

StillFast: An End-to-End Approach for Short-Term Object Interaction Anticipation

Francesco Ragusa

Giovanni Maria Farinella

Antonino Furnari

FPV@IPLab - University of Catania

Next Vision s.r.l. - Spin-off of the University of Catania

{francesco.ragusa, giovanni.farinella, antonino.furnari}@unict.it

Abstract

Anticipation problem has been studied considering different aspects such as predicting humans' locations, predicting hands and objects trajectories, and forecasting actions and human-object interactions. In this paper, we studied the short-term object interaction anticipation problem from the egocentric point of view, proposing a new end-to-end architecture named StillFast. Our approach simultaneously processes a still image and a video detecting and localizing next-active objects, predicting the verb which describes the future interaction and determining when the interaction will start. Experiments on the large-scale egocentric dataset EGO4D [17] show that our method outperformed state-of-the-art approaches on the considered task. Our method is ranked first in the public leaderboard of the EGO4D short term object interaction anticipation challenge 2022 and it is the official baseline for the 2023 one. Please see the project web page for code and additional details: <https://iplab.dmi.unict.it/stillfast/>.

1. Introduction

Anticipating human behavior in the near future from the first person (egocentric) point of view allows to build intelligent systems able to support users in different domains. Anticipation is crucial in scenarios where one needs to react before actions are executed, such as in autonomous driving, where a vehicle has to anticipate pedestrians' trajectories before they even begin crossing the street [21, 24], in kitchens where smartglasses could alert the users when they are about to touch the hot stove [2, 3] or in the industrial domain, where an intelligent helmet can improve the worker's safety alerting them in case of a dangerous interaction [29, 30].

Previous works have investigated different forms of an-

ticipation tasks, including next-active object predictions [4, 9, 17, 19, 29], predicting future actions [8, 10–12, 27, 32, 34], forecasting human-object interactions [23], predicting future hands [5, 18] or user trajectory prediction [28].

In this paper we address the problem of short-term object interaction anticipation [17] which consists in detecting and localizing the next-active objects in the scene, predicting the verb which describes the interaction and determining when the interaction will start. This task has been studied considering multiple domains [17] as well as focusing specifically on the industrial scenario [29]. We define the task as proposed in [17]. Given a video V and a timestamp t , models can process the video up to time t (denoted as $V_{:t}$) and are required to output a set of future object interaction predictions which will happen after a time interval δ . Each prediction consists in:

- A bounding box localizing the future interacted object (also referred to as next-active object);
- A noun label describing the class of the detected object (e.g., “wooden block”);
- A verb label describing the interaction which will take place in the future (e.g., “take”);
- A real number indicating the “time to contact”, i.e., the time in seconds between the current timestamp and the beginning of the interaction (e.g., 0.75s);
- A confidence score used to rank future predictions for evaluation.

Figure 1 illustrates the considered task.

Current state-of-the-art works [17, 29] have addressed this task in two steps: 1) an object detector detects and recognizes next-active objects in still frames, 2) a 3D network predicts the verb and the time to contact analyzing a video segment. Also in [6], the authors figured out the action detection task on AVA dataset considering a two steps

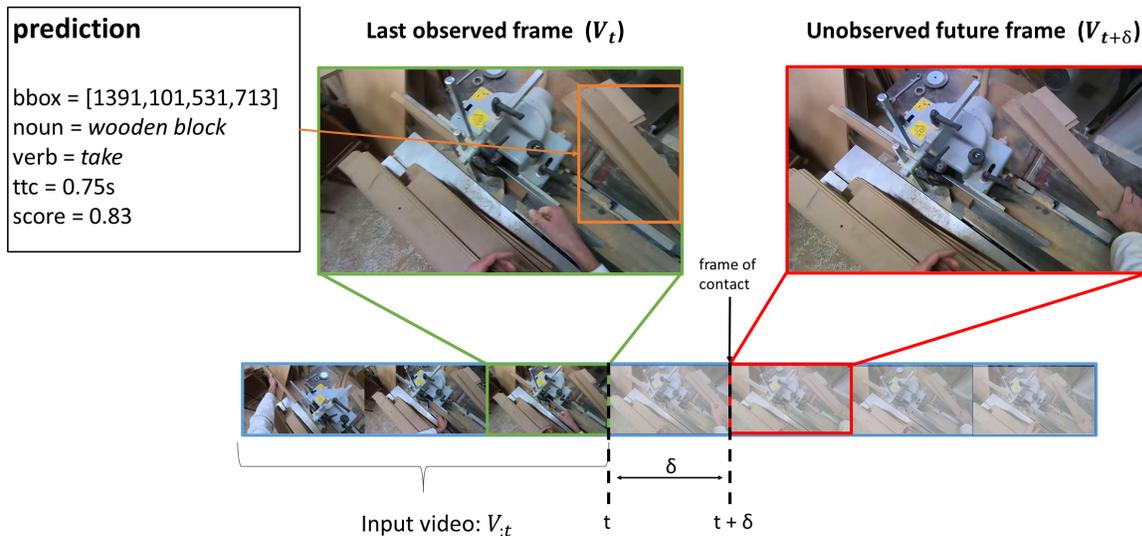


Figure 1. Short-term object interaction anticipation task. Models can process a video V up to time t (denoted as V_t) predicting the bounding box and the class related to the next-active objects, the verb which describes the future interaction, a real number which indicating when the interaction will happen ($t + \delta$) and a score. δ represents the time interval between the last observable frame V_t and the frame of contact at time $t + \delta$.

