Temporal DINO: A Self-supervised Video Strategy to Enhance Action Prediction

Izzeddin Teeti[1,*], Rongali Sai Bhargav[2], Vivek Singh[1], Andrew Bradley[1], Biplab Banerjee[2], Fabio Cuzzolin[1]

[1]VAIL, Oxford Brookes University, (iteeti, vsingh, abradley, fabio.cuzzolin)@brookes.ac.uk,

[2]Indian Institute of Technology, Bombay, (sai.bhargav, Biplab)@iitb.ac.in

Abstract

*The emerging field of action prediction - the task of forecasting action in a video sequence - plays a vital role in various computer vision applications such as autonomous driving, activity analysis and human-computer interaction. Despite significant advancements, accurately predicting future actions remains a challenging problem due to high dimensionality, complex dynamics and uncertainties inherent in video data. Traditional supervised approaches require large amounts of labelled data, which is expensive and time-consuming to obtain. This paper introduces a novel self-supervised video strategy for enhancing action prediction inspired by DINO (self-**di**stillation with **no** labels). The approach, named Temporal-DINO, employs two models; a 'student' processing past frames; and a 'teacher' processing both past and future frames, enabling a broader temporal context. During training, the teacher guides the student to learn future context by only observing past frames. The strategy is evaluated on ROAD dataset for the action prediction downstream task using 3D-ResNet, Transformer, and LSTM architectures. The experimental results showcase significant improvements in prediction performance across these architectures, with our method achieving an average enhancement of 9.9% Precision Points (PP), which highlights its effectiveness in enhancing the backbones' capabilities of capturing long-term dependencies. Furthermore, our approach demonstrates efficiency in terms of the pretraining dataset size and the number of epochs required. This method overcomes limitations present in other approaches, including the consideration of various backbone architectures, addressing multiple prediction horizons, reducing reliance on hand-crafted augmentations, and streamlining the pretraining process into a single stage. These findings highlight the potential of our approach in diverse video-based tasks such as activity recognition, motion planning, and scene understanding. Code can be found at https://github.com/IzzeddinTeeti/ssl_pred.*

1. Introduction

Computer vision techniques have advanced to the point at which they are able to outperform humans at certain object recognition tasks [55]. However, for many computer vision applications, a higher-level understanding of the scene is required. For example, achieving human-level performance in autonomous vehicles remains a formidable challenge [47]. One of the key reasons for this gap is the inherent difficulty in understanding what may happen next. Thus there is a growing recognition of the importance of *prediction*. Prediction plays a crucial role in enhancing the decision-making process of autonomous systems by anticipating the future behaviour of dynamic elements in the environment, *e.g.* other vehicles, pedestrians, and cyclists - thus ensuring safer operation. Moreover, prediction also facilitates the development of high-level understanding within autonomous systems, enabling more nuanced and contextually appropriate planning, leading to smoother interactions with other agents on the road [30].

However, prediction poses its own set of challenges, encompassing spatial, temporal, social, and stochastic dimensions [50]. Modelling these dimensions requires complex models, such as [43, 45, 61, 37], which require significant amounts of data - often scarce and costly to gather and annotate. To address this, leveraging the abundance of unlabelled data through self-supervised methods offers an enticing opportunity to enhance performance with minimal impact upon resources.

While existing self-supervised prediction methods, including [54, 63, 27], have shown promise, they have limitations that hinder their effectiveness. Firstly, their predictive capability is limited to a very short-term horizon (typically one frame ahead) which is impractical for autonomous driving scenarios requiring longer-term predictions. Secondly, these methods often involve a two-stage process [54], which is computationally expensive and time-consuming. Finally, they are typically designed for a specific architecture, lacking the ability to generalize across different architectures.

In this paper, we present a novel one-stage self-supervised representation learning strategy specifically de-

