Gated Self-Supervised Feature Fusion for Label-Efficient Hand Gesture Classification

Ziyue Lin[*]
Department of Computer Science - MScAC
University of Toronto
Toronto, Canada
ziyuelin@cs.toronto.edu

Zhengan Du[*]
Department of Electrical and Computer Engineering
University of Toronto
Toronto, Canada
zhengan.du@mail.utoronto.ca

***Abstract*—Hand Gesture Recognition (HGR) is central to VR/AR and touch-free interfaces, yet current single-image methods rely heavily on large labeled datasets and struggle under label scarcity or occlusion. Self-Supervised Learning (SSL) provides a potential solution, but its use in region-sensitive HGR remains largely unexplored. In this work, we propose the Gated Self-Supervised Feature Fusion Network, a two-stream architecture that integrates spatially structured ViT features from DINOv3 with instance-discriminative CNN features from SwAV. Our design includes (i) Segmentation-Guided Attention KL Regularization to softly guide ViT attention toward hand regions without requiring pixel-level masks, and (ii) a lightweight gated fusion module that adaptively weights complementary SSL representations to form a unified gesture descriptor. Experiments on OUHANDS show notable gains in low-label settings—e.g., 86.5% accuracy when trained with only 5% of labels, compared to other supervised methods, which are around 20% accuracy.**

***Index Terms*—Hand Gesture Recognition, Self-Supervised Learning, Low-Label Learning, Feature Fusion.**

I. Introduction

Hand Gesture Recognition (HGR) is a crucial component of Human-Computer Interaction (HCI), particularly in VR/AR systems and touch-free interfaces. The state-of-the-art HGR models are primarily based on deep learning methods. Prior work on single-image HGR has predominantly relied on transfer learning from large supervised models pretrained on large datasets with high-quality annotation. Early approaches used CNN-based classifiers trained directly on cropped RGB hand images [1]. Later, landmark-driven pipelines became popular: hand keypoint detectors such as MediaPipe Hands [2] or MMPose [3] define and detect the keypoint localization of 21 hand-knuckle coordinates, which are then fed into Multi-Layer Perceptrons (MLPs) or Graph Convolution Networks (GCNs) for classification. Hand landmarks provide compact representations of hand pose; however, this simplification miss subtle information, such as the thickness and texture of the palm and fingers, which are helpful for accurate, personalized experience for downstream applications. Moreover, they struggle at test time when images are under occlusion, motion blur, or domain shifts. Recent transformer-based models, such as ViT variants [4], have shown strong recognition performance. However, they also require substantial labeled data for training or fine-tuning. Overall, the existing approaches heavily rely on large data volumes and accurate label annotations. The domain for label-scarce HGR remains relatively under-explored.

In recent years, Self-Supervised Learning (SSL) has shown success in building many deep learning applications. SSL enables models to learn rich and meaningful visual structure and data variance as feature representations from large volumes of unlabeled data. The nature of SSL does not require labeled data, which offers a promising solution for label-scarce situations. Contrastive encoders such as SimCLR [5] learn representations by maximizing agreement between differently augmented views of the same image. SwAV [6] replaces explicit pairwise contrast with online clustering, and learns instance-discriminative features by aligning embeddings to evolving prototype assignments. Teacher–student ViT methods such as DINO [7], DINOv2 [8], and DINOv3 [9] enforce cross-view consistency between a momentum teacher and a student network to acquire emergent semantic structure. Those methods have shown strong performance on natural image benchmarks. However, their success is mainly in the medical imagary and remote sensing domains since these datasets often exhibit limited appearance variation and strong structural regularities that align well with SSL objectives. Application to region-sensitive tasks, such as single RGB image HGR is still understudied.

To address this gap, we propose the Gated Self-Supervised Feature Fusion Network[1], a two-stream architecture that integrates (i) spatially structured ViT features from DINOv3 and (ii) instance-discriminative CNN features from SwAV. Our method includes two components: (1) Segmentation-Guided Attention KL Regularization, which gently steers ViT attention toward hand regions without requiring pixel-level annotations; and (2) a lightweight gated fusion module that dynamically weights features from each SSL encoder to form a unified gesture representation.

