Лаборатория компьютерной
графики и мультимедиа
ВМК МГУ имени М.В. Ломоносова

Курс «Компьютерное зрение»

«Основы анализа видео»

Антон Конушин и Мамедов Тимур

2025 год

План лекции

1. Введение
2. Оптический поток и его оценка
3. Распознавание событий в видео
4. Отслеживание одного объекта
5. Отслеживание множества объектов

Виды видео

Видеопоток – упорядоченная последовательность изображений, полученных с одной камеры через небольшие промежутки времени/

Видеопоток подразумевает обработку на лету.

Видеопоследовательность же конечна, можно обрабатывать целиком.

- Пользовательское видео – от 3-5 кадров/сек до 30-50 кадров/сек
- Разрешение – от 320x240 до 1920*1080 (HD) (сейчас ещё 4K)
- В градациях серого (одноканальное) или цветное (3х канальное)
- Поток данных – 2Мб (один канал HD) x 3 (RGB) x 30 кадр/с = 180 Мб/с
  - Сравнимо с пропускной способностью 1Gb Ethernet сети

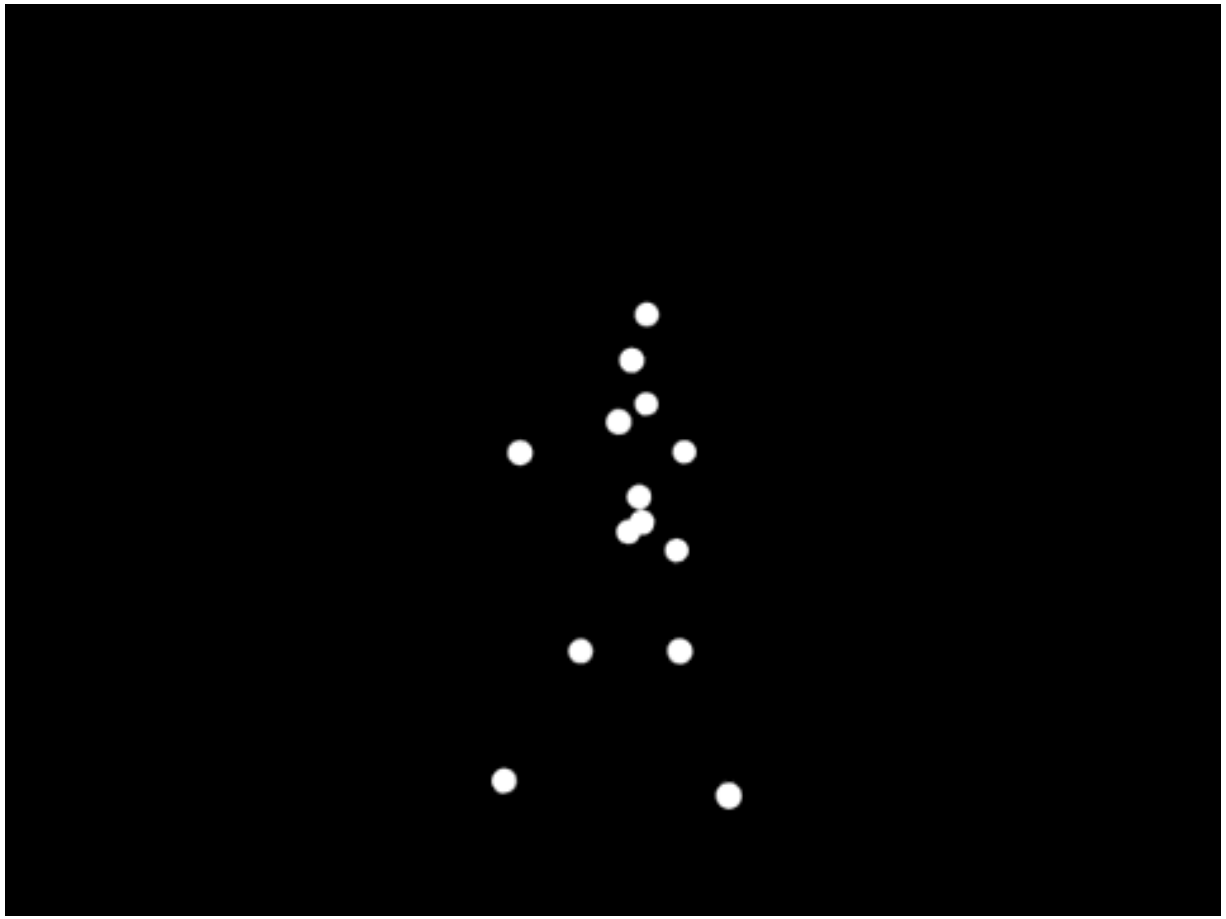
Сценарии съёмки

- Ракурс, вид наблюдаемых объектов и т.д.
- Ракурсы съёмки могут быть крайне различны
- Работающие системы удается создать, «заточившись» на определённый сценарий съёмки

