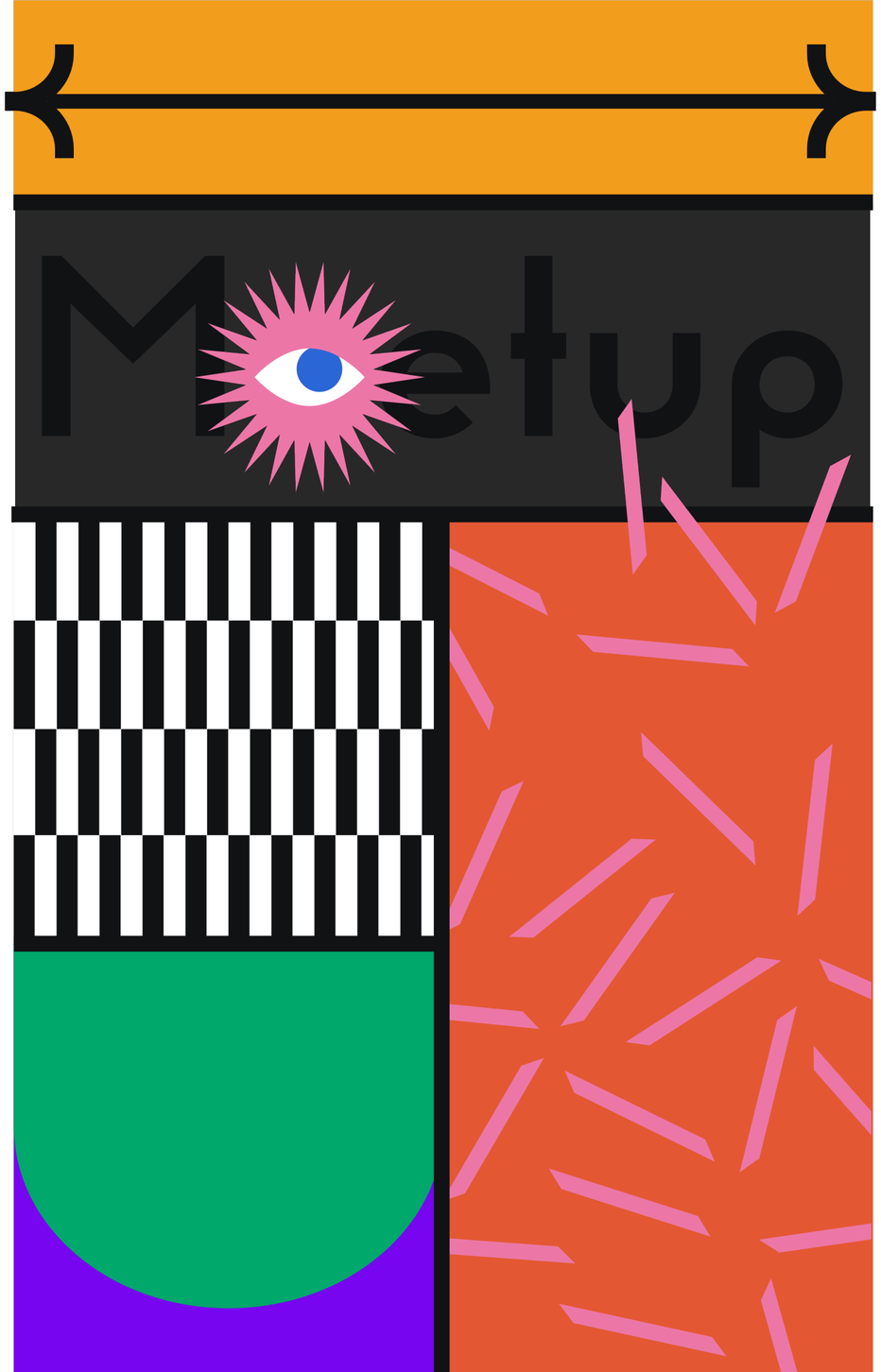
St. Petersburg
Open Data Science

Основы видеоаналитики

Александр
Весельев



Мотивация

- Систематизировать знания в области video
- Познакомить людей с различными задачами
- Подготовить слушателей к практическим применениям алгоритмов video processing

План

Обзор Video-based задач

Некоторые из video задач. Обсуждение, как они решаются

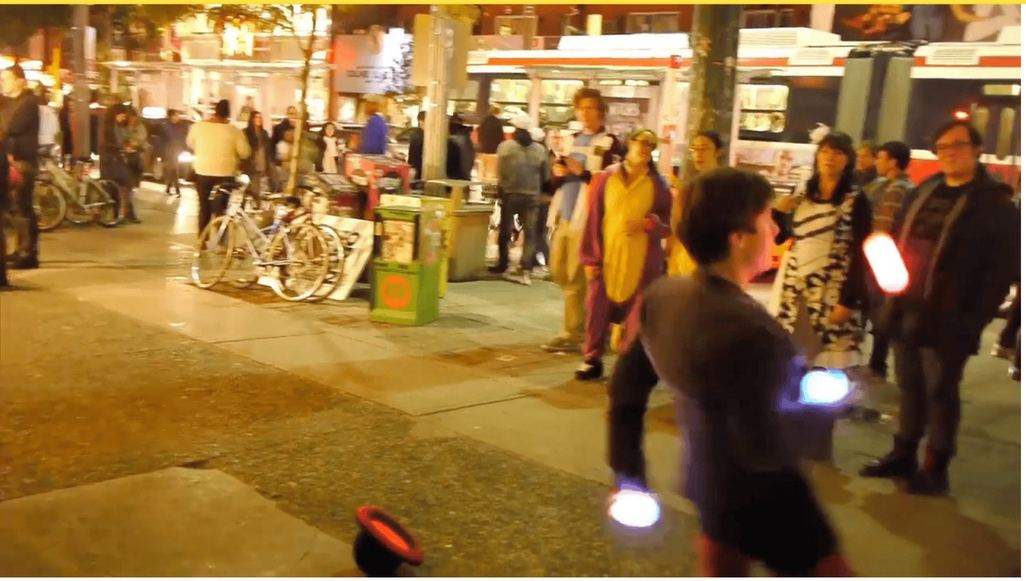
Re-Identification

Что такое задача re-identification, какие есть подходы решения

Object Tracking

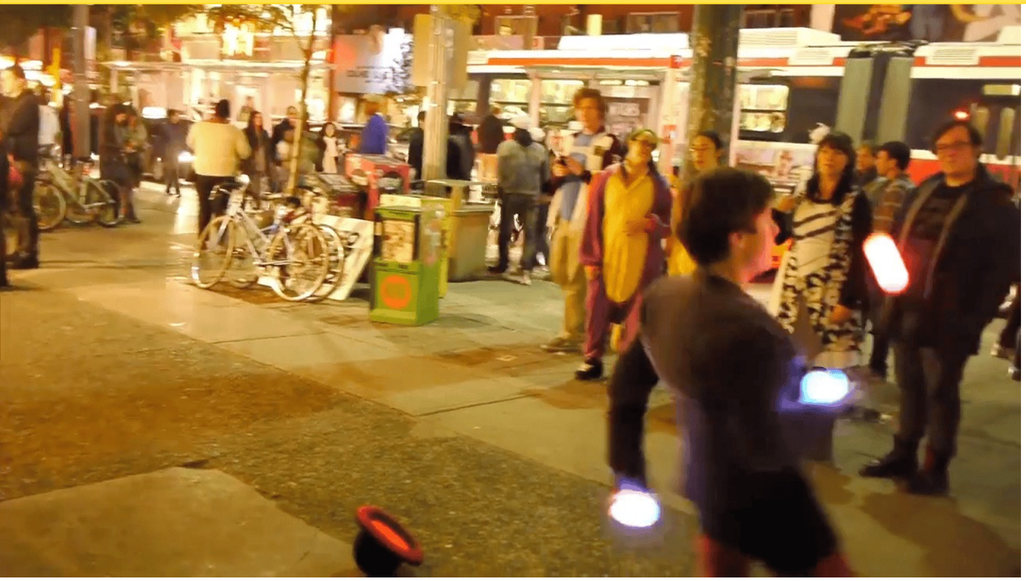
Рассмотрение алгоритмов решения задачи object-tracking

Видео



- Картинка – 3 x H x W
- Видео – 3 x T x H x W

Проблемы

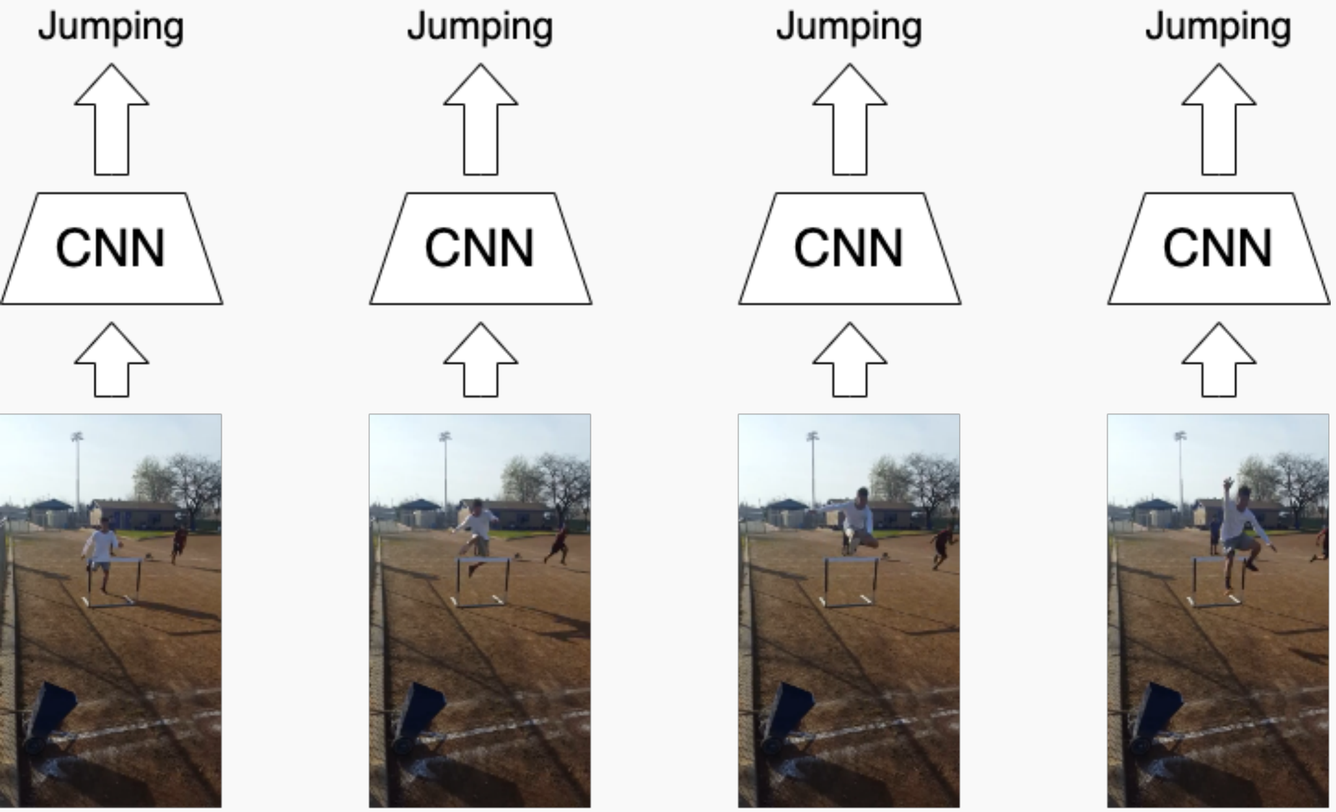


- Стандартные FPS: 24, 30
- SD видео (640x480)
  занимает ~ 1.5GB на 1 минуту
- Решение:
  - Использовать меньшее разрешение (обычно 112x112)
  - Меньший FPS (5-10)
  - Работать с короткими клипами

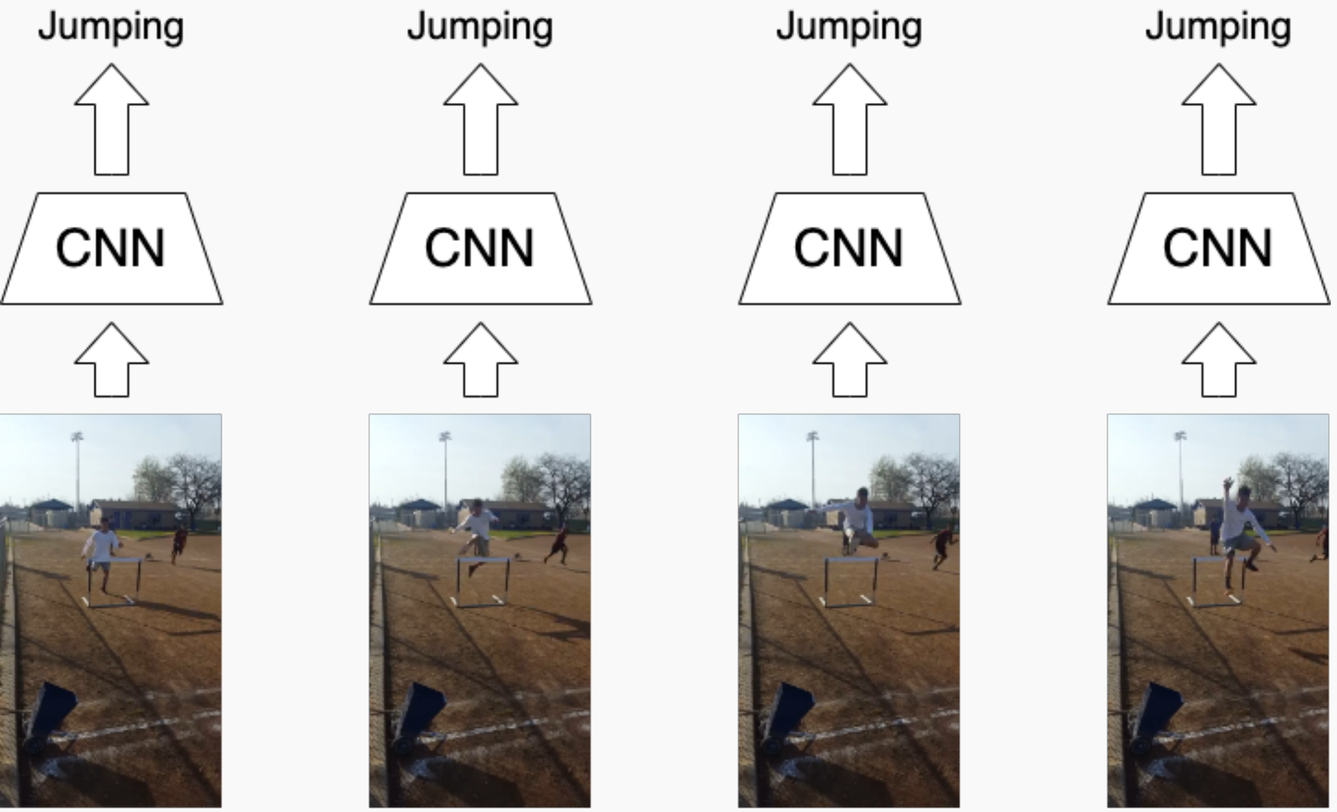
Action Classification

- Вход: Видео, 3-5 секунд
- Выход: распределение вероятностей по множеству действий