II. Related Work

A. Hand Gesture Recognition

Earlier work applied standard Convolutional Neural Networks (CNNs) on cropped RGB hand images to classify static

[*] These authors contributed equally to this work.

[1] The code for this project is publicly available on GitHub

gestures. For example, Islam and Hossain [10] proposed one of the early deep-learning pipelines for single-image HGR. They showed that a CNN trained with data augmentation could recognize static gestures from appearance alone. More recently, transformer-based models have been introduced to capture spatial structure from raw RGB inputs. Static hand gesture recognition method based on the Vision Transformer, such as [11] and ViT-HGR [12], showed that ViT architectures can outperform traditional CNNs by using patch-level self-attention to encode subtle pose cues. Alongside these raw image-based approaches, MediaPipe's 3D landmark-based pipelines for hands have become more popular, as many studies have demonstrated their effectiveness. This method typically first passes the raw RGB images to a pre-trained hand-keypoint extractor, then runs classification models on the extracted keypoint coordinates. Gil-Martín et al [13] proposed a deep learning network that uses the 21-point MediaPipe Hands skeleton as input to deep networks for efficient static gesture classification, and Li et al [14] extended this paradigm to temporal gestures by combining MediaPipe keypoints with a Transformer encoder.

B. Self Supervised Learning in Vision

1) DINOv3: The DINO (self-DIstillation with NO labels) framework, and its subsequent variants like DINOv3, represent a powerful branch of SSL based on knowledge distillation. It operates by training a ”student” network to match the output of a ”teacher” network (often a momentum-updated version of the student). By passing the global view and local view of the same image to different networks, DINO enforces global and local view consistency without the need for negative pairs. The loss compares two networks' output probabilities. The student's parameter is updated by backpropagation, while the teacher's parameter is updated through an exponential moving average of the student's. When trained via this self-distillation mechanism, ViT inherently learns to segment and localize objects by producing high-quality attention maps. The attention map visually highlights salient regions of the input image, which often correspond precisely to the object boundaries.

2) SwAV: The SwAV (Swapping Assignments between Views) is another influential SSL method that contrasts two image views by comparing clusters to which each view belongs. It bypasses the need for computationally expensive contrastive pair construction required by SimCLR [15] by using an online clustering approach. The core idea is to predict the ”prototypes” or cluster assignments of one augmented view of an image from the feature vector of another augmented view of the same image. This prediction is made against a set of trainable prototypes that partition the feature space. A key implementation strength of SwAV is its Multi-Crop strategy: generate two high-resolution views together with several low-resolution crops, which are all fed to the same backbone. It allows the model to learn from multiple different resolutions and views of the input image simultaneously. Unlike DINO's focus on structural attention, SwAV's training objective emphasizes pushing features into distinct and meaningful clusters. SwAV is usually implemented with a standard ResNet backbone, which provides instance-discriminative features that complement the structural features derived from the DINO-ViT path.

C. Feature Fusion

Feature fusion is a crucial technique for combining complementary information from different network streams, moving beyond static methods like concatenation or summation which fail to account for varying feature salience. To address this, Arevalo et al. [16] introduced a gating mechanism that learns how much information to take from each modality. Our proposed Shallow Gated Fusion mechanism adopts a similar approach, utilizing a lightweight, trainable gate to balance the contributions from two distinct SSL streams.

