SEMA: Semantic Attention for Capturing Long-Range Dependencies in Egocentric Lifelogs

Pravin Nagar[1] K.N Ajay Shastry[2*] Jayesh Chaudhari[2*] Chetan Arora[2]
[1]University of Maryland, College Park
[2]Indian Institute of Technology, Delhi

Abstract

*Transformer architecture is a de-facto standard for modeling global dependency in long sequences. However, quadratic space and time complexity for self-attention prohibits transformers from scaling to extremely long sequences ($> 10k$). Low-rank decomposition as a non-negative matrix factorization (NMF) of self-attention demonstrates remarkable performance in linear space and time complexity with strong theoretical guarantees. However, our analysis reveals that NMF-based works struggle to capture the rich spatio-temporal visual cues scattered across the long sequences resulting from egocentric lifelogs. To capture such cues, we propose a novel attention mechanism named **SEM**antic **A**ttention (SEMA), which factorizes the self-attention matrix into a semantically meaningful subspace. We demonstrate SEMA in a representation learning setting, aiming to recover activity patterns in extremely long (weeks-long) egocentric lifelogs using a novel self-supervised training pipeline. Compared to the current state-of-the-art, we report significant improvement in terms of (NMI, AMI, and F-Score) for EgoRoutine, UTE, and Epic Kitchens datasets. Furthermore, to underscore the efficacy of SEMA, we extend its application to conventional video tasks such as online action detection, video recognition, and action localization. Code is available at https://github.com/Pravin74/Semantic_attention/*

1. Introduction

Recently deep neural network models based on self-attention (referred to as transformers) [60] have shown their superiority over convolutional architecture in a variety of tasks [17, 22, 23, 39]. Motivated by this, we explore the use of transformer architecture for the task of activity clustering in extremely long egocentric videos to discover the activity patterns of the wearer. The two critical challenges while solving the mentioned problem are: (a) extremely long sequences generated over multiple days, and (b) unavailability of annotated data due to enhanced privacy concerns in egocentric settings and massive human effort required.

Multiple researchers have pointed out the inability of standard transformer architecture to scale for extremely long sequences [4, 9, 31]. This is primarily because the self-attention mechanism suffers quadratic compute and memory requirements with the sequence length. Further, transformer models typically need large supervised data, and the lack of supervision for extremely long sequential tasks makes it challenging for the application of transformers.

Earlier works addressing quadratic time complexity of self-attention can be categorized into: sparse attention-based [4, 8, 31, 57, 58, 69] and NMF-based approaches [9, 29, 49, 64]. Sparse attention-based works do not provide theoretical guarantees and typically use fixed locations to compute global attention, affecting generalization capability. On the other end, NMF-based approaches are generalizable and provide theoretical guarantees; therefore, we choose to explore NMF-based approaches in this work.

Linformer, [64] learns two projection matrices to approximate the self-attention matrix using a low-rank matrix in linear space and time complexity. Katharopoulos *et al*. [29] use a kernel-based formulation for NMF to approximate the regular quadratic-complexity of self-attention and use the associative property of matrix product to achieve linear space and time complexity for the autoregressive task. Similarly, Choromanski *et al*. [9] propose a theoretically bounded linear-complexity attention mechanism (called Performer) that projects the query and key vectors into a fixed orthogonal random subspace and the projections conceptualize the factorization of a full-rank attention matrix. One of the main observations of our work is that the predefined kernel or random subspace-based factorization is inadequate for attention modeling in long video sequences. The recent work cosFormer [49] has used a similar idea of NMF and proposed a cos-based reweighting mechanism with kernelization method to concentrate more weights on the neighboring tokens to achieve locality in attention. However, this assumption doesn't hold for the cur-

*equal contribution.

rent problem of representation learning for egocentric lifelogs, where events/activities are scattered across the days. The key contribution of this work is to suggest a semantically aware attention factorization by projecting on the subspace obtained from sample-specific representative frames.

We take motivation from the NMF framework and propose a novel linear complexity factorization method to approximate the self-attention suitable for representation learning for the extremely long sequence where patterns are repetitive in very long intervals. It has been shown that k-means clustering is a tractable approximation to the non-negative low-rank matrix factorization problem [15]. Motivated by this, instead of learning the low-rank factorization using a predefined kernel or random projections, we use a set of representative frame R to learn low-rank matrices Q and K such that self-attention matrix $A = QR^\top RK^\top$. These representative frames are computed apriori using video summarization methods. The use of representative frames allows us to integrate various semantic cues into the factorization process while ensuring theoretical guarantees on linear space and time complexity. Our idea is backed by the fact that for activity clustering task, the re-occurrence of a particular (activity) pattern is an important cue to understand its temporal boundaries. Rather than spreading attention weights over thousands of frames in a sequence, one can focus on a few exemplar frames. The proposed semantic attention, which factorizes self-attention through semantically relevant representative frames, is called SEMA.

Many representation learning works demonstrate significant performance gain when the clustering/class information is embedded in the representations [2, 6]. Motivated by this notion, we choose a self-supervised learning approach to train the SEMA-based embedding network to discover the activity patterns in egocentric lifelogs. We use self-supervised learning to generate pseudo labels and use these pseudo labels to learn cluster-centric representations. We train the framework like an EM algorithm by iterating the representation learning and self-labeling.

Contributions: The key contributions of our work are: **(1)** We propose a novel **SEM**antic **A**ttention (SEMA) based on the low-rank factorization of the self-attention matrix using representative frames. The proposed architecture can exploit sample-specific semantic cues to learn robust representation from extremely long but repetitive video sequences. **(2)** We propose a self-supervised clustering pipeline to discover activity patterns in extremely long egocentric lifelogs (recorded for up to 20 days). The approach does not rely on any priors or pre-trained networks to detect activities, objects, and/or places. **(3)** We demonstrate the performance of our contributions on the benchmark *Egoroutine*, *UTE*, and *Epic Kitchens* datasets. Compared to the current state-of-the-art approaches, the proposed technique achieves significant performance gain of $(8\%, 8\%, 15\%)$, $(2\%, 2\%, 4\%)$, and $(8\%, 8\%, 17\%)$ in terms of (NMI, AMI, F-Score) for *EgoRoutine*, *UTE*, and *Epic Kitchens* datasets, respectively. **(4)** To demonstrate the effectiveness of the proposed SEMA, we opt to apply it to SOTA works in three established video analysis tasks. We substitute their self--attention with semantic attention while keeping all other aspects unchanged. These three tasks encompass online action recognition, video recognition, and action localization. **(5)** We contribute annotations for *EgoRoutine* comprises 7 subjects of 104 days of lifelogging and *UTE* comprises 4 videos of 17 hours, to be released after publication.