План лекции

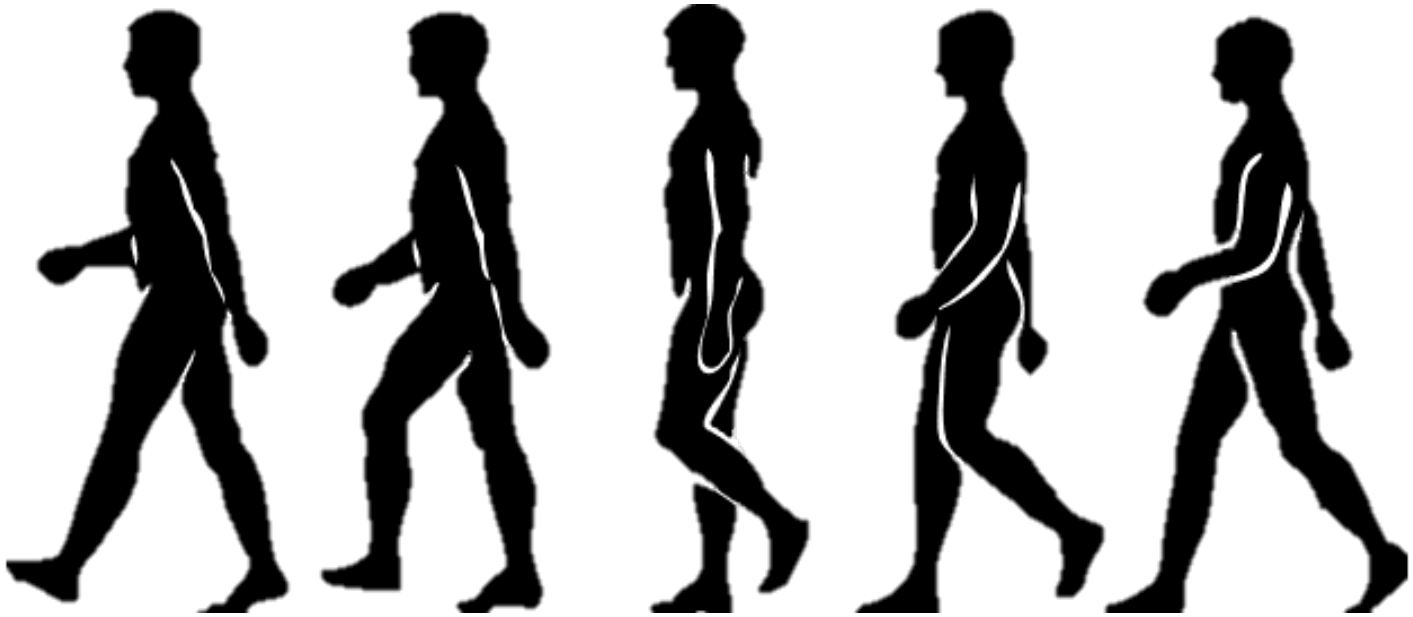
1. Введение в обработку и анализ видео
2. Оптический поток и его оценка
3. Распознавание событий в видео
4. Отслеживание одного объекта
5. Отслеживание множества объектов

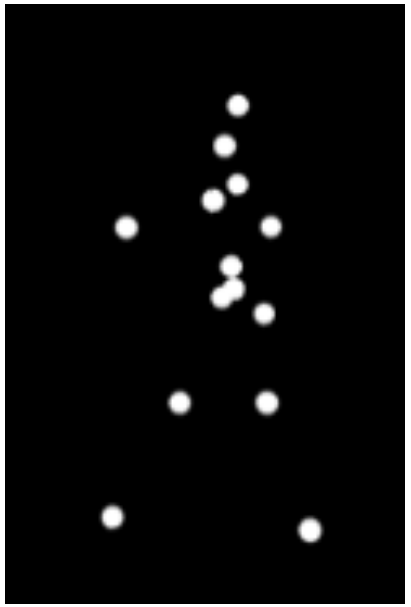
В чём главное отличие видео от изображений?



