Multimodal Egocentric Action Recognition

Giacomo Fantino s310624 | Festa Shabani s317640 | Farisan Fekri s308066

Abstract

Multimodal Egocentric Action Recognition is a specialized domain within computer vision and machine learning that focuses on identifying and classifying actions from a first-person perspective using multiple data modalities. In this paper, we propose a framework leveraging Variational Autoencoders (VAEs) to generate Electromyography (EMG) signals from RGB frames. This is achieved by training VAEs on the Epic-Kitchens and ActionNet datasets, enabling cross-modal data translation. Specifically, VAEs are employed to learn latent representations of RGB videos and EMG signals, which are then utilized to reconstruct EMG signals from RGB data. This approach enables the generation of the missing EMG modality in the Epic-Kitchens dataset, enhancing the robustness and accuracy of action recognition. Thereby, our framework highlights the potential of cross-modal data generation for improving action recognition systems. https://github.com/FestaShabani/aml23-ego/

1. Introduction

Egocentric vision, which captures the world from a first-person perspective using wearable cameras, has emerged as a promising field of study in the domain of computer vision and machine learning. This unique visual perception approach offers valuable insights about the user's activities, surroundings, and interactions, capturing dynamic and interactive scenes, often including the user's hands and objects they manipulate.

We will delve into two prominent datasets in this domain: the Epic-Kitchens [2] dataset, which is a large-scale collection of videos capturing human cooking activities in a kitchen environment, and the ActionNet dataset [3], which provides a rich, synchronized set of data streams, including eye tracking with a first-person camera, forearm muscle activity sensors, a body-tracking system using 17 inertial sensors, finger-tracking gloves, and custom tactile sensors on the hands, coupled with activity labels and with externally captured data from multiple RGB cameras, depth camera, and microphones.

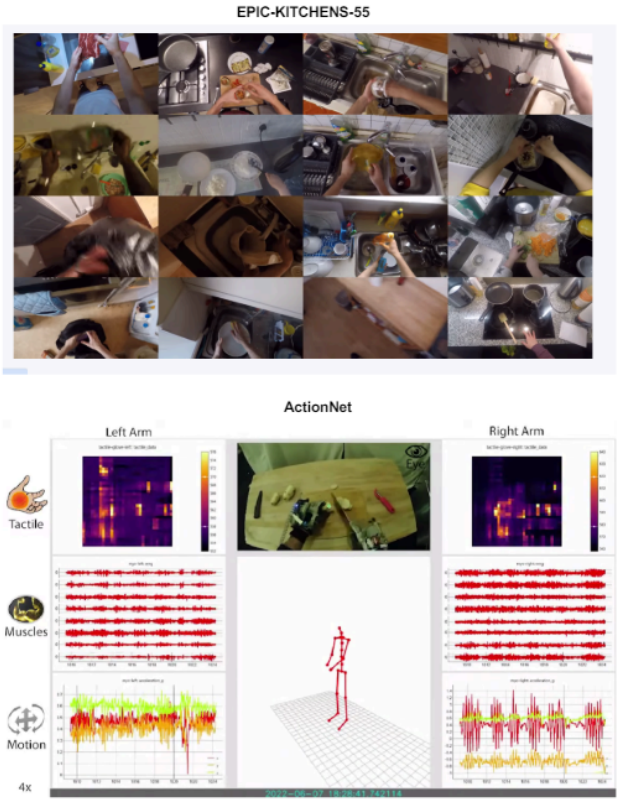

Figure 1. Datasets used in our models

For our study, we focus on the EMG and RGB modalities. To extract meaningful information from these datasets, we adopt a structured approach. Our approach begins by dividing videos into segments called clips and selecting frames using dense or uniform sampling strategies. Subsequently, we leverage a pre-trained model to extract intermediate features from the RGB streams of EPIC-Kitchens and ActionNet. We leverage the deep convolutional capabilities of the I3D [1] backbone to compress the spatial and temporal information into feature representations. These extracted features serve as inputs for training classifiers for action recognition tasks. However, before evaluating the classifiers, we analyze the features using K-means clustering to

visualize patterns within the data. This analysis helps us gain insights into the distribution and characteristics of the features. Additionally, one of our key objectives is to make use of a VAE capable of reconstructing EMG data from RGB streams. This approach involves training the VAE on the features extracted from EPIC-Kitchens and Action-Net datasets, enabling the generation of the missing EMG modality for the Epic-Kitchens dataset, ultimately leading to a more comprehensive and multimodal approach to action recognition.