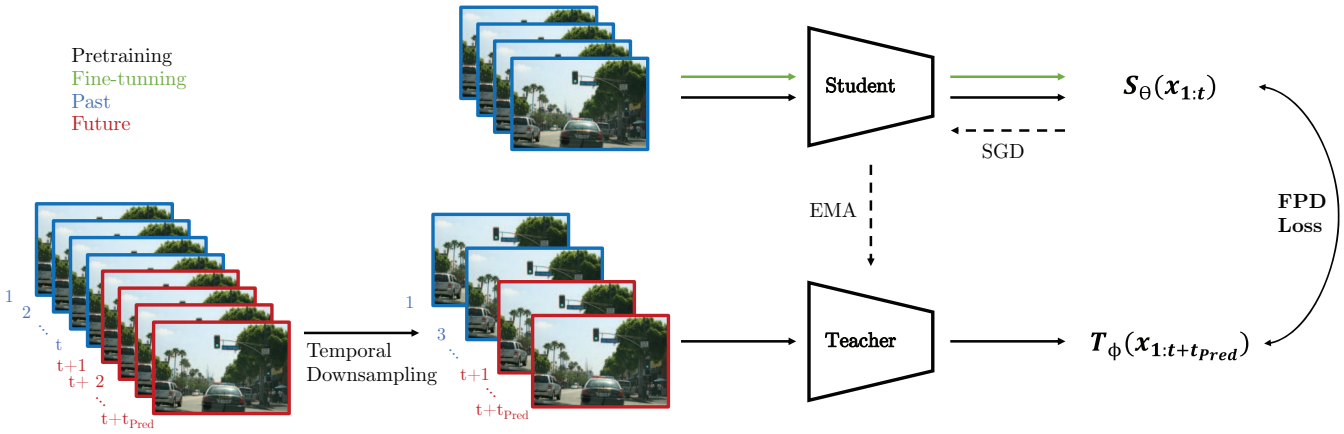


Figure 1: Overview of the proposed Temporal DINO. The *student* model processes the past frames ($x_{1:t}$), while the *teacher* processes both the past and future frames ($x_{1:t+t_{Pred}}$) . A Future-past Distillation loss is applied to their representations (S_θ and T_ϕ) to guide the *student* to capture the future temporal context from the *teacher*.

signed for videos. Our proposed approach draws inspiration from the image-based self-supervised DINO [7] model and extends its application to the temporal dimension. Leveraging a student-teacher framework, our method guides the *student* model to focus on the most informative (temporal) features, enabling accurate predictions of future events. The *student* model learns to attend to the relevant cues in the past and current moments, extracting valuable information that aids in forecasting forthcoming actions. Our approach addresses the aforementioned limitations by significantly extending the prediction range beyond one frame, enabling more practical and effective autonomous driving applications. Moreover, our proposed method eliminates the need for a two-stage process, reducing computational complexity and saving valuable time. Finally, our approach is not limited to a single architecture but is a wrapper strategy that can be used to improve prediction performance across various architectures, enhancing the generalisability and applicability of the method. The contributions of this work include:

- A novel one-stage self-supervised representation learning strategy for videos, addressing limitations in prediction range, computational complexity, and architectural generalisability.
- Proof-of-concept evaluation of the approach in both autonomous driving, and a more general task - with various deep-learning architectures (including 3D-CNN, Transformer, and LSTM), showcasing its effectiveness on real-world challenging data, and versatility to use on other models and domains.
- Identification of the optimal model architecture and loss function for capturing long-range dependencies in video data, shedding light on the most effective design choices for robust action prediction in autonomous driving scenarios.

In the following sections, we provide a comprehensive overview of related work in self-supervised learning for images and videos, and action prediction (Section 2), followed by a detailed description of our proposed method (Section 3). We then present the experimental setup, including datasets, evaluation metrics, and training details (Section 4), and discuss the results and analysis of our experiments (Section 5). Finally, we provide a thorough discussion of our findings, and highlight potential applications and future research directions (Section 6).

2. Related Work

2.1. Image-based Self-supervised Learning

Two types of self-supervised strategies are commonly employed in computer vision: pretext methods and contrastive learning. The former involves utilizing specific internal properties or tasks of the input data to learn useful representations. For instance, some models learn the contextual information of the entire scene by analyzing small image patches [10]. Other approaches focus on tasks like re-colourization of grayscale images [64], predicting transformations between different views of the same image [14], or determining the order of patches extracted from an image [38, 35]. With the introduction of Vision Transformer (ViT) [11], patch-based self-supervised methods have gained prominence in the literature [21, 3, 2]. Particularly, masked auto-encoders (MAE) [21] have emerged as a preferred mechanism for model pretraining due to their superior performance on downstream tasks.

Contrastive learning methods, on the other hand, focus on maximizing the dissimilarity between features extracted

from different samples to encourage discriminative representations [12, 5, 58]. Some discriminative self-supervised algorithms leverage instance-level discrimination to create distinct feature representations for different examples, leading to robust feature learning [22, 9, 17, 7]. Other approaches draw inspiration from clustering mechanisms [6, 1]. The DINO method [7], for example, employs self-distillation that trains a *student* model on local crops and a *teacher* model on the entire image, then find the loss between both representations. The *teacher* will push the *student* to learn global representations by seeing only the local ones. Whilst these techniques offer significant potential on images, their performance is limited on video-related tasks, such as action prediction. Thus, there is a need for advancements in video self-supervised learning methods.