- Движение – главное отличие видео от изображений
- Движение само по себе является мощной визуальной подсказкой
- Суть многих действий именно в динамике

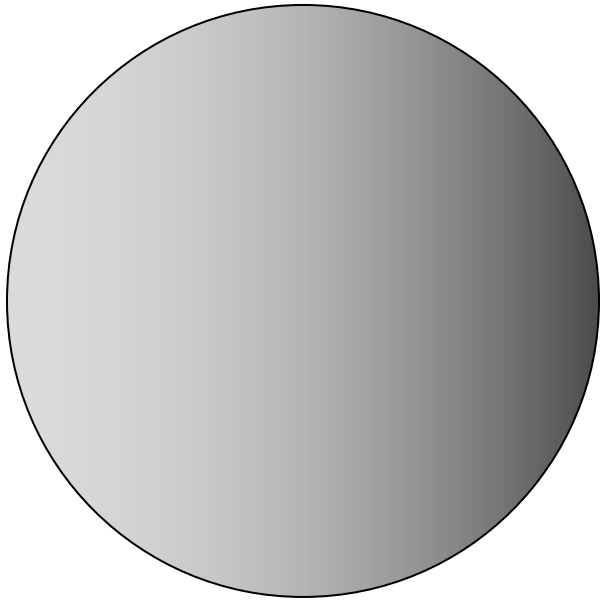
Описание движения





- Точки наблюдаемой сцены движутся относительно камеры / изображения
- Векторное поле движения 2D проекций на изображение 3D точек объектов сцены называется *полем движения* (*motion field*)
- Нужно это движение как-то формализовать, описывать и измерять

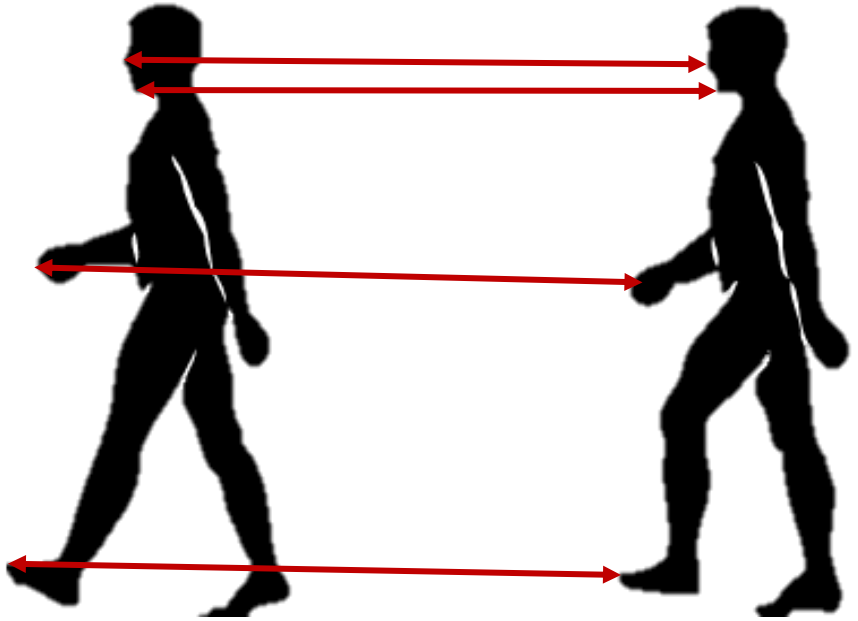
Оптический поток



- Движение точек объектов по видео увидеть можно далеко не всегда
- E.g. Серый матовый шар, освещается с одной стороны и вращается вокруг своей оси