2. Related works

First-person action recognition has gained significant attention in recent years, primarily due to the increasing availability of wearable cameras and the demand for applications in augmented reality, healthcare, and human-computer interaction. Traditional methods have predominantly relied on visual information captured by RGB cameras, but recent research suggests that combining multiple modalities can enhance the robustness and accuracy of action recognition systems.

Multimodal Approaches in Action Recognition

The work by EPIC-Fusion [5] introduces a pioneering approach to multi-modal fusion tailored for egocentric action recognition, particularly in the context of the EPIC-Kitchens dataset. By proposing a novel architecture for multi-modal temporal binding, the authors demonstrate the efficacy of combining modalities within a defined temporal window. In their study, RGB, Flow, and Audio modalities are integrated using mid-level fusion techniques, complemented by sparse temporal sampling of fused representations. Furthermore, EPIC-Fusion highlights the significance of audio cues in egocentric vision, showcasing their importance in identifying actions and interacting objects.

The paper "Multi-Modal Domain Adaptation for Fine-Grained Action Recognition" [7] tackles the issue of environmental bias in fine-grained action recognition datasets, which often leads to a drop in performance when models trained in one environment are applied to another. Traditional unsupervised domain adaptation methods have overlooked the multi-modal nature of video data. The authors propose a novel approach that combines adversarial training with multi-modal self-supervision, improving performance over source-only training. Tested on EPIC-Kitchens dataset using RGB and Optical Flow modalities, their method outperforms other unsupervised domain adaptation methods, demonstrating its effectiveness in addressing domain shift challenges.

E2(GO)MOTION's [4] use of event-based data introduces a new modality that enhances recognition in dynamic scenarios with lower computational costs.The development of the N-EPIC-Kitchens dataset, which integrates event-based information, further underscores the potential of this modality to improve action recognition accuracy in environments where traditional RGB data might fail due to rapid motions or environmental changes. All these works resonate with our focus on enhancing action recognition through multimodal integration, especially within the context of EPIC-Kitchens dataset.

Variational Autoencoders in Cross-Modal Translation

Variational Autoencoders (VAEs) have emerged as a powerful tool for learning latent representations and generating synthetic data. VAEs have been applied in various contexts, including image synthesis, anomaly detection, and cross-modal translation. For instance, Spurr et al. (2018) [8] demonstrated the efficacy of VAEs in cross-modal hand pose estimation, where they developed a coherent latent space across RGB images, keypoint detections, and 3D configurations. Their work addressed the challenge of estimating complex hand poses from images, offering a unified framework for diverse modalities.

On a similar note, Yu and Oh (2022) [10] explored the use of VAEs for anytime 3D object reconstruction, emphasizing human-robot collaboration by compressing visual data into compact latent variables. Their approach focused on enabling real-time performance in unstable environments, crucial for effective human-robot teaming. Our work extends these concepts by employing VAEs to translate RGB features into EMG signals, thus enriching the feature space for action recognition.

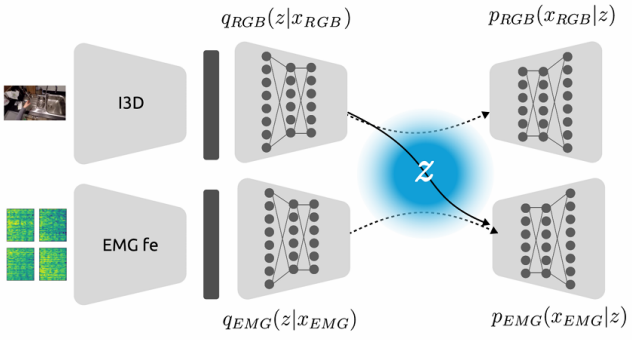

Figure 2. Visual2Signals

3. Methodology

3.1. Data Preprocessing

In action recognition tasks, the selection of frames plays a crucial role in capturing temporal dynamics and contextual information within video clips. Given that each video is partitioned into a fixed number of clips, we employ two distinct sampling strategies:

Dense Sampling: Involves sampling frames around the center of each clip.