III. Methodology

A. Overall Architecture

Our proposed Gated Self-Supervised Feature Fusion Network as illustrated in Fig. 1 is a label-efficient architecture designed for hand gesture classification. The overall framework combines two streams of SSL paradigms: DINOv3 and SwAV through feature fusion. There are three major components in the model:

1) Stream 1: DINO Feature Extraction: This stream uses a DINOv3 (ViT-S/16) model [17], a distilled version of the larger DINOv3 ViT-7B model pretrained on LVD-1689M dataset [9]. The model leverages its ViT architecture for its ability to learn superior long-range dependencies and generate highly interpretable attention maps through its 12 consecutive transformer blocks. The full input image is reshaped to 256×256 pixels and fed into the ViT encoder. The feature representation for classification is derived from the CLS token output after the final transformer block. This yields a global feature vector $h_D \in \mathbb{R}^{1024}$. During fine-tuning, the later transformer blocks' weights are selectively unfrozen to adapt to the HGC domain.

2) Stream 2: SwAV Feature Extraction: This stream utilizes a SwAV (ResNet50) encoder. Although SwAV is highly effective at learning instance-level discrimination, its traditional approach of processing the full image can dilute the feature quality when the object of interest (the hand) occupies only a small portion of the frame. The input of this stream is therefore guided by Stream 1. The output of this stream is the final global-pooled feature vector $h_S \in \mathbb{R}^{2048}$.

Attention-Guided Cropping Module

To enhance the focus of the SwAV stream, we introduce a novel Attention-Guided Cropping mechanism. During both training and inference:

1) The DINO stream processes the raw input image and generates an attention map from the final transformer layer's output.
2) This attention map is used to calculate a precise Minimum Bounding Rectangle (MBR) that tightly encapsu-

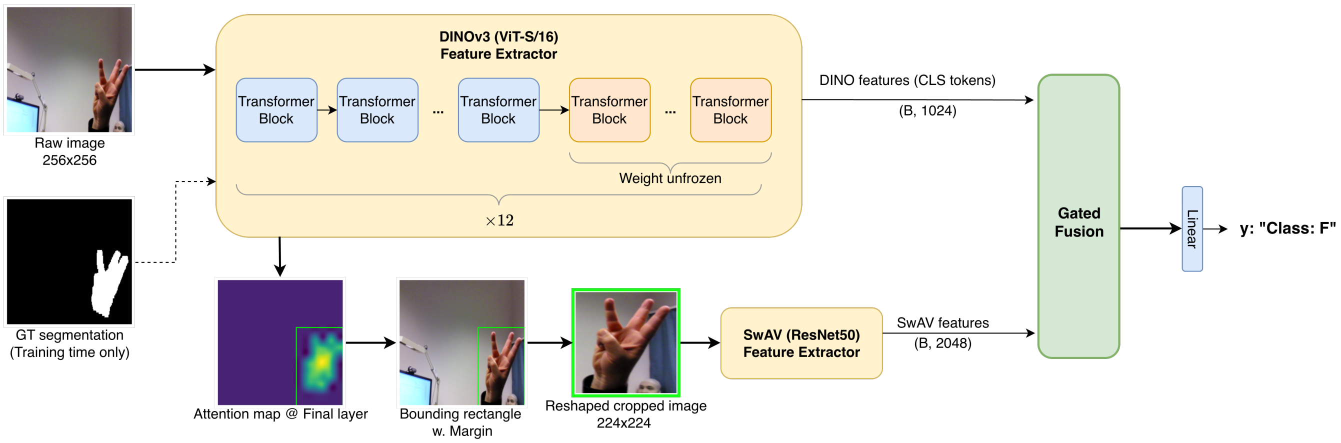


Fig. 1: System flowchart for Gated Self-Supervised Feature Fusion Network

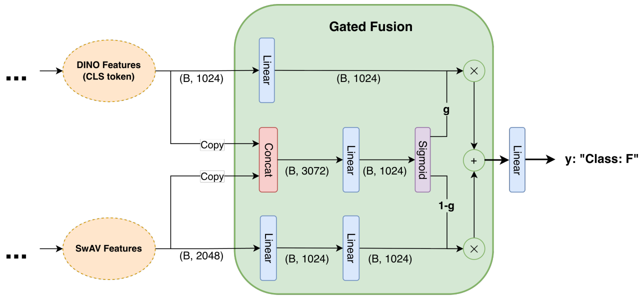


Fig. 2: Gated Fusion Module

lates the hand region, implicitly localized by the DINO attention.

