UPC, UB & URV

Recognizing Food Places in Egocentric Photo-Streams using MACNet + Self-Attention Mechanism

Artificial Vision & Pattern Recognition

Óscar Cubeles Ollé

Master in AI

Overview

1. Introduction

Health Motivation

- Obesity/overweight linked to cancer, diabetes, cardiovascular diseases
- Automated lifestyle pattern analysis for better nutrition monitoring
- Key metrics: duration, location, social context of eating behaviors

Technical Apporach

- Wearable cameras capture egocentric photo streams automatically
- Analyze food place classification and time spent in locations
- Uses EgoFoodPlaces dataset from multiple individuals with life-logging cameras

Main Contributions

- Wearable cameras capture egocentric photo streams automatically • Analyze food place classification and time spent in locations
- Uses EgoFoodPlaces dataset from multiple individuals with life-logging cameras

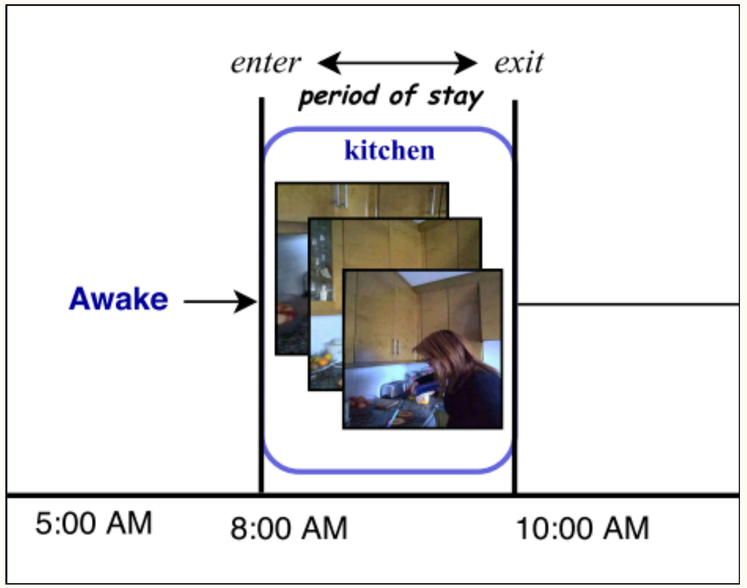


Figure showing examples of apartial daily log that shows time spent in a food place

2. Related Work

Classification Methods

- Discriminative methods: logistic regression, SVMs, boosting
- Generative models: hierarchical Bayesian systems for complex scene relationships

Neural Networks for Classification

- CNN breakthroughs:
  - AlexNet (deep architecture with ReLU activation),
  - ResNet (residual learning to prevent vanishing gradients),
  - VGG (very deep networks with small 3x3 filters)
- Key datasets: Outperforming Places2 and SUN397

Scene Recognition Challenges

- Environmental diversity for a same food place
- Complex spatial relationships using egocentric point of view
- Different camera perspectives & temporal changes.