Uniform Sampling: Frames are uniformly sampled across the entire clip.

Dense sampling is better for actions that involve rapid or detailed movements within short periods, while uniform sampling is suitable for capturing the overall temporal structure and context of the video. Throughout the experiments, various configurations have been explored, including different numbers of frames per clip and different sampling strategies. By evaluating these setups, the aim is to determine the most optimal approach for our task.

3.2. Feature Extraction from RGB Streams

For extracting intermediate features from the RGB streams, we utilize the Inflated 3D ConvNet (I3D) [1] architecture, a deep convolutional neural network designed for action recognition in videos.

I3D Backbone: The I3D model is pretrained on the Kinetics dataset, a large-scale dataset containing hundreds of thousands of video clips annotated with various human actions. The pretraining on Kinetics allows the I3D model to learn a rich set of spatiotemporal features that are highly effective for action recognition tasks.

Feature Representation: We pass each sampled frame sequence through the I3D network to obtain intermediate feature maps. These feature maps encapsulate rich spatial and temporal information from the RGB streams. Specifically, we use the output of the final convolutional layer before the classification head, which provides a compact representation of the video content.

3.3. Visualization and Clustering of Extracted Features

To gain deeper insights into the effectiveness of our feature extraction process, we employ both t-Distributed Stochastic Neighbor Embedding (t-SNE) for visualization and K-means clustering for quantitative assessment. These techniques help us understand the distribution and separability of the extracted features, providing a basis for further classification tasks.

t-SNE [9] is a powerful tool for visualizing high-dimensional data by projecting it into a lower-dimensional space while preserving the local structure of the data points.

The resulting plot helps us understand the discriminative power of the extracted features. Ideally, features from different actions should be distinct, indicating that the model has successfully captured the relevant spatiotemporal patterns.

K-means clustering is a widely-used method for partitioning data into distinct clusters based on feature similarity. The performance of K-means clustering is assessed by comparing the cluster assignments to the true action labels. A high degree of correspondence indicates that the extracted features are effective for distinguishing between different actions.

3.4. Classifier Architectures for RGB Modality

In this section, we delve into the classifier architectures employed in our study for action recognition using RGB modality data. The primary objective of training these classifiers is to determine which set of hyperparameters yields the most effective features for action recognition.

Temporal Relational Reasoning Network (TRN) [11]

The TRN architecture is adept at processing RGB modality data, enabling the fusion of information across different time steps. The TRN employs a relational reasoning module, which analyzes temporal relationships between different clips of the sample. A neural network specific for each temporal interval enables the model to capture complex dependencies and interactions. Finally, the output of the relational reasoning module is used to predict the action class for the clip. We have created a modified version that we called TRN mod, where we simplified the architecture by incorporating the classifier within the network. Similar to TRN, it computes temporal relations across scales but directly integrates the classifier into the fusion process. This modification aims to streamline the architecture while retaining the effectiveness of relational reasoning for action recognition.

Multi-Layer Perceptron (MLP)

The MLP architecture is well-suited for processing feature vectors directly, making it suitable for scenarios where RGB features have been fused into a single representation. For each clip, the MLP takes the RGB feature vector as input. This feature vector is then passed through a series of densely connected layers with non-linear activation functions. The MLP learns complex patterns in the RGB data, extracting discriminative information relevant to action recognition. The output layer of the MLP produces predictions about the action class for the clip.

Long Short-Term Memory (LSTM)

The LSTM architecture excels at handling sequential data, making it suitable for action recognition tasks involving temporal sequences of RGB frames. For each clip, the LSTM processes RGB features sequentially, considering the temporal dynamics of the input. The LSTM utilizes recurrent neural network (RNN) cells with memory units to capture long-range dependencies in the RGB data. The final hidden state of the LSTM is then used to predict the action class for the clip.