3) A small margin is added to this MBR to retain some context, and the raw input image is cropped based on this augmented MBR.
4) The cropped image is then reshaped (resized) to a fixed input dimension (e.g., 224 × 224) and fed exclusively into the SwAV (ResNet50) Feature Extractor.

This process ensures that the SwAV encoder focuses its representational power on the most relevant area of the hand, generating higher quality and instance-discriminative features (h_S).

3) Gated Fusion: The features of the two streams ($h_D \in \mathbb{R}^{1024}$ and $h_S \in \mathbb{R}^{2048}$) are combined using a lightweight, trainable Shallow Gated Fusion module, which dynamically weights the contribution of each stream:

1) Feature Projection: Both raw feature vectors are projected into a common, lower-dimensional embedding space (e.g., $\mathbb{R}^{1024}$) using two independent Linear layers: $u_D = W_D h_D$ and $u_S = W_S h_S$.
2) Gate Computation: The gate vector g is computed by concatenating the original feature vectors (h_D, h_S), passing them through a shallow network, and applying a Sigmoid activation:

$$g = \sigma\left(\text{Linear}\left([h_D; h_S]\right)\right)$$

The resulting gate vector $g \in (0,1)^{1024}$ contains stream-specific weights.
3) Dynamic Fusion: The final fused feature vector u_{fused} is calculated as a weighted sum of the projected features, using the computed gate g and its complement $(1-g)$:

$$u_{\text{fused}} = g \odot u_D + (1-g) \odot u_S$$

where $\odot$ denotes the element-wise Hadamard product. This mechanism allows the model to dynamically prioritize the DINO features when structural context is key, and the SwAV features when fine-grained details are more critical for classification.

Finally, u_{fused} is passed through a simple Linear layer classifier head to produce the final classification output y.

B. Training Algorithm

To enhance the domain specificity of the DINO attention maps, we introduce a Segmentation-Guided Attention KL Loss: L_{SG}. This loss is applied during the fine-tuning phase of the DINO stream and is computed only during training (where ground truth (GT) segmentation maps are available). It acts as a regularization term, encouraging the DINO attention map A_D to align with the provided GT segmentation S_{GT}:

$$L_{SG} = \text{KL}\left(S_{GT} \parallel A_D\right)$$

This loss ensures that attention is sharply focused on the hand region, thereby improving the quality and precision of attention-guided cropping. (e.g., The attention map in the middle of Fig 3 shows that: with the segmentation guide, DINO successfully pays more attention to the hand region instead of the person's head)

The end-to-end network is optimized using a composite loss function that balances the standard supervised classification objective with our attention regularization term. The final loss L_{final} is defined as:

$$L_{\text{final}} = L_{\text{CE}} + \lambda L_{SG}$$

where:

- L_{CE} is the standard Cross-Entropy loss computed based on the final classification output y.
- L_{SG} is the Segmentation-Guided Attention KL Loss applied to the DINO stream.
- λ is a hyperparameter used to weight the contribution of the attention regularization term.

The entire network is trained by minimizing L_{final}, which simultaneously optimizes classification performance and localization quality within the DINO feature extractor.

IV. Experiments

A. Dataset

This study utilizes the OUHANDS dataset for training and evaluating the hand gesture classification task. The dataset is specifically designed for gesture recognition, featuring diverse hand poses and backgrounds, which poses a realistic challenge for robust feature extraction. The dataset includes 10 distinct hand gesture categories, with 300 samples in each category. Each sample concludes not only the RGB image itself, but also the segmentation and bounding box of the hand. The dataset is splited into training, validation, and test sets with 160, 40 and 100 samples for each category.

Images are resized to 256×256 pixels. Standard ImageNet mean/std normalization is applied. Training data employs extensive augmentation, including random cropping, horizontal flipping ($p = 0.5$), color jittering, and rotation ($\pm 15°$).