approach. Since past works showed that composite methods have been outperformed by end-to-end methods [14, 15, 31], we believe that a similar behavior could be obtained even in the considered task. Therefore, we propose *StillFast* network. Similarly to *SlowFast* networks [6], *StillFast* has two branches which simultaneously process two versions of the input video. The “still” branch processes a high resolution still image, i.e., a video with a high spatial resolution, but low temporal resolution (a single frame), whereas the “fast” branch processes a video with a low spatial resolution, but a high temporal resolution (different frames). Our method can be trained end-to-end in a single stage increasing the training speed over traditional two-branches approaches [17].

Experiments on the large-scale dataset of egocentric videos EGO4D [17] show that the proposed method outperforms state-of-the-art approaches highlight that the proposed architecture benefits from the unified branches trained simultaneously. Additionally, we performed an ablation study to assess the impact of each component of the proposed method on the overall performance. Our method is ranked first in the public leaderboard of the EGO4D short term object interaction anticipation challenge 2022 and it is the official baseline for the 2023 one¹. To encourage future research in the field, we release the code implementing the proposed approach at: <https://iplab.dmi.unict.it/stillfast/>.

¹<https://eval.ai/web/challenges/challenge-page/1623/leaderboard/3910>

The contributions of this work are as follows: 1) we proposed a new approach which is able to simultaneously processes a *still image* and a *video*, 2) by performing experiments with state-of-the-art approaches and several ablation, we show the effectiveness of the proposed design, 3) we release the source code of the proposed approach as an extensible framework to encourage future research.

2. Related Work

Our work is related to past research on anticipation considering both third and first-person point of view.

2.1. Anticipation in Third Person Vision

Different works aimed to predict future actions before they happen from a third person point of view [8, 12, 20, 34]. The authors of [20] proposed a new hierarchical representation called *hierarchical movemes* which describes human movements considering multiple levels of granularity to infer future actions from still images or short video clips. The authors of [34] explored the task of anticipating future actions and objects learning to predict future visual representations from unlabeled videos. The authors of [12] proposed a reinforced encoder-decoder architecture (RED) for action anticipation which contains a reinforcement module to encourage the system to predict the correct action as early as possible. Also long-term action anticipation task has been explored in previous works [16] with the aim to predict minutes-long sequences of future actions. Even in the sport

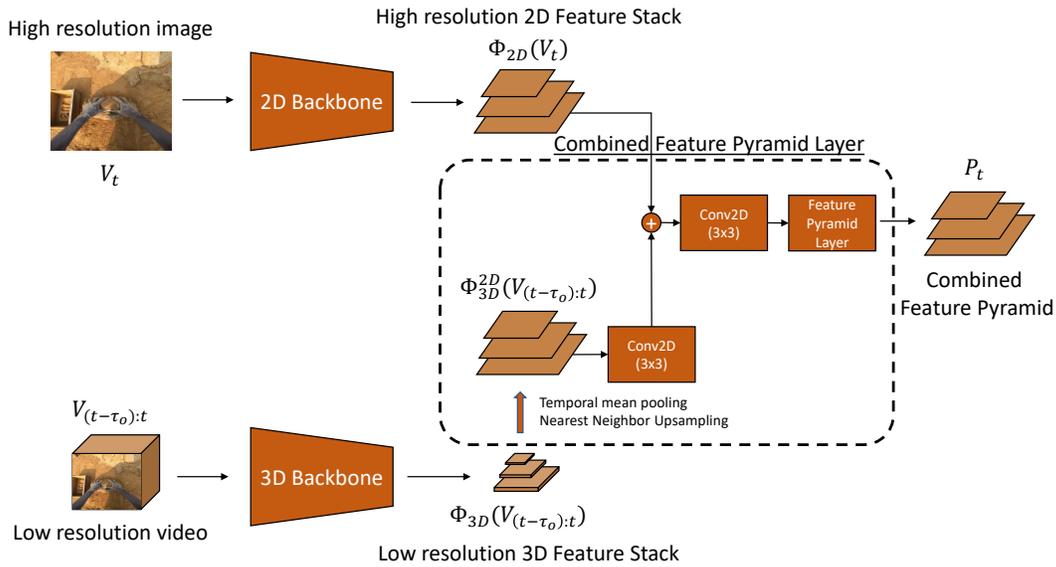


Figure 2. StillFast is composed of a two-branch backbone. Given an input video V and a timestamp t , the proposed model takes as input a high resolution frame V_t (top) and a low resolution video ($V_{(t-\tau_o):t}$) (bottom). A 2D Backbone (“still” branch) processes the high resolution frame V_t , producing a stack of 2D features $\Phi_{2D}(V_t)$. A 3D Backbone (“fast” branch), processes a low resolution video $V_{(t-\tau_o):t}$ obtaining a stack of 3D features $\Phi_{3D}(V_{(t-\tau_o):t})$. The Combined Feature Pyramid Layer is responsible to: 1) up-sample the stack of 3D features with nearest neighbor interpolation to match the spatial resolution of the 2D features and averages over the temporal dimension obtaining the $\Phi_{3D}^{2D}(V_{(t-\tau_o):t})$ features which have the same dimension of 2D features $\Phi_{2D}(V_t)$, 2) fuse these stack of features obtaining the final combined feature pyramid P_t . Before and after the sum operation we added 3x3 convolutional layers to remove artifacts introduced with the up-sampling and sum operations.