2.2. Video-based Self-supervised Learning

Video-based self-supervised learning strategies, akin to their image counterparts, encompass both pretext and contrastive approaches [46]. However, videos introduce an additional dimension, namely the temporal dimension. The inherent order of frames within a video sequence offers intrinsic properties that can be leveraged to develop effective self-supervised learning mechanisms. Despite this potential, the research in this specific domain remains relatively limited [51, 4, 36, 57].

To address this gap, recent works have explored different strategies in video-based self-supervised learning. For instance, [20] employed complementary information from RGB and optical flow streams to co-train their model using contrastive loss. [18] utilized a 3D-CNN architecture with a video transformer and trained the model using contrastive loss on positive pairs of video sequences. CVRL [42] employed a discriminative learning approach, utilizing two augmented views of the same video as positive examples and views of other videos as negatives. Another method, VideoMoCo [39], extended the MoCo [22] framework to videos by randomly dropping frames from the video and learning the same representation for each random input.

In line with these advancements, our proposed model introduces a novel approach where the *teacher* model incorporates a longer temporal sequence than the *student* model. This design choice aims to provide the *student* with a wider temporal context, encompassing future frames that the student has not yet observed. By leveraging this extended temporal context within the student-teacher framework, our model aims to enhance the *student*'s ability to learn representations that capture long-term dependencies and improve its performance in video-based self-supervised learning.

2.3. Supervised Action Prediction

Solving the action prediction task requires modelling the different dimensions of the problem, including the spatial, temporal, stochastic, and social dimensions, since the driving environment is dynamic, uncertain, and multi-agent. To model the temporal dimension, [13, 29] used 3D-CNN, [43, 16] Recurrent Neural Networks, while [61] used Transformers. Regarding the stochastic dimension, GANs [28], and CVAE models [16, 61] were utilised. To model the multi-agent aspect of the problem, different types of Graph Neural Networks of different connectivity, sparsity and homogeneity were used [15, 45, 16], semantic segmentation cues [44, 34] and social pooling [40]. However, all of those methods are supervised; they require a huge amount of labelled data which is time-consuming and expensive.

2.4. Self-supervised Action Prediction

In recent studies, limited approaches have been explored in the field of self-supervised action prediction. Zatsarynna et al. [63] employed the contrastive loss of InfoNCE [19] to encourage proximity between temporally adjacent video clips in the embedding space. To preserve the order of the clips, they complemented the InfoNCE loss with an order loss in the form of cross-entropy. Their proposed model focused on using 3D-CNN as the backbone architecture, and its evaluation was limited to this specific backbone without examining the performance on other architectures.

Another approach by Kochakarn et al. [27] utilized graph contrastive learning with the SimCLR loss function [9] to learn more informative embeddings for the prediction task. They incorporated an attention mechanism to achieve explainable action prediction, albeit limited to predicting only the next one frame. In contrast, our proposed method extends the prediction horizon to include the next 3, 6, and 12 frames, offering a more comprehensive temporal context. Furthermore, their approach specifically focused on graphs, while our method is designed to work with 3D-CNN, Transformers, and LSTMs, providing broader applicability.

It is worth noting that both of these aforementioned approaches rely heavily on the use of contrastive loss, which requires careful crafting of augmentations. This dependency on specific augmentations can limit the generalisability and robustness of the learned representations.

In contrast, Tran et al. [54] adopted a knowledge distillation approach, where a *teacher* model trained on recognition tasks transfers its knowledge to a prediction model. However, this method follows a two-stage process involving training a *teacher* model for recognition and then distilling the knowledge into a *student* model for prediction. This two-stage approach introduces additional time and computational costs, which may impact the scalability and practicality of the method, particularly in real-time or resource-constrained scenarios. Furthermore, their evaluation was focused on the I3D architecture [8] without exploring the performance on other architectures.

These related works provide valuable insights into self-supervised action prediction approaches. However, they have certain limitations in terms of the backbone architectures considered, the prediction horizons addressed, the reliance on specific augmentations, and the two-stage training process. In contrast, our one-stage proposed method aims to overcome these limitations by leveraging a different loss formulation and considering multiple backbone architectures while extending the prediction horizon, thereby contributing to the advancement of self-supervised action prediction techniques.

3. Methodology

3.1. Representation Learning for Action Prediction

In this study, our objective is to learn a mapping function f in an unsupervised manner, which takes an unlabelled and untrimmed video clip consisting of $t \in \mathbb{R}^+$ frames as input. The goal is to map the 4D input clip $x_{1:t} \in \mathbb{R}^{T \times C \times H \times W}$, to a feature vector $g(x_{1:t}) \in \mathbb{R}^d$, such that the learned features effectively transfer to the downstream task of action prediction. To achieve this, we draw inspiration from the DINO approach and adopt a student-teacher setup, as depicted in Figure 1.