B. Implementation Details

1) Hardware Configuration: All models are implemented with PyTorch. Computational experiments are conducted on a server equipped with one NVIDIA 5090 GPU for training and inference.

2) Training Setup and Optimization: We employ a targeted weight-frozen strategy for the two SSL paradigms to preserve the robust general features learned in pre-training:

- Stream 1: The weights of the initial eight Transformer Blocks in DINO ViT are frozen. Only the last four Transformer Blocks and the subsequent feature projection layers are unfrozen and fine-tuned. This allows the DINO network to adapt its higher-level features and attention patterns to the specific hand gesture domain while preserving the pre-trained knowledge.
- Stream 2: All weights in the SwAV ResNet50 backbone are unfrozen during fine-tuning.

We use AdamW optimizer with an initial learning rate LR $= 5 \times 10^{-4}$, and employ a Cosine Annealing Scheduler with 0.05 weight decay. The batch size is 32, and we train the model for 20 epochs.

C. Baseline Comparisons

To provide a comprehensive performance comparison, the proposed Gated Self-Supervised Feature Fusion model is benchmarked against several state-of-the-art supervised and self-supervised learning methods.

a) Supervised Baselines: This category includes models trained end-to-end using standard cross-entropy loss on the labeled dataset. We specifically compare against:

- Landmark-based models: One of the most widely used HGC pipelines which is based on learned hand landmarks. The raw image is first fed to a landmark extractor, such as MediaPipe Hands, and then classification models are built on top of the landmark coordinates. For this work, we compare two different methods: multi-layer perceptrons and random forests.
- Raw image models: Standard ViT, ResNet, and YOLO architectures trained directly on the full and raw input image to serve as general classification benchmarks.
- Hand-crop baselines: Another standard pipeline involving two-stage classification: 1. Hand-crop bounding boxes predicted by a separate supervised localization model (e.g., YOLO), and 2. CNN built on top of the predicted bounding box for classification.

b) Individual SSL Baselines: These models are used to gauge the quality of features learned by individual self-supervised methods compared to our fusion approach. We assess the quality of pre-trained feature representations by freezing the backbone until the last few layers and training only a simple linear classification head.

D. Evaluation Metrics

We employ two principal metrics for model evaluation: 1. Top-1 accuracy: the rate of correctly classified gestures, 2. and macro F1-score: the unweighted harmonic mean of precision and recall across all gesture classes. These two metrics give us a comprehensive and robust assessment of the performance on the multi-class classification task.

E. Label Efficiency Analysis

To validate our core hypothesis that self-supervised feature fusion is superior in data-scarce scenarios, we conduct a detailed analysis of label efficiency. This experiment involves systematically training our proposed network, as well as all baseline models, using varying percentages of the available labeled training data (e.g., 5%, 10%, 25%, 50%, and 100%). This crucial analysis directly quantifies the model's ability to transfer robust knowledge from the unlabeled pre-training stage to the downstream task. This experiment is to demonstrate the robustness of our proposed gate fusion when faced with ”low-label condition problem” compared to fully supervised approaches.

F. Qualitative Feature Analysis

We employ t-distributed Stochastic Neighbor Embedding (t-SNE) to qualitatively analyze the quality and separation of the learned feature space. t-SNE is a non-linear dimensionality

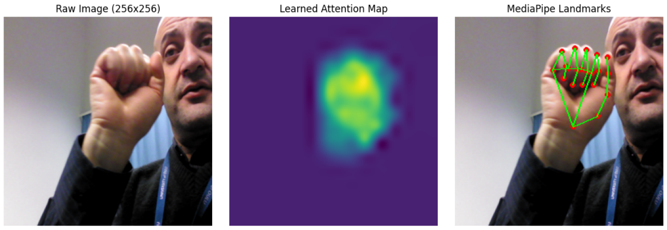


Fig. 3: Segmentation Guided DINO Attention Map vs. MediaPipe Hand Landmarks

reduction technique that maps high-dimensional data (in our case, the 1024-dimensional fused feature vectors) into a lower, two-dimensional space, while preserving local data structure. By visualizing the t-SNE plots, we can assess the distinctiveness of the final fused features, specifically looking for tight, well-separated clusters corresponding to the 10 unique gesture classes, thereby providing a visual measure of the feature representations learned by the Gated Fusion Network.