2. Related Work

Unsupervised Activity Segmentation and Clustering for Sequential data: Recently, many works have demonstrated unsupervised action segmentation [14, 16, 25, 36, 42, 65] and clustering [33, 34, 61] on video datasets comprising small video samples (a few minutes long). A few works have been done for egocentric videos such as [32] uses a stacked Dirichlet process mixture model over motion histograms, and [18, 19] use a weakly supervised technique to model the active objects for egocentric action recognition. [5] uses CNN-LSTM based autoencoders, whereas [56] uses topic modeling to learn the activity patterns performed at different time intervals over multiple days. In a nutshell, all these works fail to model the global dependencies required for egocentric lifelogs where activities are spread across days.

Self-Supervised Learning: [46] use a large network trained on a pretext task to generate pseudo labels for the target task and then train a smaller network with these pseudo labels for transferring the knowledge. For egocentric data, we do not have such large labeled data. [2] proposed a fast variant of the Sinkhorn-Knopp algorithm to generate pseudo labels for large-scale datasets. However, the equipartition assumption used is not applicable for the problem as the distribution of activity patterns is highly skewed. Recently [70] proposed a joint framework for online clustering that jointly perform clustering and features learning to deal with unstable training.

Representation Learning for Global Dependencies: Sarfraz et al. [51, 52] proposed a temporally weighted hierarchical clustering approach that uses the 1-nearest neighbor graph to cluster the semantically consistent frames present in the video. Deep representation learning using graph autoencoder is getting attention for various NLP tasks [30, 48, 63]. However, all the GCN-based works require a pre-computed adjacency matrix that implicitly assumes a particular length of an event. This is problematic in our context due to widely variable length events.

Transformers and Scaling Attention: Transformer-based approaches show remarkable performance in sequence modeling [60] but face scalability issues with long se-

quences [12,40]. The complexity of self-attention is $\mathcal{O}(N^2)$ (where N is sequence length), becoming intractable for large N. Thus an active research area has emerged to gain compute and memory efficiency by approximating self-attention. A few notable works viz `Longformer` [4], `Reformer` [31], `Fast Transformer` [62], `Routing Attention` [50], `Long-Short Transformer` [73], `Linformer` [64], `Performer` [9], and `cosFormer` [49] claim time complexities of $\mathcal{O}(N)$, $\mathcal{O}(N \log N)$, $\mathcal{O}(NCm)$, $\mathcal{O}(N^{1.5}m)$, $\mathcal{O}(Nr)$, $\mathcal{O}(Nr^2)$, $\mathcal{O}(Nrm)$, and $\mathcal{O}(Nm^2)$ respectively, where m, C, and r are the feature dimension, the number of clusters, and the dimension of the projection matrix, respectively. The long sequence problem becomes even more critical in vision problems, where researchers have used spatial [41], temporal [1], and hierarchical [17, 21] cues to scale the attention for large image and video data.

3. Proposed Approach

The objective of this work is to recover activity patterns of one's lifelog recorded over multiple days. We formulate the problem as a representation learning for a massively long temporal sequence in an unsupervised setting. The sequence representation learning formulation is motivated by the intuition that similar activity patterns should exhibit similar structures in latent space. Note that the recent representation learning works [37,38,47] demonstrated on video datasets (comprised of enormous tiny videos of non-repetitive patterns) harnessing contrastive learning are not applicable to the problem at hand. The core technical contribution of this work is to learn an *embedding network* (f_θ^{emb}) for sequence representation learning that can handle extremely long sequences (hours/days) and model the global dependencies among similar activity patterns scattered across the sequences.

3.1. Overview

Consider the photo-stream lifelog of a subject recorder over D days. We concatenate these sequences in time, $\mathbf{X} = \{X_d\}_{d=1}^{D}$, to create a single sequence per subject spanning across days. The concatenation is required to discover and link the activities happening even only once a day. Let the number of frames in $\mathbf{X}$ be denoted by N. For frame-level feature extraction, we use a BiLSTM model suggested in [20] for the *Egoroutine* dataset and 3D CNN model [59] (called *C3D* hereon), trained on the *Sports-1M* dataset for the *UTE* and the *Epic Kitchens* datasets. Then we use Principal Component Analysis (PCA) to reduce the feature dimension and generate a 512-dimensional vector for each frame. The vector for the i^{th} frame in the sequence is denoted as $\mathbf{x}_i$. Our objective is to find c activity patterns/clusters from the week-long sequence of a subject. There is no assumption on order among a pair of activities, nor are all activities necessarily performed each day. Figure 1 and Figure 2 show an overview of our pipeline, and proposed `SEMA` respectively.

3.2. Semantic Factorization of Self Attention

Self-attention in transformers: To draw global dependencies between the input sequence, we take inspiration from the transformer network [60] and borrow the self-attention mechanism in our *embedding network* (f_θ^{emb}) (see Figure 1) which generates an embedding vector for each frame in the sequence. Once the input sequence $\mathbf{X}$ of length N is linearly projected as query $\mathbf{Q} = \{\mathbf{q}_i \mid \mathbf{q}_i \in \mathbb{R}^m, i \in [N]\}$, key $\mathbf{K} = \{\mathbf{k}_i \mid \mathbf{k}_i \in \mathbb{R}^m, i \in [N]\}$, and value $\mathbf{V} = \{\mathbf{v}_i \mid \mathbf{v}_i \in \mathbb{R}^m, i \in [N]\}$, where m is the query, key, and value dimensions, then the self-attention mechanism is given as:

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{A}_{N\times N}\mathbf{V}_{N\times m}. \tag{1}$$

Here $\mathbf{A}_{N\times N} = \text{softmax}\left(\frac{\mathbf{Q}_{N\times m}\mathbf{K}^T_{N\times m}}{\sqrt{m}}\right)$ is the *attention matrix*. The vanilla self-attention has $\mathcal{O}(N^2)$ space and time complexity and does not scale to long sequences.

