Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Future Artificial Intelligence Research

A backpack full of skills: Egocentric Video Understanding with diverse task perspective

S. Peirone, F. Pistilli, A. Alliegro, G. Averta

Speaker: Francesca Pistilli

Politecnico di Torino

Napoli, 2024

GENERAL CONFERENCE
FAIR Future Artificial Intelligence Research
23|24SET2024
NAPOLI

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

FAIR Future Artificial Intelligence Research

What can we learn from a single video?

Different video tasks = different, possibly complementary, perspectives

Actions Recognition (AR)

Object State Change Classification (OSCC)

Long Term Action Anticipation (LTA)

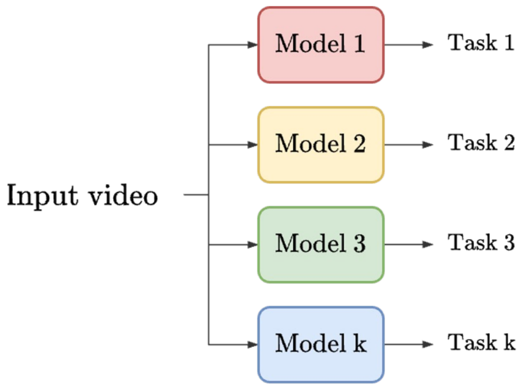
Point of No Return (PNR)

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

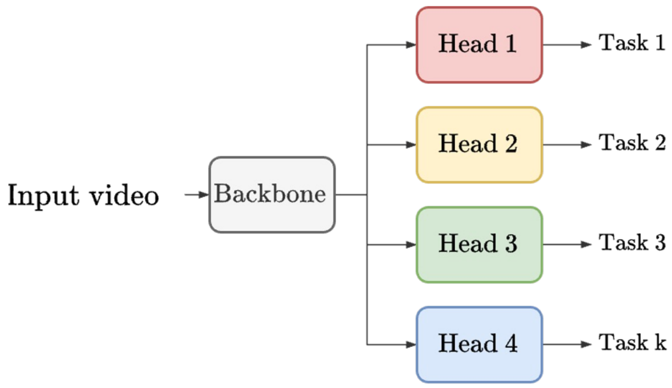
How can we learn from these perspectives?

Main approaches from the literature:

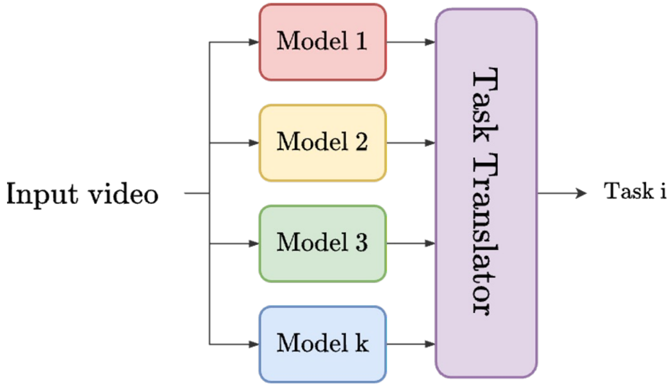
Single Task models



Multi-Task Learning



Task Translation



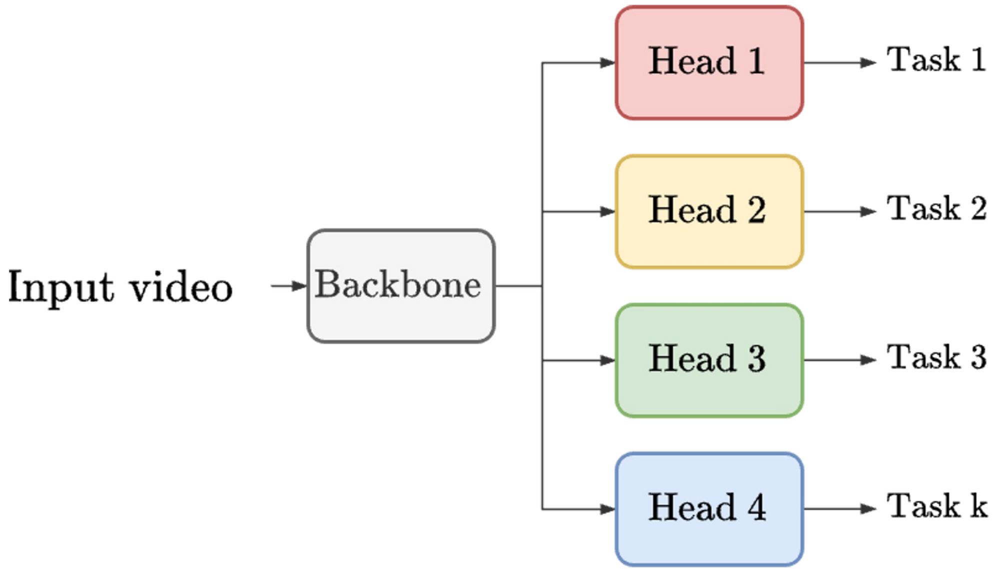
How can we learn from videos? - Multi-Task Learning

Jointly learn multiple tasks using a shared backbone and a set of task-specific heads

+ **Same model is shared across different tasks**

- **Does not *explicitly* model task sinergies**

- **May suffer of negative transfer between tasks**



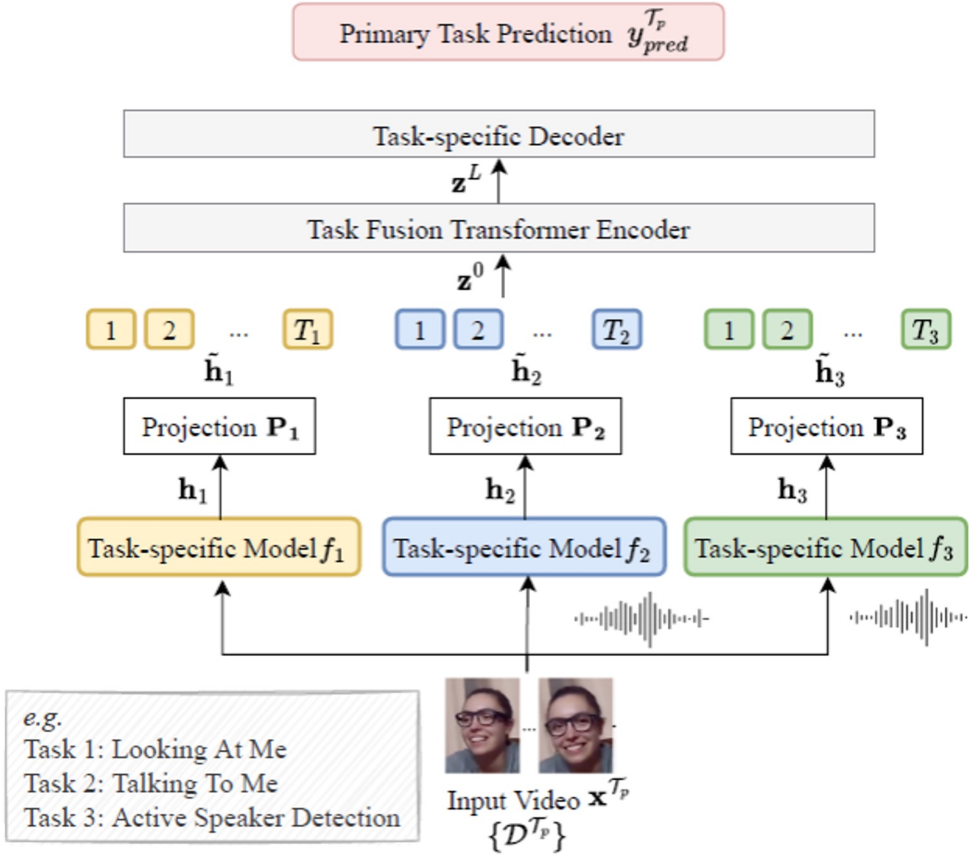
How can we learn from videos? - Cross-Task Translation