The *student* network S_θ processes only the past frames $x_{1:t}$ during both training and inference, without access to future frames. On the other hand, the *teacher* network T_ϕ processes both the past and future frames $x_{1:t+t_{Pred}}$, where t_{Pred} denotes the length of the future sequence. To ensure consistency in architecture, we downsample the sequence processed by the *teacher* network to match the number of frames processed by the *student*. The downsampling is performed by determining a sampling frequency, which is calculated as $(t+t_{Pred})/t$. For instance, if $t = 12$ frames and $t_{Pred} = 12$ frames, the *student* processes the past (12) frames, while the *teacher* processes (24 frames) with a step of 2, resulting in 12 frames. Despite processing sequences of the same sequence length, the *teacher* network has access to a wider temporal context.

3.2. Future-past Distillation Loss

During training, we introduce a knowledge distillation loss between the final embeddings of the *student* and *teacher* networks. This loss guides the *student* to distil knowledge about the future from the *teacher*, despite not having direct access to future frames. The aim is to teach the *student* to focus on the most relevant features from the past frames that contribute to predicting the future. In contrast to DINO, our Future-Past Distillation (FPD) loss is defined in the Cosine Similarity form, instead of using cross-entropy. This formulation is motivated by the findings of our ablation analysis, which indicate that the Cosine-based FPD loss yields improved performance on downstream tasks. The learning objective for the pretraining stage of future-past distillation is expressed in Equation 1. The *student* network parameters θ are updated using backpropagation optimized by stochastic gradient descent (SGD), while the *teacher* network parameters ϕ are updated using an Exponential Moving Average (EMA) based on the *student* network, with a scheduled momentum variable (m) as shown in Equation 2.

$$\theta^*, \phi^* = \arg\min_{\theta,\phi} \mathcal{L}_{FPD}\Big(S_\theta(x_{1:t}), T_\phi(x_{1:t+t_{Pred}})\Big) \quad (1)$$

$$\phi_{i+1} = m_i \times \phi_i + (1 - m_i) \times \theta_i \quad (2)$$

3.3. Downstream Task Definition

The objective of pretraining is to enhance the performance of the model in the downstream task of predicting driver's actions. Given a past (observed) clip of length $t \in \mathbb{R}^+$, the task is to predict the Ego-vehicle (driver) action in each frame in the next $t_{Pred} \in \mathbb{R}^+$ frames.

Building upon the optimal pretrained *student* model S_{θ^*}, the prediction model f will use the *student* model as a backbone and add a classification head on top of it. Subsequently, it performs further optimization (fine-tuning) on either both the backbone and the head parameters or solely the latter (we conducted experiments on both scenarios) for the prediction task. The objective is to map the learned features $S_{\theta^*}(X_{1:t})$ to the future action labels $y_{t+1:t+t_{Pred}}$, formally, $f_\psi : S_{\theta^*}(X_{1:t}) \rightarrow y_{t+1:t+t_{Pred}}$. Given the nature of a classification task, cross-entropy (CE) loss is utilized to guide the optimization process and refine the learning objective for the downstream task, as illustrated in Equation 3.

$$\theta^{**}, \psi^* = \arg\min_{\theta^*,\psi} \mathcal{L}_{CE}\left(f_\psi\Big(S_{\theta^*}(x_{1:t})\Big), y_{t+1:t+t_{Pred}}\right) \quad (3)$$

4. Experiments

4.1. Datasets

We adhered to the convention in self-supervised learning, utilizing two distinct datasets for distinct purposes: a larger dataset for pretext task pretraining and a smaller dataset for downstream task fine-tuning. Specifically, we employed the Kinetics-400 dataset [25] and the ROad event Awareness Dataset (ROAD) [48].

Kinetics-400 It is designed for action recognition and comprises over 240,000 videos. On average, each video spans 10 seconds and is assigned a single label from a pool of 400 possible action classes. The dataset's substantial video collection has facilitated its adoption in numerous

video self-supervised methods [51, 39]. For our purposes, we disregard the label information as it will not be used during pretraining.

ROAD The ROad event Awareness Dataset (ROAD) [48] is built on a fraction of Oxford RobotCar Dataset [33], and it is extended with multi-label annotations for action recognition, localisation, and prediction tasks within the context of autonomous driving. It comprises 22 videos from an egocentric view as shown in Figure 2, each with an 8-minute duration and a frame rate of 12 frames per second (fps). Importantly, it contains labels indicating the actions performed by the ego vehicle (driver). Notably, the dataset encompasses seven distinct ego-vehicle actions: *Move, Stop, Turn Left, Turn Right, Overtake, Move Left, and Move Right*. These labels serve as training data for fine-tuning our models to predict the driver's actions in future frames during the prediction task.