Baseline

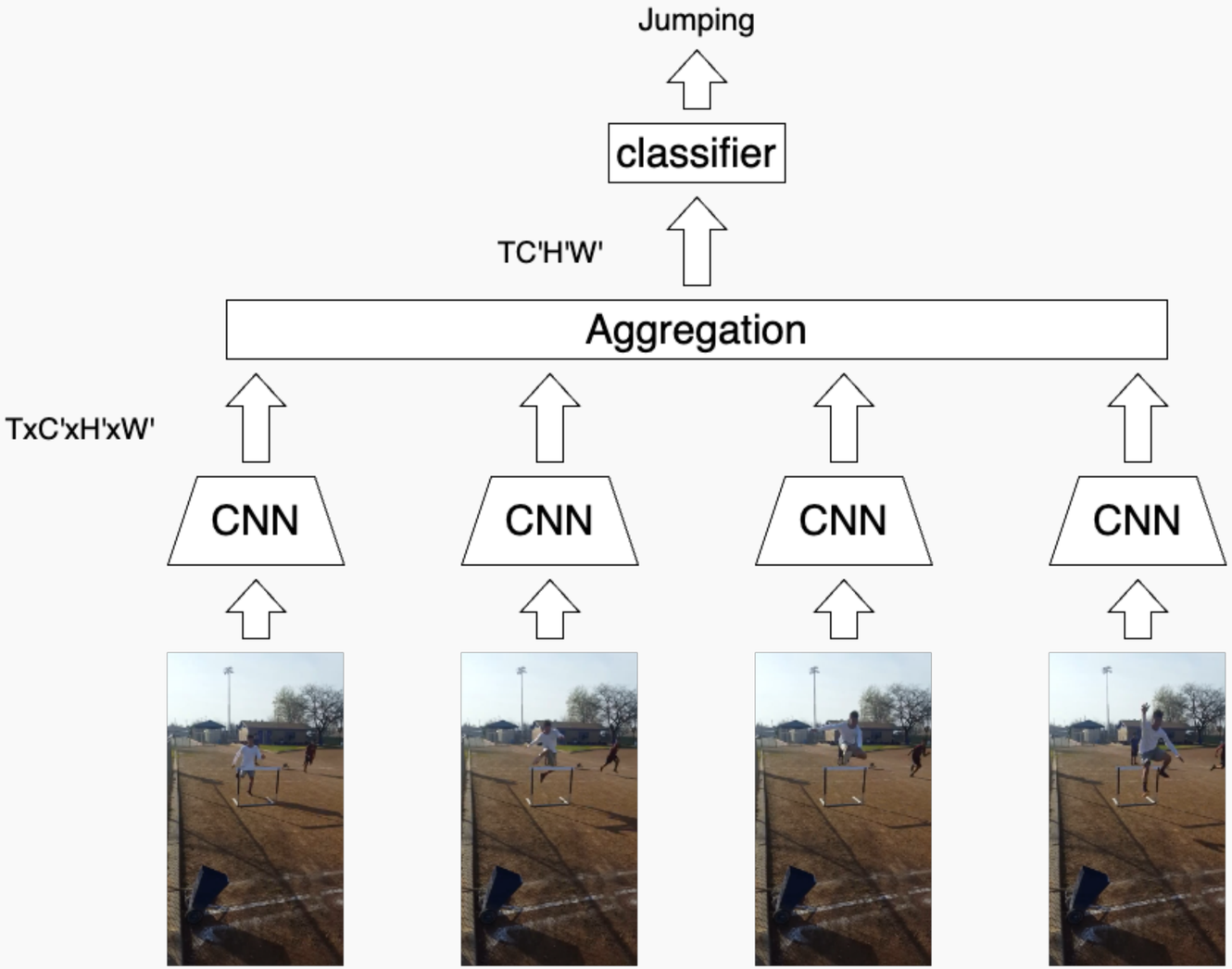


Baseline

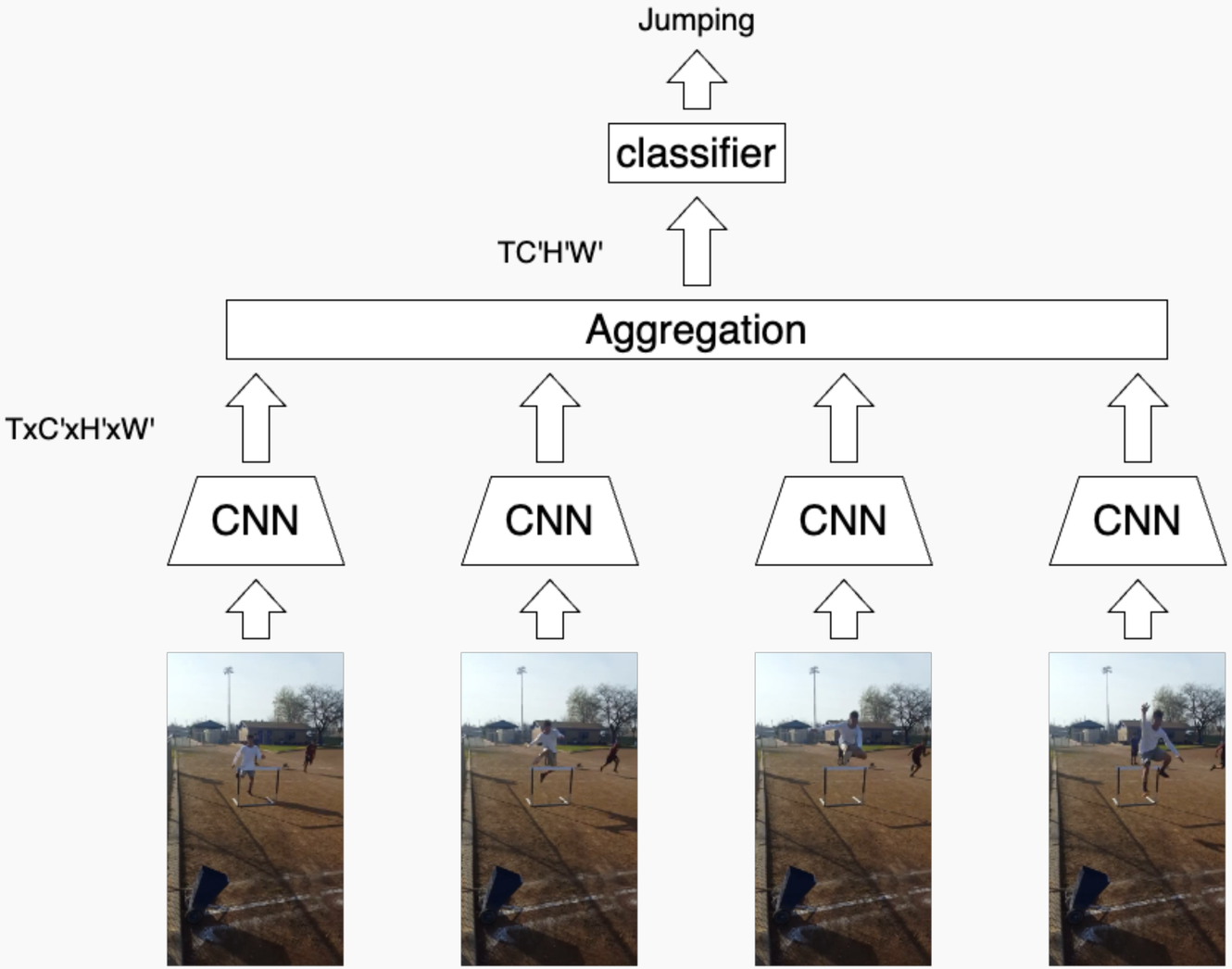


Часто является
сильным бейзлайном

Late Fusion

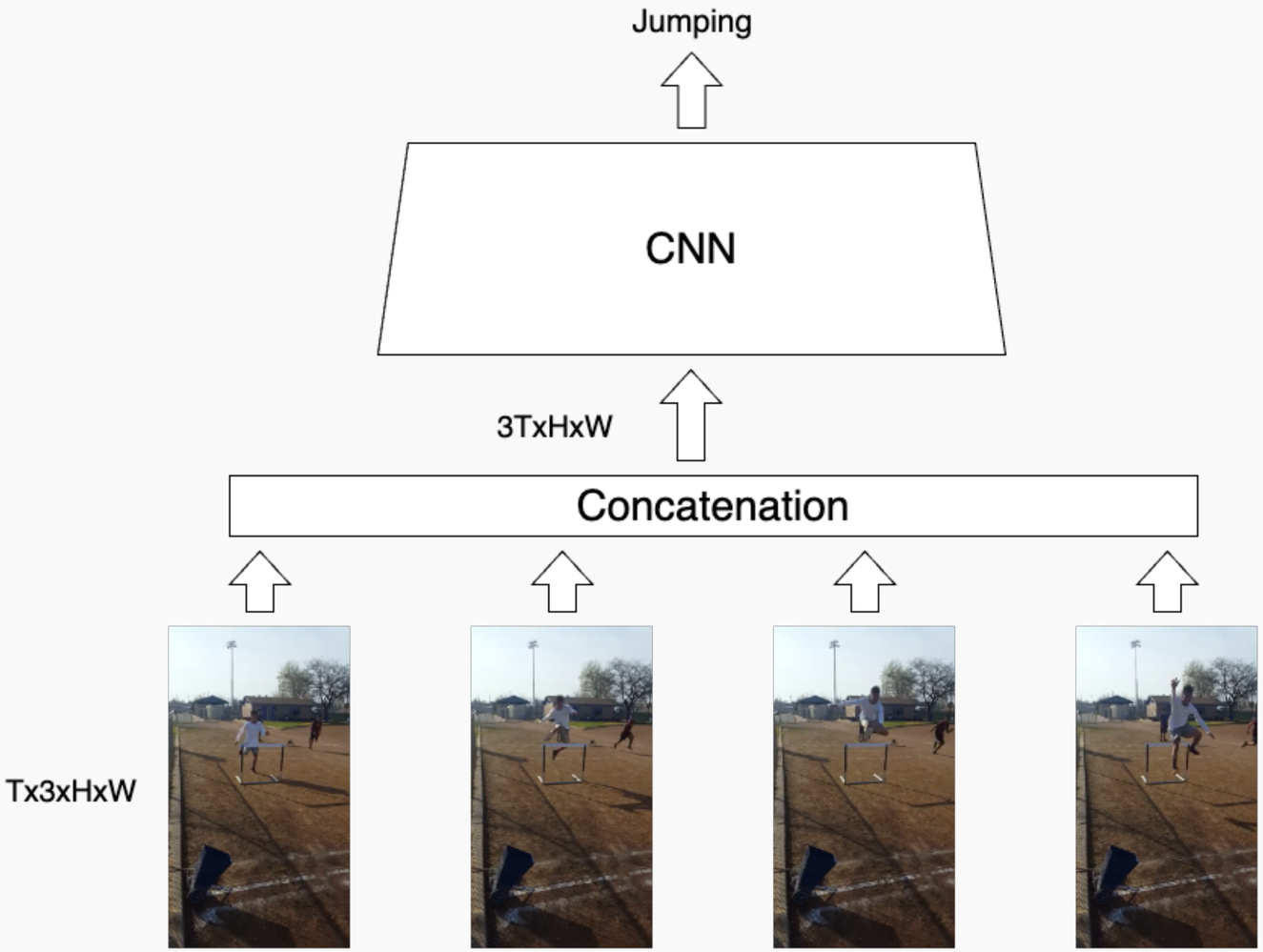

Jumping
classifier
TC'H'W'
Aggregation
TxC'xH'xW'
CNN
CNN
CNN
CNN

Late Fusion

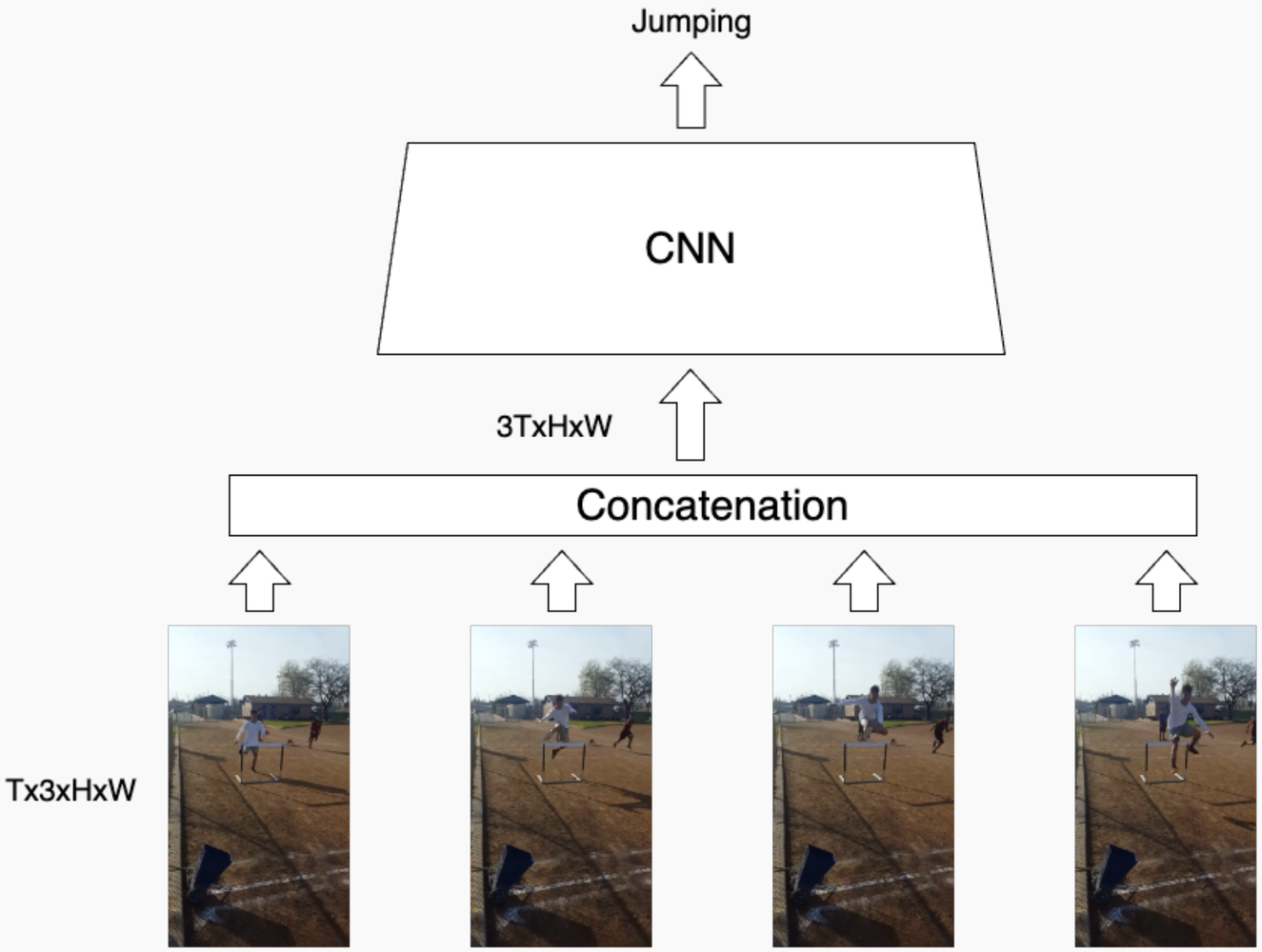


Проблема: сложно улавливать локальные изменения low-level признаков

Early Fusion

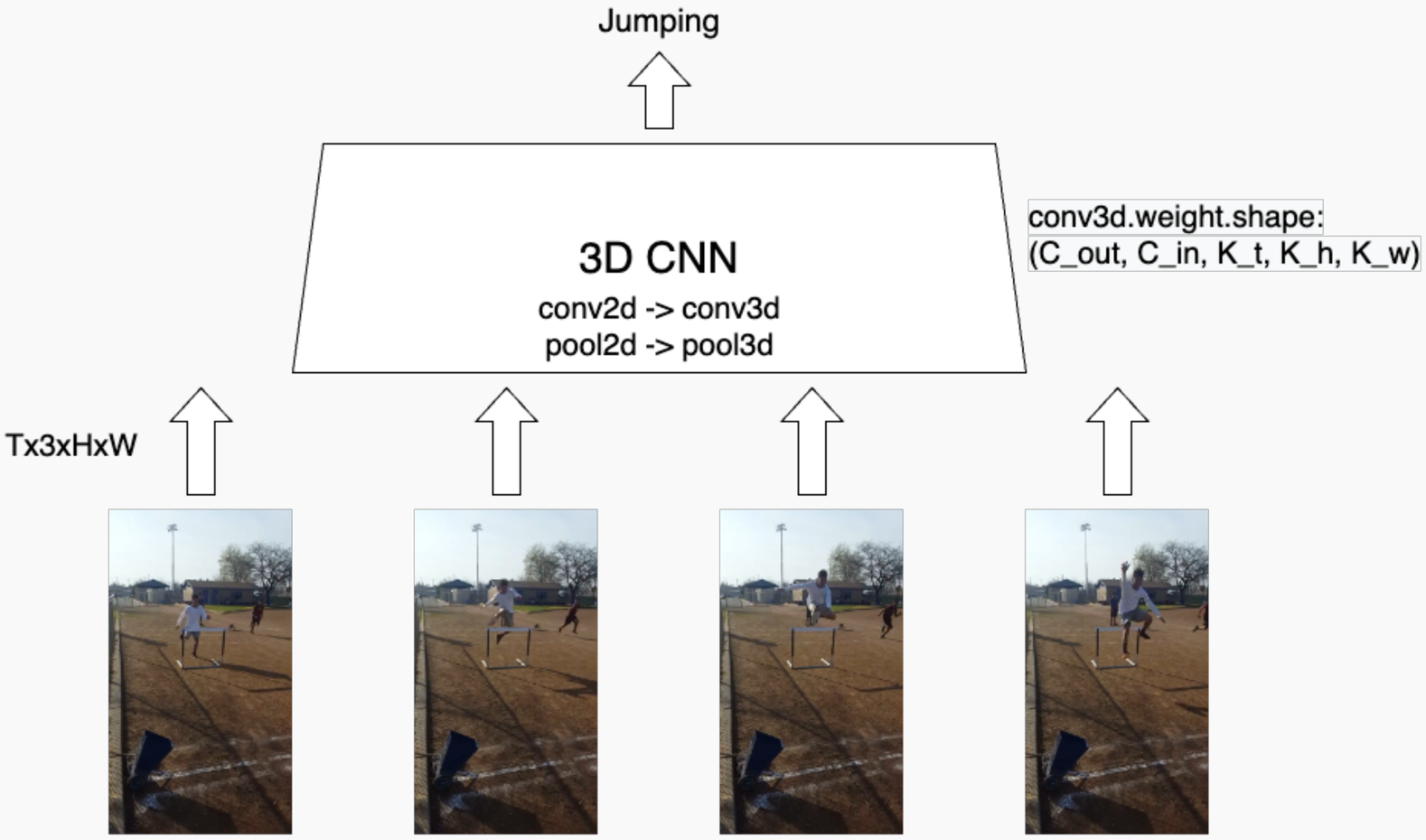


Early Fusion

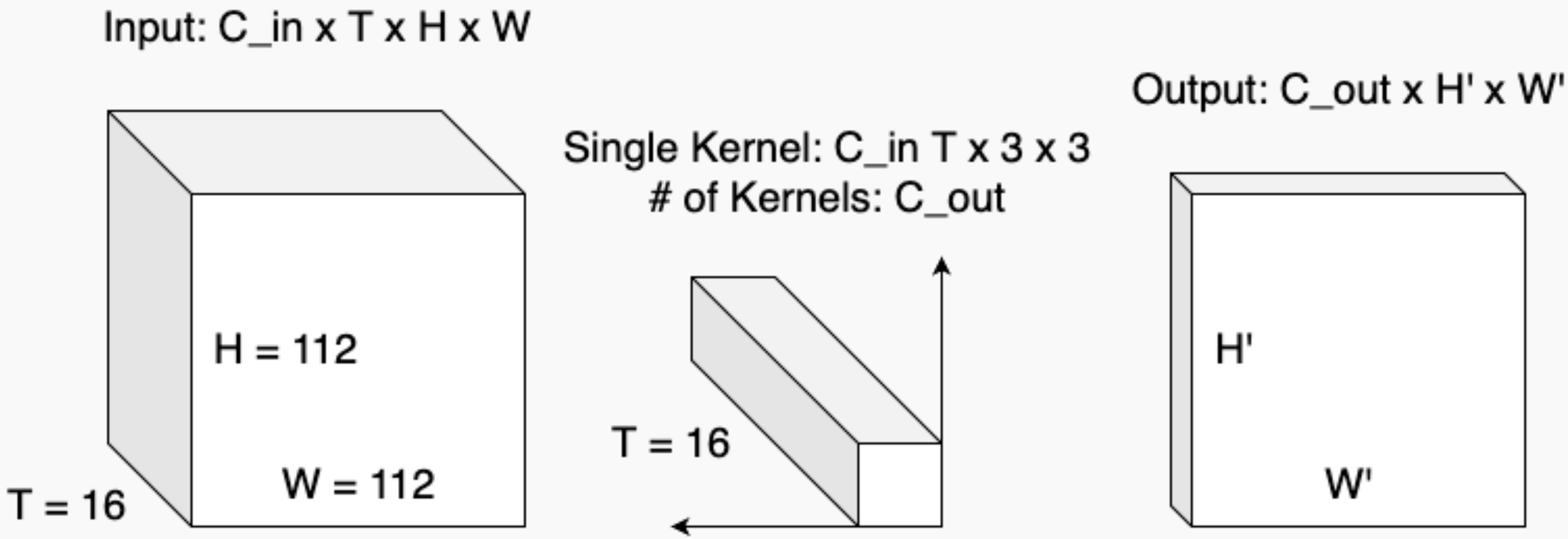


- Проблема: вся temporal агрегация происходит на 1м слое

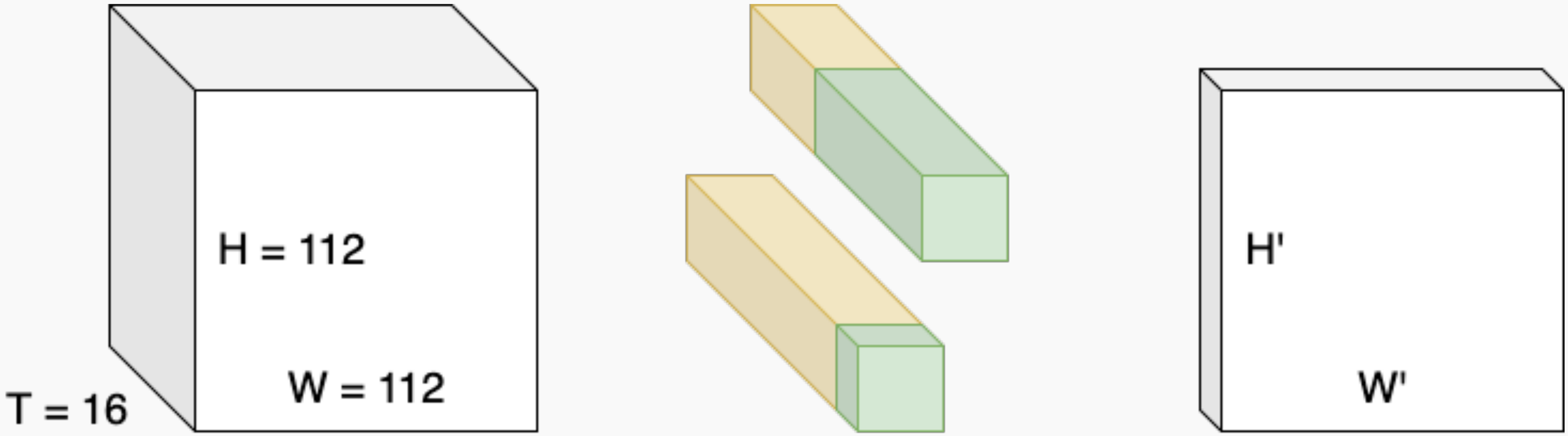
3D-CNN. Slow Fusion



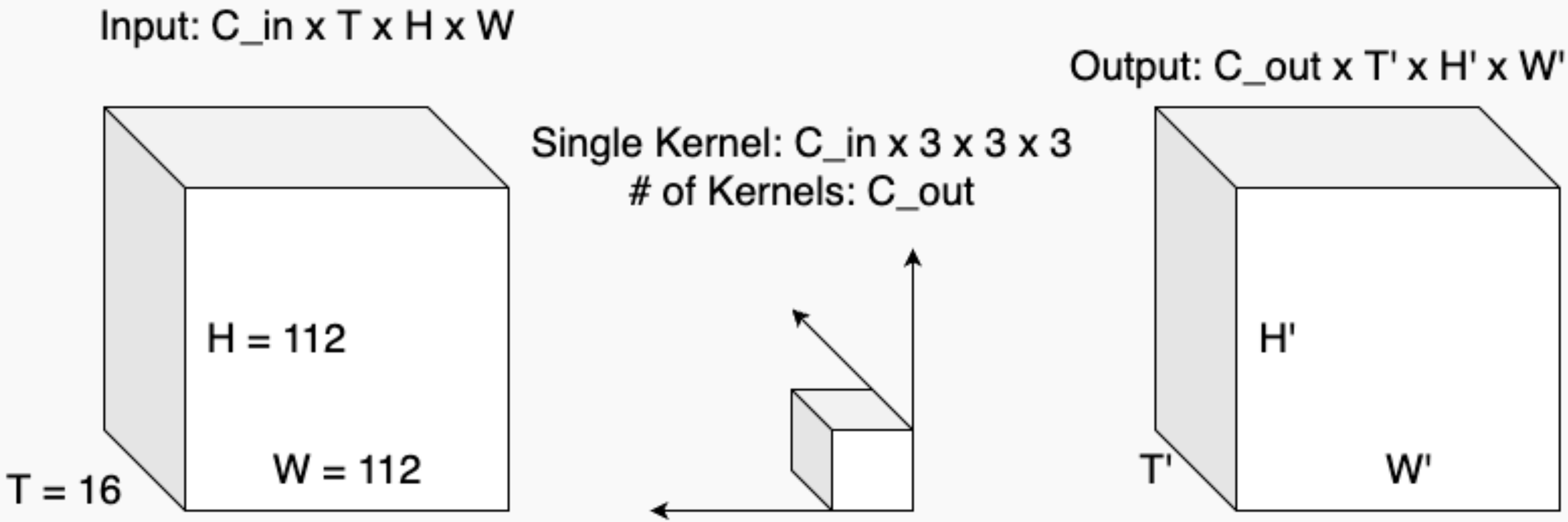
Early Fusion. Example