3.5. Feature Extraction from EMG Streams

We utilized the ActionNet's recordings of muscle activity across 8 channels from each forearm, to extract meaningful features for action recognition. The preprocessing

steps involved rectifying the signals, applying a low-pass filter, normalizing the signals, downsampling to a manageable rate and augmenting the data by dividing each action into shorter segments. The preprocessed EMG data was then converted into spectrograms to enable the extraction of time-frequency features.

CNN-based Feature Extraction: We employed a Convolutional Neural Network (CNN) to extract features from the spectrograms. CNNs are adept at capturing spatial hierarchies in the data, which is crucial for recognizing complex patterns in the EMG signals.

3.6. Variational Autoencoder for Cross-Modal Translation

In our attempt to bridge the gap between the two modalities, we employ the Variational Autoencoder (VAE) [6] as a fundamental tool for encoding high-dimensional input data such as RGB and EMG into a latent space. By leveraging the learned latent representations, we aim to generate synthetic samples of the EMG modality.

In practice, two separate VAEs are trained on RGB and EMG inputs to reconstruct each data type independently. The probabilistic nature of the VAE's latent space enables flexible and controlled sampling, which is crucial for translating between modalities. The encoder trained on RGB data is then combined with the decoder trained on EMG data, allowing the reconstruction of EMG signals from RGB inputs. However, integrating these two parts requires fine-tuning to ensure smooth operation. This is where S04 data becomes crucial: by utilizing RGB and EMG data from this specific subject, we can fine-tune the model to effectively reconstruct EMG signals from RGB inputs, capitalizing on the inherent correlation between the data types.

4. Experimental Setup and Results

4.1. Feature Extraction of RGB from Epic-Kitchens

The RGB data utilized in the first part of our study is sourced from the Epic-Kitchens [2] dataset, a comprehensive egocentric video dataset designed for action recognition. This dataset includes recordings from 32 kitchens across 4 cities. Specifically, we use the P08 (D1) subset, which comprises 1543 training and 435 test video samples. Our research focuses on the identification and classification of 8 specific actions (verbs) within this subset.

The RGB features were extracted and categorized according to:

- **Training:** Pretrained on Kinetics or further finetuned on Epic-Kitchens.
- **Sampling Strategy:** Dense or Uniform Sampling.
- **Frames per Clip:** 5, 10, 25.

Each sample was divided into a fixed length of 5 clips. To gain insights into the extracted features, we employed t-SNE and K-Means clustering techniques. We utilized t-SNE to reduce the high-dimensional feature vectors to 2D space. Using K-Means clustering with 8 clusters, we grouped similar features together based on their spatial distribution in the feature space. This approach allowed us to identify distinct clusters of activities within the dataset. We can observe from Fig. 3 that RGB features using fine-tuning, 10 frames per clip and dense sampling, demonstrate the best class separability due to the fine-tuning on the EPIC-Kitchens dataset. Dense sampling seems to be more effective, compared to uniform sampling in Fig.4 , due to I3D backbone being pretrained using a dense sampling strategy. To provide visual context, we also showed plots where images from the central frames of the video clips are overlaid on the plot. We can observe from Fig. 5 that samples seems more grouped by the environment instead of the action, thus presenting **environment bias**, a typical problem of RGB data classification.

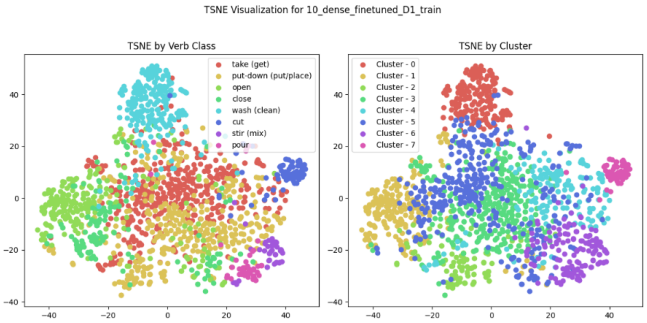

Figure 3. RGB features with fine-tuning, using 10 frames per clip and dense sampling.