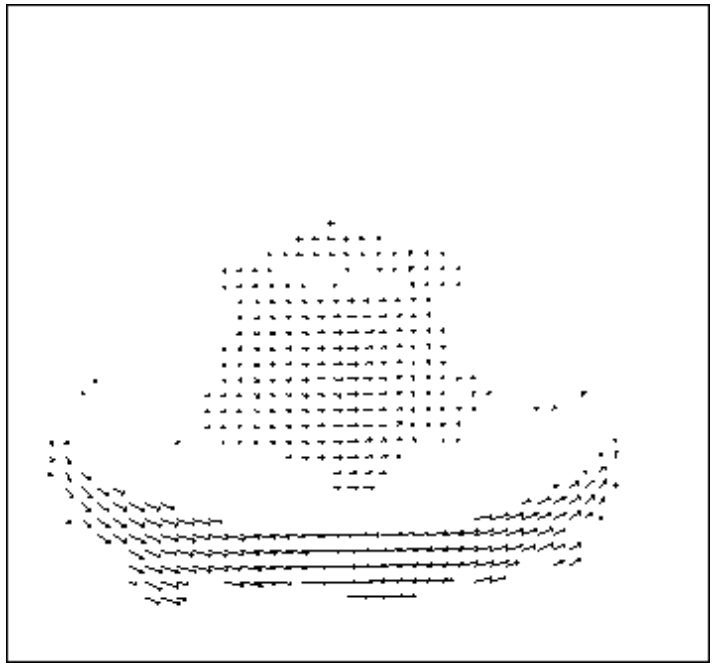
- Optical flow (оптический поток) – векторное поле видимого (apparent) движения пикселей между кадрами
- Вычисление оптического потока – одна из базовых задач анализа видео

Формализация задачи

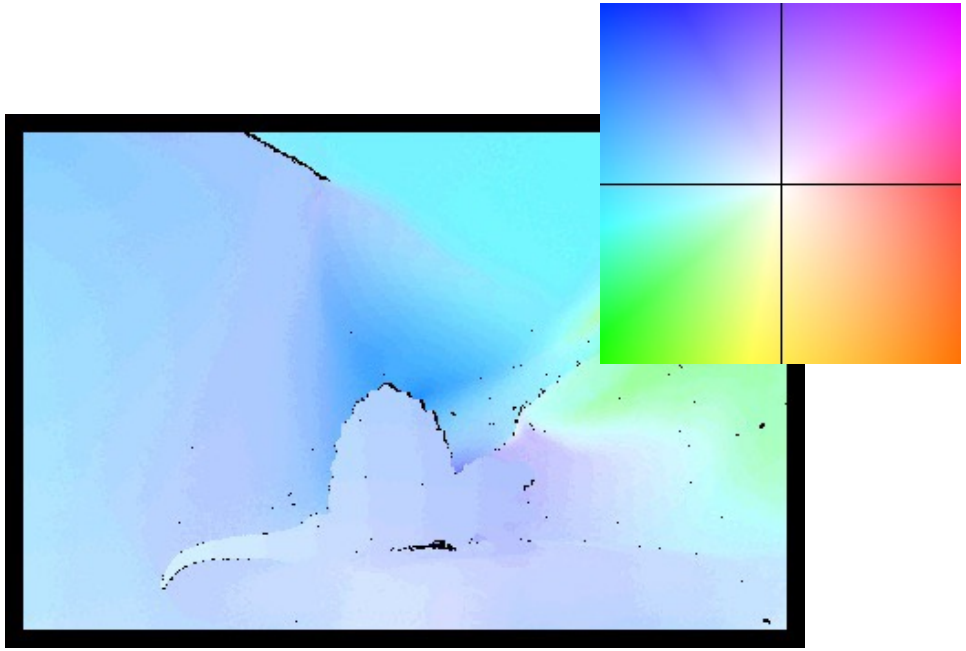


- Optical flow – векторное поле *видимого* движения пикселей между кадрами (u_{ij}, v_{ij})
- Это задача **плотного** сопоставления
- Для каждой точки (x_{ij}, y_{ij}) на первом кадре нужно найти точку на втором кадре $(x_{ij} + u_{ij}, y_{ij} + v_{ij})$

Визуализация



Вектора движения для отдельных точек или всего изображения



Цветовое кодирование вектора движения. Каждому направлению и амплитуде свой цвет и яркость

Оптический поток - векторное поле (u_{ij},v_{ij}) видимого (наблюдаемого) движения пикселей между кадрами