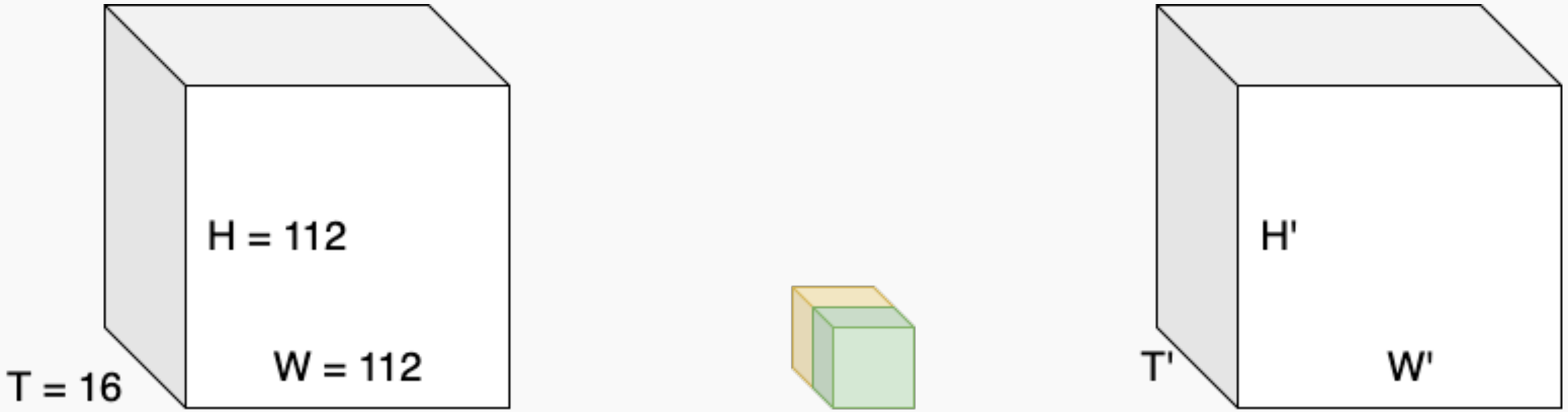
Early Fusion. Temporal Shift-Invariance



3D-CNN. Example



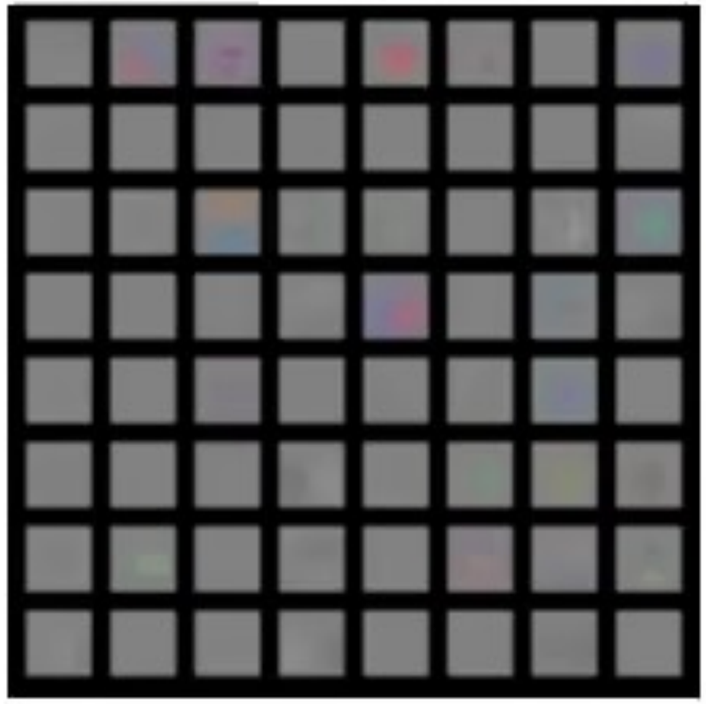
3D-CNN. Temporal Shift-Invariance



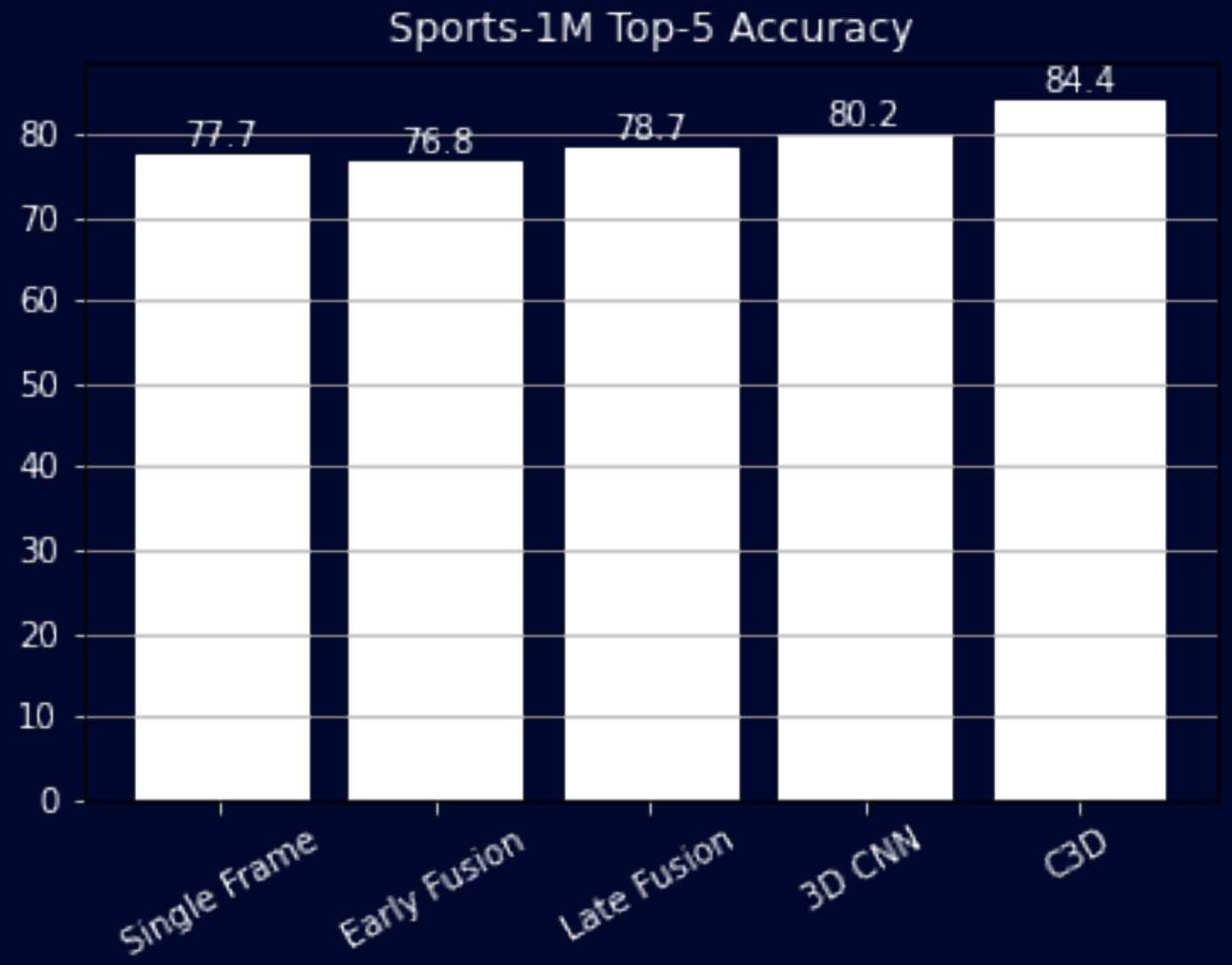
C3D

Layer	Size
Input	3 x 16 x 112 x 112
Conv1 (3x3x3)	64 x 16 x 112 x 112
Pool1 (1x2x2)	64 x 16 x 56 x 56
Conv2 (3x3x3)	128 x 16 x 56 x 56
Pool2 (2x2x2)	128 x 8 x 28 x 28
Conv3a (3x3x3)	256 x 8 x 28 x 28
Conv3b (3x3x3)	256 x 8 x 28 x 28
Pool3 (2x2x2)	256 x 4 x 14 x 14
Conv4a (3x3x3)	512 x 4 x 14 x 14
Conv4b (3x3x3)	512 x 4 x 14 x 14
Pool4 (2x2x2)	512 x 2 x 7 x 7
Conv5a (3x3x3)	512 x 2 x 7 x 7
Conv5b (3x3x3)	512 x 2 x 7 x 7
Pool5	512 x 1 x 3 x 3
FC6	4096
FC7	4096
FC8	C