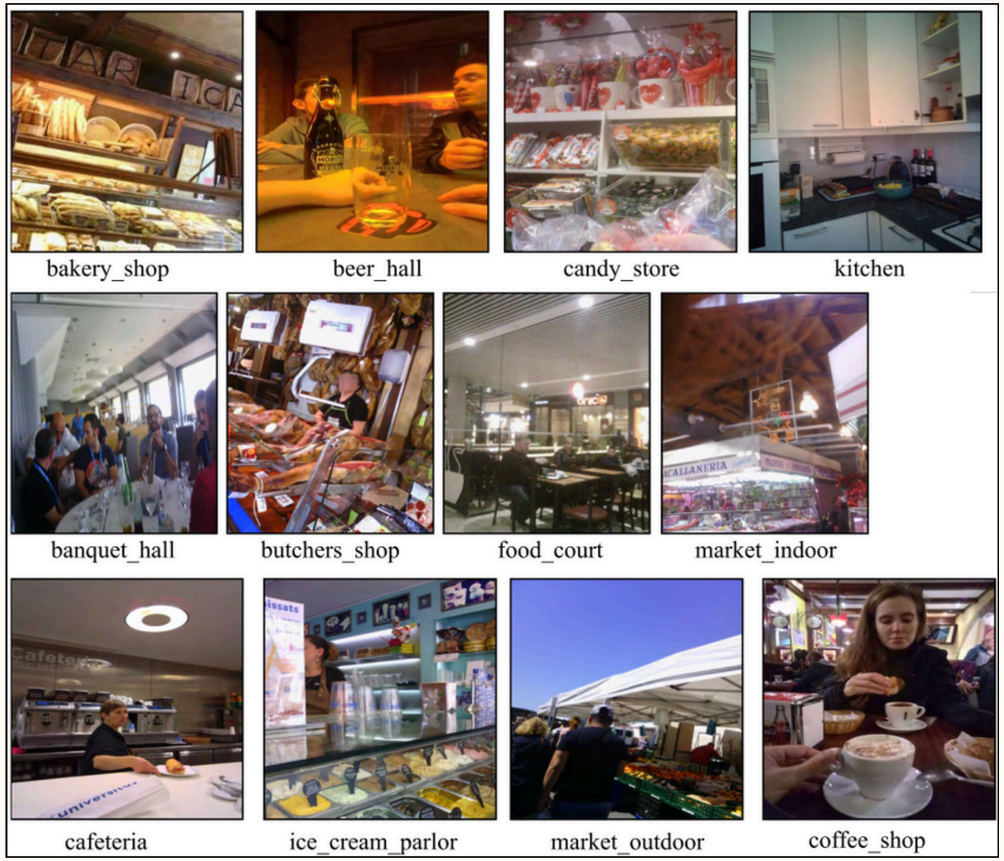


Figure showing examples of images from the EgoFoodPlaces Datase

3. Proposed Network

Attention Types

- Hard attention: looks at only part of the input
- Soft attention: weights all inputs using softmax
- Self-attention: learns what is important within the same features

Proposed Network

- **MACNet**: extracts multi-scale visual features
- **LSTM** : compute self-attention scores
- **Prediction module**: classifies the food place