domain the problem of forecasting human moves has been explored [8]. Indeed, the authors of [8] proposed a generic framework to anticipate future events in team sports videos directly from visual inputs. The anticipation problem has been explored even predicting the future location of users which allows to build an advanced surveillance systems able to predict people’s activities [21,26] or for autonomous vehicles to understand pedestrian intents to avoid accidents [24,25]. In particular, the authors of [24] tackled the problem to jointly predicting the future spatial position and the body keypoints of pedestrians to have a deeper understanding of pedestrians behavior. The authors of [25] proposed a factorized multimodal approach to predict long-term trajectories of the user focusing on RGB observations and past motion history. Recently, the authors of [26] studied the problem of using few input observations to predict accurate pedestrians position proposing a new teacher-student technique. While prediction’s tasks addressed from the third person point of view are useful in different scenarios, we focused on user-object interactions anticipation for which the egocentric point of view offers several advantages, therefore, we considered this anticipation problem relying on the

large-scale egocentric dataset EGO4D [17].

2.2. Anticipation in First Person Vision

Different past works have explored problems related to anticipation from the first person point of view. Furnari et al. [10,11] proposed a model based on LSTM networks to encode past features and predict future actions from egocentric videos. The authors of [27] extended RULSTM [11] with a novel attention-based technique to consider simultaneously “slow” and “fast” features. The authors of [37] proposed a novel transform-based fusion approach which combines multi-modal features (i.e., audio and visual) to predict future actions. To better understand future human’s behavior, past works also focused on predicting future hands and objects locations. The authors of [23] presented a two-stream fully convolutional network to forecast the presence and location of hands and objects in egocentric videos. Liu et al. [23] focused on hands movements as visual representations to predict future interaction hotspots and future actions.

While different tasks related to the anticipation problem have been deeply studied, the considered task have not been

addressed systematically.

2.3. Short-Term Object Interaction Anticipation

The problem of Short-Term Object Interaction Anticipation has been addressed in different forms. Some works studied the problem of predicting the objects which will be involved in a future human interaction (next-active objects). The authors of [9] have been the first to explore explicitly this task proposing to analyze objects trajectories. The authors of [4] focused on the prediction of hand-objects contact representations to anticipate future actions. The authors of [19] investigated the prediction of a visual attention probability map from images considering hands and objects features. The authors of [29] presented a new egocentric dataset captured in a procedural scenario specifically annotated to address the next-active object detection task, which is tackled using simple object detectors. All these works addressed different versions of the considered task, which made the proposed approaches difficult to compare and extend over. The first attempt to systematically study the problem of Short-Term Object Interaction Anticipation has been done by the authors of [17], provided a formal definition of the task which includes the prediction of next-active objects and associated verbs and time to contact (see Figure 1). Despite this effort, few approaches have been proposed so far to tackle the task in the form defined in [17]. Among them, the baseline proposed in [17] consists of two components trained independently: A Faster R-CNN [31] object detector to detect next-active objects in the last frame of the input video, and a SlowFast [7] action recognition model to attach each bounding box a verb and a time to contact prediction. The model works as a two-branch network. The 2D Faster R-CNN model processes a high resolution image (the most recent frame of the input video) to predict next-active objects and their related classes. The high resolution is needed to have enough detail in order to detect objects at different scales. The 3D SlowFast model processes a low-resolution video clip ending at the current timestamp to predict verb labels and time-to-contact values. The approach is trained in two stages. First, the Faster RCNN model is trained using all training next-active object bounding box and class labels. This object detector is hence run over the training, validation and test data examples to complement them with a set of next-active object bounding box proposals with associated noun labels and detection scores. In a separate stage, a SlowFast model is trained to predict a verb label and a positive time-to-action real number for each bounding box proposal. The authors of [1] extended this approach by employing a DINO [36] object detector to detect and recognize next-active objects in keyframes and a VideoMAE-pretrained transformer network [33] to extract features to predict verbs and time to contact for each detected bounding box.

In this work, we propose an end-to-end approach specifically designed to anticipate the location of next-active objects, the verb which describes the future interaction and how soon the interaction will take place (time to contact). The proposed method can be trained in a single stage, reducing training times and simplifying the research cycle.

3. Still Fast Network

The proposed approach simultaneously processes a still image, i.e., a video with a high spatial resolution, but a low temporal resolution (a single frame) and a video with a low spatial resolution, but a high temporal resolution (different frames). We refer to the first branch as the “still” branch, as it processes a still image, whereas we refer to the second branch as the “fast” branch, as it processes a video with high temporal resolution. We hence term our model “StillFast” network. The model processes the two inputs simultaneously and can be trained end-to-end in a single stage, which increases training speed and allows to better optimize the feature extraction process. The proposed method is comprised of two main components: a backbone which processes 2D (still image) and 3D (video) inputs and outputs a set of spatial features, and a prediction head, which takes as input the spatial features and outputs the detected next active objects together with the associated verbs and time to contact predictions.