Why Factorization of Self-attention Matrix? The quadratic time complexity of the self-attention matrix should be addressed effectively to model the global dependencies in long sequential data. Our experiments also confirm that the self-attention mechanism [60] fails miserably for long sequences and gives memory error beyond a sequence length of $14k$. Active research aims to make self-attention efficient by approximating it with heuristics [4, 31, 50], like Beltagy *et al.* [4] (`Longformer`) proposed a sparse attention mechanism that uses two types of attention- local attention for contextual representation and global attention for disseminating information across the full sequence. Kitaev *et al.* [31] (`Reformer`) have proposed a locality-sensitive hashing under the assumption that the nearby vectors assign the same hash value with high probability. In contrast, for lifelogs, we focus on linking similar activity patterns scattered across the extremely long sequence. A fundamental approach to addressing this issue without relying on any heuristics and prior information is by factorizing the attention matrix into the low-rank query and key matrix pairs and changing the order of matrix multiplication $\mathbf{Q}(\mathbf{K}^T\mathbf{V})$ for achieving linear space and time complexities [9, 54]. `Performer` [9] does the same by projecting the query-key pair onto a random subspace [9]. Our experiments reveal that a simple factorization shows moderate performance gain but is inadequate to capture repetitious visual information in extremely long egocentric videos. Hence, we propose a novel semantic factorization based on representative frames to harness the latent characteristics of the data for factorizing the attention matrix.

Semantic Factorization of Self-attention: To overcome the quadratic complexity of self-attention, we formulate the

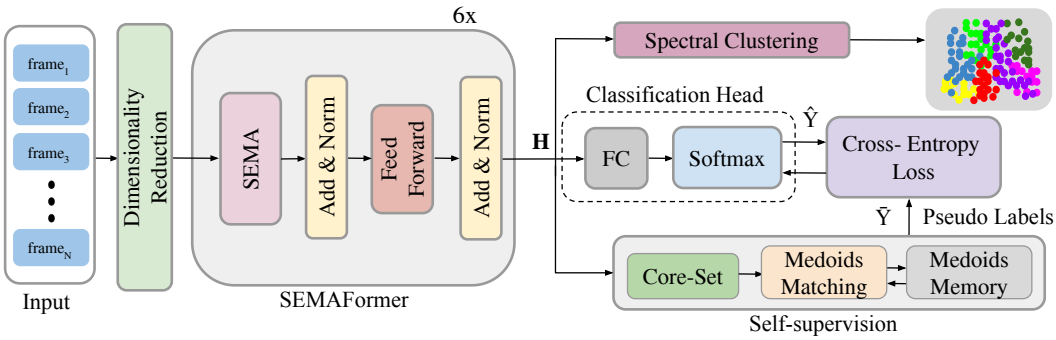

Figure 1. Illustration of the flow chart of the proposed approach. Our technique consists of a neural network f_θ parameterized by θ that is divided into two parts. The first part is an *embedding network* (`SEMAFormer`), $f_\theta^{\text{emb}} : \mathbb{R}^m \rightarrow \mathbb{R}^m$, that generates an embedding vector $\mathbf{H} \in \mathbb{R}^{N\times m}$. The second part is a *classification head*, $f_\theta^{\text{cls}} : \mathbb{R}^m \rightarrow \mathbb{R}^c$, consisting of a linear layer followed by the softmax operator, which generates the predicted labels $\hat{\mathbf{Y}} \in \mathbb{R}^{N\times c}$ corresponding to the input sequence of length N. We train the network using the pseudo labels $\tilde{\mathbf{Y}} \in \mathbb{R}^{N\times c}$ generated using the proposed self-supervised learning framework. Once the network is trained, we perform spectral clustering [45], with the number of clusters c, using the affinity matrix generated by the representations given by the *embedding network*.

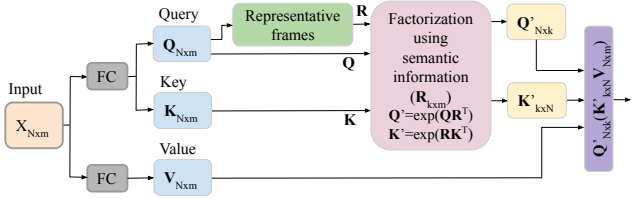

Figure 2. The figure depicts the proposed `SEMA` that factorizes the self-attention using semantically meaningful subspace by utilizing the latent characteristics of the data. It samples the representative frames from the query $\mathbf{Q}$ using apriori computed representative indices in contrast to random vectors used in the `Performer`. The resulting projections $\mathbf{Q}'$ and $\mathbf{K}'$ are low-rank decomposition of the self-attention matrix using the saliency of the data.

low-rank decomposition of attention matrix $\mathbf{A}$ as a non-negative matrix factorization (`NMF`) problem. K-means can approximate NMF, a tractable approach to non-negative low-rank matrix factorization [15]. Precisely, we factorize a full-rank attention matrix $\mathbf{A}$ to the low-rank matrices: *membership* matrix, $\mathbf{K}'$, and *reconstruction* matrix, $\mathbf{Q}'$, such that: $\mathbf{A} = \mathbf{Q}'\mathbf{K}'$. We first compute k representative frames using a separate video summarization technique. We use these representative frame indices to sample representative frames from the $\mathbf{Q}$, and stack them into a $k \times m$ matrix (refer to the Figure 2). Then we learn a $k \times N$ matrix, $\mathbf{K}'$, such that $\exp(\mathbf{R}\mathbf{K}^\top)$ can be interpreted as the distance or membership coefficient of each sample from/of each of the k clusters (represented by the corresponding representative frame). Here $\exp(\cdot)$ is applied element-wise.

We interpret multiplication with $\mathbf{Q}'$, i.e., $\mathbf{Q}'\mathbf{R}\mathbf{K}^\top$, as reconstructing a sample as the weighted sum of cluster centroids. Since conceptually we expect the reconstruction weights to be the same as the cluster membership coefficients, $\exp(\mathbf{R}\mathbf{K}^\top)$, hence we enforce $\mathbf{Q}' = \mathbf{K}'$.

Mathematical Formulation of Semantic Factorization: It is instructive to note that while our proposed factorization provides rich conceptual motivation, mathematically, we are simply factorizing $\mathbf{A} = \mathbf{Q}'\mathbf{K}'$, such that $\mathbf{Q}' = \exp(\mathbf{Q}\mathbf{R}^\top)$, and $\mathbf{K}' = \exp(\mathbf{R}\mathbf{K})$. Here, $\mathbf{R}$ is a matrix formed by stacking a set of k feature vectors corresponding to the representative frames that are chosen using video summarization. Mathematically, this is no different from `Performer`, in which the vectors are chosen as random vectors orthogonal to each other. Hence, the space and time complexity remains the same as the Performer, i.e., $\mathcal{O}(Nk + Nm + km)$ and $\mathcal{O}(Nkm)$, respectively, in addition to one-time representative frame computation. Furthermore, formal guarantees similar to the `Performer` factorization hold for the proposed `SEMA` as well (detailed proof in the supplementary).