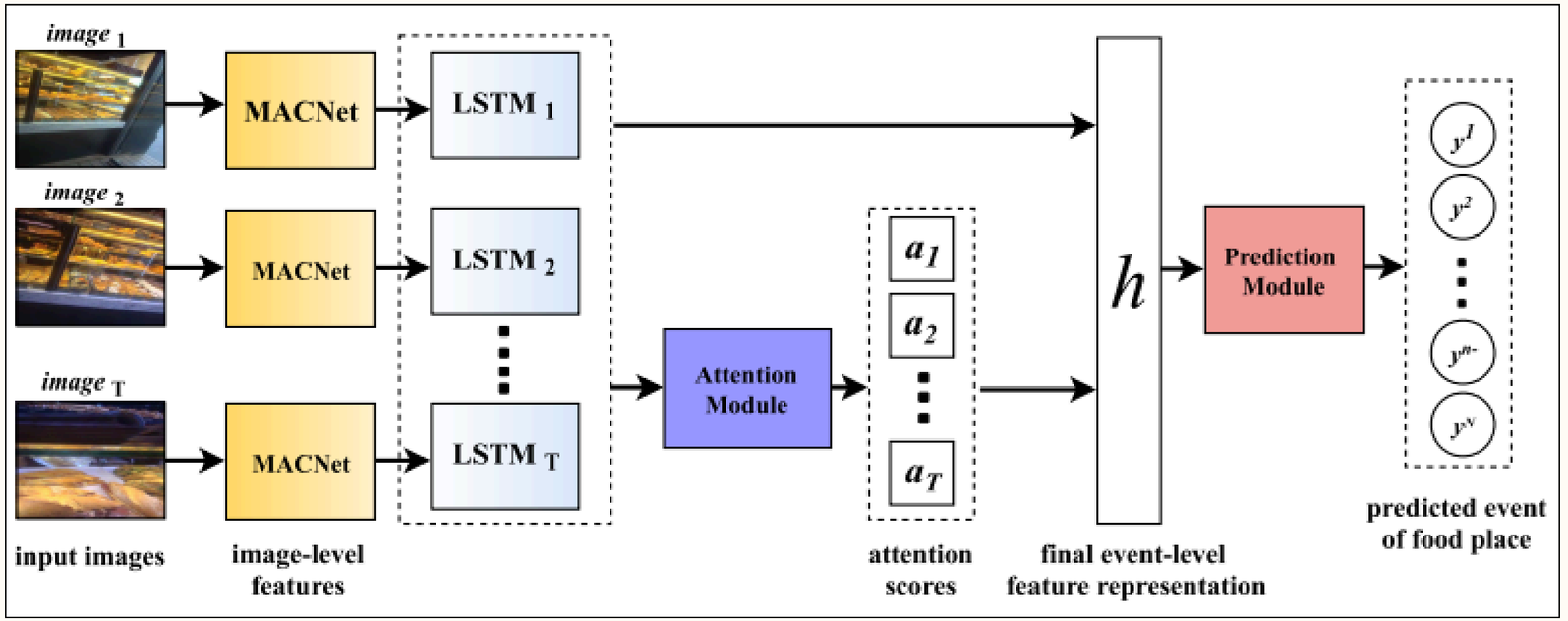


Figure of the proposed attention-based model for food places classification

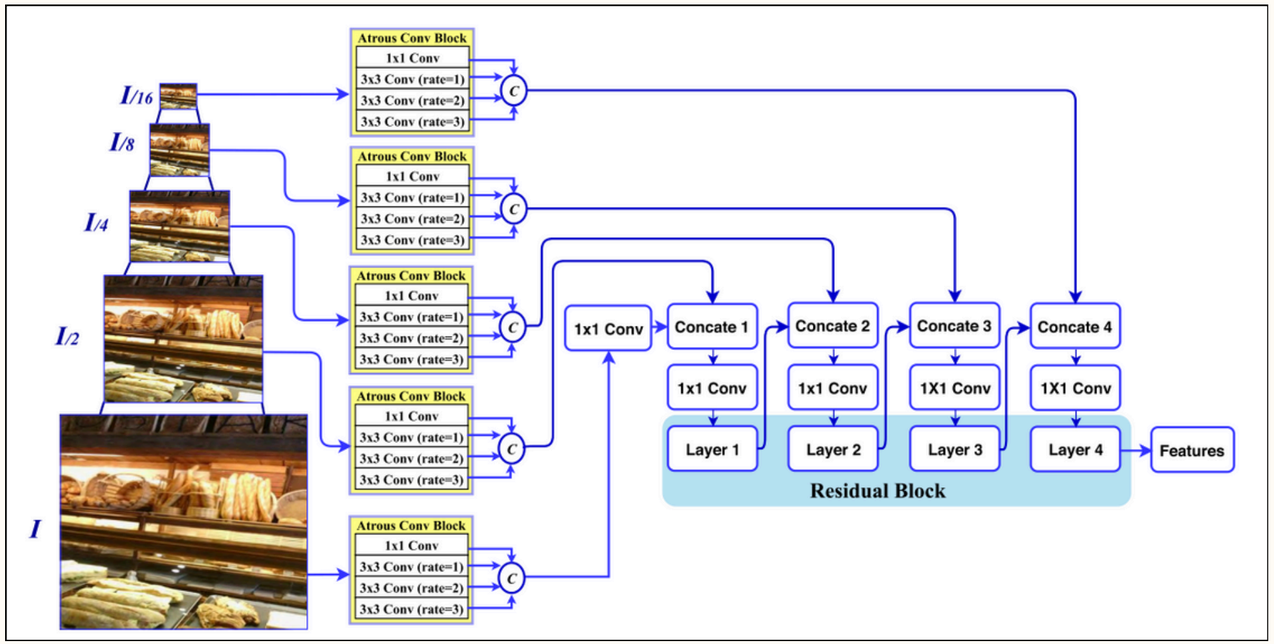
3.1 MACNet Architecture

Multi Scale Framework

- Original image size: 224 × 224 which are created 5 scaled versions each is 0.5× the previous one