The following subsections detail the main components of the proposed method.

3.1. StillFast Backbone

Figure 2 reports a diagram of the StillFast backbone. Given an input video V and a timestamp t , the proposed model takes as input a high resolution frame V_t sampled from video V at time t and a low resolution video $\mathcal{S}(V_{(t-\tau_o):t})$ which is obtained by spatially subsampling with function \mathcal{S} the video $v_{(t-\tau_o):t}$ of length τ_o (observation time) starting at time $t - \tau_o$ and ending at time t .

To process the input image and video simultaneously, the proposed model comprises a two-branch backbone.

A 2D CNN Φ_{2D} (the “still” branch) processes the high resolution frame V_t and produces a stack of 2D features at different spatial resolutions $\Phi_{2D}(V_t)$. The stack of features is obtained by collecting activations at the inner layers of the 2D network. As detailed later, we follow [22] to obtain the stack of features to produce a feature pyramid in order to enable multi-scale object detection as done in the standard [31] architecture.

A 3D CNN Φ_{3D} (the “fast” branch) processes the video $V_{(t-\tau_o):t}$ and outputs a stack of 3D features $\Phi_{3D}(V_{(t-\tau_o):t})$. We perform mid-level feature fusion by combining the 2D and the 3D feature stacks with a Combined Feature Pyramid Layer illustrated in the center part of Figure 2. This layer

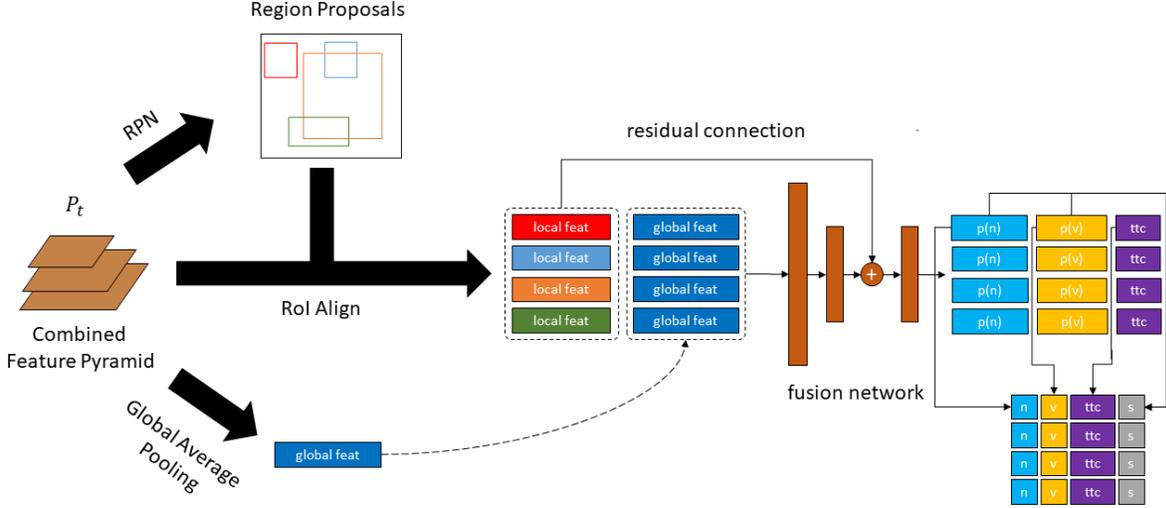


Figure 3. StillFast Prediction Head is based on the Faster R-CNN prediction head. From the Combined Feature Pyramid P_t we obtain global and local features. Local features are obtained through a Region Proposal Network (RPN) which predicts region proposals, from which we compute local features through a RoI Align layer. Global features are obtained with a Global Average Pooling operation and are concatenated with local features. These features are fed in a fusion network and then are summed to the original local features through residual connections. These local-global representations are finally used to predict object (noun) and verb probability distributions and time-to-contact (ttc) through linear layers along with the related prediction score s .

first up-samples the 3D feature map with nearest neighbor interpolation to match the spatial resolution of the 2D features and averages over the temporal dimension to obtain the $\Phi_{3D}^{2D}(V_{(t-\tau_o):t})$ features, which now have the same shape as the 2D features $\Phi_{2D}(V_t)$. These features are hence passed through a 3x3 convolutional layer, summed to the 2D features $\Phi_{2D}(V_t)$ and then passed through another 3x3 convolutional layer. The rationale behind these 2D convolutional layers is to cope with artifacts introduced with the up-sampling and sum operations. The resulting feature map is fed to a standard Feature Pyramid Layer [22] to obtain the final feature pyramid P_t .

In our experiment we use a ResNet-50 as 2D CNN and an X3D-M as 3D CNN. See the supplementary material for more details.