Finding Set of Representative Frames: Whereas the `Performer` uses random projection vectors to learn $\mathbf{Q}'$, and $\mathbf{K}'$, we enforce that $\mathbf{Q}' = \exp(\mathbf{Q}\mathbf{R}^\top)$, and $\mathbf{K}' = \exp(\mathbf{R}\mathbf{K}^\top)$, where $\mathbf{R}$ is a matrix of features of representative frames. This ensures that the factorization proceeds by first projecting to meaningful cluster centers and then reconstructing based on these projections. Our proposed pipeline allows to choose the representative frames independent of the steps for activity clustering. In our implementation, we use [43], which is a recent technique, especially for summarizing egocentric videos. This is a one-time prior computation with the number of representative frames set to 256 $(m/2)$ for all experiments. The performance may vary by changing the number of representative frames.

3.3. Activity Clustering using Self-Supervised Learning

Overview: Our `SEMAFormer` uses the proposed semantic attention-based factorization in a transformer architecture.

Our objective is to use SEMAFormer to learn embeddings for each frame in an input video which can model long-range repetitions and give similar feature embeddings for such frames. The embeddings can then be clustered to give activities. In a supervised setting, we could have trained f_θ using one of the c ground truth activity labels $y_1, ...y_N \in \{1, ..., c\}$ given for each frame. With $\widehat{\mathbf{y}}_i$ as the predicted class probability vector for a sample $\mathbf{x}_i$, the model is trained using the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c}\mathbf{y}_i[j]\,\log\,\widehat{\mathbf{y}}_i[j]. \tag{2}$$

Here $\mathbf{y}$ is the one-hot vector corresponding to label y_i. In our settings, long sequences and the privacy-sensitive nature of egocentric data prohibit the availability of the ground truth label. Hence, we adopt a self-supervised approach where we first cluster the samples into c cluster based on the learned embeddings from SEMAFormer and then use the cluster membership to generate pseudo-labels $\widetilde{\mathbf{y}}_i$ for each sample. We then train the embedding network using cross-entropy loss with respect to the pseudo-labels:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c}\widetilde{\mathbf{y}}_i[j]\,\log\,\widehat{\mathbf{y}}_i[j], \tag{3}$$

Generating Pseudo-labels: For clustering, we use the core-set algorithm [53] to generate c-medoid indices using the latest embeddings generated from SEMAFormer for each sample. The core-set algorithm is an efficient approximation of the k-center problem [53]. We give the details of the pseudolabel generation process in the supplementary.

Self-supervised Representation Learning: The proposed architecture is trained similarly to Expectation-Maximization (EM). The two steps, namely representation learning and self-labeling are as follows: (1) Freeze the current label assignment matrix $\widetilde{\mathbf{y}}$, and update the model f_θ by minimizing the Equation 3. (2) Freeze the current embedding ($\mathbf{H}$), and compute the new c medoids and pseudo labels. We run the EM for 600 iterations or until convergence, whichever happens first. After this, we use spectral clustering [44] on the most recent embeddings for each frame to detect activities. We chose spectral clustering because it is rooted in graph theory and can build global dependencies in a long sequence. However, core-set or other clustering techniques could have been equivalently used.

4. Datasets and Evaluation Methodology

Dataset: We showcase results on a publicly available *EgoRoutine* dataset [56], comprising lifelogging of seven subjects for a total of 104 days. The dataset is captured by a wearable camera fixed on the chest of a subject, capturing at 2 fpm, constituting $115,685$ captured frames in total. Compared to conventional egocentric datasets, this dataset is recorded in a highly unconstrained environment that includes a variety of indoor and outdoor scene contexts. The activities are shopping, visiting restaurants/museums/concerts, traveling on flight/bus/cab/metro, working in a lab, attending conferences, cycling, sitting at the beach, etc. The dataset does not provide activity annotations. However, we have annotated all seven subjects for our experiments. Due to the scarcity of week-long lifelogging datasets, we chose the UTE [35] dataset to demonstrate the generalization and efficacy of the proposed framework on a video dataset. The UTE dataset is not a week-long lifelogging dataset but comprises recordings of subjects performing daily activity tasks ranging up to 5 hours. Though not day-long, it still suffices to understand the key strength of the proposed model for understanding long-range repetitions. We also annotated all four videos of this dataset. Furthermore, we also synthesize a long video sequence (approx. 20k frames) using the *Epic Kitchens* dataset [11] (refer to supplementary material for details). We further demonstrate SEMA on three standard video analysis tasks. We use the *THUMOS14* [28] and *AVA* [24] datasets for online action detection and video recognition, respectively. For action localization, we employ the *ActivityNet 1.3* [27] and *THUMOS14* [28] datasets.

Evaluation and Annotations: For evaluation, we use the commonly used clustering evaluation metrics: Adjusted Mutual Information (AMI), Normalized Mutual Information (NMI), and F-score [55, 68]. These metrics range in $[0, 1]$, where larger values indicate better performance (refer to the supplementary material for more details).

Baselines: We compare with a SOTA egocentric work [13] to demonstrate the efficacy of SEMA. Dimiccoli *et al*. [13] use a threshold to control the granularity of segmentation. We tweak the threshold to generate the appropriate clusters for each subject. Due to the scarcity of recent works for activity pattern recovery, we select five Vision/NLP works aligned to our problem [2, 3, 7, 48, 51]. Part *et al*. [48] propose a novel convolutional graph autoencoder called GALA (Graph convolutional Autoencoder using LAplacian smoothing and sharpening) for representation learning. we generate a sparse adjacency matrix by considering τ closest frames for an input frame in Euclidean space under the assumption that the events are of equal length. We choose $\tau = 30$ to demonstrate the results. Similarly, Bai *et al*. [3] propose Deep Autoencoding Predictive Components (DAPC) that mask the feature dimension and temporal dimension of the input sequence and reconstruct the masked component from the latent representations. The Transformer encoder shows memory error in our case due to long sequences. Hence, we show results on bi-GRU [10] configuration of DAPC. Sar-