Middleburry optical flow dataset

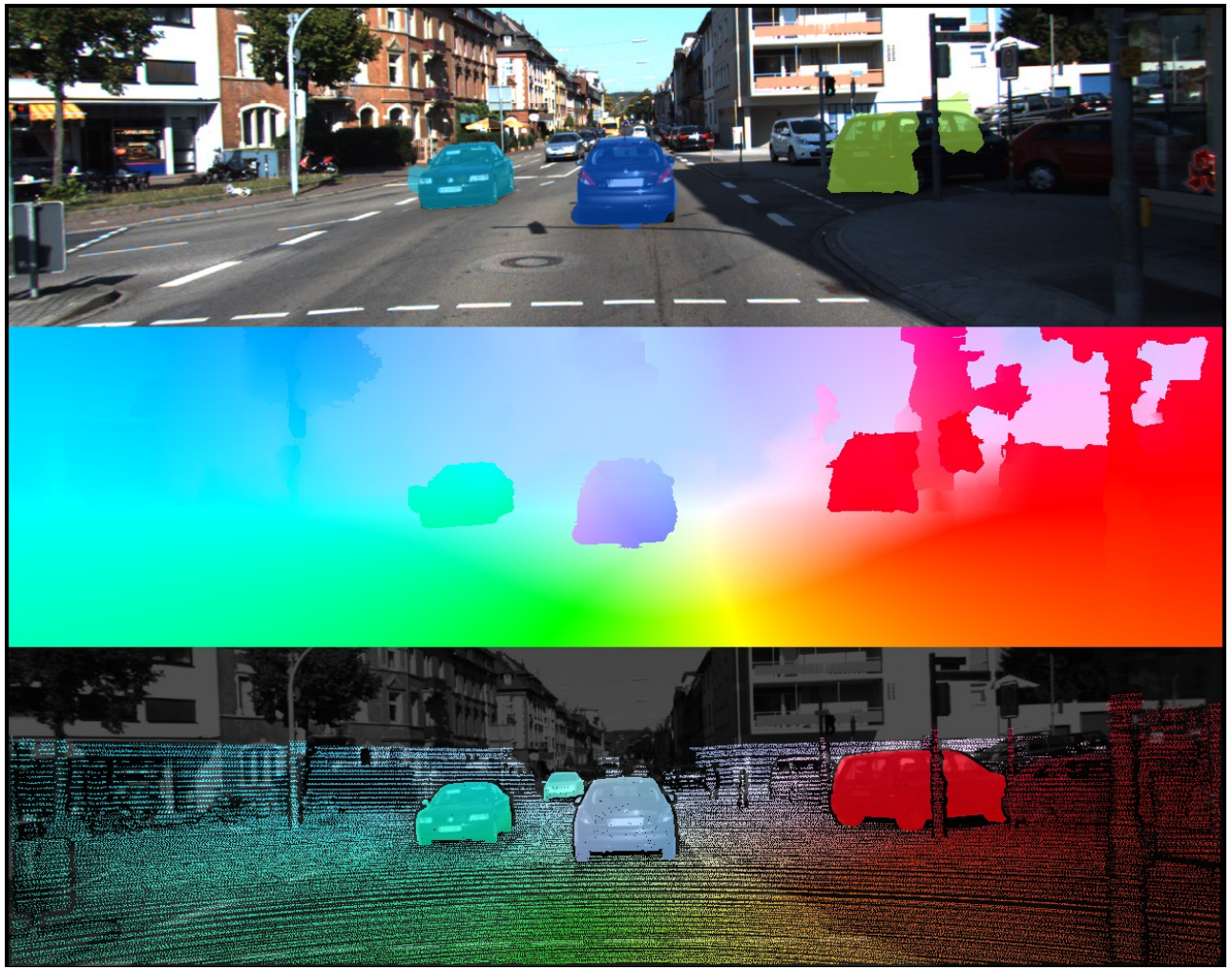
Как получить эталонные данные?

Покадровая съёмка в обычном и флуоресцентном свете

Синтетические данные

Source: S.Baker et. al. A Database and Evaluation Methodology for Optical Flow, IJCV, 2011