3.2. Prediction Head

Figure 3 shows a diagram of the proposed prediction head, which is based on the Faster R-CNN prediction head [31] as implemented in Detectron2 [35]. A Region Proposal Network (RPN) predicts region proposals from the feature pyramid P_t . A RoIAlign layer is used to extract local features from the region proposals. We found it useful to enrich these local features with global representations of the scene. To do so, we apply global average pooling to the upper layer of the combined feature pyramid and obtain a global image representation, which we concatenate to each local feature extracted from the region proposals. The concatenated features are passed through a fusion network

comprising two fully connected layers. The resulting representations are summed to the original local features through a residual connection. This allows to use global features to modulate the content of local features rather than to replace them. These fused local-global representations are used to predict object (noun) probability distributions and class-specific bounding box regression offsets using linear layers as in [31]. The predicted noun probability distributions $p(n)$ include a background class to reduce the magnitude of positive prediction probabilities in the case of uncertain predictions or proposals falling in background areas. The same features are used to predict a verb probability distribution $p(v)$ and time-to-contact TTC using linear layers. We include a background class in the verb prediction layer as well to further discard false positives when the verb cannot be reliably predicted. A softplus activation is used to predict positive TTC values. For each object proposal, we multiply the noun probability $p(n)$ by the predicted probability of the Top-1 verb, excluding the background class. This is done to make sure that the uncertainty in verb prediction influences the final prediction score $s = p(n) \cdot \max_v \{p(v)\}$. The final predictions are obtained by considering objects with a prediction score larger than or equal to 0.05. The model is trained end-to-end adding a cross entropy verb loss \mathcal{L}_v and a smooth L1 [31] time-to-contact loss \mathcal{L}_{ttc} to the standard Faster R-CNN losses. We weigh the \mathcal{L}_v with 0.1 and the \mathcal{L}_{ttc} with 0.5 in our experiments. See the supplementary material for more details.

4. Experimental Settings

In this section, we report details on the considered dataset and evaluation measures (Section 4.1) and the compared methods (Section 4.2). Please see the supplementary material for the implementation details.

4.1. Dataset and Evaluation Measures

We performed experiments on the large-scale egocentric dataset EGO4D [17]. We consider both the initial version of the dataset described in [17] (denoted as “v1” in this paper), and the recently released update of the dataset (denoted as “v2”), described at this page: <https://ego4d-data.org/docs/updates/#v20-update>, following the official training/evaluation/testing splits. We focus on the subset of the EGO4D dataset which has been explicitly labeled for the Short-Term Object Interaction Anticipation task. Version v1 of this dataset consists in 120 hours of annotated clips, including 27,801 training, 17,217 validation, and 19,780 test examples, annotated with a taxonomy of 87 noun and 74 verb classes. Version v2 consists in 243 hours of annotated clips, including 98,276 training, 47,395 validation, and 19,780 test examples, annotated with a taxonomy of 128 noun and 81 verb classes. It should be noted that v1 and v2 have different training and validation sets, but they share the same test set, which makes test results obtained on the two dataset comparable. Annotations for this shared test set are not publicly available and results can be obtained by sending predictions to an evaluation server².

The evaluation has been performed using Top-K mean Average Precision, which does not penalize methods predicting up to $K - 1$ next-active objects which are not annotated as defined in [17]. In particular, we evaluated methods with different Top-5 mAP measures, i.e., Top-5 mAP Noun, Top-5 mAP Noun+Verb, Top-5 mAP Noun+TTC and Top-5 mAP Noun+Verb+TTC which we named Top-5 mAP Overall, to assess the ability of the model to anticipate next-active object interactions considering different levels of granularity (i.e., nouns, verbs, time-to-contact). We used $K = 5$ as defined in [17].

4.2. Compared Methods

We compare our method with respect to different approaches which addressed the considered Short-Term Object Interaction Anticipation problem.

- Faster R-CNN + Random [17]: uses Faster R-CNN to detect and recognize next active objects, then predicts verbs and time to contact randomly following the distributions of training labels;

²<https://eval.ai/web/challenges/challenge-page/1623/>

Set	Method	Ver	Noun	N+V	N+TTC	Overall
Val	FRCNN+Rnd. [17]	v1	17.55	1.56	3.21	0.34
Val	FRCNN+SF [17]	v1	17.55	5.19	5.37	2.07
Val	StillFast (ours)	v1	16.20	7.47	4.94	2.48
Val	FRCNN+SF [17]	v2	21.00	7.45	7.04	2.98
Val	StillFast (ours)	v2	20.26	10.37	7.16	3.96
Test	FRCNN+Rnd. [17]	v1	20.45	2.22	3.86	0.44
Test	FRCNN+SF [17]	v1	20.45	6.78	6.17	2.45
Test	FRCNN+Feat.	v1	20.45	4.81	4.40	1.31
Test	InternVideo [1]	v1	24.60	9.18	7.64	3.40
Test	StillFast (ours)	v1	19.51	9.95	6.45	3.49
Test	FRCNN+SF [17]	v2	26.15	9.45	8.69	3.61
Test	StillFast (ours)	v2	25.06	13.29	9.14	5.12

Table 1. Results% in Top-5 mean Average Precision on the validation and test sets of EGO4D v1 and v2. In the header of the table, Ver stands for Version, V+N stands for Verb + Noun and N+TTC stands for Noun + Time to Contact. Best results per column within a section of comparable results (horizontal lines) are reported in bold.