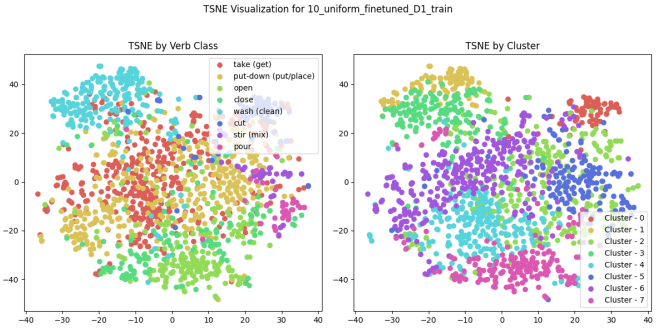

Figure 4. RGB features with fine-tuning, using 10 frames per clip and uniform sampling.

In addition to visual interpretation, classifier results provide quantitative insights into the effectiveness of the extracted features. The results from the classifiers align with our observations: the best results were obtained using a

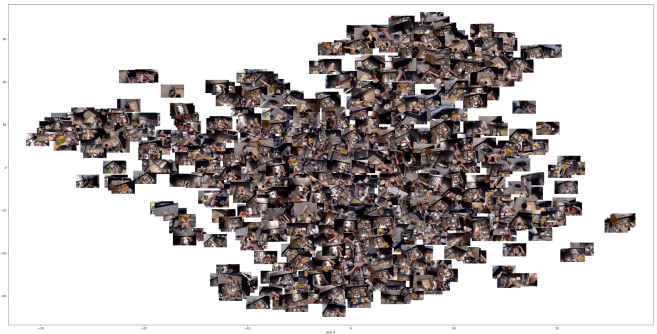

Figure 5. t-SNE visualization of RGB features using central frames, with 10 frames per clip and uniform sampling.

dense sampling approach. The best-performing model was the Temporal Relational Reasoning Network (TRN), highlighting how the feature extraction method captured all the necessary spatial and temporal information needed for effective sample classification.

# Frames	5		10		25	
Sampling	**Dense**	**Uniform**	**Dense**	**Uniform**	**Dense**	**Uniform**
MLP	29.89%	28.51%	30.11%	28.51%	28.51%	28.51%
LSTM	56.32%	53.56%	57.01%	55.63%	57.47%	55.63%
TRN	55.63%	55.17%	58.16%	57.01%	58.16%	56.09%
TRN_mod	55.17%	54.25%	58.16%	56.32%	58.16%	55.17%

Table 1. Performance of classifiers on the RGB data fine-tuned on Epic-Kitchens

We also experimented with utilizing the I3D model pretrained solely on the Kinetics dataset. As expected, we observed that the class separation decreased and the accuracy of all models diminished compared to using the fine-tuned models.

4.2. Feature Extraction of RGB and EMG from ActionNet

ActionNet [3] dataset, designed as a comprehensive multimodal dataset, includes diverse data types such as motion, tactile, muscle activity, and more. Specifically, we concentrated on the EMG (all subjects) and RGB (only subject S04). For RGB features, we extracted them using the same methodology as for the Epic-Kitchens dataset. We employed the best configuration observed in Epic-Kitchens, which involved using 10 frames per clip, dense sampling, and fine-tuning on Epic-Kitchens. Regarding EMG data, we focused our analysis on a subset of 20 verb-based action classes.

4.2.1 Preprocessing of EMG

The ActionNet dataset records muscle activity across 8 channels from each forearm. To prepare this data for feature extraction, we perform the following preprocessing steps:

Rectification: Each channel's signal is rectified by taking its absolute value.

Low-pass Filtering: A low-pass filter with a cutoff frequency of 5 Hz is applied to each rectified channel to highlight general muscle activation levels.

Normalization: The signals from all 8 channels of an armband are jointly normalized and shifted to the range [-1, 1] using the minimum and maximum values across all channels.

Downsampling: To manage the high-frequency data captured at roughly 160 Hz, we downsample the data to a more manageable rate of 10 Hz. This is done using interpolation to extract the data at the desired frequency, ensuring consistency across all sensors.

Augmentation: Given the limited number of action annotations in the ActionNet dataset, we augment the data by dividing each action into shorter segments (subactions) of 10 seconds. This step increases the number of training samples and ensures that our models can be effectively trained.

Conversion to Spectrograms

After preprocessing, the EMG data is converted into spectrograms. This transformation enables the extraction of time-frequency features that are critical for understanding muscle activity patterns. An example of the obtained spectrogram can be observed in Fig. 6.