EgoT2 proposes an innovative approach to leverage cross-task sinergies by learning to "translate" features across different tasks

+ Combine perspectives from different tasks

- Need to know all the tasks before-hand

- One model for each task



Xue, Zihui, et al. "Egocentric Video Task Translation" (CVPR 2023)

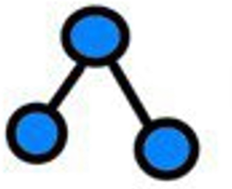
What can we learn from different perspective?

Holistic perception of video stream:

- **Correlate concepts from different tasks**
- **Collection of task-specific knowledge**
- **Exploit gained knowledge to learn novel skills**

Our Goal: Recombining Task-specific Knowledge for a Novel Task

Task-specific Knowledge



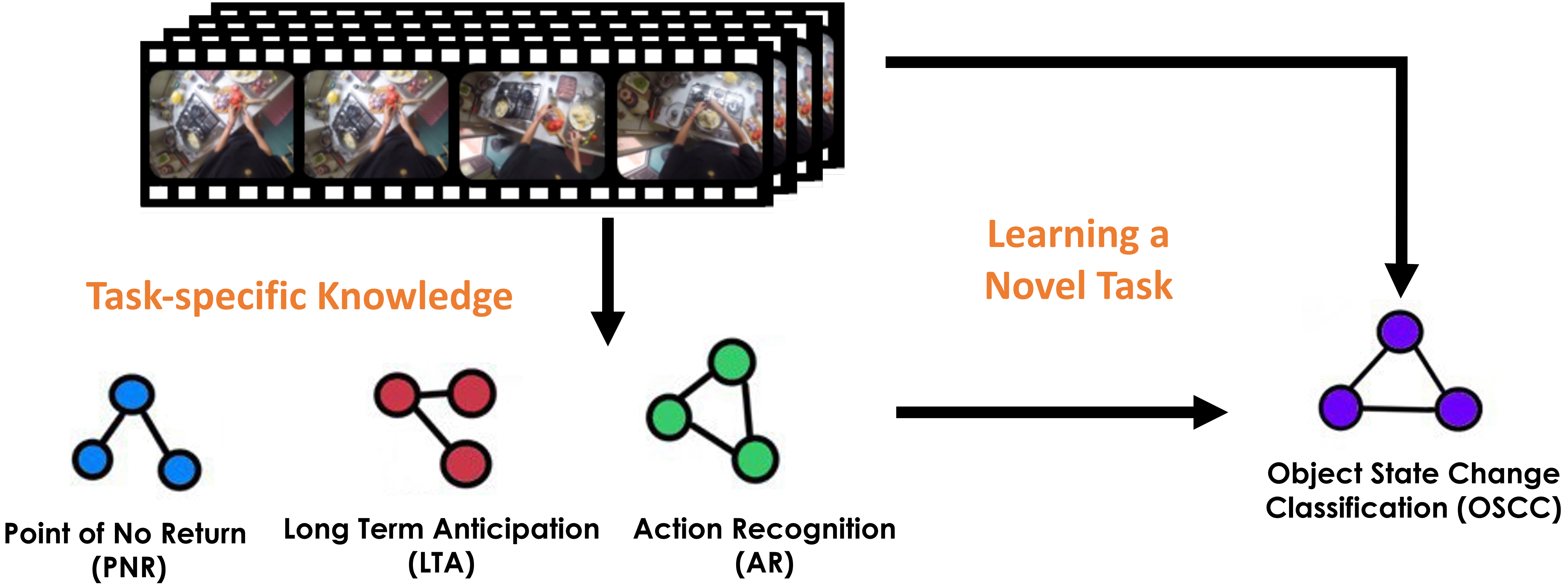
Point of No Return (PNR)

Long Term Anticipation (LTA)

Action Recognition (AR)

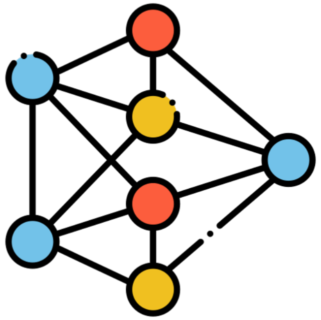
Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
Future Artificial Intelligence Research

Our Goal: Recombining Task-specific Knowledge for a Novel Task



Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

A new paradigm for Egocentric Video Understanding



Shared model
for all the tasks

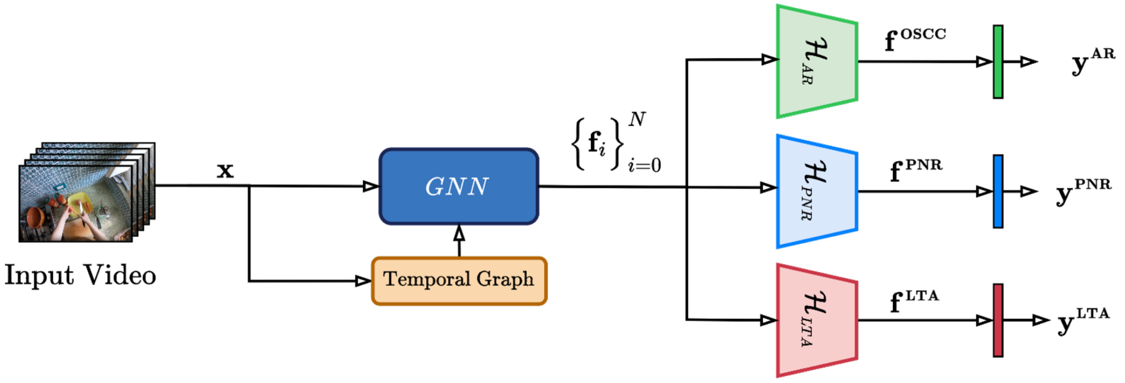
Knowledge reuse
across tasks

Outperform single and multi-task baselines

Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR
Future Artificial Intelligence Research