V. Results and Analysis

A. Overall Performance Comparison

The full model comparison, detailed in Table I, serves as the initial validation of our proposed Gated SSL Fusion approach against various baselines

1) MediaPipe Hands: The Landmark Random Forest Classifier (MediaPipe Hands + Random Forest) achieved a remarkable Top-1 Accuracy of 0.942 and a Macro F1-Score of 0.978, attributable to the quality of its landmark feature set. However, this performance is misleading regarding its general applicability. The MediaPipe extractor failed to detect any hand in 93 (5.8%) training images and 228 (22.8%) test images. The reported score is conditional on successful detection, meaning its real-world accuracy would be significantly lower due to high failure rates in varied environments.

2) Segmentation Guided DINO Features: In contrast, the DINOv3 SSL + Linear Probe demonstrates strong baseline performance, achieving a Top-1 Accuracy of 0.914. The Vision Transformer's inherent ability to capture global dependencies and generate clear attention maps allows it to locate and prioritize the hand region effectively. The gap between DINO (0.914) and our Proposed Gated SSL Fusion (0.949) demonstrates that DINO features are incomplete, lacking the instance-discriminative detail provided by the guided SwAV stream.

3) DINO Guidance for SwAV: Analyzing the individual SSL streams highlights the necessity of our dual-stream approach. The standard SwAV SSL + Linear Probe achieves the lowest performance (Top-1: 0.433), demonstrating poor representation learning for gesture classification when applied without modification. This failure stems from high background noise and the small size of the hand. This result strongly validates our architectural decision to introduce the Attention-Guided Cropping Module, which forces the CNN-based SwAV stream to process only the localized hand region.

B. Label Efficiency Analysis

As shown in Table II, the proposed Gated SSL-Fusion model consistently achieves the highest Top-1 accuracy and Macro-F1 across all labeled data regimes. Similarly, the MediaPipe Hands + Random Forest method shows highly competitive results. MediaPipe's strong performance is rooted in its geometric features, pre-trained on an extremely large hand dataset. MediaPipe first converts the image into tabular data: 21 landmark coordinates, illustrated on the right in Fig 3. A Random Forest classifier, which is highly label-efficient compared to MLPs on small tabular landmark datasets, is then applied to it. Under the extremely low-label setting of 5%, our method attains a Top-1 score of 0.865, significantly outperforming all other image-based deep learning baselines. This high-efficiency performance demonstrates that our framework provides strong label efficiency. It can maintain superior recognition capability in low-label conditions, and it is even fully competitive with the highly optimized, tabular-based landmarks approach.

C. Qualitative Feature Analysis

To visually assess the quality and discriminative power of the features extracted by different models, we utilize the t-SNE dimensionality reduction technique. The feature space is mapped to a 2D plane, and the Minimum Spanning Tree (MST) average distance between class centroids is calculated. A higher MST distance indicates better separation between different categories in the feature space, suggesting a stronger feature representation capability of the model.

We compare three representative models: the ResNet50 model, the supervised Vision Transformer model, and our Gated SSL-Fusion model.

1) ResNet-50 (Avg MST Dist: 18.61): As shown in Figure 4a, the feature space extracted by ResNet-50 exhibits significant overlap, particularly in the central region of the plot, where many class clusters converge. The low MST average distance (18.61) confirms this visual observation, indicating that the model struggles to learn highly discriminative hand gesture features when processing raw images contaminated by background noise.

2) ViT (Avg MST Dist: 22.95): The ViT model, trained with supervised learning (Figure 4b), achieves notably better feature separation compared to ResNet-50. The clusters are more compact, and the boundaries between categories are generally clearer. The increase in MST average distance to 22.95 suggests that the Vision Transformer architecture captures superior global features than the CNN, enhancing the feature space's inherent discriminative quality.