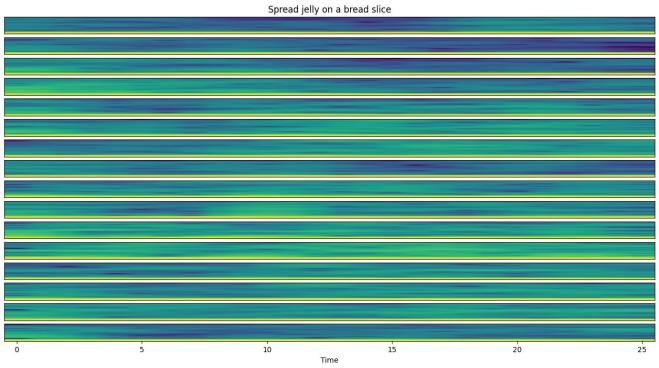

Figure 6. Spectrogram example showing muscle activity for both forearms

We experimented with various configurations, including a higher frequency rate (80 Hz) and different sample durations (5 and 20 seconds). Our findings indicate that shorter segments contain less information, making it challenging for the network to extract sufficient information for action classification. Conversely, longer samples suffer from a smaller training dataset size, posing challenges for classifier training.

4.2.2 Feature Extraction Using CNNs

Convolutional Neural Networks (CNNs) are employed to extract features from the spectrograms. CNNs are well-suited for this task as they can capture spatial hierarchies in the data, making them effective for recognizing complex patterns in the EMG signals.

CNN architecture:

The provided CNN architecture is designed for feature extraction in a 20-class classification problem. It starts with two convolutional layers: the first with 2 filters and the second with 4 filters, both using 3x3 kernels and ReLU activation. This is followed by max-pooling layers to reduce dimensionality. The resulting feature maps are flattened and passed through a dropout layer for regularization. Finally, two fully connected layers are used: the first with 832 neurons and the second with 20 neurons, corresponding to the number of classes. The model outputs logits for classification and feature vectors for further processing.

This preprocessing and feature extraction pipeline enables the effective use of EMG data for multimodal egocentric action recognition, facilitating the generation of simulated EMG samples from RGB inputs through our VAE-based framework.

4.2.3 Visualization

Following feature extraction, we visualized the data to gain insights into the underlying patterns, as depicted in Fig. 8. Notably, we observed instances where different labels denoted the same action but were distinguished by the object involved (e.g., *Slice a potato* and *Slice bread*), making it challenging to differentiate based solely on muscle activity. To address this, we merged samples belonging to classes that referenced the same action. This approach effectively highlighted how samples pertaining to the same actions clustered together. While we maintained separate classes for actions like *get item from cabinet* and *get item from refrigerator*, it is evident that samples within these classes share significant similarities.

4.3. Visual2Signals: Cross-modal translation

We employed Variational Autoencoders (VAEs) to learn latent representations and facilitate reconstruction tasks on the extracted features from both RGB data of Epic-Kitchens and EMG+RGB data from ActionNet. Specifically:

1. **VAE for RGB Reconstruction:** The first VAE was trained on the extracted RGB features from the Epic-Kitchens dataset. This VAE was designed to reconstruct RGB frames, leveraging variational inference for effective feature extraction and reconstruction.
2. **VAE for EMG Reconstruction:** The second VAE was utilized for reconstructing EMG signals extracted from the ActionNet dataset. This VAE architecture was tailored to handle the time-frequency features derived from EMG spectrograms, providing insights into muscle activity patterns.
3. **Hybrid VAE for Subject S04 in ActionNet:** To explore cross-modal learning, we constructed a hybrid VAE using the encoder from the RGB VAE and the decoder from the EMG VAE. This hybrid VAE was fine-tuned specifically on subject S04 of ActionNet, leveraging both RGB and EMG modalities to enhance action recognition capabilities.