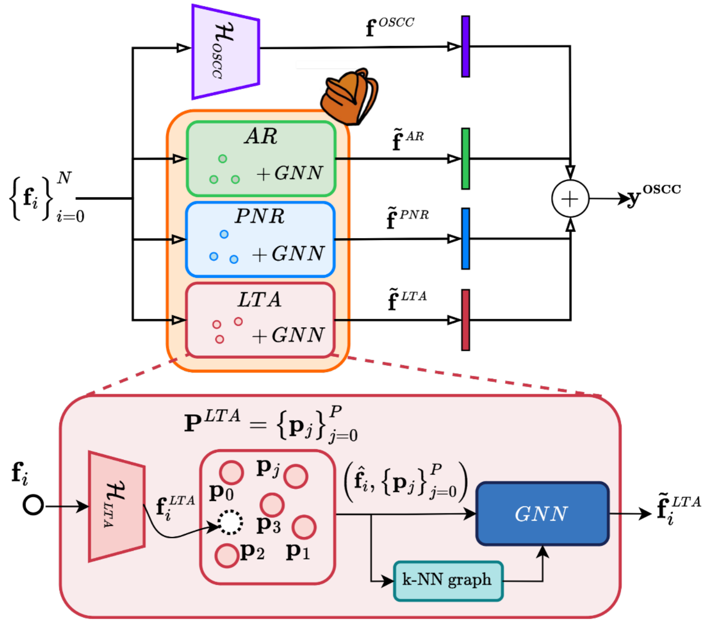
The EgoPack approach

Step 1: MTL Pre-training step



Multi-task pre-training
on a set of known task

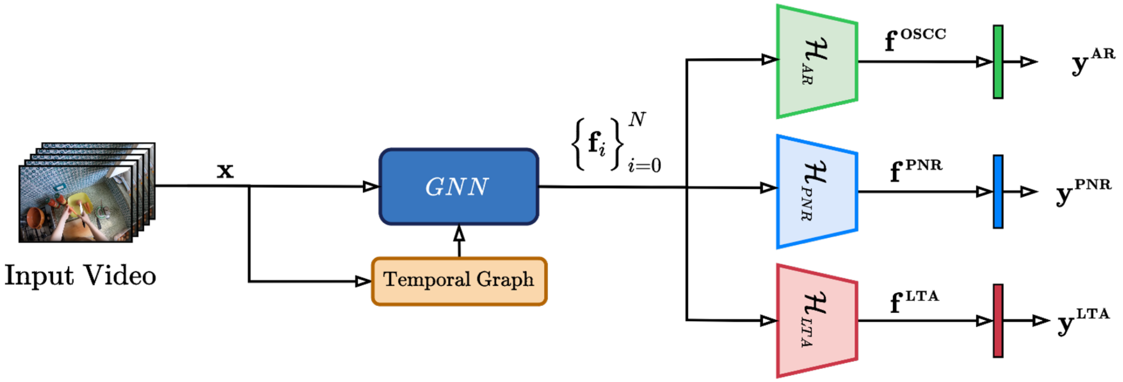
Step 2: Novel Task Learning



Fine-tuning on a novel task
with EgoPack's cross-task interaction

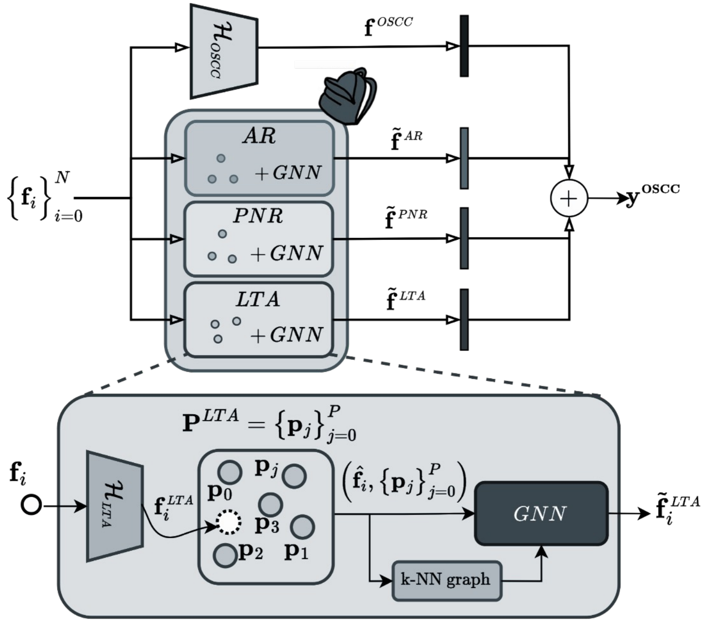
The EgoPack approach

Step 1: MTL Pre-training step



Multi-task pre-training
on a set of known task

Step 2: Novel Task Learning



Fine-tuning on a novel task
with EgoPack's cross-task interaction

Finanziato dall'Unione europea
NextGenerationEU

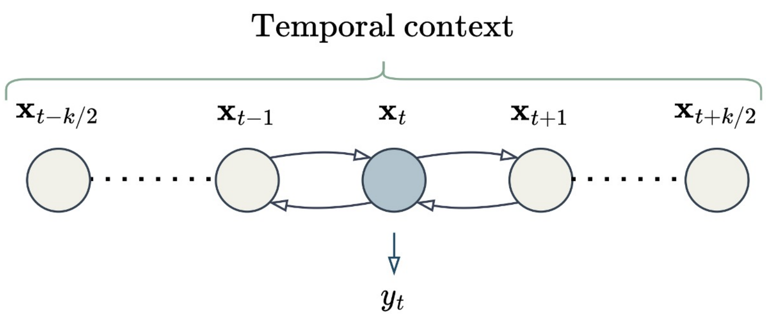
Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

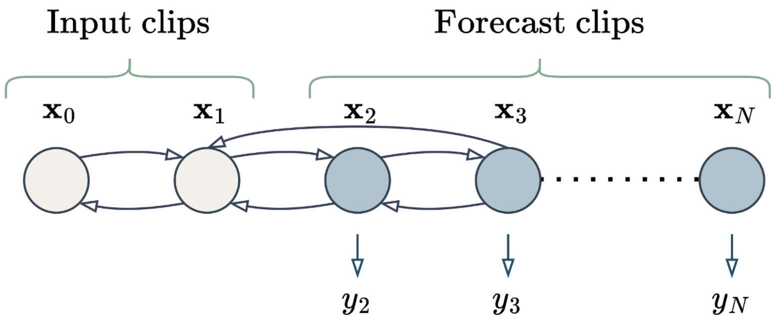
Future Artificial Intelligence Research

Step 1: A graph-based temporal model

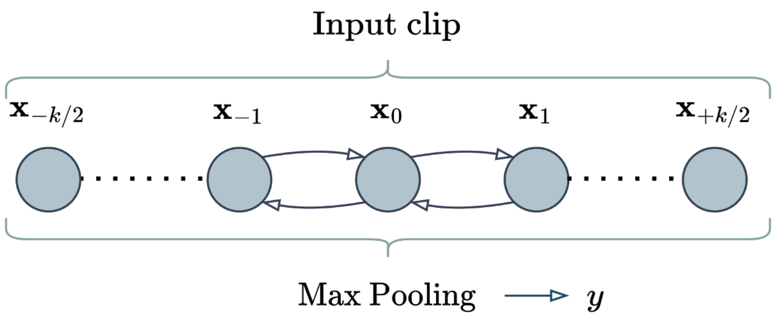
We can model many egocentric vision tasks with a shared graph-based structure…



Node Classification
(AR, PNR)