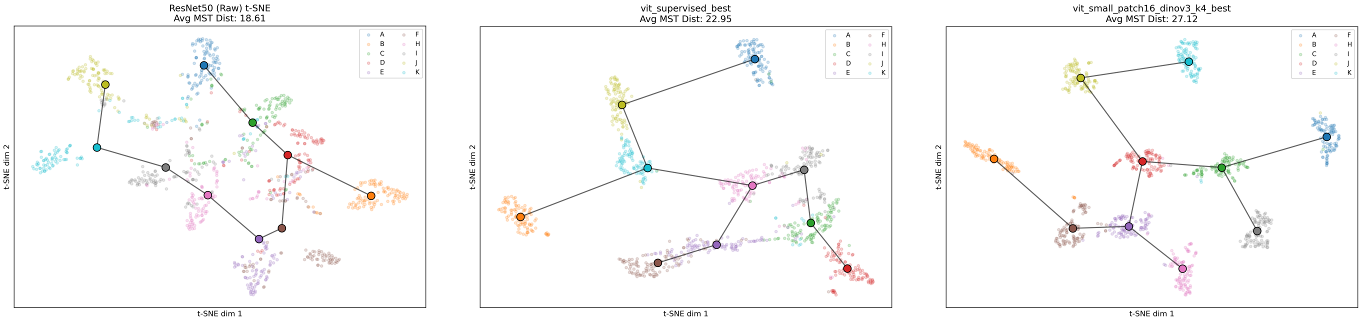
3) Gated SSL-Fusion (Avg MST Dist: 27.12): Among all models compared (Figure 4c), our Gated SSL-Fusion model demonstrates the most outstanding feature separation, achieving the highest MST average distance of **27.12**. Visually, all 10 class clusters are tightly grouped internally but maximally separated in space, with negligible overlap. This strongly validates that DINO forces the model to learn features focused on semantic consistency, such as subtle hand details and pose

TABLE I: Comprehensive Performance Comparison

Method Description	**Backbone(s)**	**Top-1**	**Macro-F1**
Landmark MLP Classifier	MediaPipe Hands + MLP	0.846	0.781
Landmark Random Forest Classifier	MediaPipe Hands + Random Forest	0.942	**0.978**
CNN Features + Linear Probe	ResNet-50	0.759	0.758
CNN Features + Linear Probe	YOLOv8	0.886	0.897
ViT Features + Linear Probe	ViT-small	0.870	0.866
Two-Stage Detection & Classification	YOLOv8 + ResNet-50	0.871	0.899
DINOv3 SSL + Linear Probe	DINOv3	0.914	0.915
SwAV SSL + Linear Probe	SwAV	0.433	0.431
Gated SSL-Fusion (Ours)	DINOv3 + SwAV	**0.949**	0.946

TABLE II: Label efficiency analysis: Accuracy and F1-score comparison under low-data regimes

Percentage of Labeled Data	**MediaPipe Hands + Random Forest**		**ViT**		**ResNet50**		**Gated SSL-Fusion (Ours)**	
	Top-1	Macro-F1	Top-1	Macro-F1	Top-1	Macro-F1	Top-1	Macro-F1
1%	0.300	0.303	0.127	0.084	0.102	0.058	**0.431**	**0.390**
5%	0.734	0.736	0.236	0.199	0.204	0.173	**0.865**	**0.863**
10%	**0.873**	0.874	0.434	0.436	0.167	0.167	0.872	**0.878**
25%	0.922	0.921	0.720	0.719	0.264	0.250	**0.947**	**0.947**



(a) ResNet-50 t-SNE

(b) ViT t-SNE

(c) Gated SSL-Fusion t-SNE

Fig. 4: t-SNE plots for three models with mean MST distances.

variations, thereby maximizing the distinctiveness between categories in the feature space.