- Faster R-CNN + SlowFast [17]: the baseline defined in [17] composed of two components trained independently. A Faster R-CNN [31] object detector which detects and recognizes next-active objects and a SlowFast [6] 3D network to predict for each bounding box a verb and a time to contact;
- InternVideo [1]: follows the same two-stage recipe of Faster R-CNN + SlowFast. A DINO [36] object detector is used to predict next-active objects and a VideoMAE-pretrained transformer video backbone [33] to predict verb and time to contact for each detected next-active object.
- Faster R-CNN + Features: a Faster R-CNN object detector is used to detect and recognize next active objects, then pre-extracted Ominvove features [13] are used to predict the associated verb and time to contact. This baseline has been proposed as a quickstart approach with the EGO4D dataset. In our experiments, we consider the results reported in this document: https://colab.research.google.com/drive/10k_6F1O6K8kX1S4sEnU62HoOBw_CPngR.

5. Results

This section compares the proposed StillFast method to competitors (Section 5.1) and analyses the contribution of the different components of the approach to performance (Section 5.2).

5.1. Comparison with the State of the Art

Table 1 reports the results of the compared methods the validation and test sets of both v1 and v2 of the EGO4D

dataset [17] using the aforementioned Top-5 mAP evaluation measure. As can be noted from Table 1, the proposed method improves over the FRCNN+Rnd. and FRCNN+SF baselines by significant margins in the v1 validation set on verb-related metrics: 7.47% vs 5.19% Noun+Verb Top-5 mAP (+2.28%), and 2.48% vs 2.07% Overall Top-5 mAP (+0.41%). These results suggest that the ability of the proposed approach to process images and video with a unified backbone and accounting for the uncertainty in verb prediction as described in Section 3.2 is beneficial to performance. Indeed, it should be noted that both are based on components with similar performance: a ResNet-50 2D backbone in both cases, a SlowFast 3D backbone for FRCNN+SF and an X3D-M backbone for StillFast. On the downside, the proposed StillFast approach achieves worse performance on Noun Top-5 mAP (16.20% vs 17.55%, hence -1.35%) and on Noun + TTC Top-5 mAP (4.94% vs 5.37%, hence -0.43%). We speculate that this is due to training instabilities due to the multi-task nature of our training procedure, as compared with FRCNN+SF, which is trained in two separate stages (Section 5.2 further analyses performance when multi-tasking is reduced within our architecture). This seems to be mitigated when more training data is available. Indeed, when training on v2 (fourth and fifth row of Table 1), StillFast consistently outperforms FRCNN+SF on Noun + Verb, Noun + TTC and Overall Top-5 mAP, while the loss of performance on Noun Top-5 mAP is much smaller (20.26% vs 21.00%, hence only -0.74%).

The advantages of StillFast are more evident in the test set, which suggests better generalization of the proposed approach. In v1, the proposed method outperforms FRCNN+SF by +3.17% (9.95% vs 6.78%) according to Noun+Verb Top-5 mAP, +0.28% (6.45% vs 6.17%) according to Noun+TTC Top-5 mAP, and +1.04% (3.49% vs 2.45%) according to Overall Top-5 mAP. Noun Top-5 mAP results are still lower, but closer to the ones of FRCNN+SF (19.51% vs 20.45%, hence -0.95%). Also in the test set, using more training data brings larger performance margins to StillFast. In v2 test set, the proposed approach surpasses FRCNN+SF by +3.84% (13.29% vs 9.45%) according to Noun + Verb Top-5 mAP, +0.45% (9.14% vs 8.69%) according to Noun + TTC Top-5 mAP, and +1.51% (5.12% vs 3.61%) according to Overall Top-5 mAP. If we consider that v1 and v2 share the same test set, StillFast jumps from a 9.95% of v1 to 13.29% of v2 Noun + Verb Top-5 mAP and from 3.49% of v1 to 5.12% of v2 Overall mAP, by merely adding more training data, which suggests that the proposed approach can scale in the presence of larger datasets.

In the v1 test set, we are able to also compare StillFast with other methods for which results are publicly available, including InternVideo [1] and FRCNN+Feat. StillFast outperforms FRCNN+Feat. on all evaluation measures except Noun Top-5 mAP, where it achieves slightly worse perfor-

Method	Noun	N+V	N+TTC	Overall
FRCNN+SF [17]	17.55	5.19	5.37	2.07
Nouns Only	19.69	-	-	-
Standard Head	18.42	6.39	5.28	2.17
Proposed Head	16.20	7.47	4.94	2.48

Table 2. Comparison of the performance of different heads.

Method	Noun	N+V	N+TTC	Overall
Proposed Head	16.20	7.47	4.94	2.48
-global features	16.90	5.77	4.60	1.94
-res. connections	15.44	6.40	4.80	2.25
-verb-noun product	14.95	6.29	4.28	1.78
Sum Fusion	15.26	6.68	4.93	2.49

Table 3. Ablation study of the proposed head.

mance (19.51% vs 20.45%). Note that FRCNN+Feat. uses the same Faster RCNN object detector component as FRCNN+SF and is trained in two stages, hence similar considerations as the ones made while comparing the proposed approach with FRCNN+SF apply here as well. StillFast still outperforms InternVideo with respect to Noun+Verb Top-5 mAP (9.95% vs 9.18%) and Overall Top-5 mAP (3.49% vs 3.40%) despite being based on less advanced components (StillFast relies on a Faster R-CNN object detector, while InternVideo relies on DINO [36], StillFast relies on an X3D video backbone, while InternVideo relies on VideoMAE [33] pre-trained transform-based video backbones) and being trained end-to-end (InternVideo is trained in two stages as FRCNN+SF). We leave the integration of higher performing components in StillFast to future works.