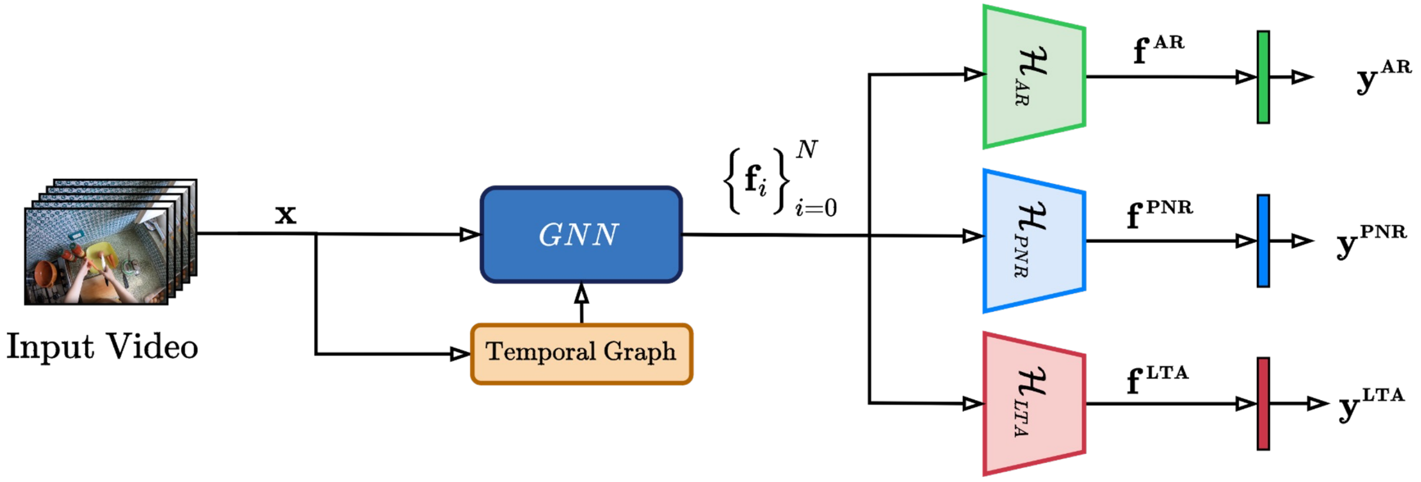
Future Node Classification (LTA)



Graph Classification
(OSCC)

Each node is a temporal segment and
egocentric video tasks become different graph operations

Step 1: Temporal Multi-Task Pre-Training



The output of the Temporal Model is specialized into task-specific features using a set of **task-specific heads**

The output are the task logits $\mathbf{y}_i^k \in \mathbb{R}^{D_o^k}$

Finanziato dall'Unione europea
NextGenerationEU

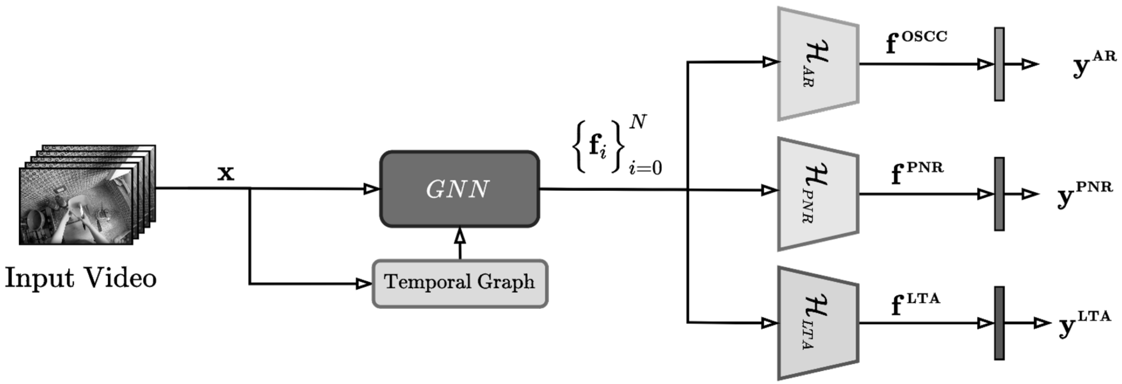
Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Future Artificial Intelligence Research

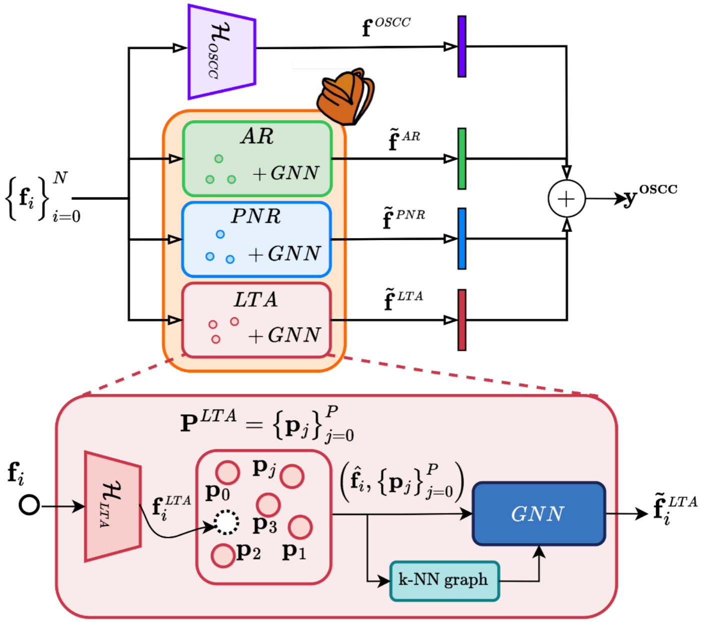
The EgoPack approach

Step 1: MTL Pre-training step



Multi-task pre-training
on a set of known task

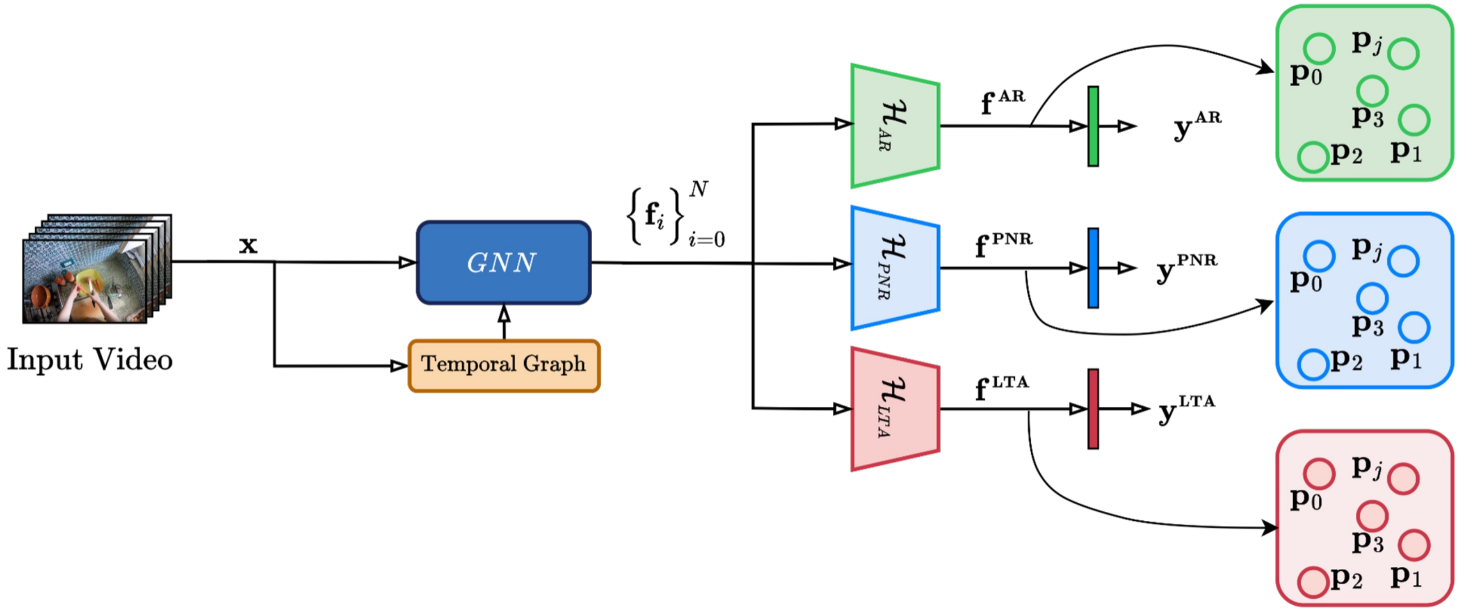
Step 2: Novel Task Learning



Fine-tuning on a novel task
with EgoPack's cross-task interaction

Step 2: Novel Task Learning with EgoPack

Given as input the same video, the model's heads express **different and complementary perspectives** on the content of the video