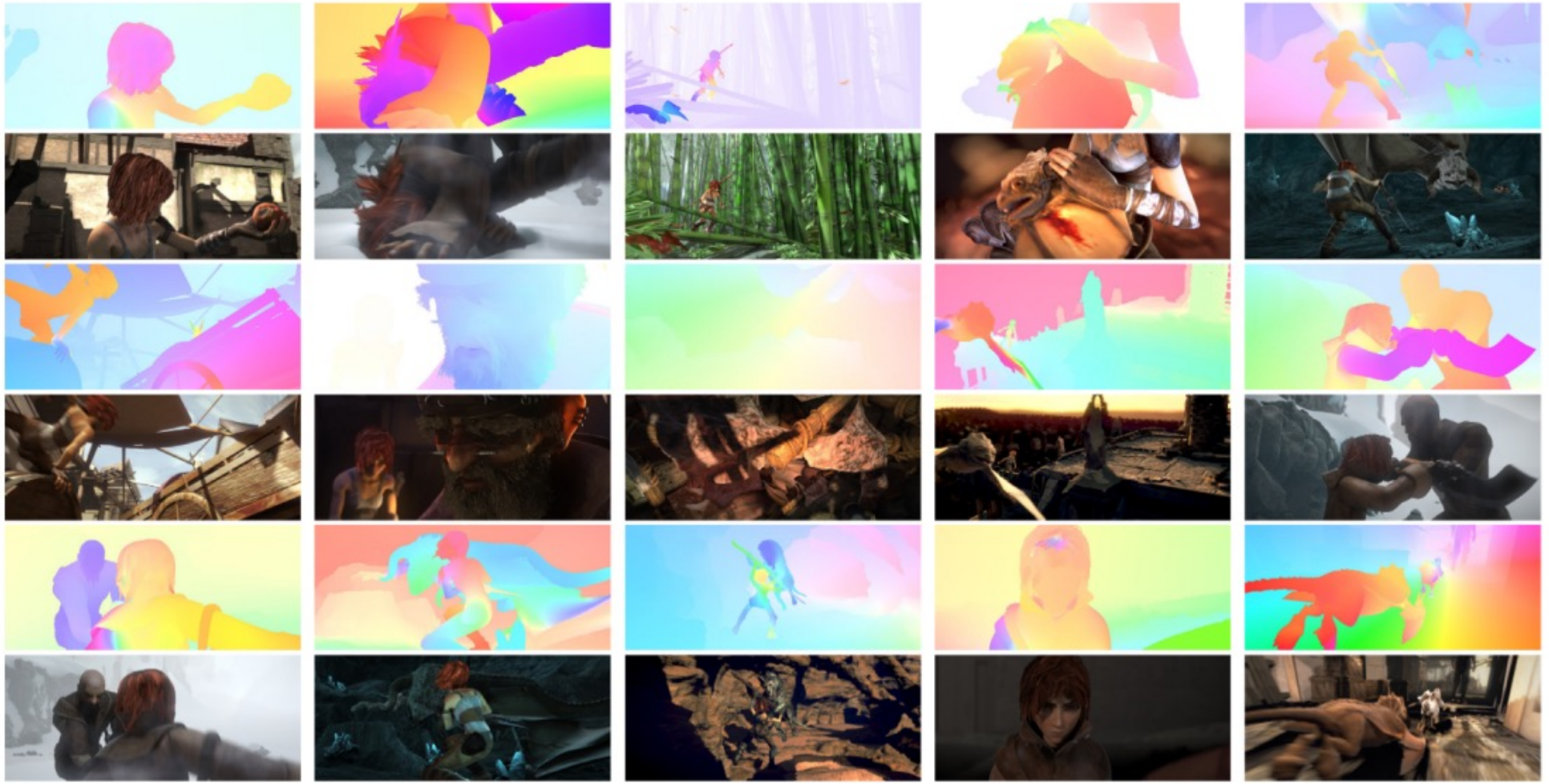
KITTI



- 3D сканирование
- Отдельно фон и движущиеся объекты
- Вписывание 3D моделей объектов в движущиеся объекты