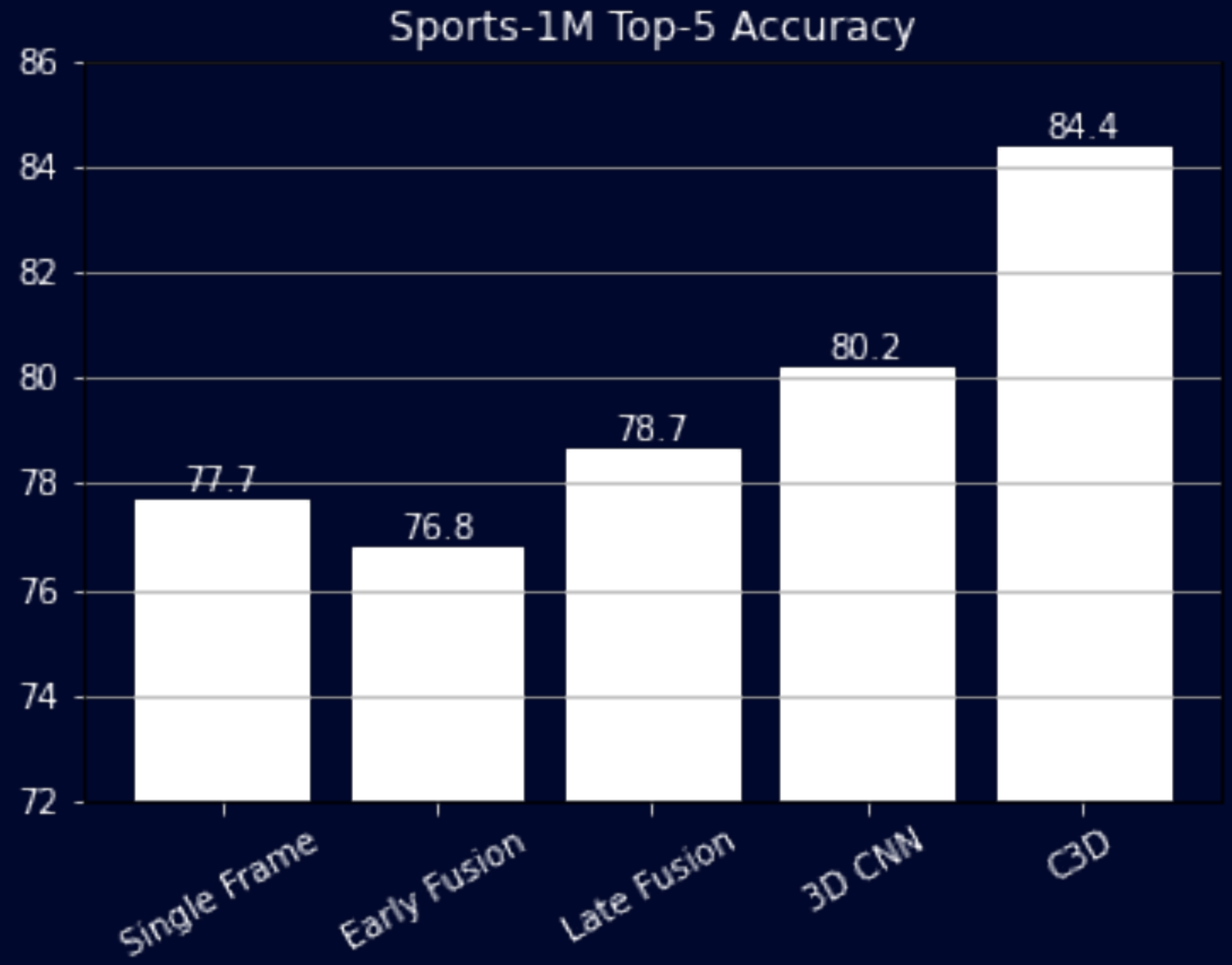
C3D



Sports-1M benchmark



Sports-1M benchmark

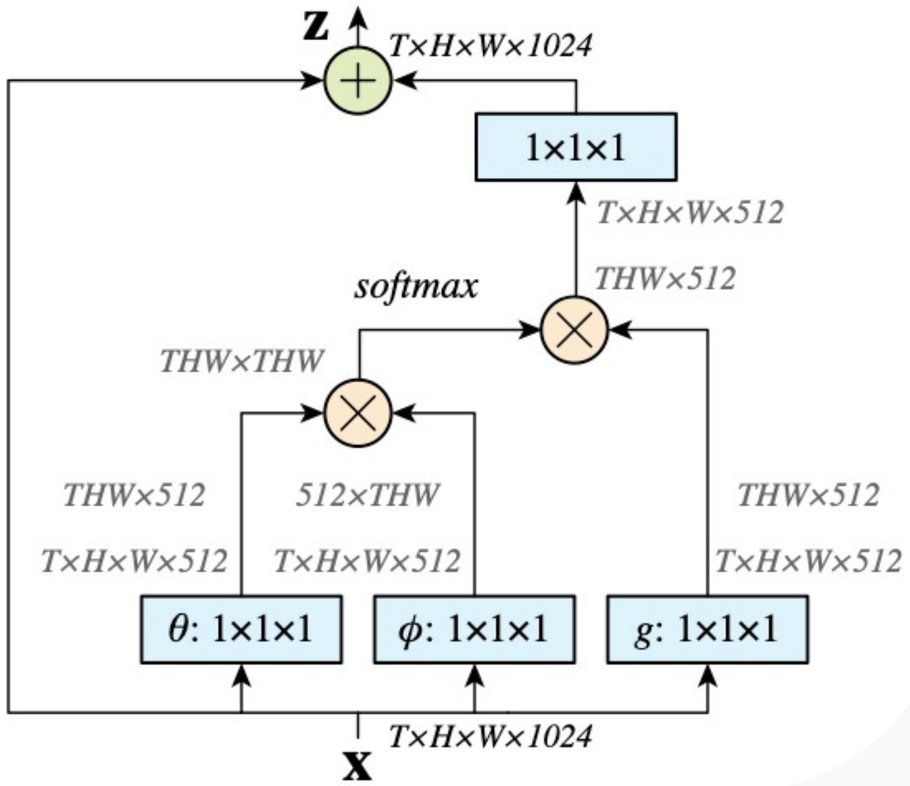


Inflating 2D to 3D (I3D)

- Идея: Взять 2d архитектуру и заменить все 2d свертки и пулинги на 3d аналоги

- Также можно переиспользовать веса
  - Копировать веса свертки T раз (temporal dimension)
  - Нормализовать полученный 3D тензор: разделить веса на T

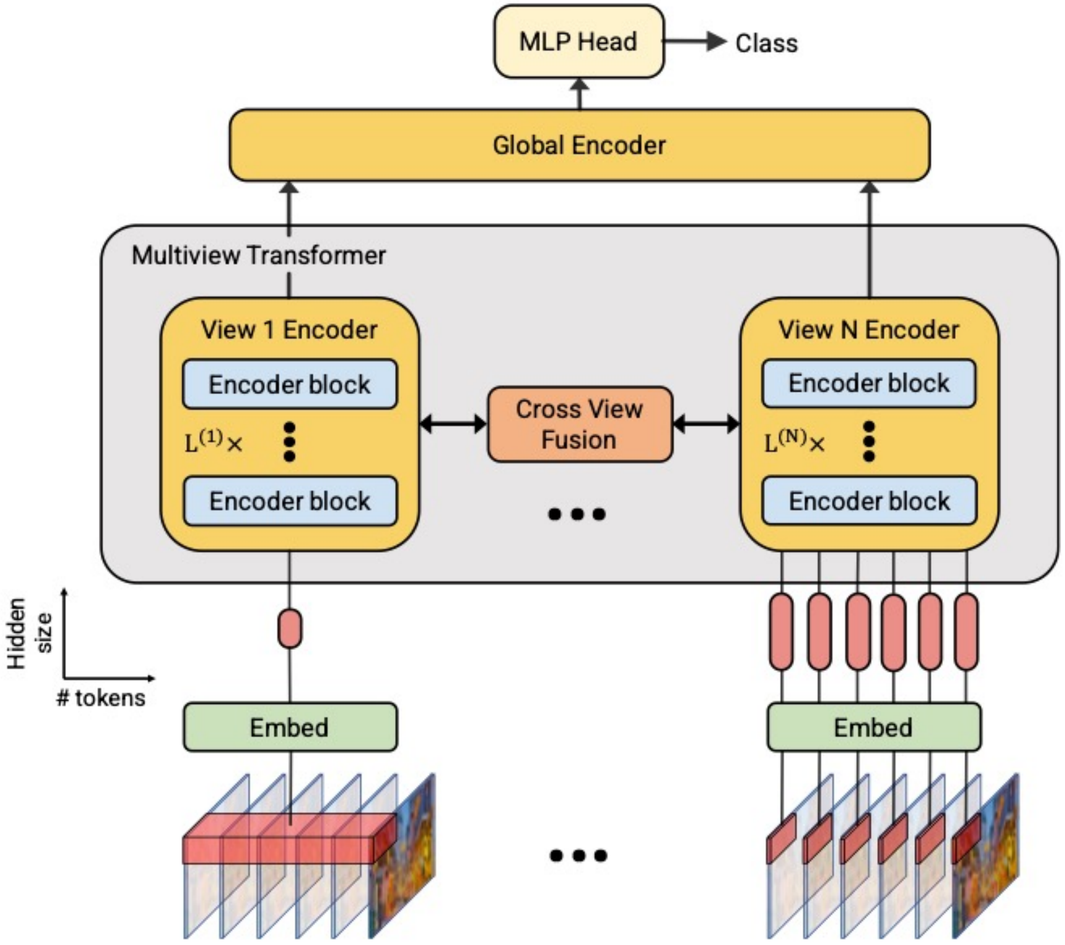
Nonlocal block



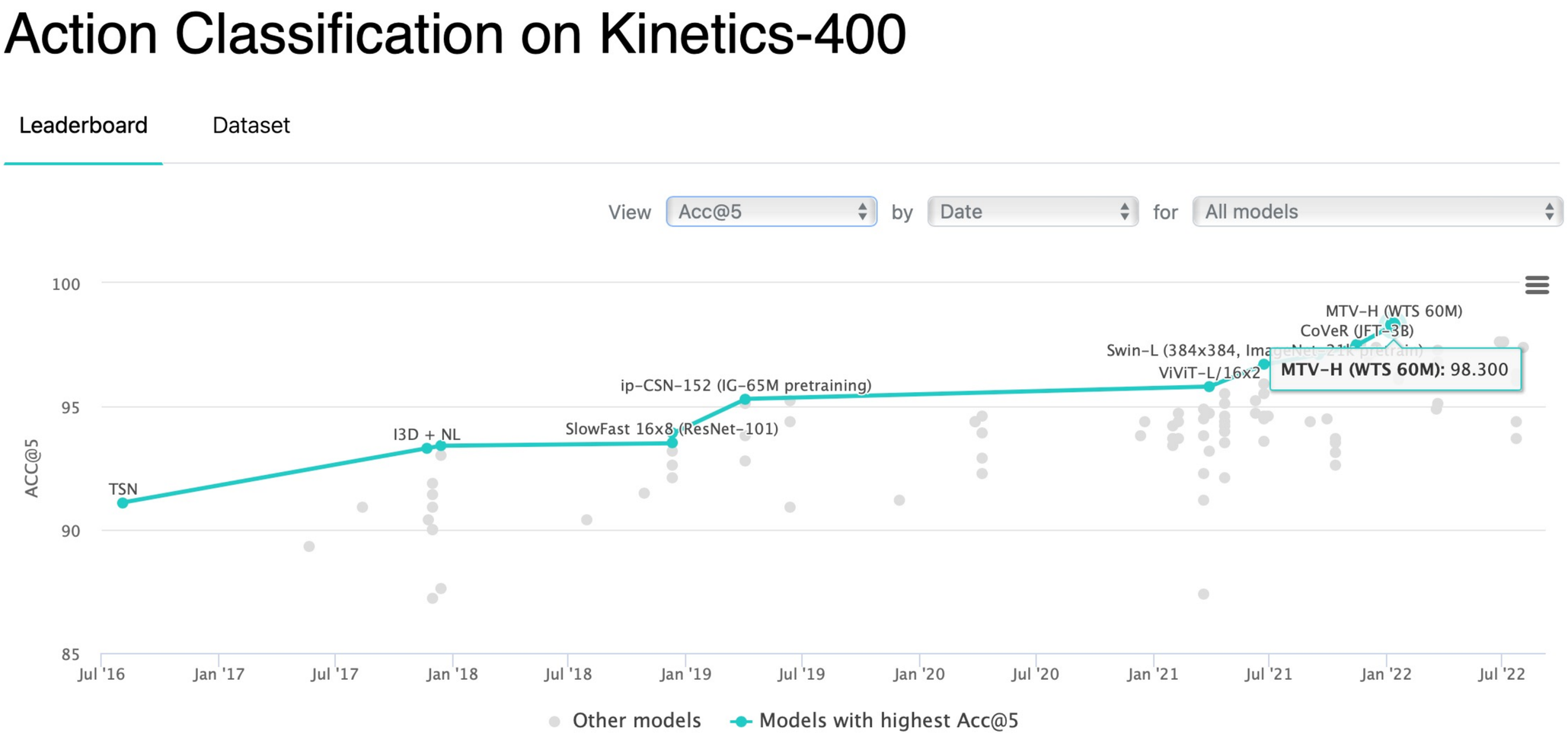
- θ $-$ $queries$
- ϕ $-$ $keys$
- g $-$ $values$

- Если инициализоровать последнюю свертку нулями, то весь блок будет Identity (из-за residual connection). Т.о. можно вставить в уже существующую архитектуру

Multiview Transformes



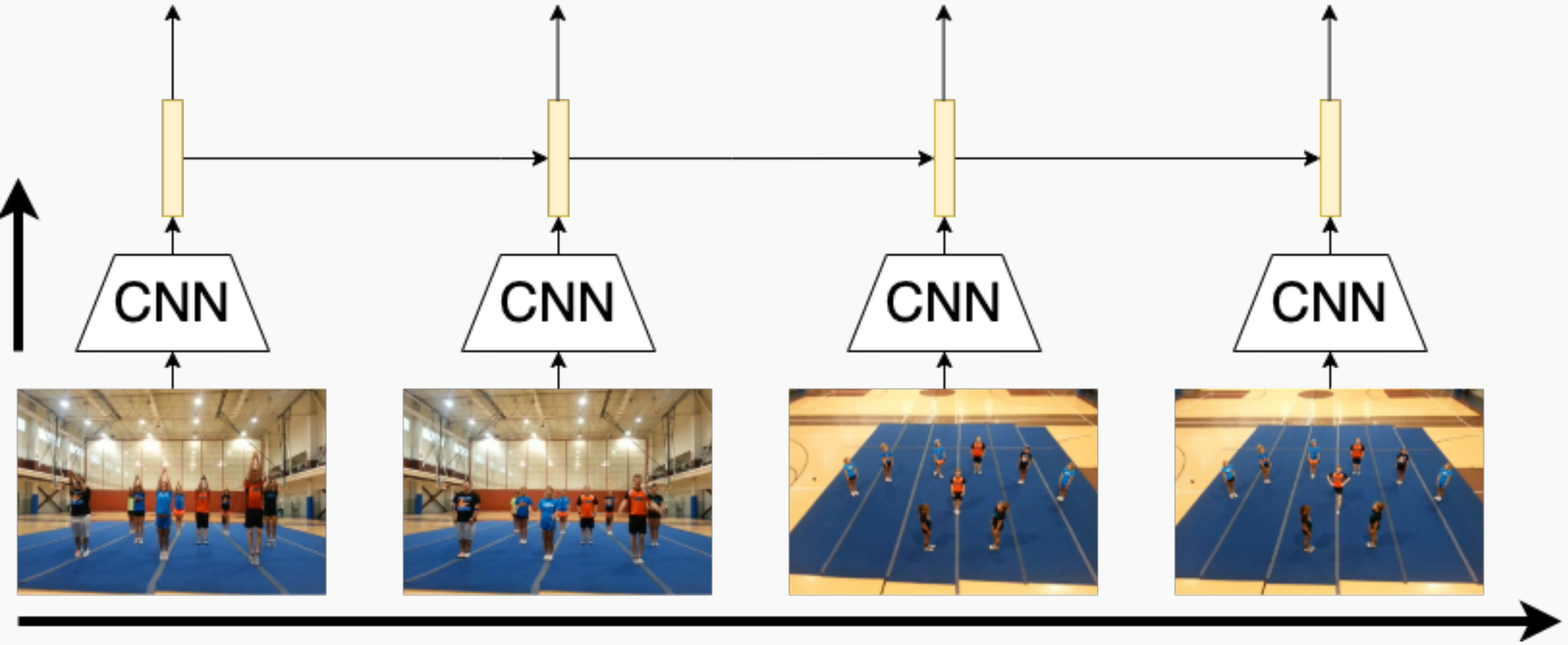
Multiview Transformes



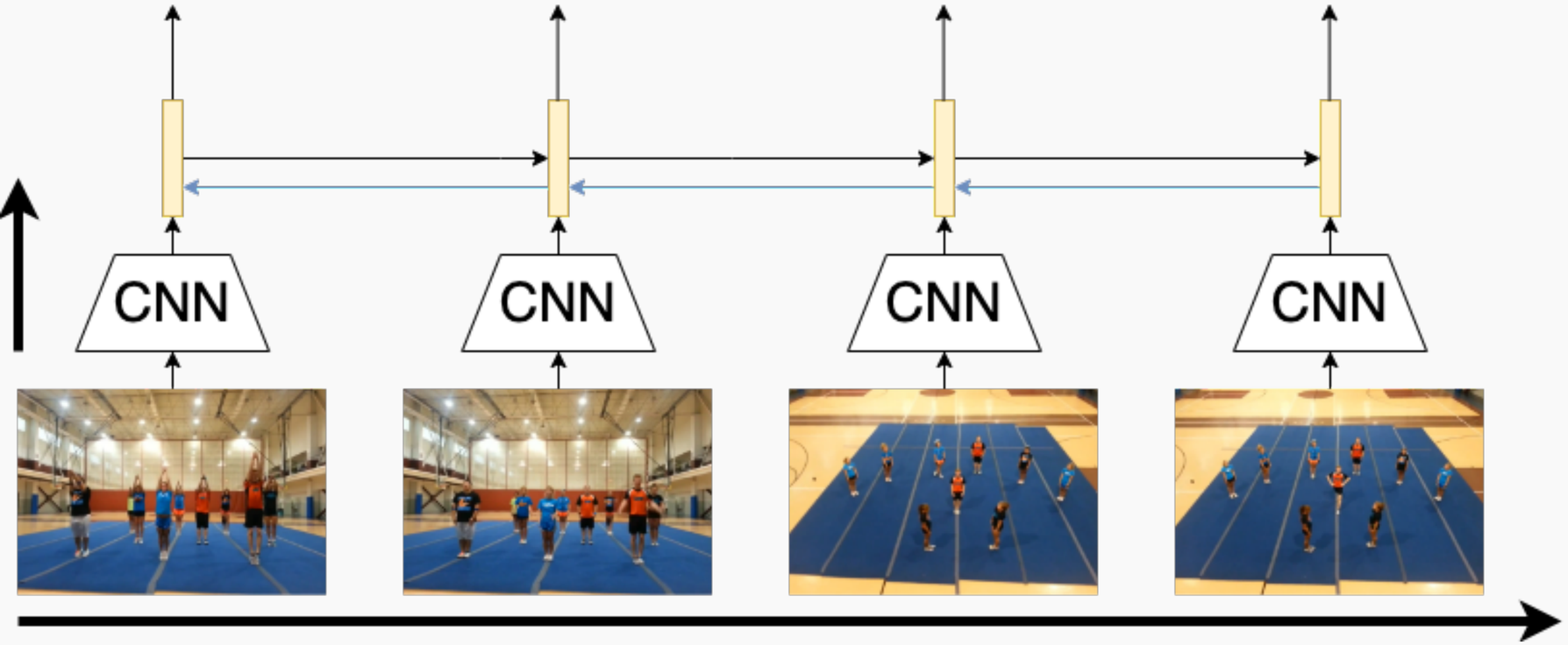
Temporal Action Localization

- Вход: Длинное видео (>5 секунд)
- Задача: для каждого действия на видео найти начало и конец действия