Methods	c = 12			c = 13			c=15		
	F1↑	AMI↑	NMI↑	F1↑	AMI↑	NMI↑	F1↑	AMI↑	NMI↑
SR-clustering [13]	0.3044	0.0913	0.0924	0.2697	0.1294	0.1312	0.2614	0.1537	0.1557
TW-FINCH [51]	0.3132	0.1548	0.1603	0.3259	0.1649	0.1655	0.3072	0.1530	0.1545
SeLa [2]	0.6642	0.6291	0.6299	0.6662	0.6150	0.6158	0.5855	0.5954	0.5963
DAPC + bi-GRU [3]	0.7135	0.6129	0.6135	0.6152	0.6040	0.6048	0.6343	0.6080	0.6089
GALA [48]	0.6357	0.6079	0.6085	0.6458	0.6084	0.6093	0.5381	0.5932	0.5941
CARL [7]	0.5551	0.5219	0.5253	0.5847	0.5258	0.5262	0.5721	0.5139	0.5144
Ours+naive⋆ [60]	0.2262	0.1651	0.1674	0.2257	0.1749	0.1769	0.2292	0.1423	0.1451
Ours+Long [4]	0.5576	0.5989	0.5995	0.6212	0.6066	0.6073	0.6575	0.5982	0.5990
Ours+Perf [9]	0.6955	0.6219	0.6224	0.6001	0.5938	0.5944	0.6842	0.5996	0.6006
Ours+SeLa [2]	0.6478	0.5991	0.6025	0.6573.	0.6152	0.6160	0.7185	0.6276	0.6286
Ours+cos [49]	0.7233	0.6299	0.6305	0.6965	0.6328	0.6335	0.5739	0.5875	0.5885
Ours+SEMA	**0.7482**	**0.6510**	**0.6515**	**0.7976**	**0.6837**	**0.6842**	**0.7960**	**0.6806**	**0.6814**

Table 1. Comparison between various SOTA approaches for subject S1 in *EgoRoutine* dataset. For $c = 13$, we merge 'in cab' and 'in metro' to 'transportation' class and 'in lab kitchen' to 'walking in lab and chitchatting' class in the ground truth annotations. For $c = 12$, we further merge the 'food in lab' to 'at restaurant' class. $\star$ represents that the self-attention gives memory error after 14000 sequence length, the results are evaluated for less than 14000 sequence length.

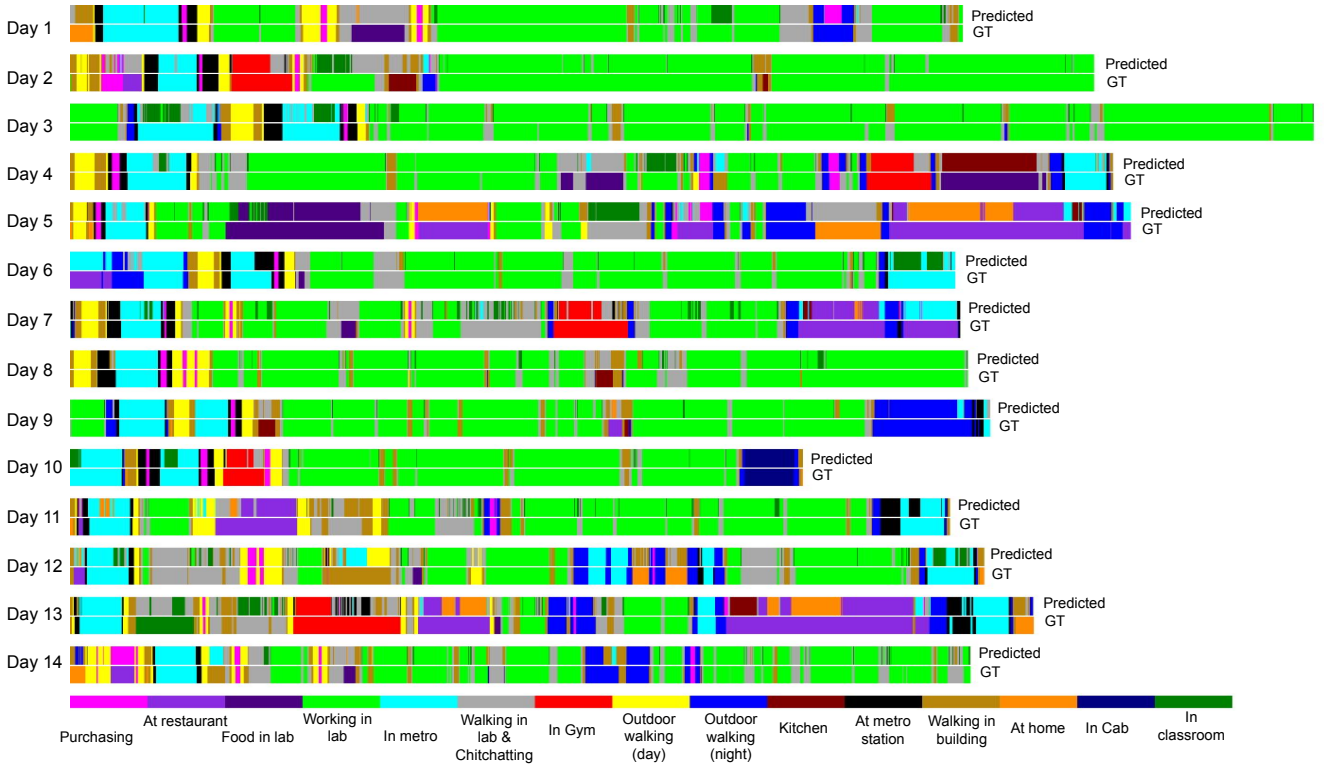

Figure 3. The figure demonstrates the visualization of a comparison between the predicted class and ground truth for different days (for better visualization, we have divided the concatenated sequence into multiple days). We use Hungarian matching for a one-to-one mapping between ground truth and predicted clusters. Figure best visible in color.

fraz *et al*. [51] proposed a temporally-weighted hierarchical clustering (TW-FINCH) algorithm that groups semantically related frames of a video using a 1-nearest neighbor graph. The algorithm partitioned the data at multiple granularities. We picked the partition closest to ground-truth clusters for comparison. Chen *et al*. [7] proposed a contrastive action representation learning (CARL) framework that uses a novel sequence contrastive loss. We trained the architecture on our datasets and used spectral clustering on the frame-wise representations generated. Furthermore, to prove the efficacy of the proposed SEMA, we replace it with four SOTA attention mechanisms, namely self-

attention, Longformer, Performer, and cosFormer in the proposed pipeline, and call them as Ours+naive (i.e. Transformer), Ours+Long, Ours+Perf, and Ours+cos, respectively. We also integrate the SOTA self-supervised method name, SeLa [2] in the proposed pipeline and replace the fully connected layer with the proposed SEMAFormer and named it Ours+SeLa.