Figure 4 reports two success (left) and two failure examples (right). The model struggles with uncertain future actions (“put cement” vs “mold cement”) and unusual actions (“clean book” vs “put book”).

5.2. Ablation study

In this section, we further investigate other aspects of the proposed StillFast architecture including the verb/noun prediction trade-off introduced by the prediction head described in Section 3.2, and the contribution of different architectural choices to the overall performance. The tables in this section report validation results of experiments performed on EGO4D v1 data. All results are in Top-5 mAP%.

Prediction Head Table 2 compares the results obtained by the proposed model when different prediction heads are used. The results of FRCNN+SF are also reported for reference. “Nouns Only” predicts only next-active objects and does not predict any verb or time to contact, “Standard Head” refers to the standard prediction head proposed in [17] which uses a SlowFast model to attach a verb and a time to contact prediction to each detected bounding box. “Proposed Head” refers to the head described in 3.2, in-



Figure 4. Two success examples (left) and two failure cases (right).

	Noun	N+V	N+TTC	Overall
Proposed backbone	16.20	7.47	4.94	2.48
w/o 3D backbone	14.54	6.17	4.10	1.82
w/o conv. block post sum	15.13	6.79	4.80	2.25
post-pyramid fusion	15.01	6.74	4.69	2.34

Table 4. Ablation study considering the contribution of each component of the proposed architecture.

cluding global features and accounting for verb prediction uncertainty. As can be noted, predicting only nouns outperforms the FRCNN+SF baseline by +2.05% (19.69% vs 17.55%) with respect to Noun Top-5 mAP. Adding verb and time to contact prediction with a standard head decreases Noun Top-5 mAP performance to 18.42%, which still outperforms FRCNN+SF (17.55%), while obtaining better or comparable performance on the other metrics. This suggests that the proposed approach is penalized by the training instabilities caused by multi-tasking (i.e., predicting simultaneously nouns, verbs and time to contact). The proposed head achieves the best results in terms of Noun+Verb Top-5 mAP and Overall performance thanks to the use of global features and verb-noun score product, which effectively accounts for the uncertainty in the prediction of verbs. Table 3 assesses the impact of the main components of the proposed head, comparing the performance of the overall module with versions in which we remove global features, residual connections, and the verb-noun score product module. We also compare the head with a version which replaces the proposed global-local fusion mechanism by a simple sum (last row). As can be observed, the proposed architecture obtains overall the best results. Sum fusion achieves slightly better overall Top-5 mAP (+0.01%), but lower results according to the other metrics, for which we prefer the proposed concatenation + residual connection design.

Backbone Table 4 shows the impact of the different components in the backbone when the proposed head is used. We observe that without a 3D branch (second row) the performance of the method drops according to all mAP measures by significant margins (e.g., 1.82% vs 2.48% Overall Top-5 mAP). This suggests that the proposed backbone design successfully combines 3D and 2D features to anticipate future interactions. The third row shows that re-

moving 2D convolutional blocks after the sum operation in the combined feature pyramid layer negatively affect performance (e.g., 2.25% vs 2.48% Top-5 Overall mAP). We speculate that the use of these layers allows to deal with artifacts potentially introduced in the upsampling and sum operations. The last row reports the performance of an alternative design of the network which attaches a 2D feature pyramid to the 2D network and a 3D feature pyramid to the 3D network and then fuses the resulting feature pyramids. As can be noted, this design is less effective than the proposed one, with a Top-5 Overall mAP score equal to 2.34% (vs 2.48% of the proposed design).

6. Conclusion

In this paper, we have presented StillFast, an end-to-end approach to tackle the Short Term Object Interaction Anticipation task. Differently from previous methods, StillFast is designed to process video and image inputs simultaneously and encourage future research in this field by providing a practical framework designed to be easily extensible. While the proposed method achieves promising performance EGO4D, we are aware of some of the limitations of the approach, which we leave to future works. Specifically, the current implementation of the method is based on convolutional 2D and 3D backbones, while benefits may arise from the use of more recent transformer-based components as done in previous works [1]. Moreover, the current end-to-end training procedure seems to be penalized by the multi-tasking arising from predicting noun, verb and time to contact within a single model. We trust that future works will be able to address such limitations and extend on the proposed work.

ACKNOWLEDGEMENTS

This research is supported by Next Vision³ s.r.l., by the project Future Artificial Intelligence Research (FAIR) – PNRR MUR Cod. PE0000013 - CUP: E63C22001940006 and by the project MISE - PON I&C 2014-2020 - Progetto ENIGMA - Prog n. F/190050/02/X44 – CUP: B61B19000520008.