LSTM

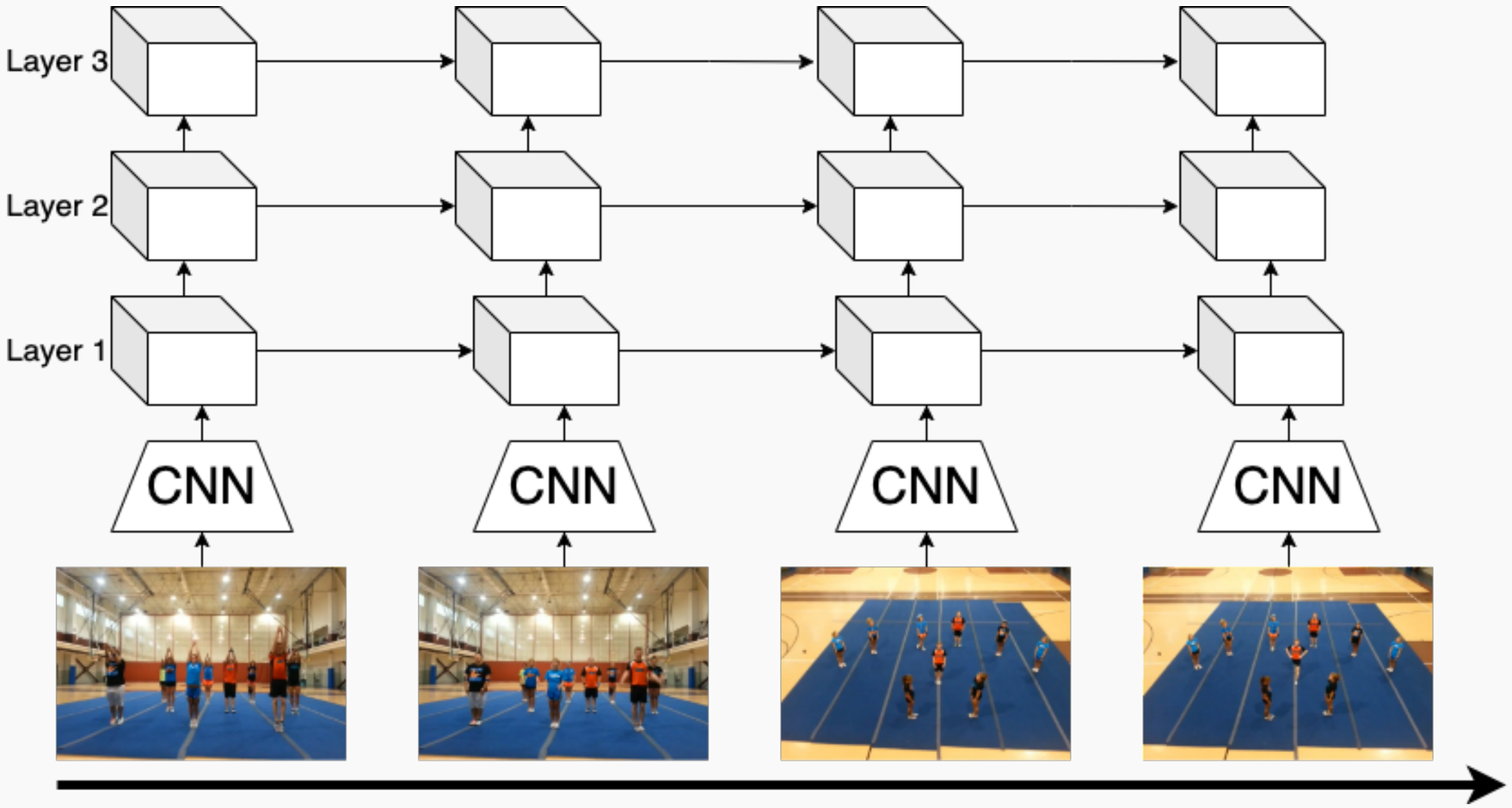


LSTM

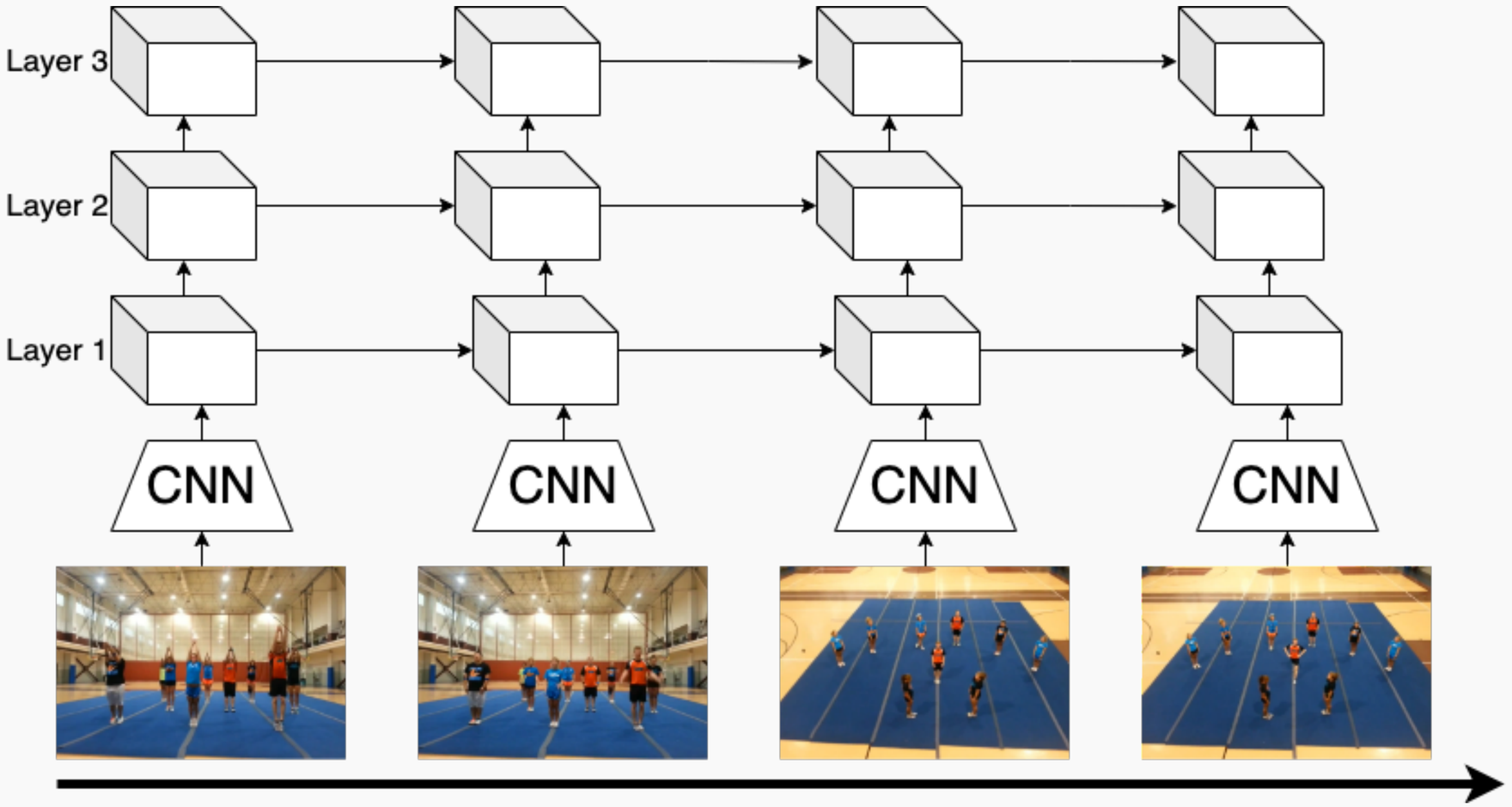


Использовать CNN в качестве экстрактора, обучать только LSTM

Recurrent Convolutional Network

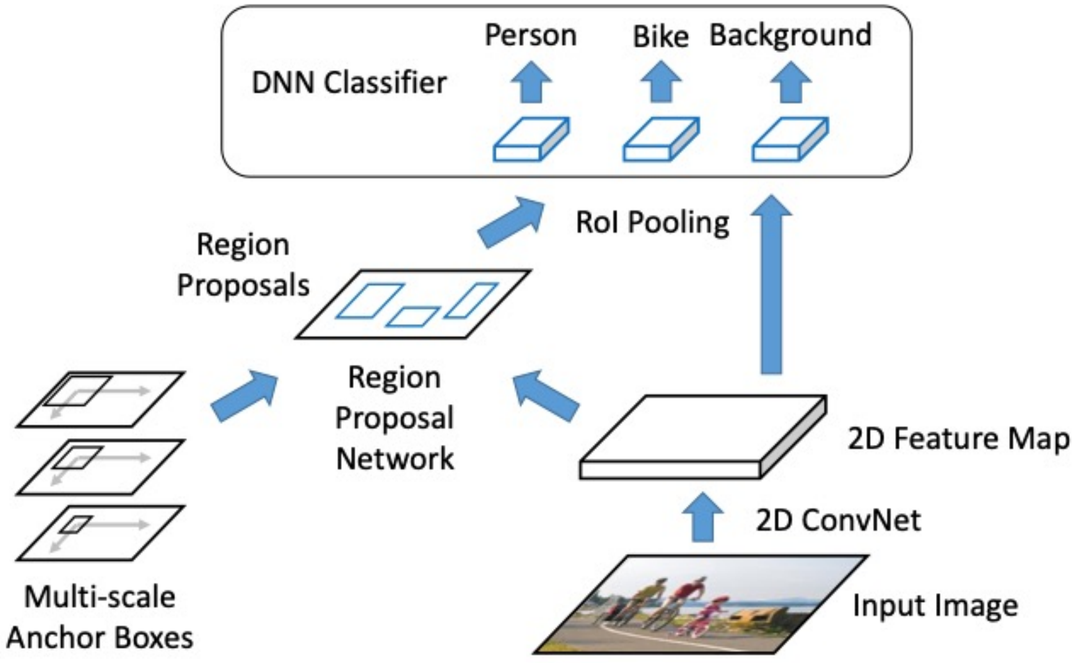


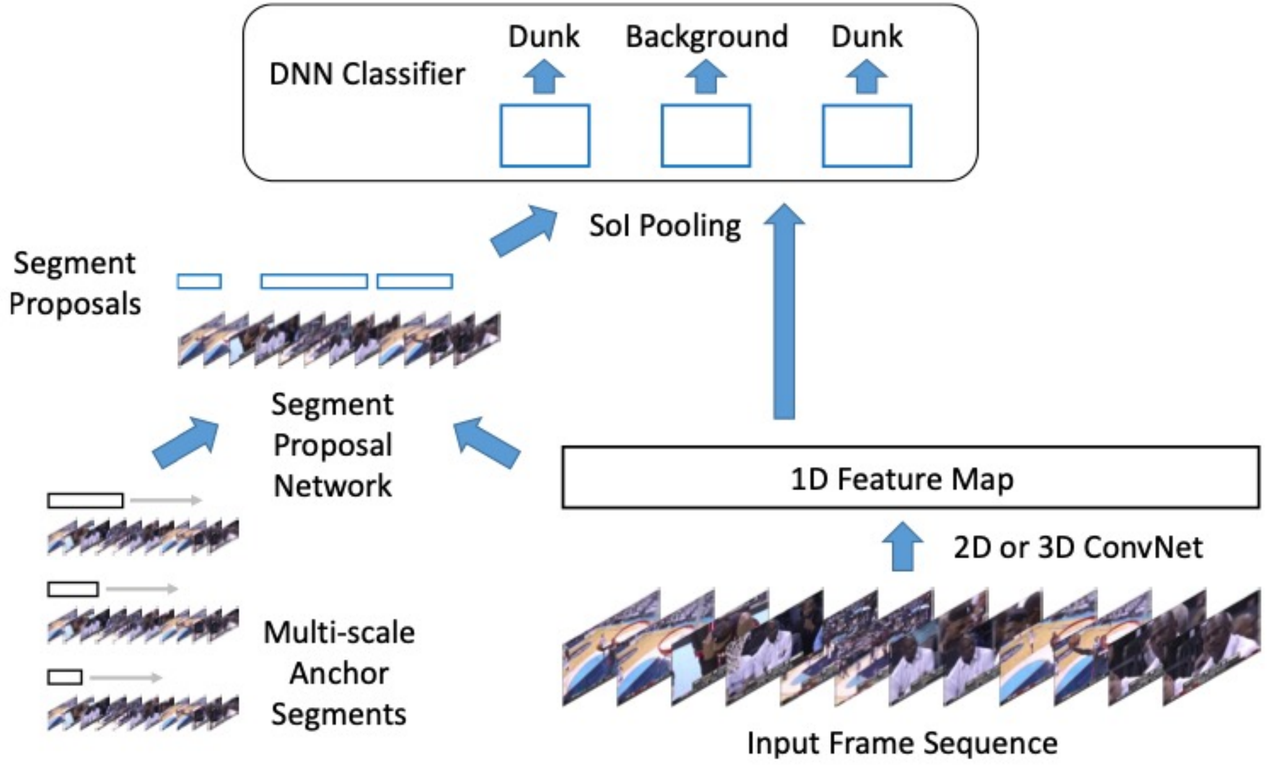
Recurrent Convolutional Network



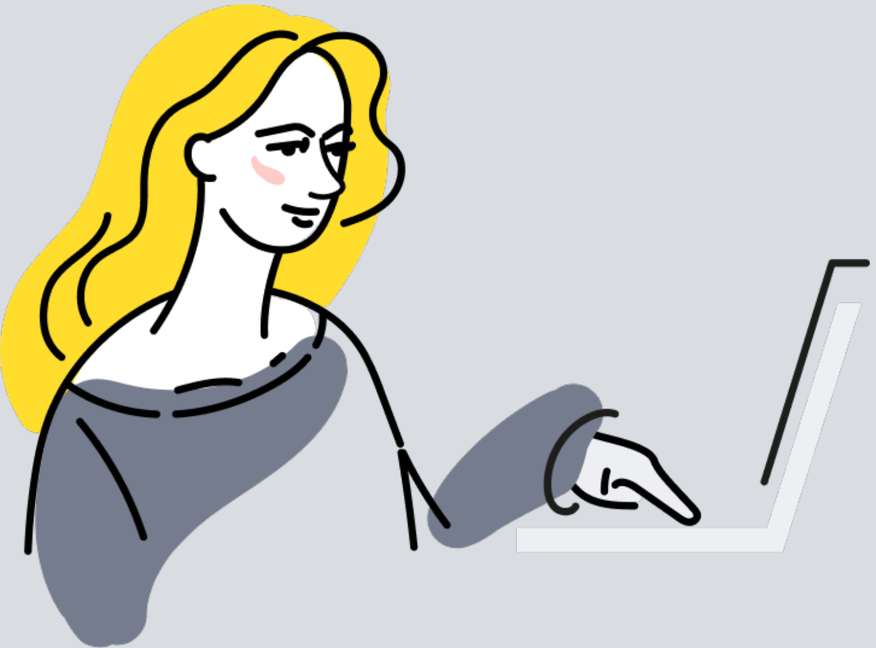
- $h_{t+1} = \tanh(W_h h_t + W_x x_t)$
- $Linear \rightarrow Convolution$

Faster RCNN


DNN Classifier
Person
Bike
Background
RoI Pooling
Region Proposals
Region Proposal Network
2D Feature Map
2D ConvNet
Multi-scale Anchor Boxes
Input Image

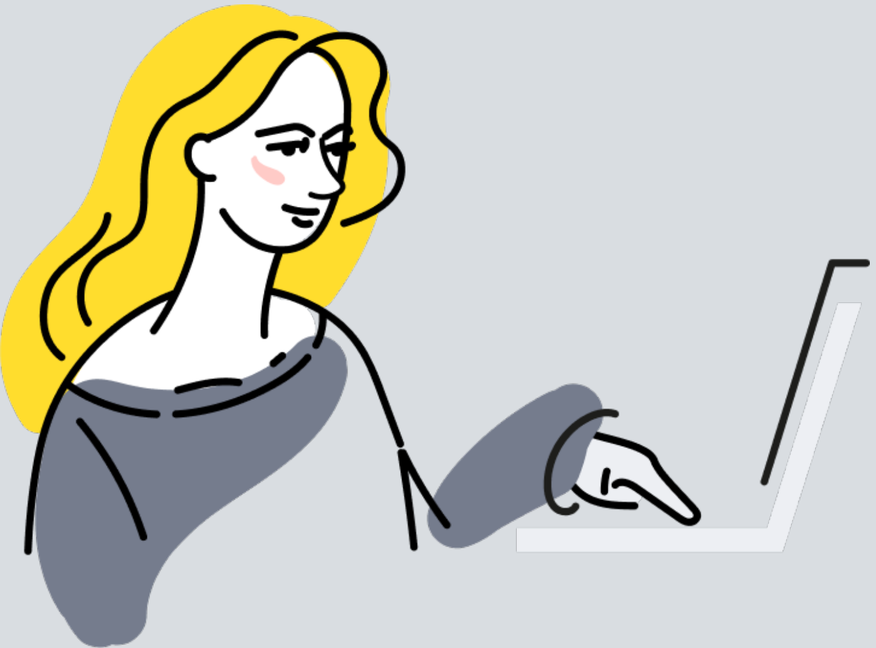

DNN Classifier
Dunk
Background
Dunk
SoI Pooling
Segment Proposals
Segment Proposal Network
1D Feature Map
2D or 3D ConvNet
Multi-scale Anchor Segments
Input Frame Sequence