4.2. Models

We conducted a series of experiments utilizing four diverse deep-learning architectures.

R3D [52] This architecture employs a 3D convolutional neural network (3D-CNN) backbone for processing video data. It is based on the ResNet-18 [23] architecture, which has proven to be successful in image recognition tasks.

Swin [31] This architecture utilizes a Transformer-based 3D backbone for video processing. It is based on the Transformer architecture, originally introduced in the context of image recognition as Vision Transformer (ViT) [11].

ResNet-LSTM This architecture combines a convolutional neural network (CNN)-based 2D backbone using the ResNet-50 architecture for image processing, along with a Long Short-Term Memory (LSTM) layer for capturing temporal dependencies.

ViT-LSTM Same as the previous but replacing the ResNet with a Transfomer-based 2D backbone.

The first two models use spatio-temporal backbones, which extract spatial and temporal features simultaneously, while the last two use a spatial backbone and connect it using a temporal one. These models exhibit differences in their depth and working mechanisms, offering a diverse range of approaches to our experiments. By leveraging these varied architectures, we can comprehensively investigate our proposed representation learning strategy, examine their performance, and compare their effectiveness according to the experiments outlined in the subsequent section.

4.3. Experimental Protocol

In accordance with the conventions of self-supervised learning experiments, we evaluated the performance of our models on the downstream task of action prediction under three distinct protocols:
1) *Full-Supervised:* This protocol represents the results obtained from the model without employing the pretraining strategy. The model was trained solely on the labelled data of the prediction task.
2) *Linear Probing:* In this protocol, the proposed strategy was applied, and the model was fine-tuned by solely updating the parameters of the prediction head. The backbone of the model was kept frozen during this fine-tuning process.
3) *Fine-tuning:* This protocol involved applying the proposed strategy and performing full fine-tuning of the entire network, including both the backbone and the prediction head. All parameters of the model were updated during the fine-tuning process.

By examining the model's performance across these three protocols, we can gain insights into the effectiveness and impact of the pretraining strategy on action prediction.

4.4. Implementation Details

We used Pytorch framework [41] to implement the models, and the Precision (P) metric was used to evaluate their performance on the action prediction task. All experiments were conducted on NVIDIA A30 graphics cards. Below are the details of the pretraining and fine-tuning.

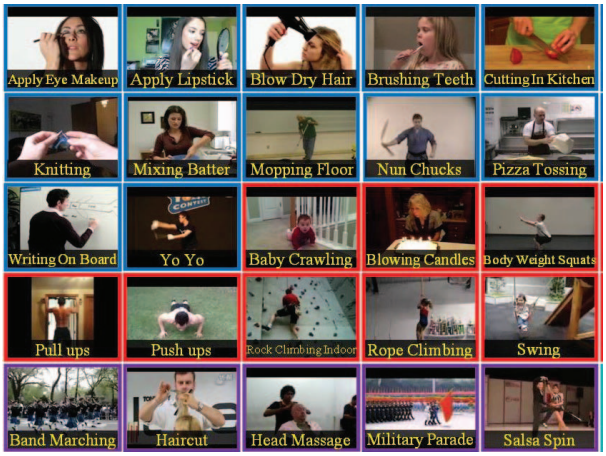
Pretraining For pretraining, we utilized either the full dataset of Kinetics-400 or the training split of ROAD (depending on the experiment). The pretraining was optimised for 1000 (50) epochs using SGD optimiser with a learning rate of 0.005 (0.001) and a batch size of 64 (32) for Kinetics-400 and ROAD, respectively. Additionally, a cosine scheduler was employed for updating the momentum variable of the exponential moving average (EMA). Concerning the models, the LSTM architecture had a hidden dimension of 512, and the small structure of ViT (ViT-s) was used.

Fine-tuning Fine-tuning was performed on the ROAD dataset, with a data split of 60% for training, 20% for validation, and 20% for testing. Cross-entropy loss was employed as the objective function for fine-tuning, and it was optimised for 10 epochs using the SGD optimizer with a learning rate of 0.001, and a batch size of 32.

4.5. Ablations

In this section, we present different ablations to study the effects of different architectural components on the model

(a) ROAD dataset



(b) UCF101 dataset

Figure 2: Sample images from ROAD and UCF101 datasets.

performance. Specifically, we investigate the effects of variations in input sequence temporal length, self-supervised objective, backbone selection, and the strategy's performance on another downstream task (action recognition). Detailed explanations of each ablation study are provided below.

Backbone and Temporal Length The choice of model backbone plays a critical role in the model's ability to learn effective features. Similarly, the length of the input sequence contributes to the model's capability to capture temporal dynamics and visual changes within the scene. Longer sequence lengths generally provide more visual data, but they may also introduce challenges in modelling long-term dependencies and potentially lead to performance degradation, as shown in Table 1.