³Next Vision: <https://www.nextvisionlab.it/>

References

- [1] Guo Chen, Sen Xing, Zhe Chen, Yi Wang, Kunchang Li, Yizhuo Li, Yi Liu, Jiahao Wang, Yin-Dong Zheng, Bingkun Huang, Zhiyu Zhao, Junting Pan, Yifei Huang, Zun Wang, Jiashuo Yu, Yanan He, Hongjie Zhang, Tong Lu, Yali Wang, Liming Wang, and Yu Qiao. Internvideo-ego4d: A pack of champion solutions to ego4d challenges. *ArXiv*, abs/2211.09529, 2022. [4](#), [6](#), [7](#), [8](#)
- [2] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. [1](#)
- [3] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Rescaling egocentric vision. *CoRR*, abs/2006.13256, 2020. [1](#)
- [4] E. Dessalene, C. Devaraj, M. Maynard, C. Fermuller, and Y. Aloimonos. Forecasting action through contact representations from first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. [1](#), [4](#)
- [5] Chenyou Fan, Jangwon Lee, and M. Ryoo. Forecasting hand and object locations in future frames. *ArXiv*, abs/1705.07328, 2017. [1](#)
- [6] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2018. [1](#), [2](#), [6](#)
- [7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. [4](#)
- [8] Panna Felsen, Pulkit Agrawal, and Jitendra Malik. What will happen next? forecasting player moves in sports videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3362–3371, 2017. [1](#), [2](#), [3](#)
- [9] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *J. Vis. Commun. Image Represent.*, 49:401–411, 2017. [1](#), [4](#)
- [10] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *International Conference on Computer Vision (ICCV)*, 2019. [1](#), [3](#)
- [11] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2020. [1](#), [3](#)
- [12] J. Gao, Zhenheng Yang, and Ramakant Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. *ArXiv*, abs/1707.04818, 2017. [1](#), [2](#)
- [13] Rohit Girdhar, Mannat Singh, Nikhil Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16081–16091, 2022. [6](#)
- [14] R. Girshick. Fast R-CNN. In *ICCV*, 2015. [2](#)
- [15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. [2](#)
- [16] Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, and Minsu Cho. Future transformer for long-term action anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3052–3061, June 2022. [2](#)
- [17] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrahm Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Meryem Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022. [1](#), [2](#), [3](#), [4](#), [6](#), [7](#)
- [18] Wenqi Jia, Miao Liu, and James M. Rehg. Generative adversarial network for future hand segmentation from egocentric video. In *European Conference on Computer Vision*, 2022. [1](#)
- [19] Jingjing Jiang, Zhixiong Nan, Hui Chen, Shitao Chen, and Nanning Zheng. Predicting short-term next-active-object through visual attention and hand position. *Neurocomputing*, 433:212–222, 2021. [1](#), [4](#)
- [20] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. volume 8691, pages 689–704, 09 2014. [2](#)
- [21] Lihuan Li, Maurice Pagnucco, and Yang Song. Graph-based spatial transformer with memory replay for multi-future pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2231–2241, June 2022. [1](#), [3](#)
- [22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. [4](#), [5](#)
- [23] Miao Liu, Siyu Tang, Yin Li, and James M. Rehg. Forecasting human-object interaction: Joint prediction of motor

- attention and actions in first person video. In *Computer Vision – ECCV 2020*, pages 704–721, Cham, 2020. Springer International Publishing. 1, 3
- [24] Karttikeya Mangalam, Ehsan Adeli, Kuan-Hui Lee, Adrien Gaidon, and Juan Carlos Niebles. Disentangling human dynamics for pedestrian locomotion forecasting with noisy supervision. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2773–2782, 2019. 1, 3
- [25] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15233–15242, October 2021. 3
- [26] Alessio Monti, Angelo Porrello, Simone Calderara, Pasquale Coscia, Lamberto Ballan, and Rita Cucchiara. How many observations are enough? knowledge distillation for trajectory forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6553–6562, June 2022. 3
- [27] Nada Osman, Guglielmo Camporese, Pasquale Coscia, and Lamberto Ballan. Slowfast rolling-unrolling lstms for action anticipation in egocentric videos. *CoRR*, abs/2109.00829, 2021. 1, 3
- [28] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4697–4705, 2016. 1
- [29] Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain, 2022. 1, 4
- [30] Francesco Ragusa, Antonino Furnari, Antonino Lopes, Marco Moltisanti, Emanuele Ragusa, Marina Samarotto, Luciano Santo, Nicola Picone, Leo Scarso, and Giovanni Maria Farinella. Enigma: Egocentric navigator for industrial guidance, monitoring and anticipation. In *International Conference on Computer Vision Theory and Applications (VIS-APP)*, 2023. 1
- [31] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 2, 4, 5, 6
- [32] Debaditya Roy and Basura Fernando. Action anticipation using latent goal learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2745–2753, January 2022. 1
- [33] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 4, 6, 7
- [34] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 98–106, 2015. 1, 2
- [35] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 5
- [36] Hao Zhang, Feng Li, Siyi Liu, Lei Zhang, Hang Su, Jun-Juan Zhu, Lionel Ming shuan Ni, and Heung yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *ArXiv*, abs/2203.03605, 2022. 4, 6, 7
- [37] Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelbogen, and Jürgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation, 2022. 3