Feature Extraction

- Each scaled image goes into an atrous CNN block of dilation rates 1, 2, and 3
- Then a sequential concatenation and ResNet blocks are used to create the final feature vector.



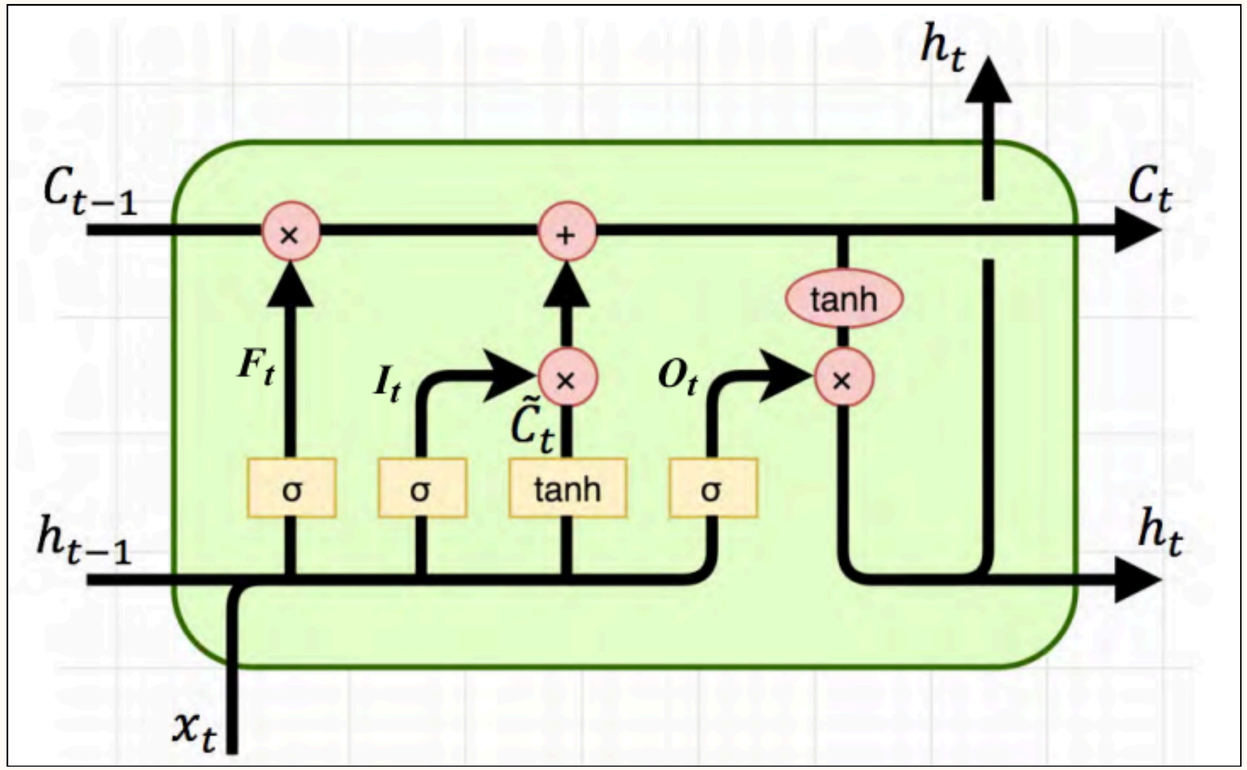
Architecture of their previous work, MACNet, for the image-level feature extraction.