In this specific ablation study, We conducted experiments with the four backbones mentioned in Section 4.2, and we varied the temporal depth by using sequence lengths of 3, 6, and 12 frames for each backbone. We trained these models using the three experimental protocols outlined in Section 4.3. The prediction performance of the four models with different input lengths under the three protocols is summarized in Table 1.

Loss Function Existing literature demonstrates that the choice of learning objective in self-supervised algorithms significantly influences the performance of models on downstream tasks [46]. In our experiments, we employed three widely used loss functions: Cross-entropy, Cosine Similarity, and Mean Squared Error (MSE) loss. The evaluation of these loss functions was performed on R3D and Swin backbones, and the corresponding results are summarized in Table 2.

Action Recognition In addition, we performed supplementary experiments to assess the impact of our proposed strategy on a different video-based downstream task, namely Action Recognition. For this purpose, the models underwent pretraining on the Kinetics-400 dataset and subsequent fine-tuning on the UCF101 dataset [49], shown in Figure 2. Table 4 presents the state-of-the-art (SOTA) performance for action recognition.

5. Results

Observing Table 1, it is evident that our proposed strategy yielded notable enhancements in the prediction performance of all backbone architectures, albeit to varying extents. Notably, the impact of the strategy is particularly pronounced on the 2D-based backbones, which initially exhibited comparatively lower results in the fully-supervised setup. This observation aligns with expectations since these backbones lack inherent spatio-temporal modelling capabilities. However, when trained with T-DINO, the ResNet-LSTM and ViT-LSTM backbones witnessed substantial improvements, with average increases of 22.5 and 10.1 in terms of precision points (PP), respectively. Additionally, the video backbones, R3D and Swin, experienced PP gains of 5.2 and 1.8, respectively. Table 3 shows the comparison of our strategy to other SOTA methods, showing that it surpasses the performance of both supervised and self-supervised approaches on action prediction on ROAD. Within the same table, the column 'Supervised' indicates a fully supervised method, the opposite denotes a self-supervised approach, and 'Frozen' means the fine-tuning process exclusively updates the classification head while leaving the pre-trained backbone untouched, the opposite indicates that the entire model was updated during the fine-tunning.

Table 1: The precision of different backbones with varying input lengths under the three protocols mentioned in Section 4.3.

Backbone	Pretrained on	Interval (frames)	Linear Probe	Fine-tunning	Supervised	Improvement
R3D	ROAD	3	36.6	77.7	74.1	3.6
		6	29.7	69.2	64.9	4.3
		12	36.2	53.3	45.7	7.6
Swin	ROAD	3	84.7	87.2	86.4	0.8
		6	75.8	82.6	81.7	0.9
		12	60.1	60.9	57.1	3.8
ResNet+LSTM	Kinetics-400	3	77.6	84.6	62.9	21.7
		6	70.3	81.8	58.3	23.5
		12	58.3	76.7	54.3	22.4
ViT+LSTM	Kinetics-400	3	73.5	77.7	69.3	8.4
		6	70.2	76.7	65.5	11.2
		12	64.21	66.2	55.5	10.7

Table 2: Evaluation of three common loss functions on R3D and Swin backbones.

Backbone	Loss	Linear Probe	Fine-tunning	Supervised	Improvment
R3D	MSE	32.0	37.3	45.7	-8.4
	Cosine	36.2	**53.3**	45.7	7.6
	Cross-entropy	30.9	50.4	45.7	4.7
Swin	MSE	57.2	57.4	57.1	0.3
	Cosine	60.1	**60.9**	57.1	3.8
	Cross-entropy	49.6	58.4	57.1	1.3

To have a better understanding of the generalisation capability of the proposed model, we compare the performance of T-DINO with SOTA methods on another task, human action recognition (on UCF101), summarised in Table 4. The results highlight the effectiveness of the enhanced temporal modelling offered by T-DINO when applied to the R3D backbone,

Furthermore, examining the results for varying input sequence intervals across each backbone, it becomes apparent that greater improvements are observed at longer input sequences. This suggests that pretraining with T-DINO equips the backbones with enhanced abilities to capture and model long-term dependencies in the data. Notably, the Swin-transformer-based models demonstrate higher accuracy, attributed to the superior representation capabilities offered by Transformers. We observed that transformer-based spatial feature extractor (ViT) combined with LSTM-based temporal sequence modelling results in a huge improvement in the downstream task. Analyzing the results obtained from the longest input configuration consisting of 12 frames, it is evident that models pretrained on the larger dataset, Kinetics-400, exhibit superior performance compared to those pretrained on the ROAD dataset. Moreover, the models that incorporate separate modelling of spatial and temporal relationships, such as ViT+LSTM and ResNet+LSTM, outperform the models that jointly model these relationships, namely R3D and Swin, by a margin of 16.6 and 5.7 percentage points, respectively.

