., Huang, X., Zheng, J., Liu, Y., Li, H., 2023. Mixmae: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6252–6261.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Ni, S., Al-qaness, M. A., Hawbani, A., Al-Alimi, D., Abd Elaziz, M., Ewees, A. A., 2024. A Survey on Hand Gesture Recognition based on Surface Electromyography: Fundamentals, Methods, Applications, Challenges and Future Trends. *Applied Soft Computing*, 166, 112235. doi.org/10.1016/j.asoc.2024.112235.

Ojeda-Castelo, J. J., Capobianco-Uriarte, M. d. L. M., Piedra-Fernandez, J. A., Ayala, R., 2022. A Survey on Intelligent Gesture Recognition Techniques. *IEEE Access*, 10, 87135-87156. doi.org/10.1109/ACCESS.2022.3199358.

Papadimitriou, K., Potamianos, G., 2023. Sign language recognition via deformable 3d convolutions and modulated graph convolutional networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 1–5.

Qi, J., Ma, L., Cui, Z., Yu, Y., 2024. Computer Vision-based Hand Gesture Recognition for Human-Robot Interaction: A Review. *Complex & Intelligent Systems*, 10(1), 1581–1606. doi.org/10.1007/s40747-023-01173-6.

Ryumin, D., Ivanko, D., Axyonov, A., 2023a. Cross-Language Transfer Learning using Visual Information for Automatic Sign Gesture Recognition. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 209–216. doi.org/10.5194/isprs-archives-xlviii-2-w3-2023-209-2023.

Ryumin, D., Ivanko, D., Ryumina, E., 2023b. Audio-Visual Speech and Gesture Recognition by Sensors of Mobile Devices. *Sensors*, 23(4), 2284. doi.org/10.3390/s23042284.

Sara, U., Akter, M., Uddin, M. S., 2019. Image Quality Assessment through FSIM, SSIM, MSE and PSNR - a Comparative Study. *Journal of Computer and Communications*, 7(3), 8–18. doi.org/10.4236/JCC.2019.73002.

Sincan, O. M., Keles, H. Y., 2020. AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset and Baseline Methods. *IEEE Access*, 8, 181340–181355. doi.org/10.1109/ACCESS.2020.3028072.

Vostrikov, S., Anderegg, M., Benini, L., Cossettini, A., 2024. Unsupervised Feature Extraction from Raw Data for Gesture Recognition with Wearable Ultra Low-Power Ultrasound. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 71(7), 831–841. doi.org/10.1109/TUFFC.2024.3404997.

Xu, Z., Dai, Y., Liu, F., Chen, W., Liu, Y., Shi, L., Liu, S., Zhou, Y., 2023. Swin MAE: Masked Autoencoders for Small Datasets. *Computers in Biology and Medicine*, 161, 107037. doi.org/10.1016/j.compbiomed.2023.107037.

Zhao, W., Hu, H., Zhou, W., Mao, Y., Wang, M., Li, H., 2024. Masa: Motion-Aware Masked Autoencoder with Semantic Alignment for Sign Language Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11), 10793–10804. doi.org/10.1109/TCSVT.2024.3409728.

Zou, Y., Cheng, L., 2021. A Transfer Learning Model for Gesture Recognition based on the Deep Features Extracted by CNN. *IEEE Transactions on Artificial Intelligence*, 2(5), 447–458. doi.org/10.1109/TAI.2021.3098253.