3.2 Long Short-Term Memory (LSTM) Cell

Purpose: Handle sequential data while avoiding the vanishing gradient problem of standard RNNs.

Gated Structure which controls flow of information through:

- **Forget Gate:** Decides which past information to discard.
- **Input Gate:** Decides what new information to add to memory.
- **Cell State:** Acts as long-term memory, carrying relevant info across time steps.
- **Output Gate:** Determines what information to send to the next layer.



3.3 Attention Module

Input features:

- Extracted from MACNet for each image: x_0, x_1, ...xt

LSTM processing:

- Sequentially feed features into LSTMs: $LSTM_1$, $LSTM_2$ ..., LSTMt
- Each output: LSTMt caputres contextual dependencies across images

Attention module:

- Each LSTM output (LSTMt) scored with global attention vector V
- Score for t-th image: $S_t =< V, LSTM_t > .$
- Normalize scores via softmax → attention weights:

$$\alpha_t = \frac{exp(S_t)}{\sum_{t=1}^{T} exp(S_t)},$$

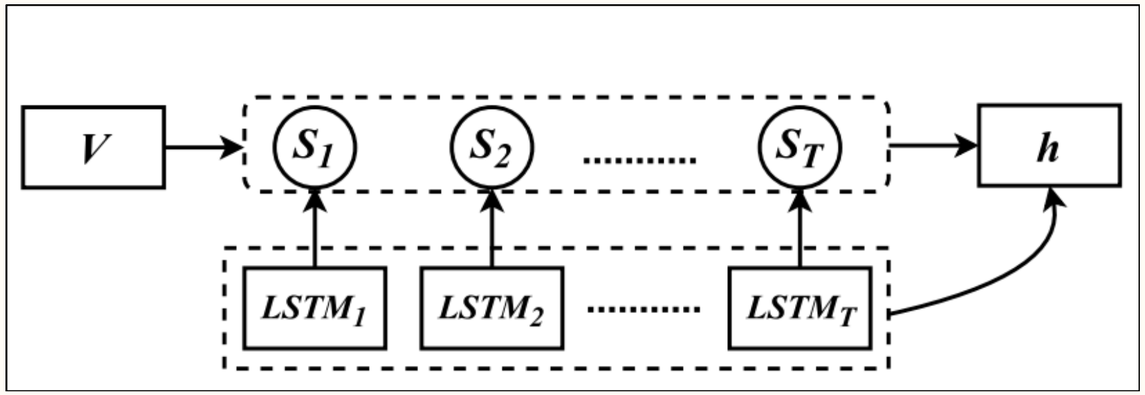
Weighted aggregation:

- Compute event-level vector
- Higher attention → more contribution to h

$$h = \sum_{t=1}^{T} \alpha_t LSTM_t,$$

Final representation:

- h summarizes the sequence: captures both sequential context and image importance
- Used to train the prediction module for event classification



Global self-attention mechanism for final event-level feature representation

3.4 Prediction Module

Prediction Module

- Fully Connected Neural Network as prediction Module for a multi-label event prediction module
- Where:
  - $\hat{y}_n$: Predicted Label
  - y_n: Ground Truth
  - w_n and b_n are the classification weight and bias parameters

Loss Function

- Where:
  - E is the cross-entropy function

$$\hat{y}^n = p(y^n|h) = \frac{1}{1 + e^{-}(w^n h + b^n)} \in [0, 1],$$

Prediction Label Function

$$\ell = -\frac{1}{N}\sum_{n=1}^{N} E(y^n, \hat{y^n}),$$

Label Classification Loss Function

4.1 Experimental Results- Datset

Dataset Description

- **EgoFoodPlaces Dataset**: Data was collected using a lifelogging camera worn by **16 users** on theri chest
- Instead of still images, each class is composed of events
- **Challenges**: Occlusion, Blurriness, Dark images

Preprocessing

- Variance of Laplacian used to **remove blurry images**, those with variance below 500 were considered blurry
- **K-Means clustering** with K=3 was applied. **Uninformative images** = one cluster with more than 90% of the pixels

Class Distribution & Split

- The dataset is composed of **22 unbalanced classes**, reflecting real-life visiting frequencies of locations.
- Split in **70% training, 20% testing & 10% validation**
- **An event-based split** was used to ensure that all frames from the same event belong to the same subset.