Рекап и для дальнейшего изучения



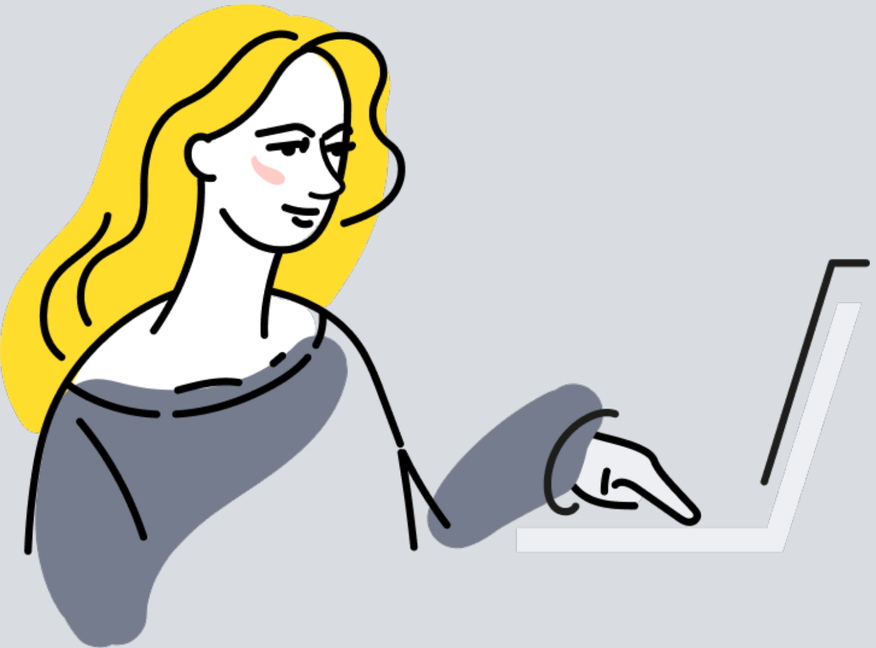
- Action Recognition:
  - Plain 2D-CNN – отличный baseline
  - Late Fusion Early Fusion – часто не дают большого прироста, по сравнению с baseline
  - Slow Fusion – 3D свертки. Temporal-invariant
  - C3D, I3D – Перенос архитектур из 2D в 3D. Transfer Learning

Рекап и для дальнейшего изучения



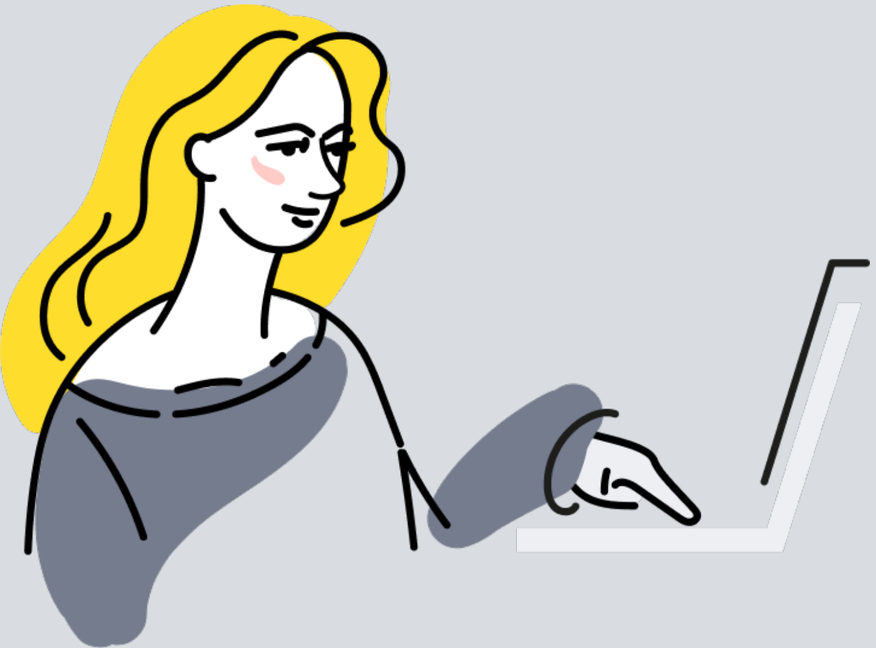
- Action Recognition:
  - Nonlocal blocks - Self-Attention. Можно вставлять блоки в уже обученные сетки и учить дальше
  - Multiview Transformer - Transformers (SOTA)

Рекап и для дальнейшего изучения



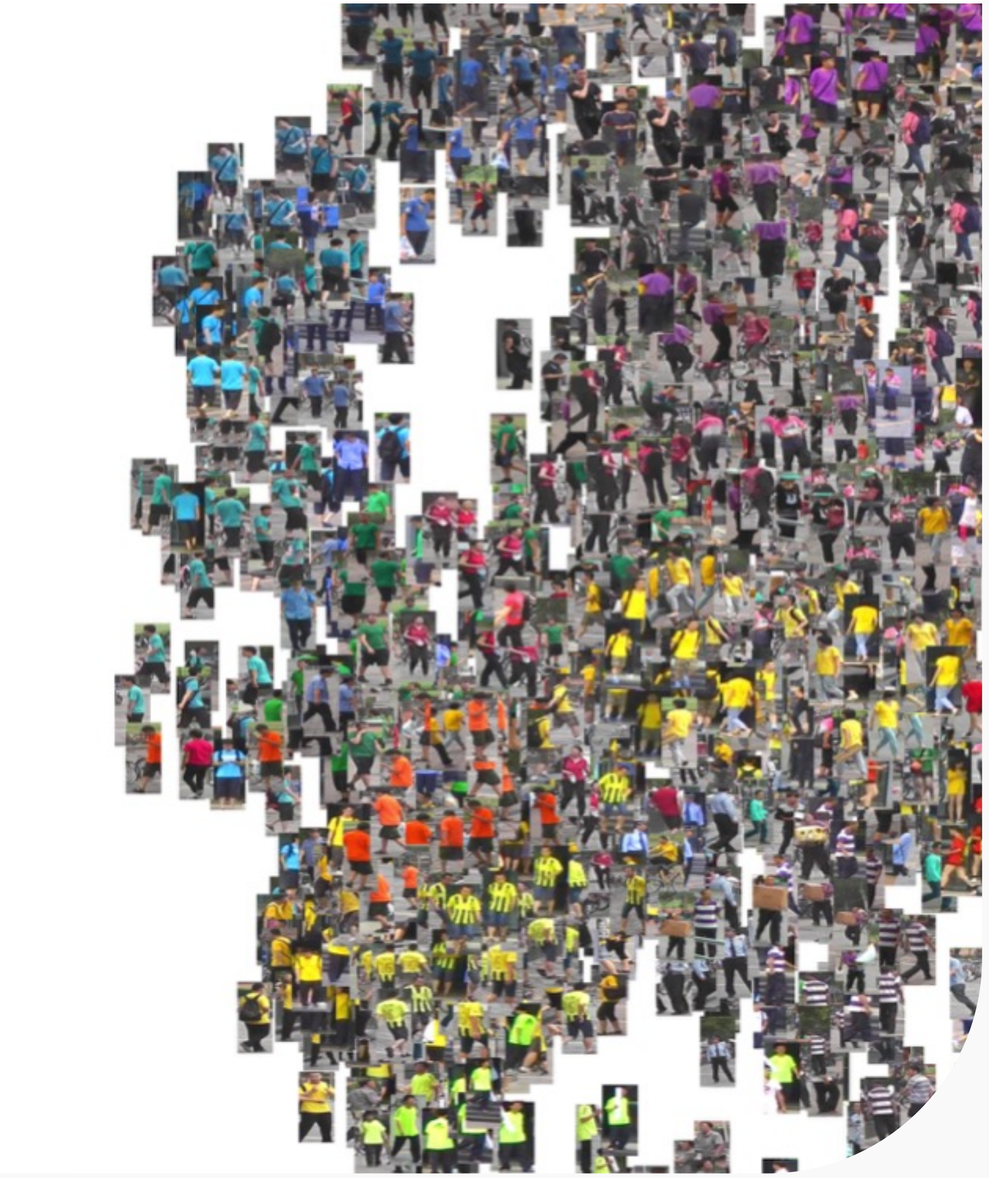
- Temporal Action Localization:
  - Plain 2D-CNN – также хороший baseline
  - LSTM – RNN поверх 3d-cnn эмбеддингов. Замораживаем CNN, учим LSTM
  - Recurrent Convolutional Networks - Multi-layer RNN со свертками, вместо линейных слоев
  - Faster-RCNN - адаптация на случай 3D. Segment Proposal вместо Region

Рекап и для дальнейшего изучения



- Slow-Fast Networks
- Two-stream convolutional networks
- Temporal Segmentation Networks
- VideoMAE

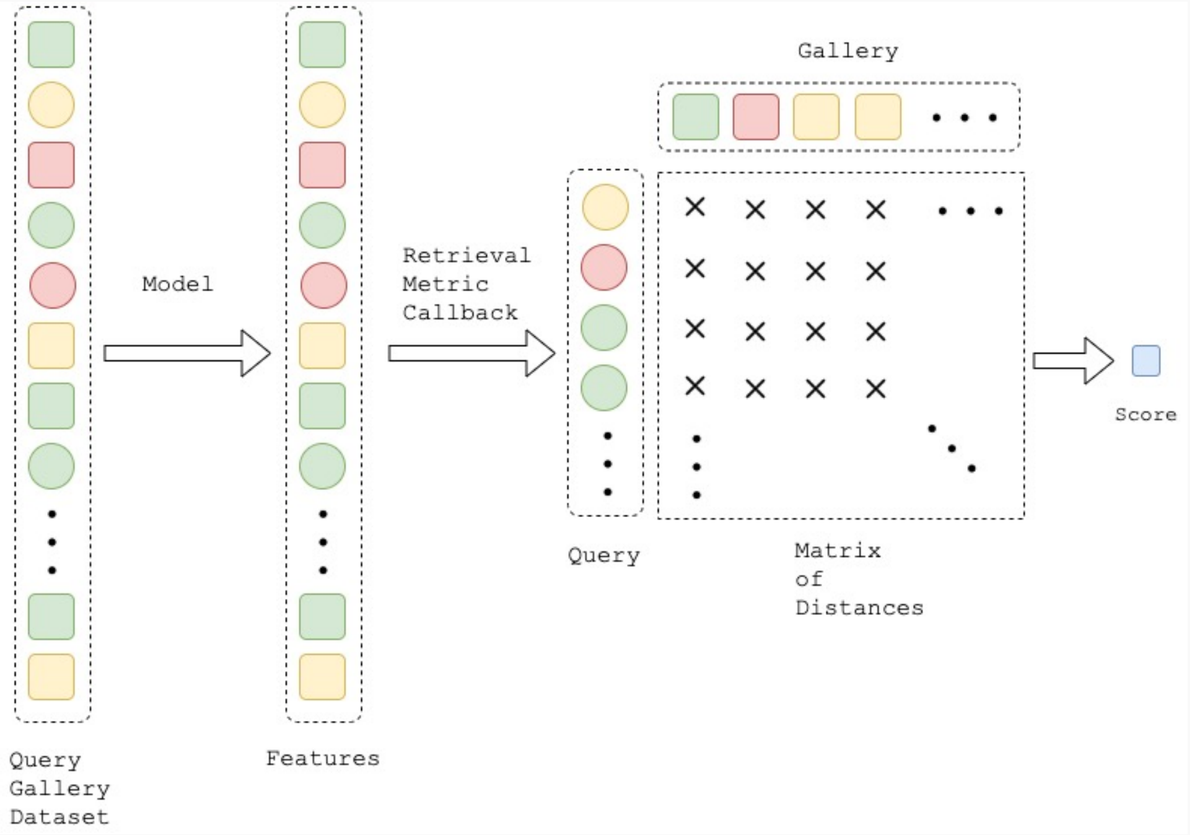
Re-Identification



Metric Learning

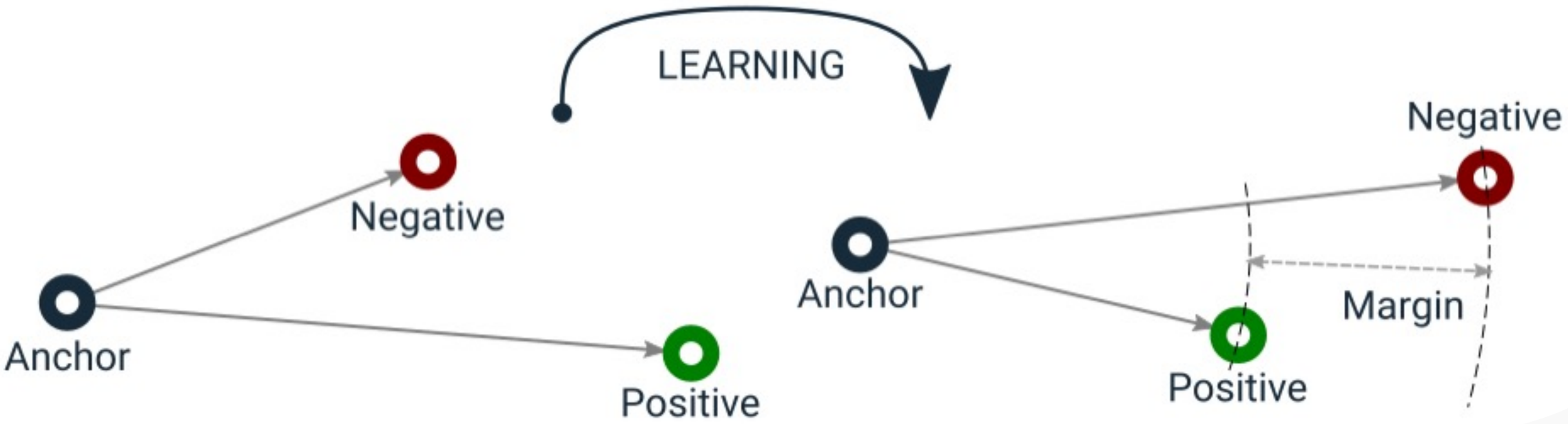
Задача: выучить репрезентации изображений таким образом, что похожие изображения будут близки в пространстве представлений.

Метрики



- Accuracy@k
- Recall@k
- Precision@k
- mAP

Triplet Loss

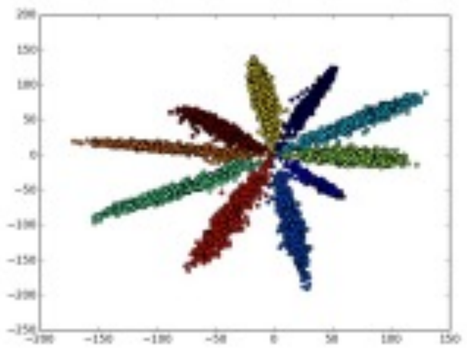


Triplet Loss

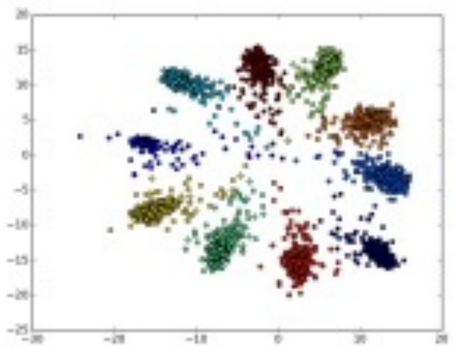
- $f_\theta(x)\colon \mathbb{R}^F \rightarrow \mathbb{R}^D$
- $D(x, y)\colon \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$
- $D\left(f_\theta(x_i), f_\theta\left(x_j\right)\right) = |f_\theta(x_i) - f_\theta\left(x_j\right)|_2^2$
- $\mathcal{L} = \sum_{a,p,n\colon y_a = y_p \neq y_n} [0, \mathrm{m} + D_{a,p} - D_{a,n}]_+$

- Mining *hard* triplets

Center Loss



(a) softmax loss

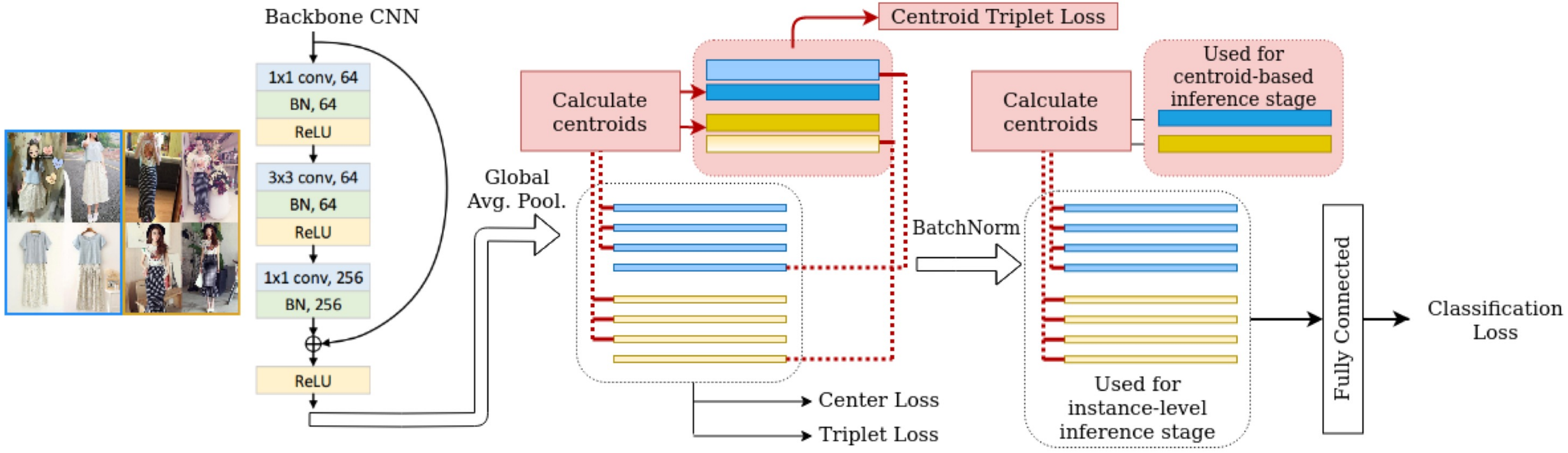


(b) center loss

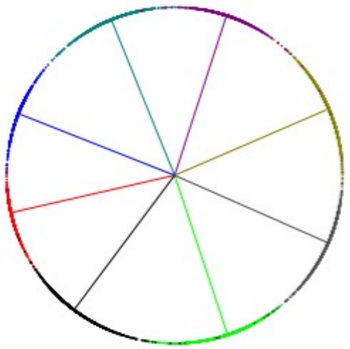
- $\mathcal{L}_{Center} = \frac{1}{2}\sum_{j=1}^{B} |f_{t_j} - c_{y_j}|_2^2$

Centroids in Image Retrieval

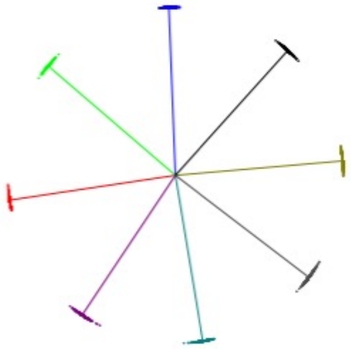
M. Wieczorek et al.