Architecture of the Variational Autoencoder (VAE): The VAE architecture consists of an encoder and a decoder. The encoder module utilizes sequential layers of linear transformations, batch normalization, ReLU activation, and dropout of 0.2 for regularization. It outputs mean (μ) and standard deviation (σ) parameters. The decoder module reconstructs the latent representations back into the original feature space using sequential layers of linear transformations, ReLU activation, batch normalization, and dropout. This end-to-end framework, featuring a bottleneck size of 256, supports variational inference through the reparameterization trick, facilitating effective feature reconstruction and fine-tuning tasks across different modalities and datasets.

The VAE models were trained over 100 epochs with an initial learning rate (α) set at 0.001. The learning rate was adjusted using a step decay schedule, decreasing by a factor of 0.01 every 30 epochs. The beta (β) parameter for the VAE was set to 0.00001 to balance the reconstruction loss and the Kullback-Leibler divergence during training.

We conducted a comparative analysis of the extracted EMG features against those reconstructed by the RGB (Fig. 7) and EMG (Fig. 9) VAEs. Both VAEs successfully generated a latent space that accurately reconstructed the original data. Fine-tuning on subject S04 of ActionNet further enabled us to compute synthetic EMG data, resulting in promising outcomes (Fig. 10).

Starting with the models using only EMG, we obtained the following results:

- TRN: 53%
- TRN_mod: 29%
- LSTM: 54%
- MLP: 29.66%

Similar to the results for RGB, these models effectively learned the temporal relationships within the data, enabling accurate sample classification.

Subsequently, we constructed a multimodal model using late fusion, leveraging the best-performing models for each modality. The combined results are shown in Tab. 2.

RGB Model	EMG Model	Accuracy
TRN	LSTM	57%
TRN	TRN_mod	56.55%
TRN_mod	LSTM	56.1%

Table 2. Multimodal accuracy

Class	RGB	EMG	RGB+EMG
take	54%	**62%**	57%
put	**62%**	36%	59%
open	**67%**	62%	65%
close	**41%**	35%	38%
wash	**70%**	68%	65%
cut	**69%**	62%	62%
stir	60%	**75%**	70%
pour	**27%**	23%	23%

Table 3. Results for different modalities

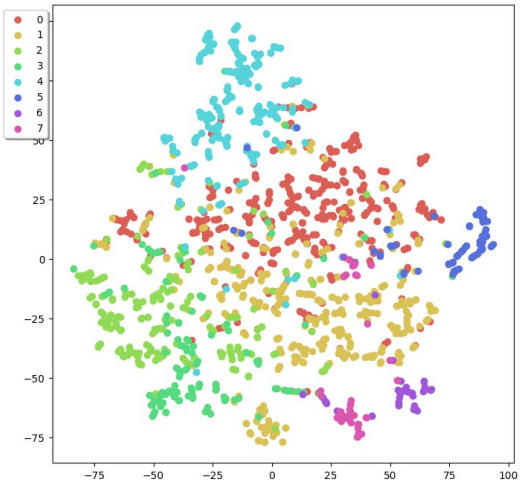

Figure 7. Reconstructed RGB features

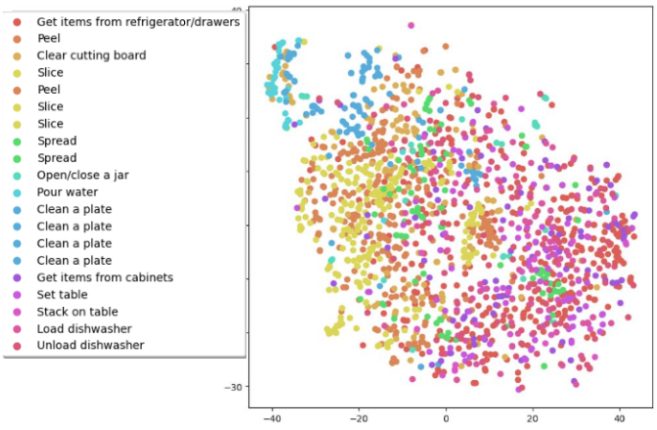

Figure 8. EMG features

4.4. Analysis of the results

Although the newly trained classifiers did not outperform the RGB-only modality, we conducted an analysis focusing on the precision for each class.