http://www.cvlibs.net/datasets/kitti/eval_flow.php

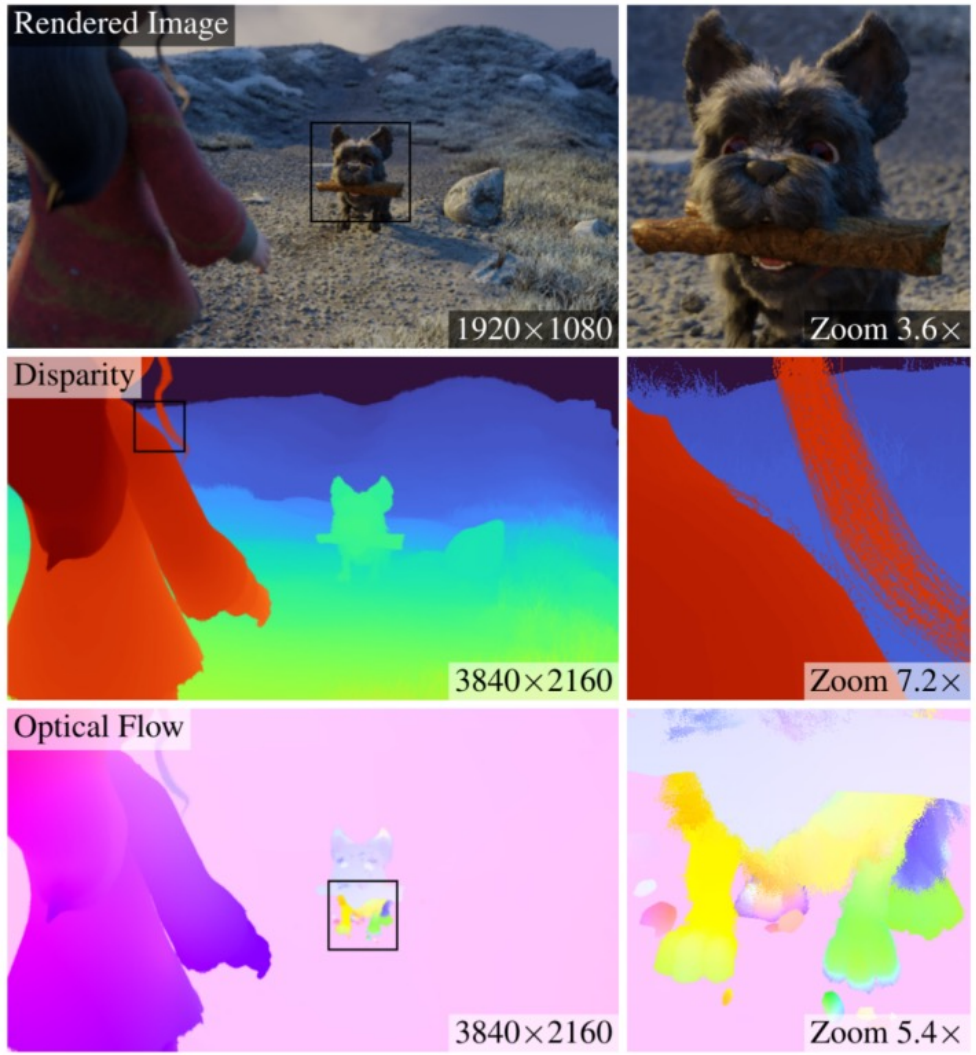
MPI Sintel



1064 training and 564 test frames, 1024 x 436 разрешение

D. Butler et. al. A naturalistic open source movie for optical flow evaluation. ECCV 2012

Spring



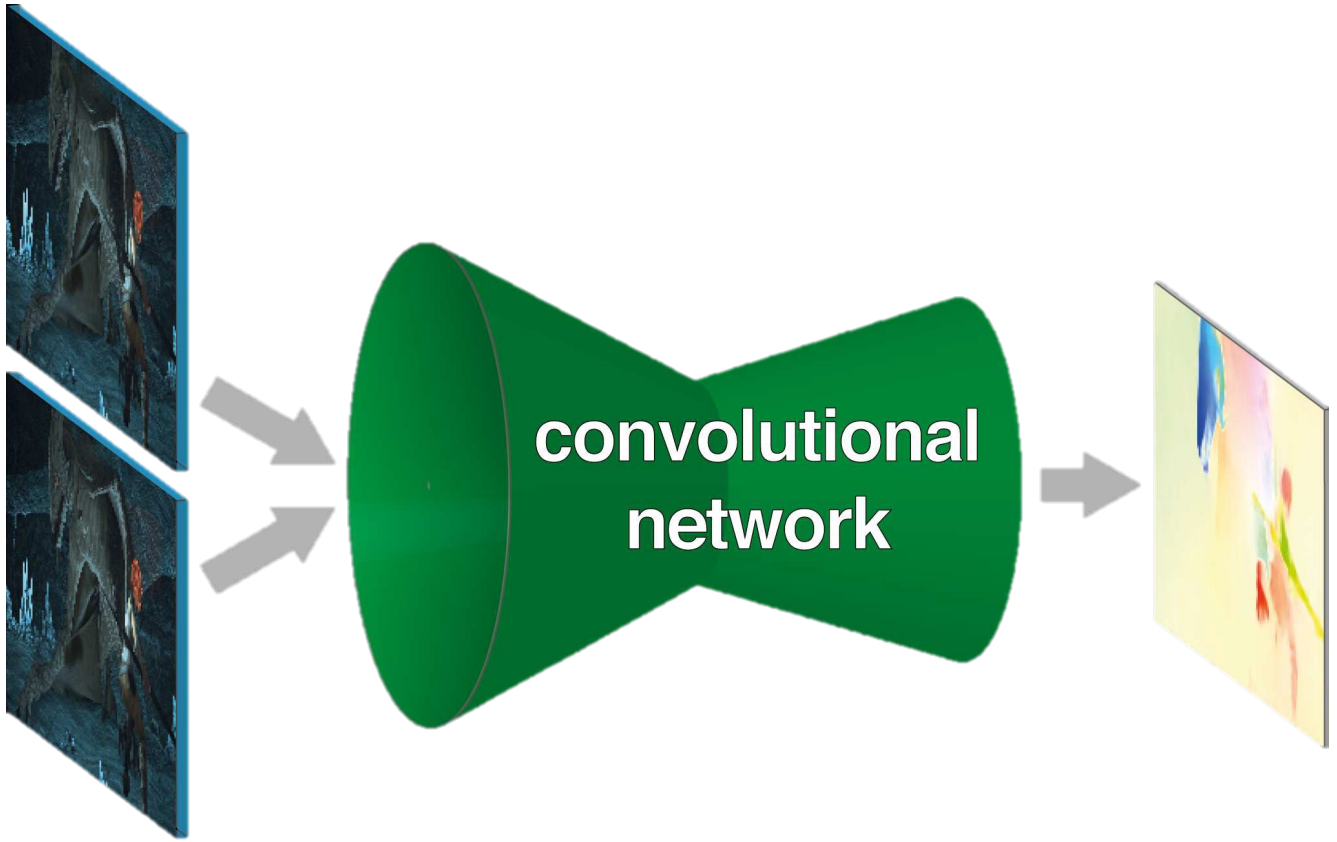
6000 FullHD кадров

4K ground truth

Mehl et al. Spring: A High-Resolution High-Detail Dataset and Benchmark for Scene Flow, Optical Flow and Stereo. CVPR 2023