	Train		Val		Test		Total	
Classes	images	events	images	events	images	events	images	events
bakery shop	356	36	108	11	128	13	592	60
banquet hall	420	42	150	15	146	15	716	72
bar	1320	132	410	41	730	73	2460	246
beer hall	600	60	110	11	344	35	1054	106
butchers shop	261	27	60	6	50	5	371	38
cafeteria	1443	145	200	20	370	37	2013	202
candy store	360	36	80	8	90	9	530	53
coffee shop	2060	206	260	26	590	59	2910	291
delicatessen	680	68	80	8	50	5	810	81
dining room	3020	302	420	42	930	93	4370	437
fastfood restaurant	920	92	150	15	330	33	1400	140
food court	200	20	90	9	40	4	330	33
ice cream parlor	160	16	50	5	60	6	270	27
kitchen	3300	330	400	40	990	99	4690	469
market indoor	800	80	150	15	210	21	1160	116
market outdoor	1313	132	60	6	250	25	1623	163
picnic area	667	67	140	14	260	26	1067	107
pizzeria	1120	112	370	37	600	60	2090	209
pub indoor	372	38	60	6	150	15	582	59
restaurant	4551	456	550	55	1120	112	6222	623
supermarket	3812	382	862	87	1423	143	6097	612
sushi bar	1270	127	340	34	426	43	2036	204
Total	29005	2909	5100	511	9287	932	**43392**	**4352**

The distribution of images per class in the EgoFoodPlaces dataset.

4.3 Experimental Results - Evaluation

Benchmark Evaluation

The performance of MACNet+SA was evaluated and compared with **VGG16, ResNet50, InceptionV3**, and **MACNet** without self-attention (SA) using **F1 score, Top-1,** and **Top-5 accuracy.**

MACNet+SA outperformed other models with the highest average F1 scores, with 0.86 on the validation set and 0.80 on the test set

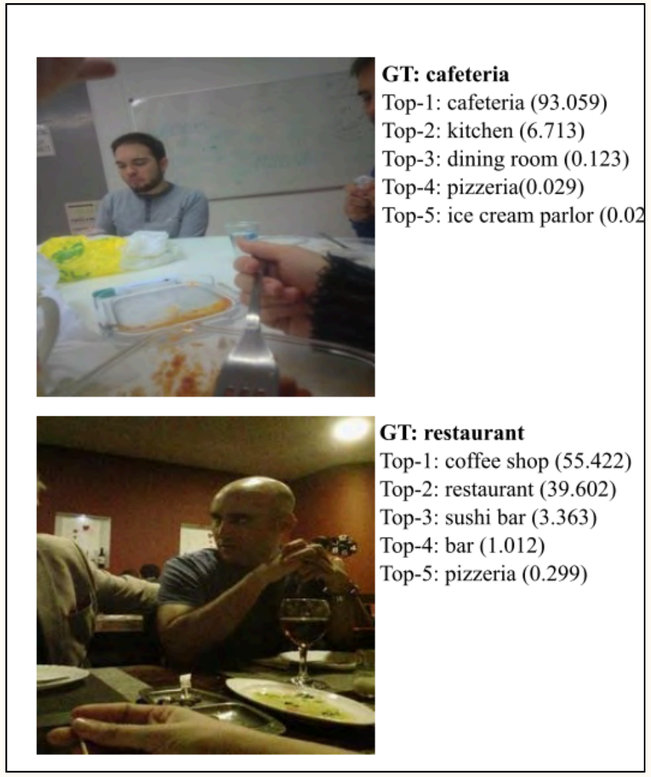
For some classes, such as butcher's shop, dining room, and indoor/outdoor markets, **MACNet without SA** achieved **better results**, likely because these scenes can be **effectively recognized from single images** rather than requiring sequential information.