ArcLoss



(a) Softmax



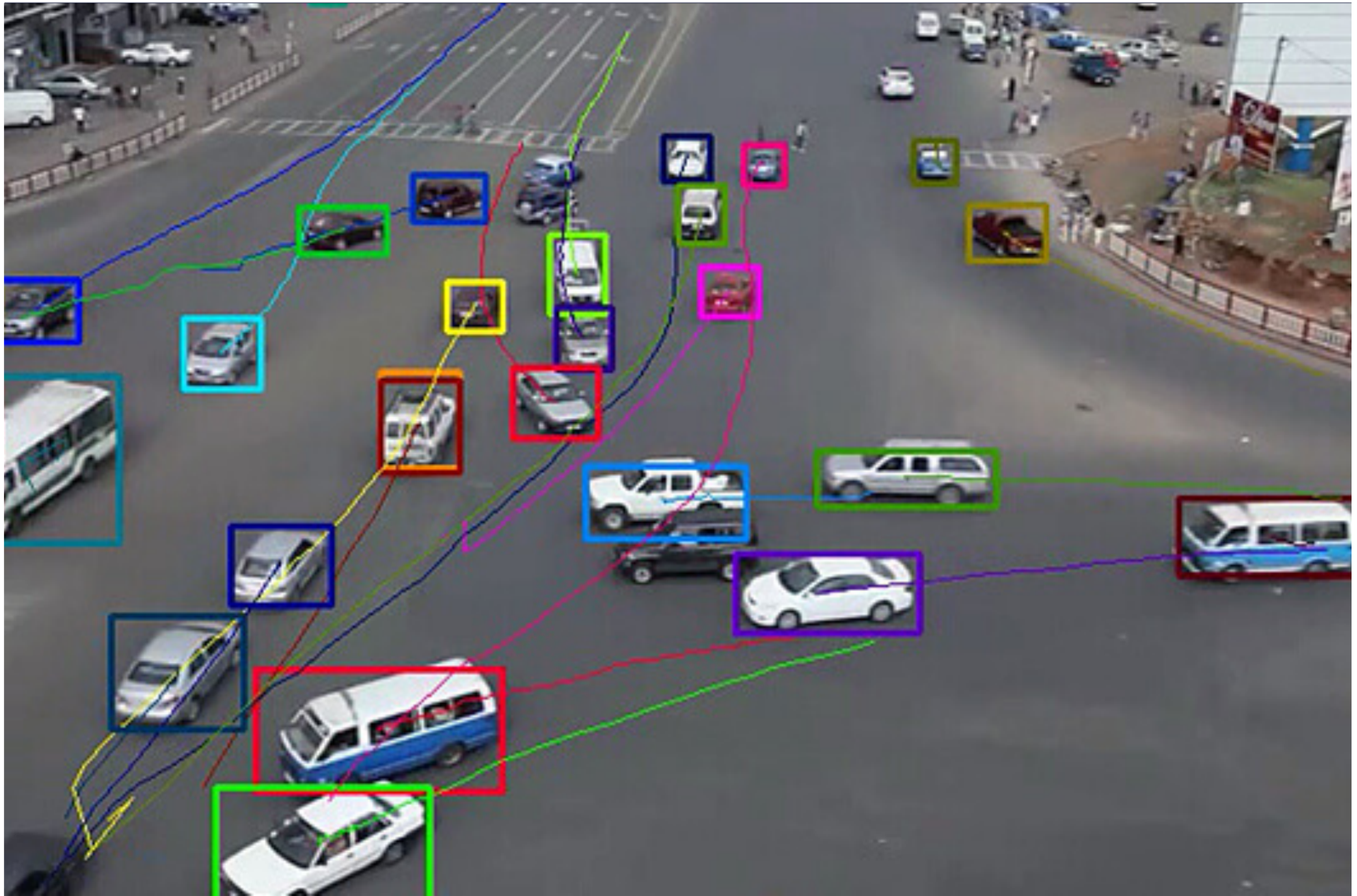
(b) ArcFace

- $W \in \mathbb{R}^{feat \times Nclass}$
- $|W_{:,j}|_2^2 = 1$
- $W_j^T x_i = |W_{:,j}|_2^2 \, |x_i|_2^2 \cos(\theta_j)$
- $|x|_2^2 = s$
- $\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{e^{s\left(\cos\left(\theta_{y_i} + m\right)\right)}}{e^{s\left(\cos\left(\theta_{y_i} + m\right)\right)} + \sum_{j=1, j \neq i}^{n} e^{s * \cos\left(\theta_{y_j}\right)}} \right)$

Рекап

- TripletLoss – популярный подход в Metric-Learning. Из минусов: сложно обучать из-за подбора триплетов
- CenterLoss – Добавляем к classification loss, чтобы размещать классы более компактно
- Centroids – использование центроид классов для обучения/инференса (SOTA)
- ArcLoss – Идея как у CenterLoss, но поумнее (часто работает лучше, чем TripletLoss)
- CosineLoss – Похожая на ArcLoss идея, другая реализация

Object Tracking



Метрики

- MOTA – Multi-object tracking accuracy
- $MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t}$
- MOTP – Multi-object tracking precision
- $MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t}$
- MT – mostly tracked trajectories
- ML – mostly lost trajectories
- IDsw – number of times an ID switches to a different previously tracked object

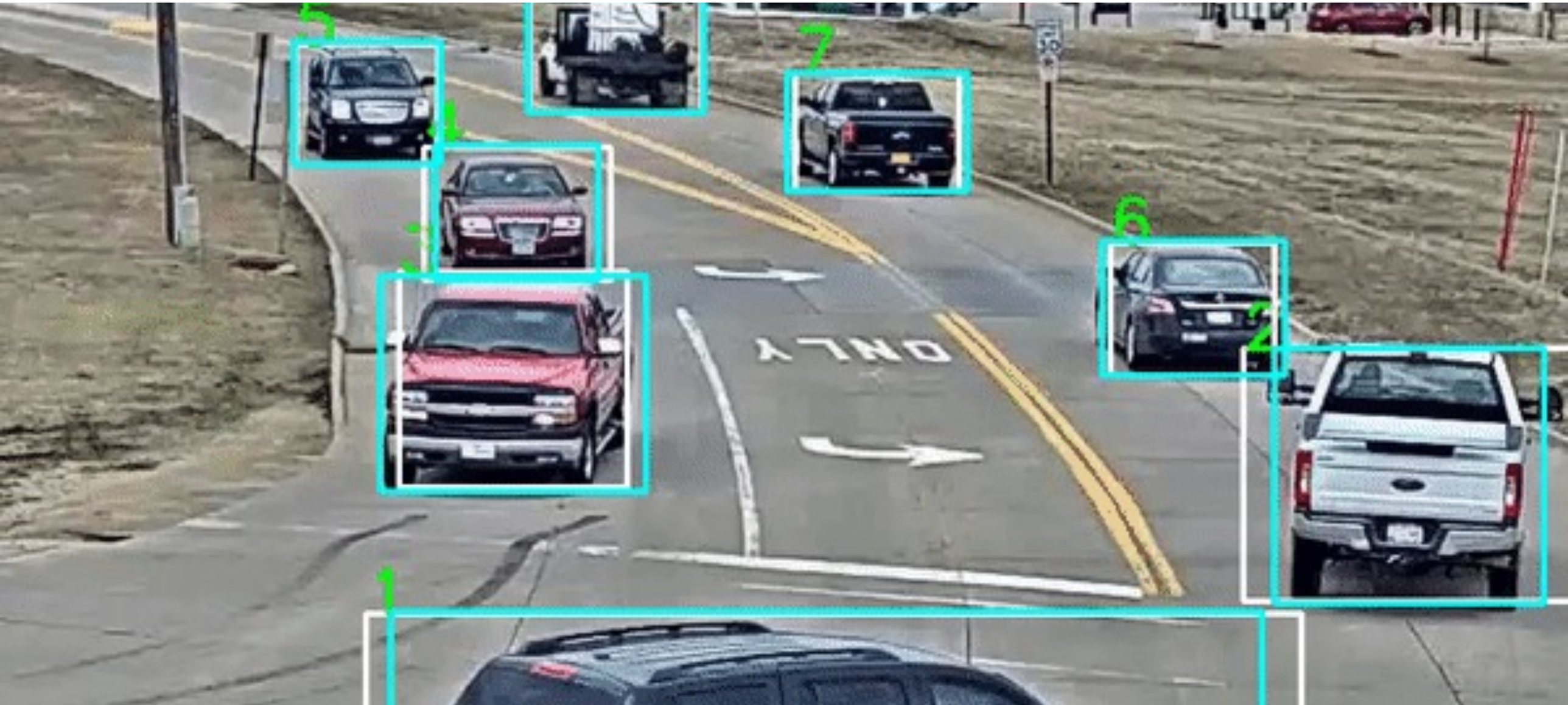
arxiv

Analytical Solutions

```
!pip install opencv-contrib-python
import cv2


OPENCV_OBJECT_TRACKERS = {
    "csrt": cv2.legacy.TrackerCSRT_create,   # best accuracy
    "kcf": cv2.legacy.TrackerKCF_create,
    "boosting": cv2.legacy.TrackerBoosting_create,
    "mil": cv2.legacy.TrackerMIL_create,
    "tld": cv2.legacy.TrackerTLD_create,
    "medianflow": cv2.legacy.TrackerMedianFlow_create,
    "mosse": cv2.legacy.TrackerMOSSE_create
}
```

Tracking By Detection



SORT

- Kalman Filter
- Hungarian Algorithm

kalman filter explained

hungarian algorithm explained

arxiv

git

SIMPLE ONLINE AND REALTIME TRACKING

Alex Bewley†, *Zongyuan Ge*†, *Lionel Ott*°, *Fabio Ramos*°, *Ben Upcroft*†

Queensland University of Technology†, University of Sydney°

ABSTRACT

This paper explores a pragmatic approach to multiple object tracking where the main focus is to associate objects efficiently for online and realtime applications. To this end, detection quality is identified as a key factor influencing tracking performance, where changing the detector can improve tracking by up to 18.9%. Despite only using a rudimentary combination of familiar techniques such as the Kalman Filter and Hungarian algorithm for the tracking components, this approach achieves an accuracy comparable to state-of-the-art online trackers. Furthermore, due to the simplicity of our tracking method, the tracker updates at a rate of 260 Hz which is over 20x faster than other state-of-the-art trackers.

***Index Terms*—** Computer Vision, Multiple Object Tracking, Detection, Data Association

1. INTRODUCTION

This paper presents a lean implementation of a tracking-by-detection framework for the problem of multiple object tracking (MOT) where objects are detected each frame and represented as bounding boxes. In contrast to many batch based tracking approaches [1, 2, 3], this work is primarily targeted towards online tracking where only detections from the previous and the current frame are presented to the tracker. Additionally, a strong emphasis is placed on efficiency for facilitating realtime tracking and to promote greater uptake in applications such as pedestrian tracking for autonomous vehicles.

The MOT problem can be viewed as a data association problem where the aim is to associate detections across frames in a video sequence. To aid the data association process, trackers use various methods for modelling the motion [1, 4] and appearance [5, 3] of objects in the scene. The methods employed by this paper were motivated through observations made on a recently established visual MOT benchmark [6]. Firstly, there is a resurgence of mature data association techniques including Multiple Hypothesis Tracking (MHT) [7, 3] and Joint Probabilistic Data Association (JPDA) [2] which occupy many of the top positions of the MOT benchmark. Secondly, the only tracker that does not use the Aggregate Channel Filter (ACF) [8] detector is also

Thanks to ACARP for funding.

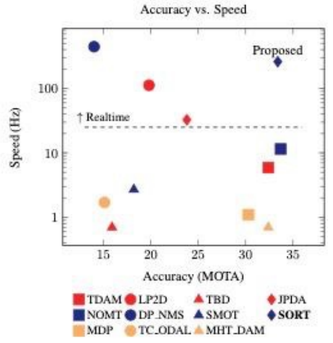
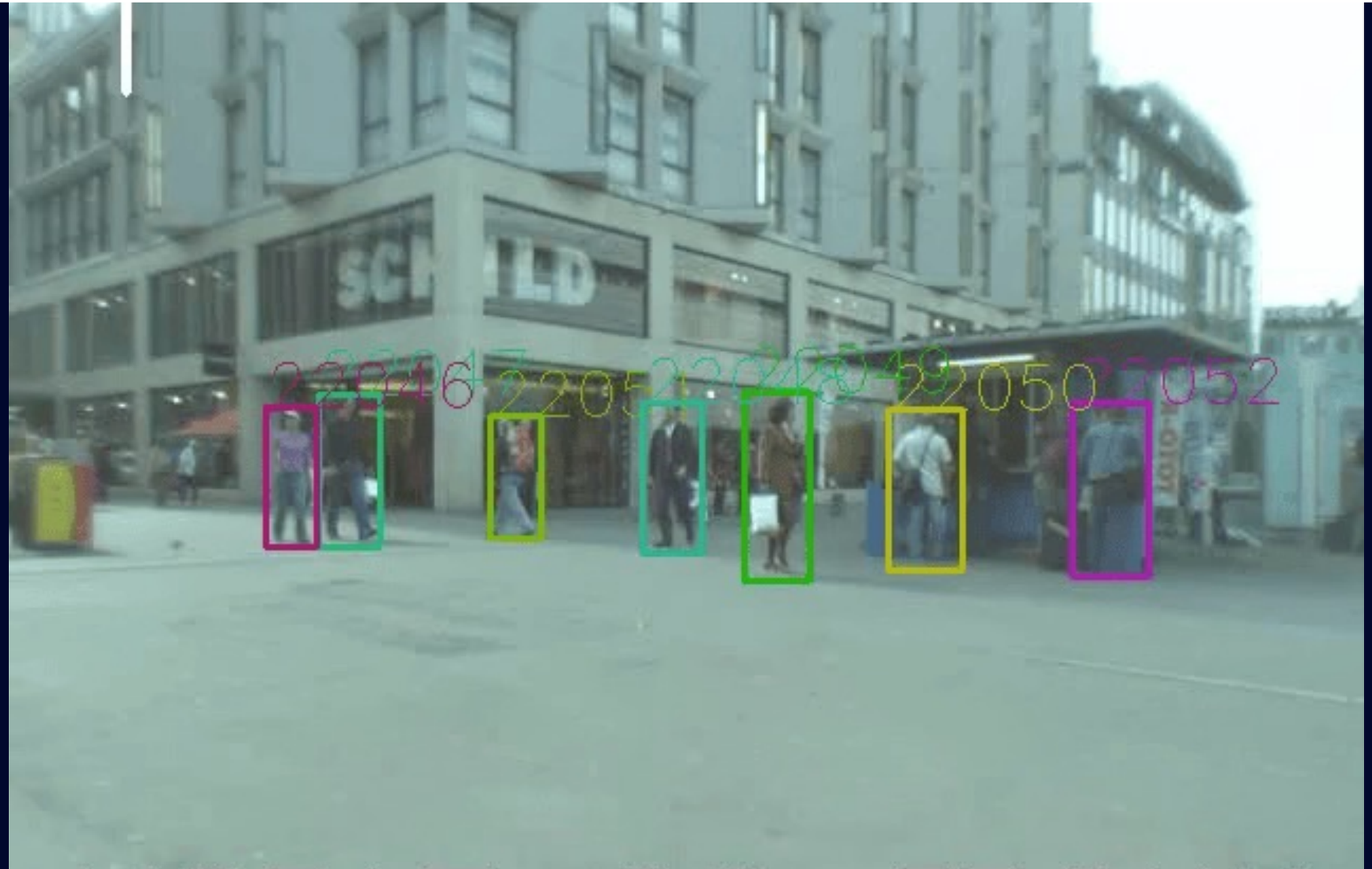


Fig. 1. Benchmark performance of the proposed method (**SORT**) in relation to several baseline trackers [6]. Each marker indicates a trackers accuracy and speed measured in frames per second (FPS) [Hz], i.e. higher and more right is better.

the top ranked tracker, suggesting that detection quality could be holding back the other trackers. Furthermore, the trade-off between accuracy and speed appears quite pronounced, since the speed of most accurate trackers is considered too slow for realtime applications (see Fig. 1). With the prominence of traditional data association techniques among the top online and batch trackers along with the use of different detections used by the top tracker, this work explores how simple MOT can be and how well it can perform.

Keeping in line with Occam's Razor, appearance features beyond the detection component are ignored in tracking and only the bounding box position and size are used for both motion estimation and data association. Furthermore, issues regarding short-term and long-term occlusion are also ignored, as they occur very rarely and their explicit treatment intro-

SORT



DeepSort

SIMPLE ONLINE AND REALTIME TRACKING WITH A DEEP ASSOCIATION METRIC

Nicolai Wojke[†], *Alex Bewley*[◇], *Dietrich Paulus*[†]

University of Koblenz-Landau[†], Queensland University of Technology[◇]

ABSTRACT

Simple Online and Realtime Tracking (SORT) is a pragmatic approach to multiple object tracking with a focus on simple, effective algorithms. In this paper, we integrate appearance information to improve the performance of SORT. Due to this extension we are able to track objects through longer periods of occlusions, effectively reducing the number of identity switches. In spirit of the original framework we place much of the computational complexity into an offline pre-training stage where we learn a deep association metric on a large-scale person re-identification dataset. During online application, we establish measurement-to-track associations using nearest neighbor queries in visual appearance space. Experimental evaluation shows that our extensions reduce the number of identity switches by 45%, achieving overall competitive performance at high frame rates.