In regard to the selection of the most optimal loss function, the findings presented in Table 2 indicate that T-DINO pretrained using the Cosine Similarity loss surpasses the performance of models trained with MSE or Cross-entropy loss.

Of significant importance, self-supervised models typically require extensive datasets and a high number of pretraining epochs to achieve satisfactory generalization on downstream tasks. However, our proposed strategy, pretrained on the ROAD dataset, exhibits a relatively comparable level of performance to models pretrained on the larger Kinetics-400 dataset. Notably, the ROAD dataset possessed a substantially smaller size and was pretrained with a lower number of epochs. These findings demonstrate T-DINO's resource and time efficiency in achieving desirable results.

Table 3: Comparison of the proposed methods with SOTA on ROAD for the action prediction task when the input length is 12 frames

Method	Pretrained on	Backbone	Supervised	Frozen	Precision
R3D [53]	Road	R3D	✓	✗	45.7
Swin [32]	Road	Swin	✓	✗	57.1
Resnet+LSTM [62]	Road	ResNet50	✓	✗	54.3
ViT+LSTM	Road	ViT	✓	✗	55.5
VideoMAE [51]	Kinetics	ViT	✗	✗	75.1
T-DINO (ours)	Road	R3D	✗	✓	53.3
T-DINO (ours)	Kinetics	ResNet+LSTM	✗	✓	58.3
T-DINO (ours)	Kinetics	ResNet+LSTM	✗	✗	**76.7**
T-DINO (ours)	Kinetics	ViT+LSTM	✗	✓	64.2
T-DINO (ours)	Kinetics	ViT+LSTM	✗	✗	66.2

Table 4: Comparison of the proposed methods with SOTA on UCF101 for action recognition task.

Method	Year	Pretrained on	Arch.	Supervised	Frozen	Acc.
ClipOrder [59]	2019	UCF101	R3D	✗	✗	72.40
3D ST-puzzle [26]	2019	Kinetics	C3D	✗	✗	65.80
Wang et al. [56]	2019	UCF101	C3D	✗	✗	61.20
PRP [60]	2020	Kinetics	R3D	✗	✗	72.10
SpeedNet [4]	2020	Kinetics	S3D-G	✗	✗	81.10
CSJ [24]	2021	K+UCF101	R(2+3)D	✗	✗	79.50
VideoMoCo [39]	2021	Kinetics	R(2+1)D	✗	✗	78.7
CACL [18]	2022	UCF101	R(2+1)D	✗	✗	82.5
VideoMAE [51]	2022	Kinetics	ViT	✗	✗	**96.1**
T-DINO (ours)	2023	Kinetics	ResNet+LSTM	✗	✗	80.26
T-DINO (ours)	2023	Kinetics	R3D	✗	✗	83.9
T-DINO (ours)	2023	Kinetics	ViT+LSTM	✗	✗	85.86

6. Conclusion

This study represents the first attempt to leverage future information in a 'past training' model, and the promising results indicate that this teacher-student approach could provide a significant performance improvement in various prediction tasks across a number of ubiquitous model architectures in a variety of different domains - without the requirement for any additional training data. Our proposed strategy, called Temporal-DINO, leverages a teacher-student self-distillation architecture to guide the student model to learn future temporal context by observing the past only. Unlike other approaches that involve a two-stage process or rely on hand-crafted augmentations with limited prediction horizons, our one-stage strategy overcomes these limitations. Additionally, ablations highlight the strategy's generalisability, efficiency, and feasibility in hardware-constrained applications.

In terms of future directions, several avenues can be pursued to further enhance and expand our proposed approach. Firstly, the inclusion of Graph Neural Networks (GNNs) as additional architectural variations could be explored to examine the strategy's ability to enhance the social dimension modelling. Secondly, expanding the evaluation of our approach to encompass a broader range of datasets from diverse domains.

References

[1] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019. 3

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *arXiv preprint arXiv:2301.08243*, 2023. 2

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[4] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9922–9931, 2020. 3, 8

[5] Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *International Conference on Machine Learning*, pages 517–526. PMLR, 2017. 3

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 3

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[10] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 5

[12] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014. 3

[13] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. Dada: Driver attention prediction in driving accident scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 3

[14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[15] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Gohome: Graph-oriented heatmap output for future motion estimation. *arXiv preprint arXiv:2109.01827*, 2021. 3

[16] Harshayu Girase, Haiming Gang, Srikanth Malla, Jiachen Li, Akira Kanehara, Karttikeya Mangalam, and Chiho Choi. Loki: Long term and key intentions for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9803–9812, 2021. 3