In terms of **Top-1 and Top-5 accuracy**, **MACNet+SA achieved the best results** on both validation and test sets, followed by MACNet.

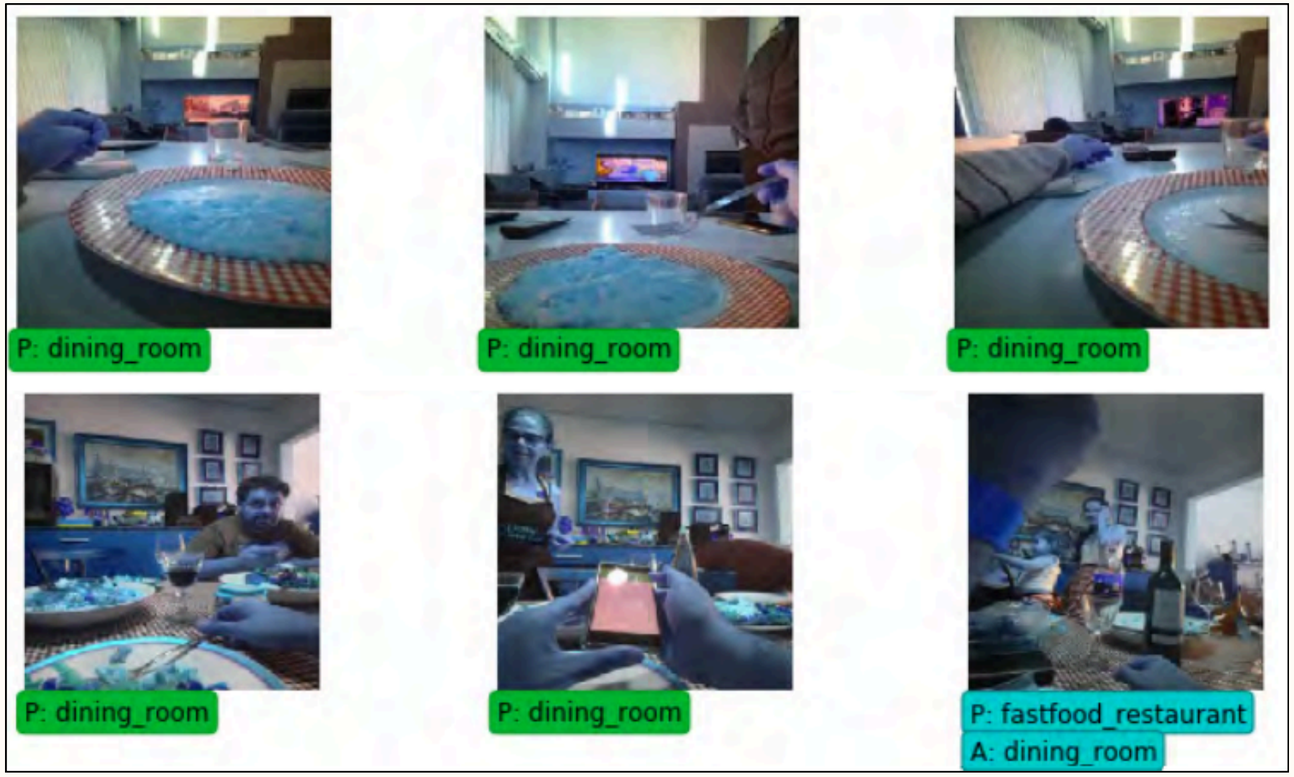
Categories	VGG16		ResNet50		InceptionV3		MACNet		MACNet+SA	
dataset	val	test	val	test	val	tests	val	test	val	test
bakery shop	0.77	0.59	0.72	0.65	0.77	0.75	0.74	0.68	**0.85**	**0.84**
banquet hal	0.71	0.48	0.62	0.38	0.73	0.50	0.64	0.51	**0.75**	**0.70**
bar	0.66	0.52	0.37	0.36	0.74	0.56	0.65	0.58	**0.85**	**0.73**
beer hall	0.77	0.48	0.92	0.45	0.91	0.40	0.95	**0.51**	**0.96**	0.44
butchers shop	0.71	0.83	0.72	0.91	0.72	0.89	**0.79**	**0.92**	0.73	0.88
cafeteria	0.61	0.47	0.64	0.60	0.70	0.59	0.78	0.63	**0.94**	**0.78**
candy store	0.65	0.59	**0.66**	0.63	0.65	0.57	0.63	**0.64**	0.64	0.58
coffee shop	0.45	0.71	0.57	0.71	0.66	0.68	0.89	0.75	**0.93**	**0.87**
delicatessen	0.52	0.62	0.55	0.73	0.50	0.64	**0.69**	0.56	0.59	**0.75**
dining room	0.62	0.67	0.71	0.74	0.73	0.75	**0.92**	**0.87**	0.87	0.86
fastfood restaurant	0.33	0.44	0.33	0.49	0.32	0.50	**0.77**	0.56	0.68	**0.63**
food court	0.64	0.66	0.63	0.69	0.70	0.63	0.82	0.63	**0.86**	**0.73**
ice cream parlor	0.65	0.64	0.64	0.60	**0.72**	**0.69**	0.66	0.64	0.67	0.65
kitchen	0.79	0.85	0.91	0.89	0.88	0.87	0.90	0.89	**0.93**	**0.92**
market indoor	0.53	0.44	0.56	0.60	0.40	0.48	**0.81**	0.64	0.76	**0.82**
market outdoor	0.42	0.53	0.37	0.77	0.39	0.70	**0.61**	0.69	0.48	**0.78**
picnic area	0.51	0.44	0.59	0.47	0.49	0.45	0.68	0.46	**0.80**	**0.67**