VI. Conclusion

In this report, we introduce the Gated Self-Supervised Feature Fusion Network (GSSL-Fusion), a novel two-stream architecture designed to address the challenges of accurate and label-efficient single-image Hand Gesture Recognition (HGR) by harnessing diverse Self-Supervised Learning (SSL) representations. The efficacy of our architecture lies in three critical design choices: (1) Guided Feature Extraction: We fused complementary SSL encoders: DINOv3 (ViT) for structural features and SwAV (ResNet50) for granulated, instance-discriminative features. Since standard SwAV failed on full images due to noise, we implement an Attention-Guided Cropping mechanism, using DINO's attention map to automatically detect the hand and feed a precise crop to the SwAV stream. This drastically improves SwAV's feature quality. (2) Attention Regularization: We incorporate a Segmentation-Guided Attention KL Loss during fine-tuning. This stabilizes and sharpens the DINO encoder's attention maps, forcing the network to focus on true hand boundaries. This guided attention is essential for both DINO's feature quality and providing the highly accurate bounding box for the SwAV input. (3) Dynamic Fusion: We use a Shallow Gated Fusion layer to dynamically weigh the contributions from the projected DINO and SwAV features. This mechanism allows the model to intelligently prioritize structural context on a per-sample basis, avoiding information loss common with static concatenation.

The integrated GSSL-Fusion framework achieves a state-of-the-art Top-1 Accuracy of 0.949 on the fully labeled OUHANDS dataset, outperforming all comparable image-based deep learning baselines. Furthermore, the label efficiency analysis demonstrates superior robustness: under the extremely low-label condition of 5% training data, GSSL-Fusion still achieved a Top-1 score of 0.865.

In summary, this work proves that guided fusion of complementary SSL method is a powerful and label-efficient strategy for deep learning systems in region-sensitive tasks.

References

[1] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3d convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 1–7.

[2] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand tracking," 2020. [Online]. Available: https://arxiv.org/abs/2006.10214

[3] M. Contributors, "Openmmlab pose estimation toolbox and benchmark," https://github.com/open-mmlab/mmpose, 2020.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[6] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," 2021. [Online]. Available: https://arxiv.org/abs/2104.14294

[7] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," 2021. [Online]. Available: https://arxiv.org/abs/2006.09882

[8] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2024. [Online]. Available: https://arxiv.org/abs/2304.07193

[9] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski, "Dinov3," 2025. [Online]. Available: https://arxiv.org/abs/2508.10104

[10] M. Z. Islam, M. S. Hossain, R. ul Islam, and K. Andersson, "Static hand gesture recognition using convolutional neural network with data augmentation," in *2019 Joint 8th International Conference on Informatics, Electronics Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision Pattern Recognition (icIVPR)*, 2019, pp. 324–329.

[11] Y. Zhang, J. Wang, X. Wang, and et al., "Static hand gesture recognition method based on the vision transformer," *Multimed Tools Appl*, vol. 82, pp. 31 309–31 328, 2023.

[12] M. Montazerin, S. Zabihi, E. Rahimian, A. Mohammadi, and F. Naderkhani, "Vit-hgr: Vision transformer-based hand gesture recognition from high density surface emg signals," 2022. [Online]. Available: https://arxiv.org/abs/2201.10060

[13] M. Gil-Martín, M. R. Marini, I. Martín-Fernández, S. Esteban-Romero, and L. Cinque, "Hand gesture recognition using mediapipe landmarks and deep learning networks," in *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART), Volume 3*. SciTePress, 2025, pp. 24–30.

[14] H.-H. Li and C.-C. Hsieh, "Dynamic hand gesture recognition using mediapipe and transformer," *Engineering Proceedings*, vol. 108, no. 1, 2025. [Online]. Available: https://www.mdpi.com/2673-4591/108/1/22

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[16] J. Arevalo, T. Solorio, M. M. y Gómez, and F. A. González, "Gated multimodal units for information fusion," 2017. [Online]. Available: https://arxiv.org/abs/1702.01992

[17] R. Wightman, "Pytorch image models," https://github.com/huggingface/pytorch-image-models, 2019.