Implementation Details: The SEMAFormer comprises six transformer layers, wherein the conventional self-attention has been replaced by the proposed SEMA mechanism. We use Principal Component Analysis for dimensionality reduction for all the experiments, which resulted in a 512-dimensional feature vector. We utilize $m/2$ frames to compute the representative loss at each layer and set c to the number of activity patterns in the input sequence (from GT). For medoids matching, we use bipartite matching between the previously generated medoids (extract the current embedding corresponding to the previously generated indices stored in medoids memory) and current medoids in Euclidean space. We generate pseudo labels for every 50^{th} epoch. We set the learning rate as 0.01, the number of neurons at the feedforward network as 2048, and the adam optimizer with a 40 epoch of warmup [60]. We use $f = \text{ReLU}$ for better generalization, similar to Performer. We remove the positional encoding as the sequence of the events is stochastic for the problem at hand.

5. Experiments & Results

Id	Score	SeLa [2]	DAPC [3]	GALA [48]	CARL [7]	Ours+ Perf	Ours+ cos	Ours+ SEMA
P01	AMI	0.5036	0.219	0.5068	0.4024	0.4999	0.5108	**0.5116**
	NMI	0.5056	0.223	0.5089	0.4080	0.5019	0.5128	**0.5136**
	F1	0.5517	0.1239	0.5557	0.4415	0.5325	0.5522	**0.5562**
P02	AMI	0.5449	0.2445	0.5432	0.4501	0.5413	0.5601	**0.5603**
	NMI	0.5455	0.2481	0.5438	0.4529	0.5419	0.5607	**0.5608**
	F1	0.519	0.2179	0.5813	0.4821	0.5716	0.5714	**0.5870**
P03	AMI	0.4149	0.1876	0.4051	0.2534	0.4261	**0.4426**	0.4400
	NMI	0.4166	0.1892	0.407	0.2589	0.4277	0.4441	**0.4461**
	F1	0.5503	0.2814	0.6093	0.2816	0.5923	0.6186	**0.6198**
P04	AMI	0.3038	0.1123	0.4253	0.2981	0.3632	0.4164	**0.4339**
	NMI	0.3052	0.1156	0.4264	0.2962	0.3644	0.4176	**0.4351**
	F1	0.3328	0.3581	0.5559	0.3287	0.4488	0.5311	**0.6870**

Table 2. Performance comparison with SOTA in terms of F1 score, AMI, and NMI for all the subjects of the *UTE* dataset.

Quantitative Comparison for Different Number of Clusters: Table 1 shows the quantitative evaluation based on AMI, NMI, and F-score for different granularities of clusters. We demonstrate that Ours+SEMA outperforms all the SOTA frameworks with a huge margin for 14 days long sequence of subject S1. When we replace the SEMA with SOTA attention mechanisms, the performance drops considerably as the SOTA mechanism fails to harness the rich semantic information. Furthermore, Ours+SeLa use the equipartition assumption for generating the pseudo labels hence underperforms compared to Ours+SEMA as the equipartition assumption used in [2] does not hold for the highly skewed activity patterns in egocentric lifelogs.

Score	SeLa [2]	DAPC [3]	GALA [48]	CARL [7]	Ours+ Perf	Ours+ cos	Ours+ SEMA
AMI	0.3229	0.0267	0.3900	0.3158	0.3884	0.4102	**0.4710**
NMI	0.3234	0.0271	0.3904	0.3140	0.3887	0.4105	**0.4713**
F1	0.3161	0.2051	0.3154	0.2992	0.4543	0.3644	**0.4830**

Table 3. Performance comparison with SOTA in terms of F1 score, AMI, and NMI for the *Epic Kitchens* dataset.

Id	Score	TW-FINCH [51]	SeLa [2]	DAPC [3]	GALA [48]	CARL [7]	Ours+ Perf	Ours+ cos	Ours+ SEMA
S1	AMI	0.1530	0.5954	0.6080	0.5932	0.5139	0.5939	0.5875	**0.6806**
	NMI	0.1545	0.5963	0.6089	0.5941	0.5144	0.5948	0.5885	**0.6814**
	F1	0.3072	0.5855	0.6343	0.5381	0.5721	0.6423	0.5739	**0.7960**
S2	AMI	0.3489	0.4832	0.4794	**0.4901**	0.3811	0.4765	0.4829	**0.4901**
	NMI	0.3551	0.4889	0.4852	0.4932	0.3820	0.4824	0.4887	**0.4957**
	F1	0.2541	0.4497	0.4504	0.4901	0.3729	0.4395	0.4383	**0.4960**
S3	AMI	0.1038	0.4704	0.5083	0.5262	0.4334	0.4891	0.4787	**0.5756**
	NMI	0.1055	0.4717	0.5096	0.5275	0.4347	0.4905	0.4800	**0.5768**
	F1	0.2227	0.4885	0.5546	0.5965	0.5018	0.5208	0.4499	**0.7202**
S4	AMI	0.4640	0.5474	0.5518	0.5630	0.4939	0.5663	0.5704	**0.5750**
	NMI	0.4699	0.5513	0.5557	0.5668	0.4955	0.5699	0.5740	**0.5786**
	F1	0.2882	0.4200	0.4415	0.5117	0.5192	0.4575	0.4390	**0.5821**
S5	AMI	0.4722	0.5845	0.5868	0.5658	0.4932	0.5787	0.5865	**0.5913**
	NMI	0.4769	0.5870	0.5892	0.5685	0.4983	0.5812	0.5890	**0.5937**
	F1	0.3230	0.4808	0.4907	0.4707	0.4594	0.4671	0.4912	**0.6074**
S6	AMI	0.1801	0.5371	0.5078	0.5838	0.4866	0.5277	0.5839	**0.6252**
	NMI	0.1823	0.5392	0.5101	0.5857	0.4823	0.5297	0.5860	**0.6272**
	F1	0.2645	0.5453	0.4213	0.6720	0.5681	0.4928	0.6470	**0.6813**
S7	AMI	0.3057	0.5510	0.5625	0.5630	0.4034	0.5569	0.5620	**0.5833**
	NMI	0.3078	0.5553	0.5667	0.5675	0.4068	0.5612	0.5662	**0.5873**
	F1	0.3584	0.4764	0.4953	0.5093	0.3979	0.5264	0.5488	**0.5745**

Table 4. Performance comparison with SOTA in terms of F1 score, AMI, and NMI for all the subjects of the *EgoRoutine* dataset.