As we can observe from Tab. 3 for the classes *take* and *stir* the EMG and EMG+RGB classifiers achieve superior results compared to RGB. Interestingly for the EMG data the class *put* has a drop in precision, but using EMG+RGB data we were able to improve and get a result closer to RGB. This observation suggests that while the overall accuracy might be lower with EMG data alone, advanced multimodal models can effectively leverage both modalities, enhancing performance beyond what is achievable with RGB data alone.

5. Conclusion

In this study, we explored the application of Variational Autoencoders (VAEs) for enhancing multimodal egocentric action recognition through cross-modal translation. By leveraging VAEs trained on EpicKitchen and ActionNet datasets, we successfully generated Electromyography (EMG) signals from RGB frames. This approach enabled us to bridge the gap between different modalities, demonstrating the feasibility of synthesizing missing EMG data in the EpicKitchen dataset.

Our results underscore the potential of modality translation as a robust technique for data augmentation in multimodal tasks. Models trained on translated data achieved comparable accuracy to those trained on original modalities, affirming the efficacy of VAEs in capturing and leveraging latent representations across modalities.

Moreover, while multimodal learning exhibited performance comparable to using RGB data alone, our findings highlight ongoing opportunities for advancing action recognition systems. Future research should focus on refining multimodal fusion techniques and exploring additional modalities to further enhance model performance and generalize across diverse environments.

References

[1] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017. 1, 3

[2] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, and W. Price. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 1, 4

[3] J. DelPreto. ActionNet: A Multimodal Dataset for Human Activities Using Wearable Sensors in a Kitchen Environment. Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, n.d. 1, 5

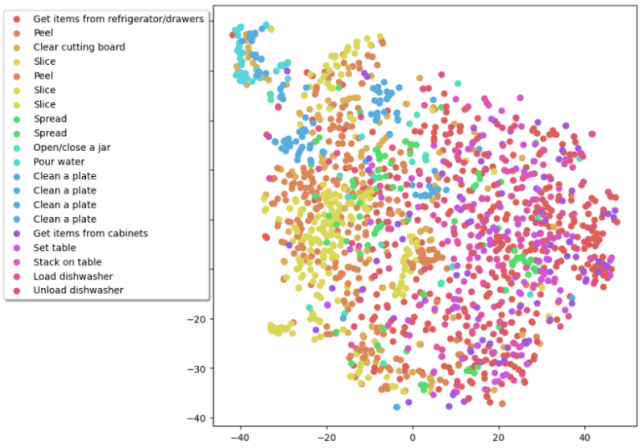

Figure 9. Reconstructed EMG features

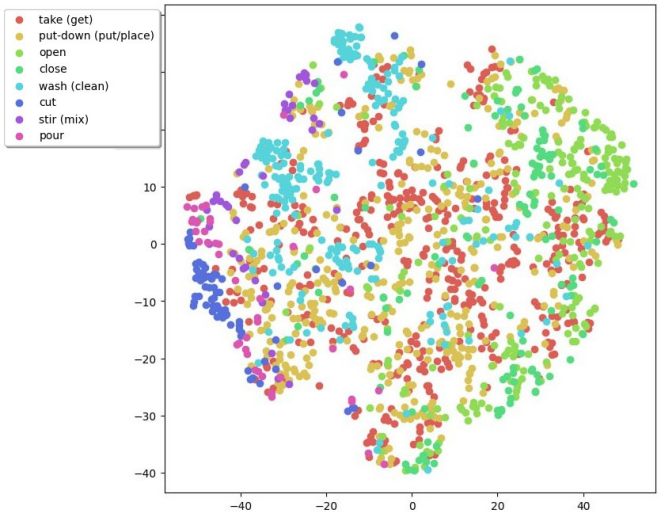

Figure 10. Synthetic EMG data for EK

[4] G. Goletto, M. Cannici, E. Gusso, M. Matteucci, B. Caputo, C. Plizzari, and M. Planamente. E(GO)^{2}MOTION: Motion augmented event stream for egocentric action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 2

[5] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 2

[6] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. 4

[7] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[8] Adrian Spurr et al. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 2

[9] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 3

[10] H. Yu and J. Oh. Anytime 3d object reconstruction using multi-modal variational autoencoder. *IEEE Robotics and Automation Letters*, 7(2):2162–2169, 2022. 2

[11] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 831–846, Cham, 2018. Springer International Publishing. 3