Step 2.1: Prototypes collection

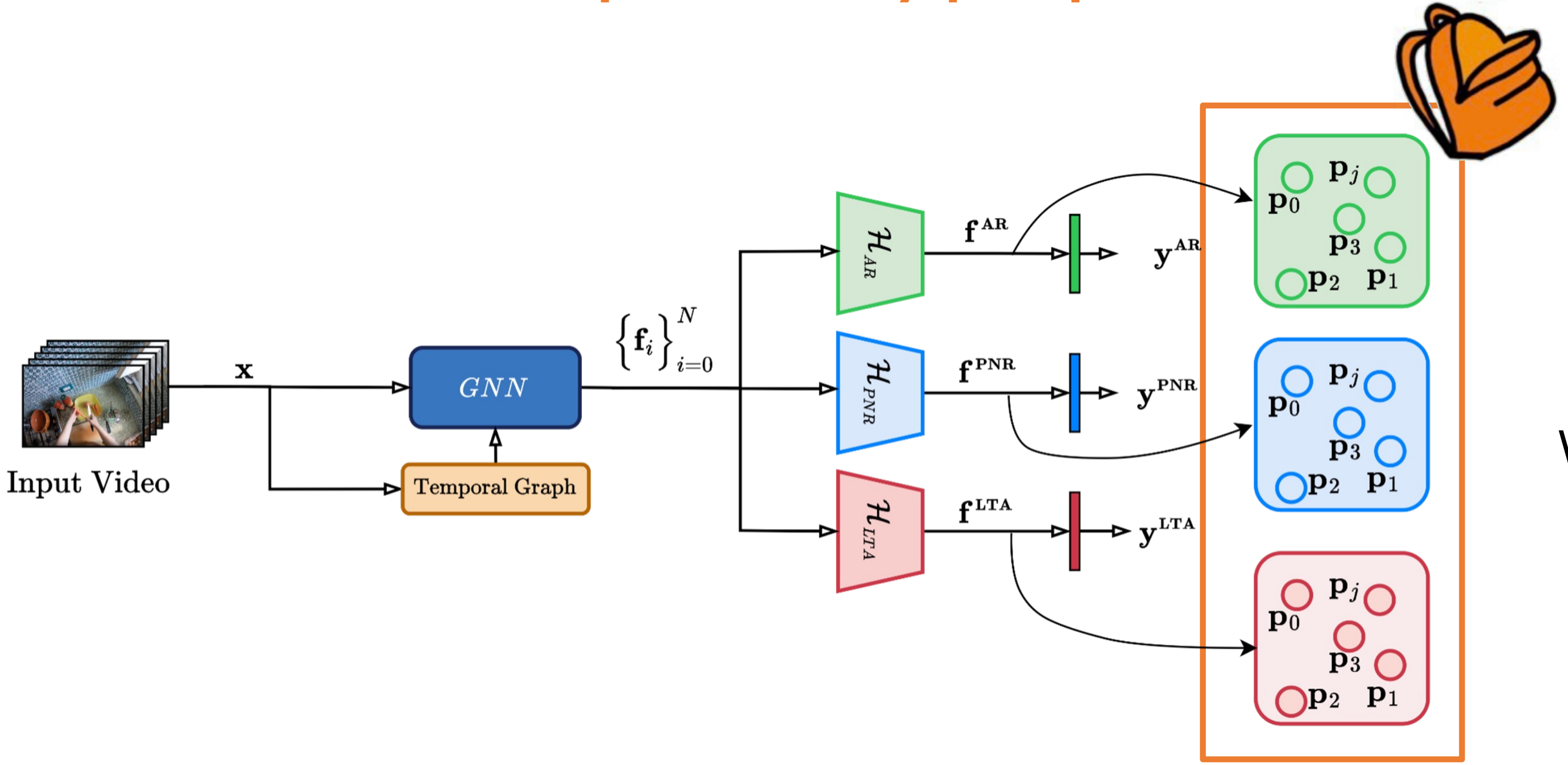
We collect action-wise **task-specific prototypes** by feeding the model with AR videos

$$\mathbf{P}^k = \{\mathbf{p}_0^k, \mathbf{p}_2^k, \ldots, \mathbf{p}_P^k\} \in \mathbb{R}^{P \times D_k}$$

for each task

Step 2: Novel Task Learning with EgoPack

Given as input the same video, the model's heads express **different and complementary perspectives** on the content of the video