Qualitative Results: Figure 3 demonstrates a visualization of the results obtained for the sequence corresponding to subject S1 (all 14 days concatenated sequentially). The figure shows that SEMA performs robustly for all activity patterns. We observed that the most repetitious activity pattern, 'working in lab' is handled and significantly recovered. Furthermore, the SEMA is robust for minority classes as well and precisely recovers 'in cab' (appeared once on day 10, refer Figure 3) and 'at metro station'. However, we observe misclassifications due to high overlap among the context and the objects involved in the activity patterns. For example, 'food in lab' is frequently misclassified as 'walking in lab and chitchatting' or 'kitchen' as the former shares the common context (the lab) and the latter shares common objects (the food). Furthermore, 'walking in lab and

Model	SharedQK	F1	AMI	NMI
Ours+Perf	NA	0.6842	0.5996	0.6006
Ours+SEMA	✗	0.7235	0.6319	0.6328
Ours+SEMA	✓	**0.7960**	**0.6806**	**0.6814**

Table 5. Performance comparison the proposed framework Ours+SEMA with various desing choises for subject 'S1' for 'c' =15. SharedQK and NA represent the linear layer shared for the query and the key and not applicable, respectively.

chitchatting' shows confusion with 'walking in building' and 'working in lab' at the boundaries due to the smooth transition between the activity patterns. We also demonstrate similar visualization for the *UTE* and *Epic Kitchens* datasets in the supplementary material.

Quantitative Comparison for All Subjects: Table 4, Table 2, and Table 3 demonstrate the quantitative comparison with the SOTA techniques for *EgoRoutine*, *UTE*, and *Epic Kitchens* datasets, respectively. We show significant performance improvement in terms of F1-score, AMI, and NMI for all three datasets. We observe that the GALA [72] performs comparably to SEMA for subject $S2$ of the *EgoRoutine* dataset as it uses a sparse adjacency matrix with τ closest frames, and the choice of τ seems best for this subject. Similarly, Ours+cos also demonstrates comparable performance for the *UTE* dataset and perform marginally better for subject $P03$ in term of AMI. CARL [7] performs poor for all three egocentric datasets as it uses frame-level *ResNet-50* [26] features followed by Transformer for harnessing the local temporal context (of 240 frames).

Ablation Study: Table 5 presents a comprehensive ablation analysis that illustrates the contribution of various design choices in Ours+SEMA. Initially, we replace the SEMA with Performer attention [9]. The result indicates that SEMA outperforms Performer attention by a significant margin. Additionally, we introduce a constraint where $\mathbf{Q}'$ is set to be equal to $\mathbf{K}'$. This constraint not only aligns with the conceptual framework but also leads to significant performance enhancements compared to the setting when $\mathbf{Q}'$ and $\mathbf{K}'$ are allowed to differ. It's important to note that the bandwidth of the representative frames (set as $m/2$ for all subjects) remains a latent characteristic of the data and is influenced by the diversity of lifelogs, hence resulting performance might change. By employing the shared $\mathbf{Q}'$ and $\mathbf{K}'$ approach, the proposed attention mechanism surpasses the SOTA frameworks by a substantial margin. For a more detailed exploration, the supplementary material provides visualizations and a comparative analysis of the attention maps generated by SEMA and Performer.

Standard video tasks with SEMA: To prove the generalizability of SEMA, we demonstrate its application on three standard video analysis tasks: online action detection, video recognition, and action localization. In each of these tasks, we choose a SOTA and substitute the self-attention block with SEMA, keeping intact the rest of the architecture and hyperparameters. The results showcased in Table 6 and Table 7 illustrate that when working with 15% representative frames, SEMA delivers performance on par with that of self-attention across all the tasks. It is noteworthy, however, that SEMA exhibits significantly lower GPU memory consumption compared to self-attention, particularly when considering a batch size of 1. This characteristic indicates its potential utility in edge AI scenarios.

Task	Action Detection		Video Recognition	
Method	LSTR [67]		MeMViT [66]	
Dataset	THUMOS14 [28]		AVA [24]	
Eval. Measure	mAP% ↑	GPU mem ↓	mAP%	GPU mem
Self-Atten	69.5	1195MB	**24.5**	2205 MB
SEMA	**69.6**	**813** MB	21.65	**1986** MB

Table 6. Performance comparison of SEMA with SOTA for online action detection and video recognition tasks.

Task	Action Localization (ActionFormer [71])			
Dataset	ActivityNet 1.3 [27]		THUMOS14 [28]	
Eval. Measure	Avg. mAP ↑	GPU mem ↓	Avg. mAP	GPU mem
Self-Atten	**36.06**	1608 MB	**66.33**	2820 MB
SEMA	35.30	**1162** MB	65.65	**2316** MB

Table 7. The table demonstrates the performance comparison of SEMA for action localization task on two standard datasets.

6. Conclusion

We focus on the problem of activity pattern clustering from the week-long recordings of a subject from an egocentric camera in a completely unsupervised setting. Current transformer models could not handle such long sequences, and hence, we have introduced a novel semantic attention transformer that can exploit the redundancy present in the lifelogs for scaling to such long sequences. We propose to factorize the attention matrix into the low-rank query and key matrices using learnable and parameter-free semantic attention. Our results on the *EgoRoutine*, *UTE*, and *Epic Kitchens* datasets, demonstrate the efficacy of SEMA on the focused task. The proposed semantic attention-based factorization is a generic idea and can also be used for other video analysis requiring long-range contextual cues. We demonstrate the same on three tasks, viz. action recognition, video recognition, and action localization.

Acknowledgments

This work was supported in part by DST, Government of India, under project id T-138. Pravin is supported by Visvesvaraya Ph.D. fellowship from Government of India.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. 3

[2] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2020. 2, 5, 6, 7

[3] Junwen Bai, Weiran Wang, Yingbo Zhou, and Caiming Xiong. Representation learning for sequence data with deep autoencoding predictive components. In *ICLR*, 2021. 5, 6, 7

[4] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 1, 3, 6

[5] Bharat Lal Bhatnagar, Suriya Singh, Chetan Arora, and CV Jawahar. Unsupervised learning of deep feature representation for clustering egocentric actions. In *IJCAI*, 2017. 2

[6] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018. 2

[7] Minghao Chen, Fangyun Wei, Chong Li, and Deng Cai. Frame-wise action representations for long videos via sequence contrastive learning. In *CVPR*, 2022. 5, 6, 7, 8