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 3

[18] Sheng Guo, Zihua Xiong, Yujie Zhong, Limin Wang, Xiaobo Guo, Bing Han, and Weilin Huang. Cross-architecture self-supervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19270–19279, 2022. 3, 8

[19] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Y.W. Teh and M. Titterington, editors, *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9 of *JMLR WCP*, pages 297–304. Journal of Machine Learning Research - Proceedings Track, 2010. 3

[20] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020. 3

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[24] Yuqi Huo, Mingyu Ding, Haoyu Lu, Ziyuan Huang, Mingqian Tang, Zhiwu Lu, and Tao Xiang. Self-supervised video representation learning with constrained spatiotemporal jigsaw. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 751–757. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track. 8

[25] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 4

[26] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8545–8552, 2019. 8

[27] Pawit Kochakarn, Daniele De Martini, Daniel Omeiza, and Lars Kunze. Explainable action prediction through self-supervision on scene graphs. *arXiv preprint arXiv:2302.03477*, 2023. 1, 3

[28] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian D. Reid, Seyed Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *NeurIPS*, 2019. 3

[29] Iuliia Kotseruba, Amir Rasouli, and John K. Tsotsos. Benchmark for evaluating pedestrian action prediction. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1257–1267, 2021. 3

[30] Peng Li, Xiaofei Pei, Zhenfu Chen, Xingzhen Zhou, and Jie Xu. Human-like motion planning of autonomous vehicle based on probabilistic trajectory prediction. *Applied Soft Computing*, 118:108499, 2022. 1

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 5

[32] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 8

[33] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 5

[34] 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*, pages 15233–15242, 2021. 3

[35] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020. 2

[36] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 527–544. Springer, 2016. 3

[37] Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple & efficient attention networks. *arXiv preprint arXiv:2207.05844*, 2022. 1

[38] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 69–84. Springer, 2016. 2

[39] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11205–11214, 2021. 3, 5, 8

[40] Bo Pang, Tianyang Zhao, Xu Xie, and Ying Nian Wu. Trajectory prediction with latent belief energy-based model. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11809–11819, 2021. 3

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[42] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 3

[43] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6261–6270, 2019. 1, 3

[44] Amir Rasouli, Mohsen Rohani, and Jun Luo. Bifold and semantic reasoning for pedestrian behavior prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15600–15610, 2021. 3

[45] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 683–700. Springer, 2020. 1, 3

[46] Madeline C Schiappa, Yogesh S Rawat, and Mubarak Shah. Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 2022. 3, 6

[47] Nirajan Shiwakoti, Peter Stasinopoulos, and Francesco Fedele. Investigating the state of connected and autonomous vehicles: a literature review. *Transportation Research Procedia*, 48:870–882, 2020. Recent Advances and Emerging Issues in Transport Research – An Editorial Note for the Selected Proceedings of WCTR 2019 Mumbai. 1

[48] G. Singh, S. Akrigg, M. Maio, V. Fontana, R. Alitappeh, S. Khan, S. Saha, K. Jeddisaravi, F. Yousefi, J. Culley, T. Nicholson, J. Omokeowa, S. Grazioso, A. Bradley, G. Gironimo, and F. Cuzzolin. Road: The road event awareness dataset for autonomous driving. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(01):1036–1054, jan 2023. 4, 5

[49] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[50] Izzeddin Teeti, Salman Khan, Ajmal Shahbaz, Andrew Bradley, and Fabio Cuzzolin. Vision-based intention and trajectory prediction in autonomous vehicles: A survey. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5630–5637. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Survey Track. 1

[51] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 3, 5, 8

[52] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. *CoRR*, abs/1711.11248, 2017. 5

[53] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 8

[54] Vinh Tran, Yang Wang, Zekun Zhang, and Minh Hoai. Knowledge distillation for human action anticipation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2518–2522. IEEE, 2021. 1, 3

[55] Leonard Elia van Dyck, Roland Kwitt, Sebastian Jochen Denzler, and Walter Roland Gruber. Comparing object recognition in humans and deep convolutional neural networks—an eye tracking study. *Frontiers in Neuroscience*, 15:1326, 10 2021. 1

[56] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4006–4015, 2019. 8

[57] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015. 3

[58] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 3

[59] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 8

[60] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6548–6557, 2020. 8

[61] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 3

[62] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. 8

[63] Olga Zatsarynna, Yazan Abu Farha, and Juergen Gall. Self-supervised learning for unintentional action prediction. In *Pattern Recognition: 44th DAGM German Conference, DAGM GCPR 2022, Konstanz, Germany, September 27–30, 2022, Proceedings*, pages 429–444. Springer, 2022. 1, 3

[64] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016. 2