pizzeria	0.77	0.62	0.39	0.48	0.81	0.67	0.68	0.67	**0.99**	**0.95**
pub indoor	0.86	0.49	**0.96**	0.88	0.93	0.70	0.95	**0.92**	0.94	0.83
restaurant	0.51	0.47	0.62	0.46	0.60	0.51	0.72	0.55	**0.85**	**0.66**
supermarket	0.80	0.81	0.81	0.86	0.83	0.84	0.71	0.88	**0.91**	**0.89**
sushi bar	0.78	0.44	0.88	0.44	0.76	0.43	0.95	0.73	**0.99**	**0.88**
Avg. F_1 score	0.66	0.62	0.68	0.65	0.72	0.66	0.79	0.72	**0.86**	**0.80**

Average F1 score of VGG16, ResNet50, InceptionV3, MACNet and MACNet+SA model using both validation and test sets from EgoFoodPlaces dataset.

4.4 Experimental Results & Discussions



Top 5 Accuracy Correct (top) and Incorrect (bottom) examples



Examples of Properly predicted images (Row1) & Improperly predicted frames (Row2)

5. Conclusions

Purpose

- MACNet+SA enables automatic food place classification from egocentric photo-streams.
- Supports the generation of dietary reports to monitor and improve eating habits.

Key Findings

- Combines temporal modeling and self-attention to process sequences of images (events).
- Outperforms state-of-the-art models (VGG16, ResNet50, InceptionV3, MACNet) on the EgoFoodPlaces dataset.
- Achieves high F1 scores and Top-1/Top-5 accuracies, demonstrating effective sequence-based scene understanding.

Future Work

- Develop a mobile application for dietary reporting with egocentric cameras.
- Provide personalized, real-time insights into daily food-related behaviors.