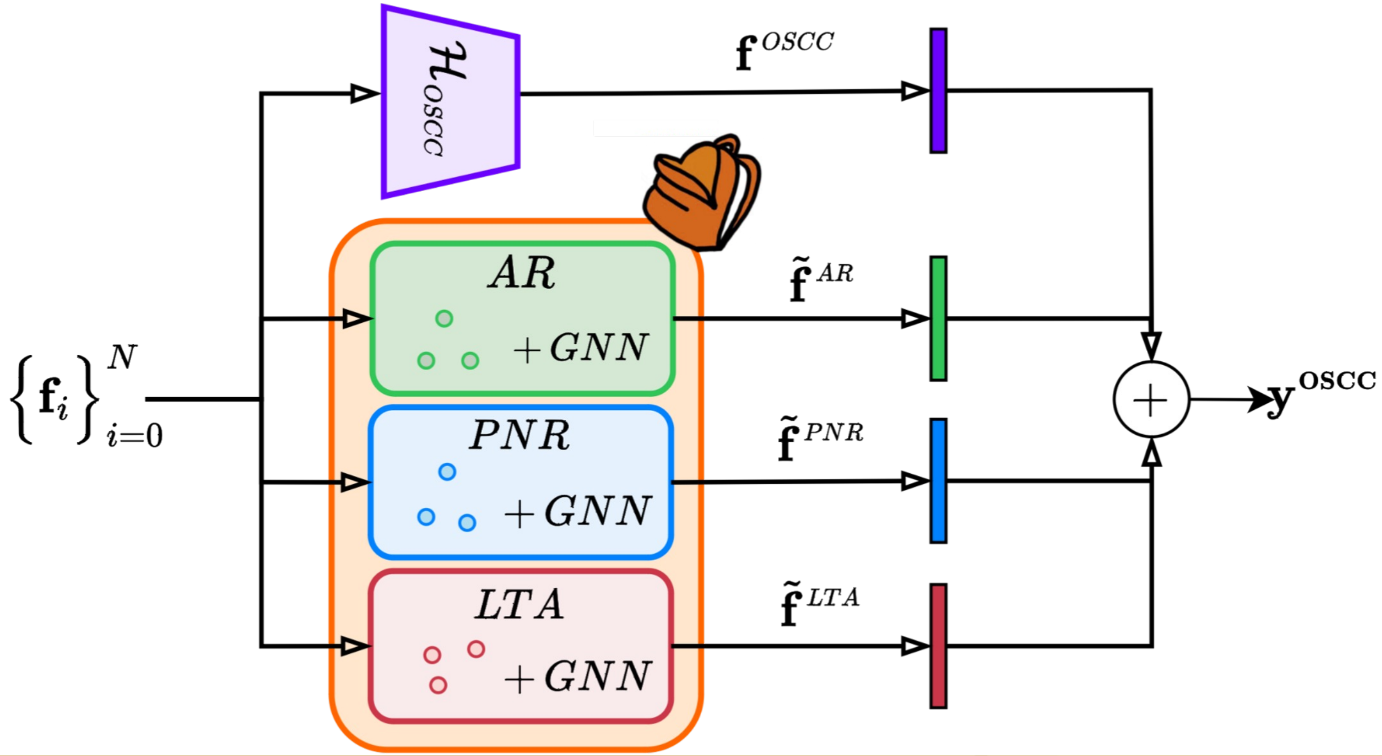


Step 2.1: Prototypes collection

We call these prototypes a ***"backpack of skills"***

Step 2: Novel Task Learning with EgoPack

To learn a **novel task**, e.g., **Object State Change Classification**,
we add the corresponding head and exploit the synergies with the previous tasks.

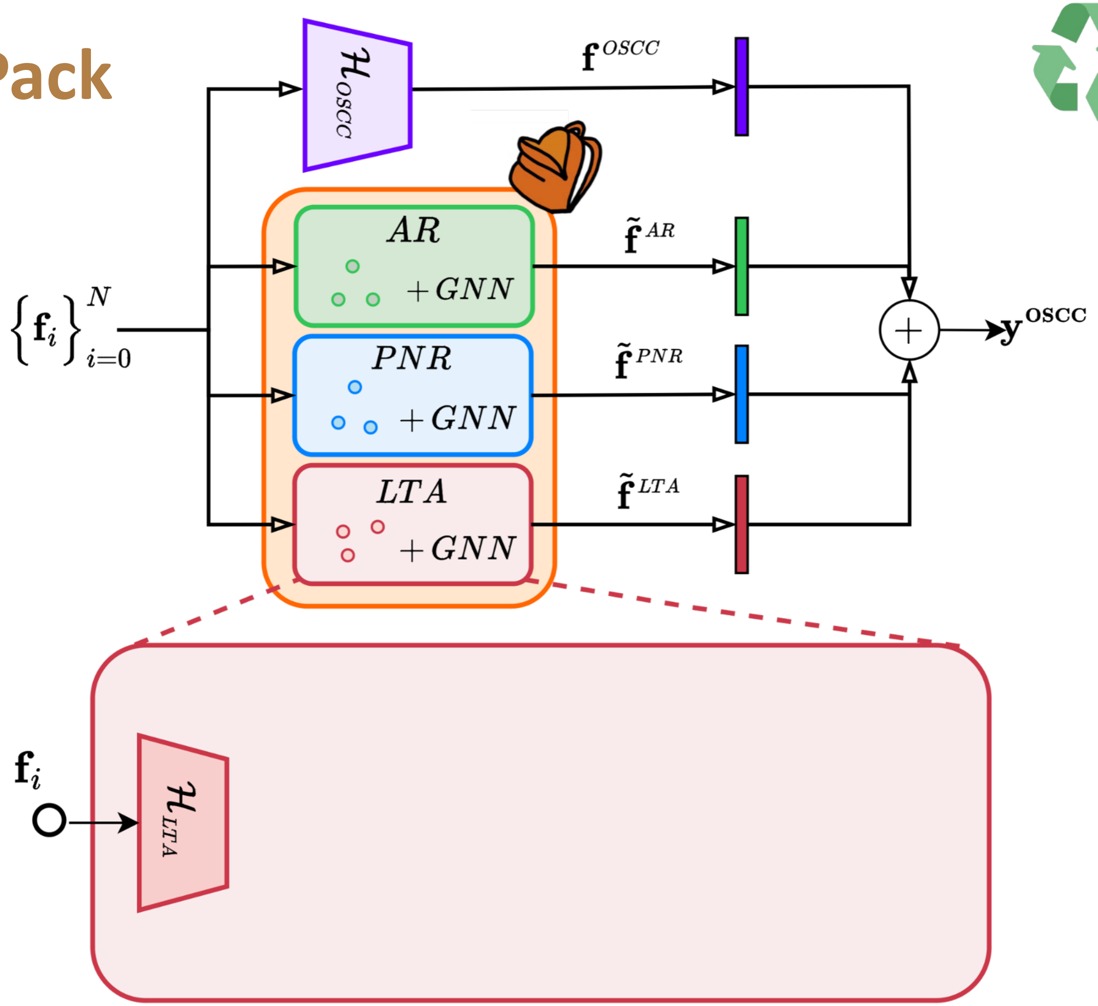


New head trained for the novel task (OSCC)

Knowledge reuse from previous tasks using the prototypes from AR, PNR and LTA

Step 2: Novel Task Learning with EgoPack

We feed the temporal features through the task-specific heads of the pre-training tasks to obtain $\mathbf{f}_i^k$.