[8] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. 1

[9] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020. 1, 3, 6, 8

[10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 5

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. *IJCV*, 2022. 5

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[13] Mariella Dimiccoli, Marc Bolanos, Estefania Talavera, Maedeh Aghaei, Stavri G Nikolov, and Petia Radeva. Sr-clustering: Semantic regularized clustering for egocentric photo streams segmentation. *CVIU*, 2017. 5, 6

[14] Mariella Dimiccoli and Herwig Wendt. Learning event representations for temporal segmentation of image sequences by dynamic graph embedding. *IEEE Transactions on Image Processing*, 2020. 2

[15] Chris Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SIAM*, 2005. 2, 4

[16] Zexing Du, Xue Wang, Guoqing Zhou, and Qing Wang. Fast and unsupervised action boundary detection for action segmentation. In *CVPR*, 2022. 2

[17] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 1, 3

[18] Alireza Fathi, Ali Farhadi, and James M Rehg. Understanding egocentric activities. In *ICCV*, 2011. 2

[19] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011. 2

[20] Ana Garcia del Molino, Joo-Hwee Lim, and Ah-Hwee Tan. Predicting visual context for unsupervised event segmentation in continuous photo-streams. In *ACMMM*, 2018. 3

[21] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *NIPS*, 2020. 3

[22] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, 2019. 1

[23] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *ICCV*, 2021. 1

[24] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 5, 8

[25] Chaobo He, Yulong Zheng, Xiang Fei, Hanchao Li, Zeng Hu, and Yong Tang. Boosting nonnegative matrix factorization based community detection with graph attention auto-encoder. *IEEE Transactions on Big Data*, 2021. 2

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 8

[27] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 5, 8

[28] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *CVIU*, 2017. 5, 8

[29] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, 2020. 1

[30] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016. 2

[31] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020. 1, 3

[32] Kris M Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011. 2

[33] Anna Kukleva, Hilde Kuehne, Fadime Sener, and Jurgen Gall. Unsupervised learning of action classes with continuous temporal embedding. In *CVPR*, 2019. 2

[34] Sateesh Kumar, Sanjay Haresh, Awais Ahmed, Andrey Konin, M Zeeshan Zia, and Quoc-Huy Tran. Unsupervised action segmentation by joint representation learning and online clustering. In *CVPR*, 2022. 2

[35] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 5

[36] Jun Li and Sinisa Todorovic. Action shuffle alternating learning for unsupervised action segmentation. In *CVPR*, 2021. 2

[37] Yuanze Lin, Xun Guo, and Yan Lu. Self-supervised video representation learning with meta-contrastive network. In *CVPR*, 2021. 3

[38] Weizhe Liu, Bugra Tekin, Huseyin Coskun, Vibhav Vineet, Pascal Fua, and Marc Pollefeys. Learning to align sequential actions in the wild. In *CVPR*, 2022. 3

[39] Xin Liu, Silvia L Pintea, Fatemeh Karimi Nejadasl, Olaf Booij, and Jan C van Gemert. No frame left behind: Full video action recognition. In *CVPR*, 2021. 1

[40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[41] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 3

[42] Jonghwan Mun, Minchul Shin, Gunsoo Han, Sangho Lee, Seongsu Ha, Joonseok Lee, and Eun-Sol Kim. Bassl: Boundary-aware self-supervised learning for video scene segmentation. In *ACCV*, 2022. 2

[43] Pravin Nagar, Anuj Rathore, CV Jawahar, and Chetan Arora. Generating personalized summaries of day long egocentric videos. *IEEE Transactions on PAMI*, 2021. 4

[44] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *NIPS*, 2001. 5

[45] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2002. 4

[46] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *CVPR*, 2018. 2

[47] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Probabilistic representations for video contrastive learning. In *CVPR*, 2022. 3

[48] Jiwoong Park, Minsik Lee, Hyung Jin Chang, Kyuewang Lee, and Jin Young Choi. Symmetric graph convolutional autoencoder for unsupervised graph representation learning. In *CVPR*, 2019. 2, 5, 6, 7

[49] Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. In *ICLR*, 2022. 1, 3, 6

[50] Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021. 3

[51] Saquib Sarfraz, Naila Murray, Vivek Sharma, Ali Diba, Luc Van Gool, and Rainer Stiefelhagen. Temporally-weighted hierarchical clustering for unsupervised action segmentation. In *CVPR*, 2021. 2, 5, 6, 7

[52] Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *CVPR*, 2019. 2

[53] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 5

[54] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *WACV*, 2021. 3

[55] Suriya Singh, Chetan Arora, and CV Jawahar. First person action recognition using deep learned descriptors. In *CVPR*, 2016. 5

[56] Estefania Talavera, Carolin Wuerich, Nicolai Petkov, and Petia Radeva. Topic modelling for routine discovery from egocentric photo-streams. *Pattern Recognition*, 2020. 2, 5

[57] Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. Sparse sinkhorn attention. In *ICML*, 2020. 1

[58] Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*, 2020. 1

[59] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 3

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 1, 2, 3, 6, 7

[61] Rosaura G VidalMata, Walter J Scheirer, Anna Kukleva, David Cox, and Hilde Kuehne. Joint visual-temporal embedding for unsupervised learning of actions in untrimmed sequences. In *WACV*, 2021. 2

[62] Apoorv Vyas, Angelos Katharopoulos, and François Fleuret. Fast transformers with clustered attention. *NIPS*, 2020. 3

[63] Chun Wang, Shirui Pan, Guodong Long, Xingquan Zhu, and Jing Jiang. Mgae: Marginalized graph autoencoder for graph clustering. In *CIKM*, 2017. 2

[64] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 1, 3

[65] Zhe Wang, Hao Chen, Xinyu Li, Chunhui Liu, Yuanjun Xiong, Joseph Tighe, and Charless Fowlkes. Sscap: Self-supervised co-occurrence action parsing for unsupervised temporal action segmentation. In *WACV*, 2022. 2

[66] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *CVPR*, 2022. 8

[67] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *NIPS*, 2021. 8

[68] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *CVPR*, 2016. 5

[69] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *NIPS*, 2020. 1

[70] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised representation learning. In *CVPR*, 2020. 2

[71] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *ECCV*, 2022. 8

[72] Jiaming Zhou, Kun-Yu Lin, Haoxin Li, and Wei-Shi Zheng. Graph-based high-order relation modeling for long-term action recognition. In *CVPR*, 2021. 8

[73] Chen Zhu, Wei Ping, Chaowei Xiao, Mohammad Shoeybi, Tom Goldstein, Anima Anandkumar, and Bryan Catanzaro. Long-short transformer: Efficient transformers for language and vision. *NIPS*, 2021. 3