Index Terms— Computer Vision, Multiple Object Tracking, Data Association

1. INTRODUCTION

Due to recent progress in object detection, tracking-by-detection has become the leading paradigm in multiple object tracking. Within this paradigm, object trajectories are usually found in a global optimization problem that processes entire video batches at once. For example, flow network formulations [1, 2, 3] and probabilistic graphical models [4, 5, 6, 7] have become popular frameworks of this type. However, due to batch processing, these methods are not applicable in online scenarios where a target identity must be available at each time step. More traditional methods are Multiple Hypothesis Tracking (MHT) [8] and the Joint Probabilistic Data Association Filter (JPDAF) [9]. These methods perform data association on a frame-by-frame basis. In the JPDAF, a single state hypothesis is generated by weighting individual measurements by their association likelihoods. In MHT, all possible hypotheses are tracked, but pruning schemes must be applied for computational tractability. Both methods have recently been revisited in a tracking-by-detection scenario [10, 11] and shown promising results. However, the performance of these methods comes at increased computational and implementation complexity.

Simple online and realtime tracking (SORT) [12] is a much simpler framework that performs Kalman filtering in image space and frame-by-frame data association using the Hungarian method with an association metric that measures bounding box overlap. This simple approach achieves favorable performance at high frame rates. On the MOT challenge dataset [13], SORT with a state-of-the-art people detector [14] ranks on average higher than MHT on standard detections. This not only underlines the influence of object detector performance on overall tracking results, but is also an important insight from a practitioners point of view.

Fig. 1: Exemplary output of our method on the MOT challenge dataset [15] in a common tracking situation with frequent occlusion.

While achieving overall good performance in terms of tracking precision and accuracy, SORT returns a relatively high number of identity switches. This is, because the employed association metric is only accurate when state estimation uncertainty is low. Therefore, SORT has a deficiency in tracking through occlusions as they typically appear in frontal-view camera scenes. We overcome this issue by replacing the association metric with a more informed metric that combines motion and appearance information. In particular, we apply a convolutional neural network (CNN) that has been trained to discriminate pedestrians on a large-scale person re-identification dataset. Through integration of this network we increase robustness against misses and occlusions while keeping the system easy to implement, efficient, and applicable to online scenarios. Our code and a pre-trained CNN model are made publicly available to facilitate research experimentation and practical application development.

- Kalman Filter
- Hungarian Algorithm
- Deep Appearance Descriptor
- Matching Cascade

arxiv

git

Сравнение методов (Online)

Название	MOTA	MOTP	IDsw	Runtime
EAMTT	52.5	78.8	910	12Hz
POI	**66.1**	79.5	805	10Hz
SORT	59.8	**79.6**	1423	**60Hz**
DeepSort	61.4	79.1	**781**	40Hz

ByteTrack

ByteTrack: Multi-Object Tracking by Associating Every Detection Box

Yifu Zhang[1], Peize Sun[2], Yi Jiang[3], Dongdong Yu[3], Fucheng Weng[1], Zehuan Yuan[3], Ping Luo[2], Wenyu Liu[1], Xinggang Wang[1†]

[1]Huazhong University of Science and Technology [2]The University of Hong Kong [3]ByteDance Inc.

Abstract

Multi-object tracking (MOT) aims at estimating bounding boxes and identities of objects in videos. Most methods obtain identities by associating detection boxes whose scores are higher than a threshold. The objects with low detection scores, e.g. occluded objects, are simply thrown away, which brings non-negligible true object missing and fragmented trajectories. To solve this problem, we present a simple, effective and generic association method, tracking by associating almost every detection box instead of only the high score ones. For the low score detection boxes, we utilize their similarities with tracklets to recover true objects and filter out the background detections. When applied to 9 different state-of-the-art trackers, our method achieves consistent improvement on IDF1 score ranging from 1 to 10 points. To put forwards the state-of-the-art performance of MOT, we design a simple and strong tracker, named ByteTrack. For the first time, we achieve 80.3 MOTA, 77.3 IDF1 and 63.1 HOTA on the test set of MOT17 with 30 FPS running speed on a single V100 GPU. ByteTrack also achieves state-of-the-art performance on MOT20, HiEve and BDD100K tracking benchmarks. The source code, pre-trained models with deploy versions and tutorials of applying to other trackers are released at https://github.com/ifzhang/ByteTrack.

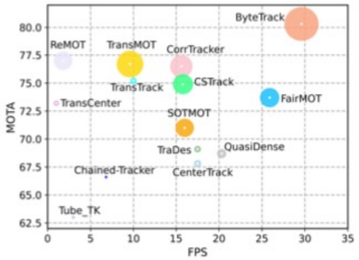


Figure 1. MOTA-IDF1-FPS comparisons of different trackers on the test set of MOT17. The horizontal axis is FPS (running speed), the vertical axis is MOTA, and the radius of circle is IDF1. Our ByteTrack achieves 80.3 MOTA, 77.3 IDF1 on MOT17 test set with 30 FPS running speed, outperforming all previous trackers. Details are given in Table 4.

1. Introduction

Was vernünftig ist, das ist wirklich; und was wirklich ist, das ist vernünftig.

—— *G. W. F. Hegel*

Tracking-by-detection is the most effective paradigm for multi-object tracking (MOT) in current. Due to the complex scenarios in videos, detectors are prone to make imperfect predictions. State-of-the-art MOT methods [1–3, 6, 12, 18, 45, 59, 70, 72, 85] need to deal with true positive / false positive trade-off in detection boxes to eliminate low confidence detection boxes [4, 40]. However, is it the right way to eliminate all low confidence detection boxes? Our answer is NO: as Hegel said "What is reasonable is real; that which is real is reasonable." Low confidence detection boxes sometimes indicate the existence of objects, *e.g.* the occluded objects. Filtering out these objects causes irreversible errors for MOT and brings non-negligible missing detection and fragmented trajectories.

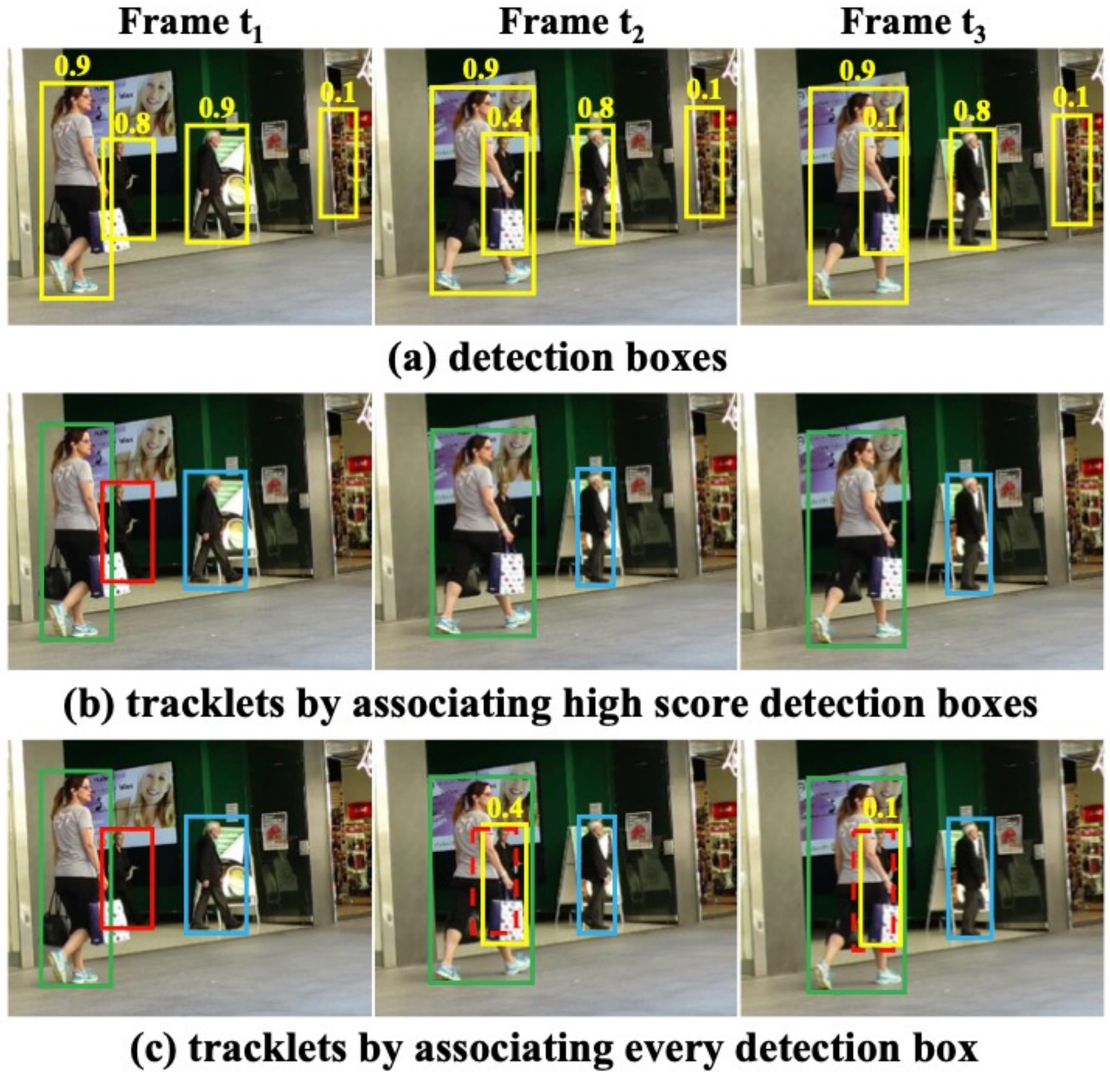
Figure 2 (a) and (b) show this problem. In frame t_1, we initialize three different tracklets as their scores are all higher than 0.5. However, in frame t_2 and frame t_3 when occlusion happens, red tracklet's corresponding detection score becomes lower *i.e.* 0.8 to 0.4 and then 0.4 to

† Corresponding author.
Part of this work was performed while Yifu Zhang worked as an intern at ByteDance.

arxiv

git

ByteTrack


Frame t_1
Frame t_2
Frame t_3
0.9
0.8
0.9
0.1
0.9
0.4
0.8
0.1
0.9
0.1
0.8
0.1
0.4
0.1
(a) detection boxes
(b) tracklets by associating high score detection boxes
(c) tracklets by associating every detection box

Сравнение методов (MOT17)

Название	MOTA	IDsw	FPS
SORT	74.6	291	30.1
DeepSort	75.4	239	13.5
ByteTrack	76.6	159	29.6

FairMOT

Noname manuscript No.
(will be inserted by the editor)

FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking

Yifu Zhang[1†] · **Chunyu Wang**[2†] · **Xinggang Wang**[1*] · **Wenjun Zeng**[2] · **Wenyu Liu**[1]

Received: date / Accepted: date

Abstract Multi-object tracking (MOT) is an important problem in computer vision which has a wide range of applications. Formulating MOT as multi-task learning of object detection and re-ID in a single network is appealing since it allows joint optimization of the two tasks and enjoys high computation efficiency. However, we find that the two tasks tend to compete with each other which need to be carefully addressed. In particular, previous works usually treat re-ID as a secondary task whose accuracy is heavily affected by the primary detection task. As a result, the network is biased to the primary detection task which is not *fair* to the re-ID task. To solve the problem, we present a simple yet effective approach termed as *FairMOT* based on the anchor-free object detection architecture CenterNet. Note that it is not a naive combination of CenterNet and re-ID. Instead, we present a bunch of detailed designs which are critical to achieve good tracking results by thorough empirical studies. The resulting approach achieves high accuracy for both detection and tracking. The approach outperforms the state-of-the-art methods by a large margin on several public datasets.

Yifu Zhang
E-mail: yifuzhang@hust.edu.cn
Chunyu Wang
E-mail: chnuwa@microsoft.com
Xinggang Wang
E-mail: xgwang@hust.edu.cn
Wenjun Zeng
E-mail: wezeng@microsoft.com
Wenyu Liu
E-mail: liuwy@hust.edu.cn

[1] Huazhong University of Science and Technology, Wuhan, China
[2] Microsoft Research Asia, Beijing, China
* Corresponding Author
† Yifu Zhang and Chunyu Wang have contributed equally.

The source code and pre-trained models are released at https://github.com/ifzhang/FairMOT.

Keywords FairMOT · Multi-Object Tracking · One-Shot · Anchor-Free · Real-Time Inference

1 Introduction

Multi-Object Tracking (MOT) has been a longstanding goal in computer vision (Bewley et al., 2016; Wojke et al., 2017; Chen et al., 2018a; Yu et al., 2016). The goal is to estimate trajectories for objects of interest presented in videos. The successful resolution of the problem can immediately benefit many applications such as intelligent video analysis, human computer interaction, human activity recognition (Wang et al., 2013; Luo et al., 2017), and even social computing.

Most of the existing methods such as (Mahmoudi et al., 2019; Zhou et al., 2018; Fang et al., 2018; Bewley et al., 2016; Wojke et al., 2017; Chen et al., 2018a; Yu et al., 2016) attempt to address the problem by two separate models: the *detection* model firstly detects objects of interest by bounding boxes in each frame, then the *association* model extracts re-identification (re-ID) features from the image regions corresponding to each bounding box, links the detection to one of the existing tracks or creates a new track according to certain metrics defined on features.

There has been remarkable progress on object detection (Ren et al., 2015; He et al., 2017; Zhou et al., 2019a; Redmon and Farhadi, 2018; Fu et al., 2020; Sun et al., 2021b,a) and re-ID (Zheng et al., 2017a; Chen et al., 2018a) respectively recently which in turn boosts the overall tracking accuracy. However, these two-step methods suffer from scalability issues. They cannot achieve real-time inference speed when there are a large number of objects in the environment because the two models do not share features and they need

- Unfairness cased by anchors
- Unfairness cased by features
- Unfairness cased by feature dimension

arxiv

git

FairMOT

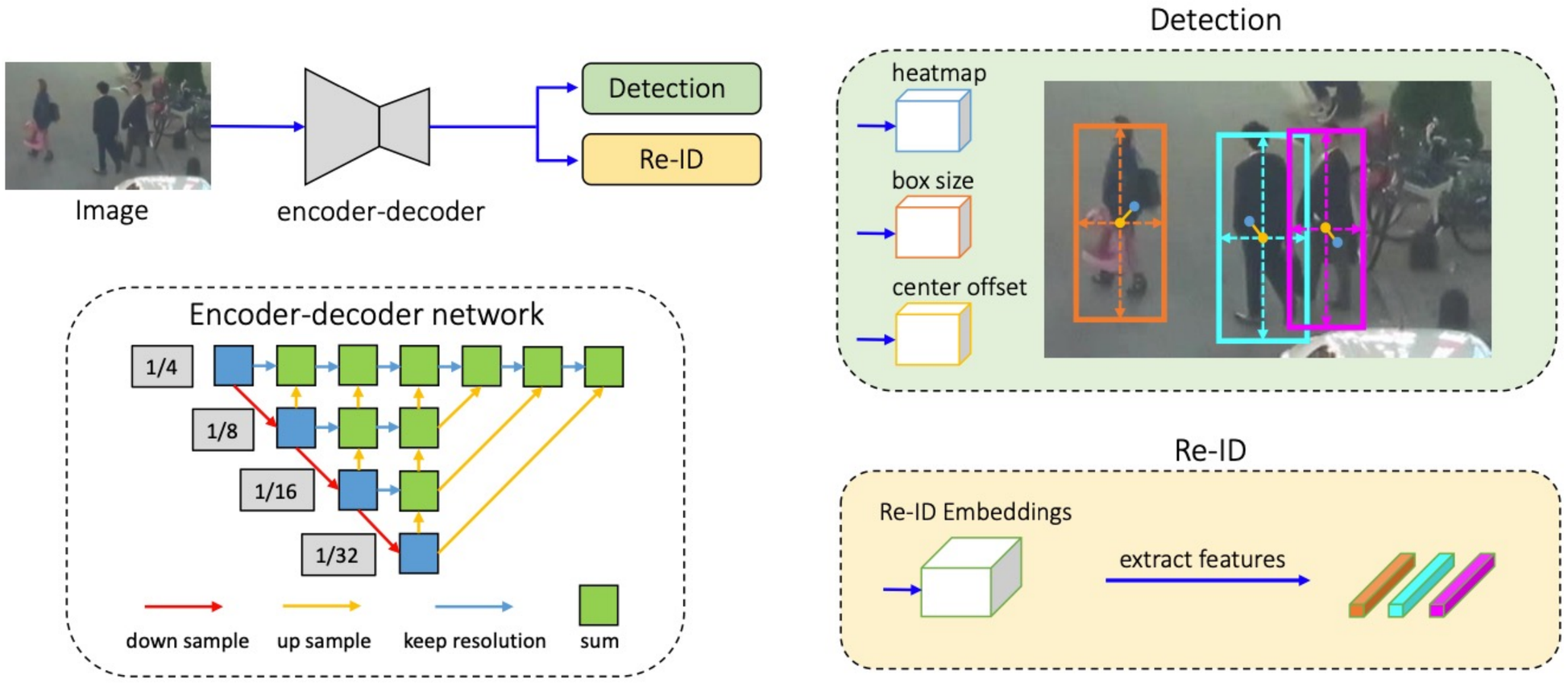


Fig. 1 Overview of our one-shot tracker *FairMOT*. The input image is first fed to an encoder-decoder network to extract high resolution feature maps (stride=4). Then we add two homogeneous branches for detecting objects and extracting re-ID features, respectively. The features at the predicted object centers are used for tracking.

Рекап и для дальнейшего изучения

- SORT – Быстрый и простой Tracking by Detection. Из минусов много Identity switches
- DeepSORT – SORT + ReID. Меньше Identity switches
- ByteTrack – Использование всех детекций (даже с маленьким confidence) довольно сильно улучшает общее качество
- FairMOT – End2End Detector + ReID. Используется как вход для tracking алгоритмов (например, DeepSORT)

Рекап и для дальнейшего изучения

- OC-SORT
- BoT-SORT
- StrongSort
- mm-tracking

IT's TINKOFF

Александр Весельев | DL engineer

a.veselyev@tinkoff.ai

@podidiving