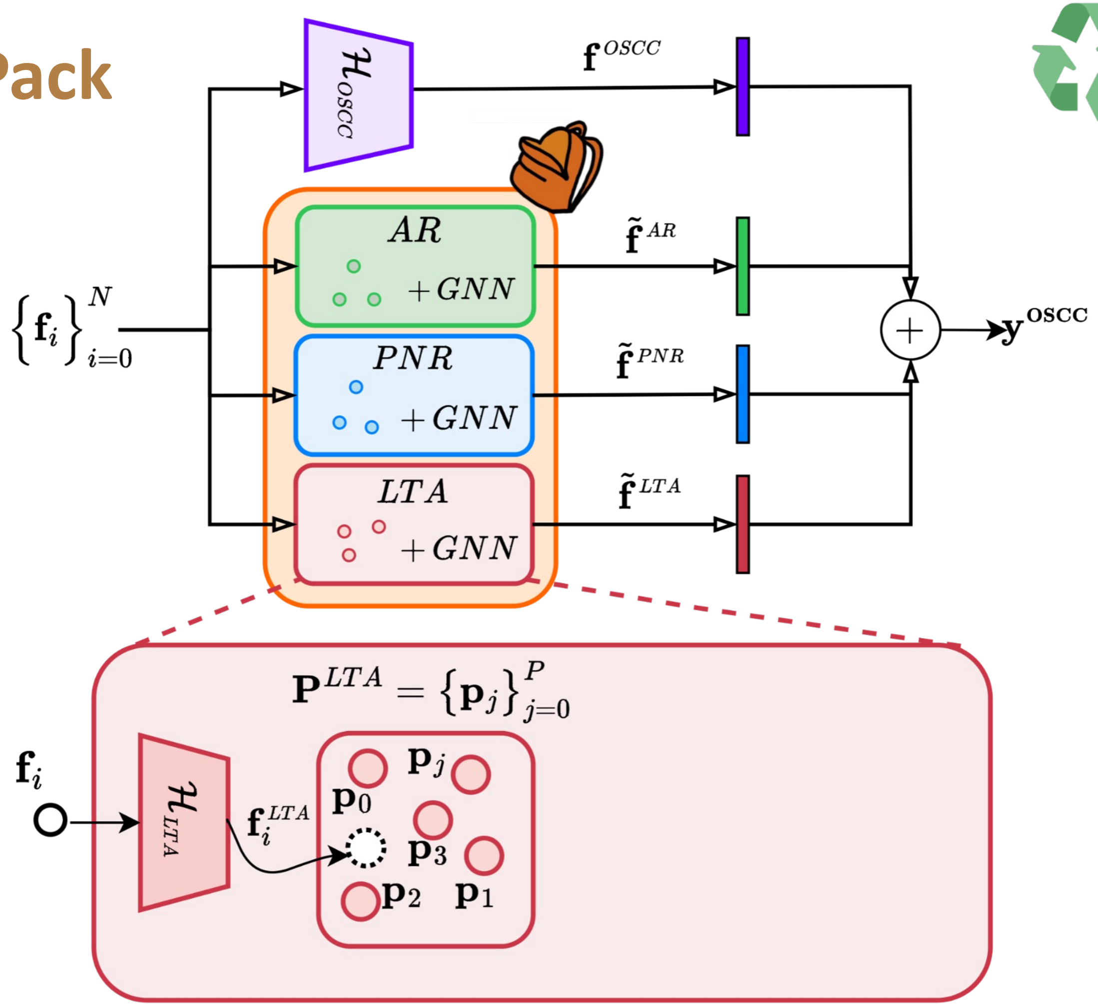


Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA | FAIR Future Artificial Intelligence Research

Step 2: Novel Task Learning with EgoPack

We feed the temporal features through the task-specific heads of the pre-training tasks to obtain $\mathbf{f}_i^k$.

These features act as queries to look for the **closest matching prototypes** using k-NN in the features space.



Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

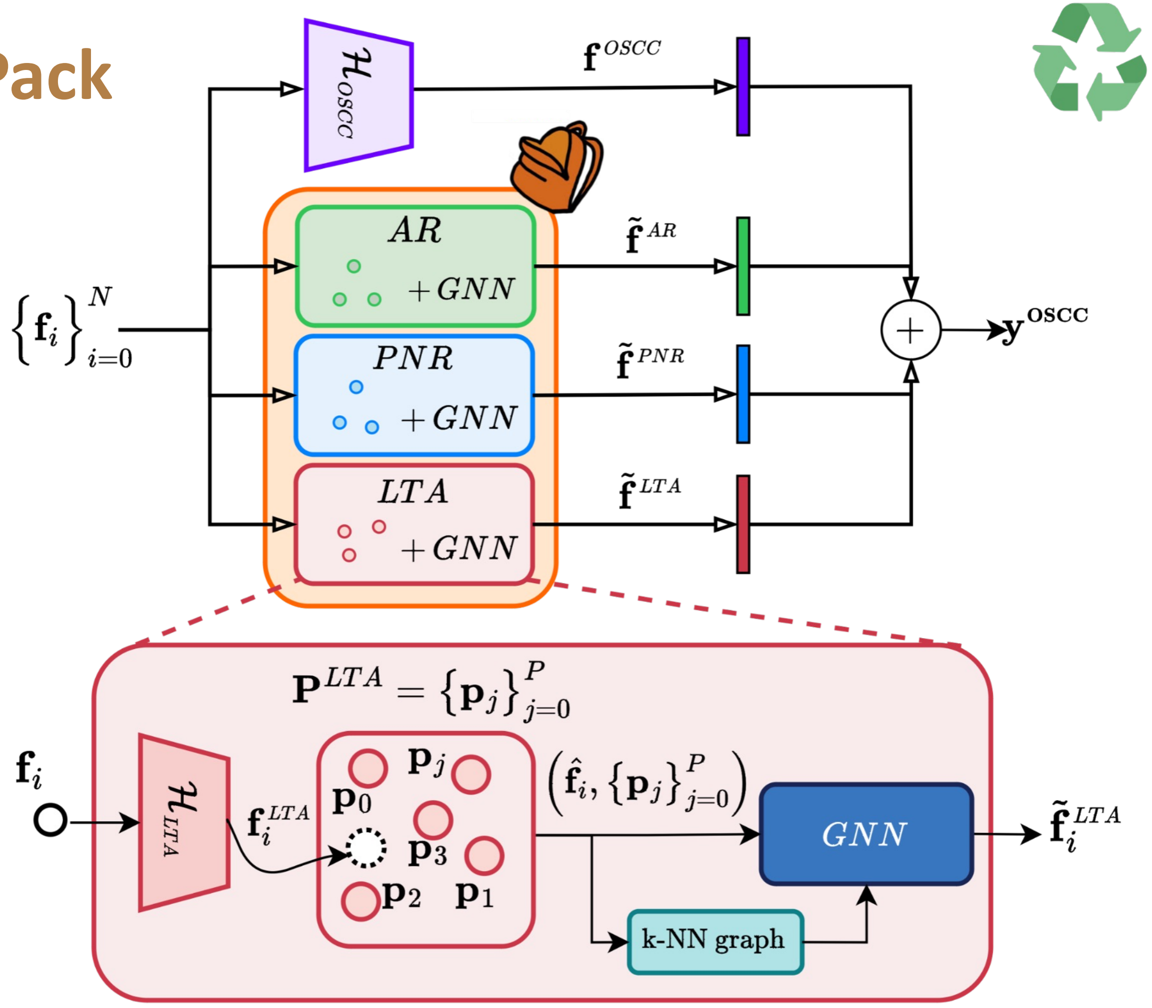
Step 2: Novel Task Learning with EgoPack

We feed the temporal features through the task-specific heads of the pre-training tasks to obtain $\mathbf{f}_i^k$.

These features act as queries to look for the closest matching prototypes using k-NN in the features space.

We refine the task features using **Message Passing with task prototypes.**

$$\mathbf{f}_i^{(l+1),k} = \mathbf{W}_r^{(l)}\mathbf{f}_i^{(l),k} + \mathbf{W}^{(l)} \cdot \max_{\mathbf{p}_j^k \in \mathcal{N}_i^{(l),k}} \mathbf{p}_j^k$$



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Experimental Results - Ego4D HOI Tasks

We validate EgoPack on AR, OSCC, PNR and LTA from Ego4D.

	Trained on frozen features	AR		OSCC	LTA		PNR
		Verbs Top-1 (%)	Nouns Top-1 (%)	Acc. (%)	Verbs ED (↓)	Nouns ED (↓)	Loc. Err. (s) (↓)
Ego4D Baselines [25]	✗	22.18	21.55	68.22	0.746	0.789	0.62
EgoT2s [68]	✗	23.04	23.28	**72.69**	0.731	0.769	**0.61**
MLP	✓	24.08	30.45	70.47	0.763	**0.742**	1.76
Temporal Graph	✓	24.25	30.43	71.26	0.754	0.752	**0.61**
Multi-Task Learning	✓	22.05	29.44	71.10	0.740	0.746	0.62
Task Translation†	✓	23.68	28.28	71.48	0.740	0.756	**0.61**
EgoPack	✓	**25.10**	**31.10**	71.83	**0.728**	0.752	**0.61**

MLP: graph nodes are processed individually, with no temporal modelling

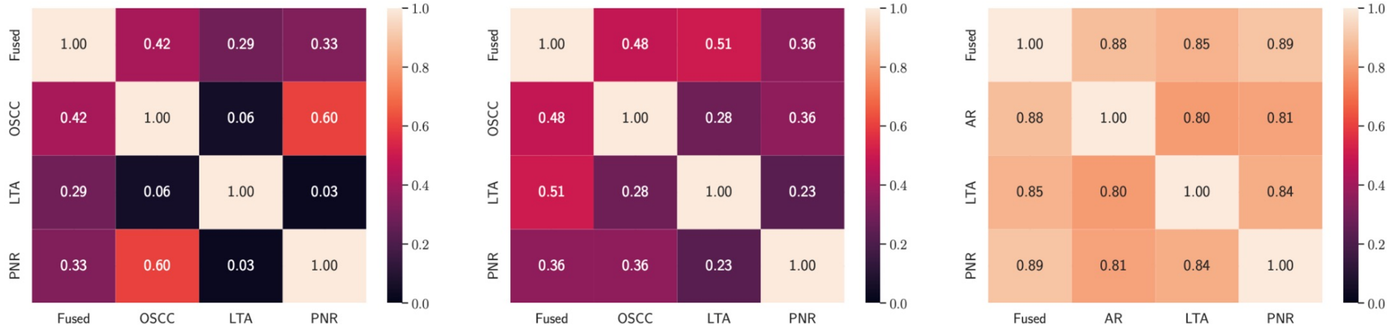
Temporal Graph: MLP + Temporal Graph model of EgoPack

Multi-Task Learning: all tasks are trained together with our Temporal Graph model

Task Translation: re-implementation of EgoT2 using pre-extracted (frozen) features

Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Cross-tasks agreement ratio

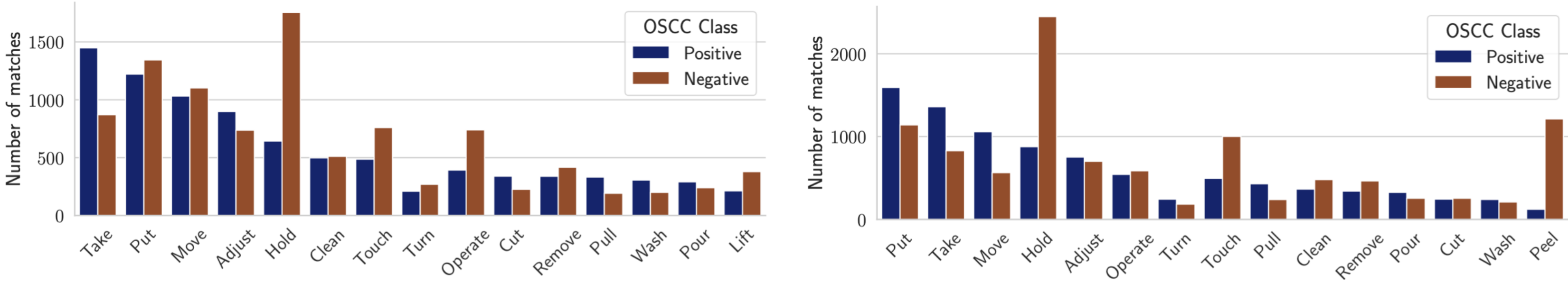


(a) AR Verb (b) AR Noun (c) OSCC

How much different "*perspectives*" bring complementary information?

Queried prototypes

When the novel task is OSCC, what are the closest prototypes from the AR and PNR tasks?



(a) AR Task Prototypes (b) PNR Task Prototypes

Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

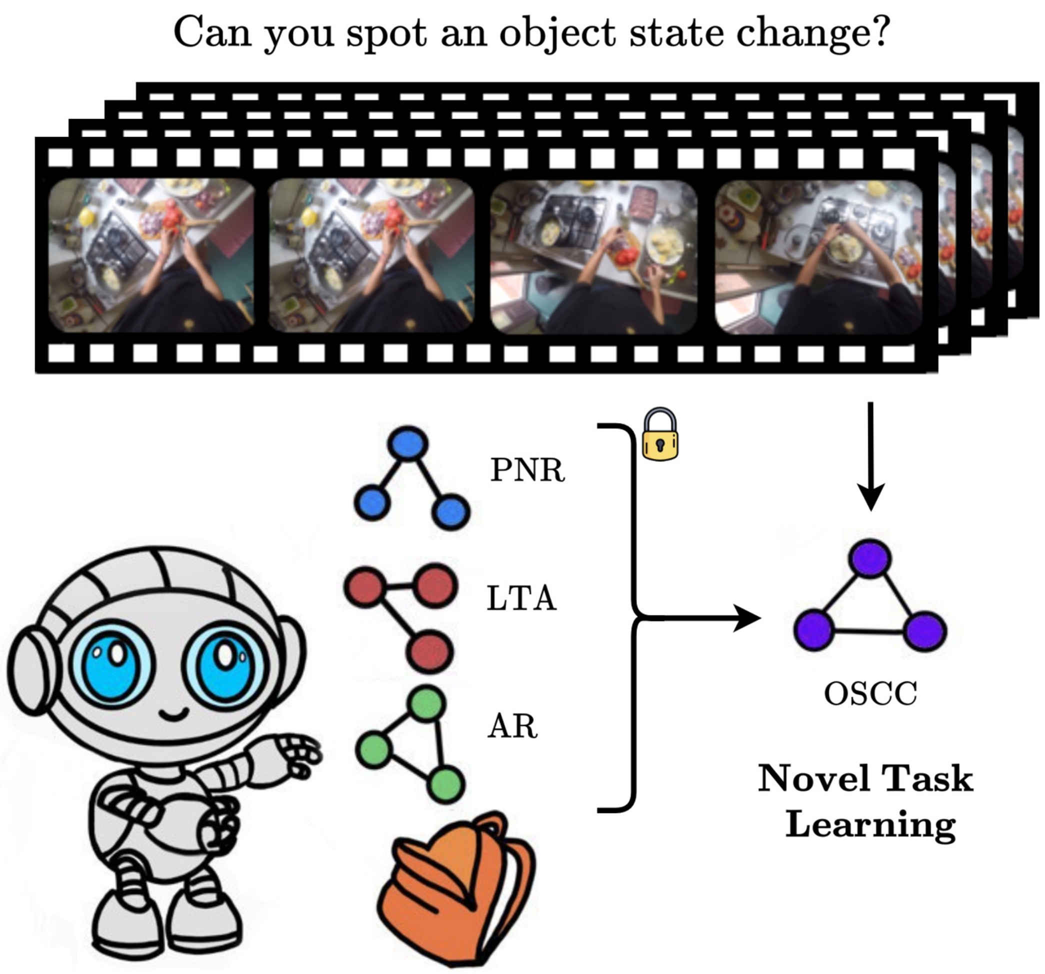
Thank you for your attention!

francesca.pistilli@polito.it

Egocentric Video Understanding with EgoPack

Shared model: model all tasks using a shared graph-based structure

Knowledge Reuse: collect the knowledge learnt from a set of tasks in a backpack of skills ready to be reused when learning a new task



Simone Alberto Peirone, Francesca Pistilli, Antonio Alliegro, Giuseppe Averta "A backpack full of skills: Egocentric Video Understanding with diverse task perspective" (CVPR 2024)