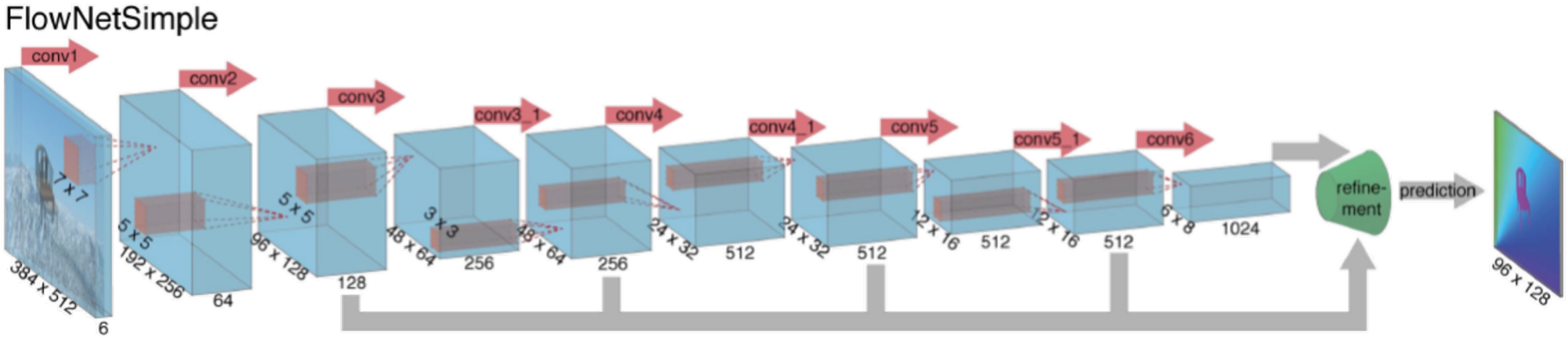
FlowNet

Прямолинейное применение свёрточных нейросетей к вычислению оптического потока



A. Dosovitskiy et. al. FlowNet: Learning Optical Flow with Convolutional Networks. 2015

FlowNet



- Объединяем 2 кадра в 6и канальное изображение
- Применяем нейросеть
- Можно сделать сложнее, объединив признаки с 2х картинок с помощью специального слоя сравнения патчей

A. Dosovitskiy et. al. FlowNet: Learning Optical Flow with Convolutional Networks. 2015

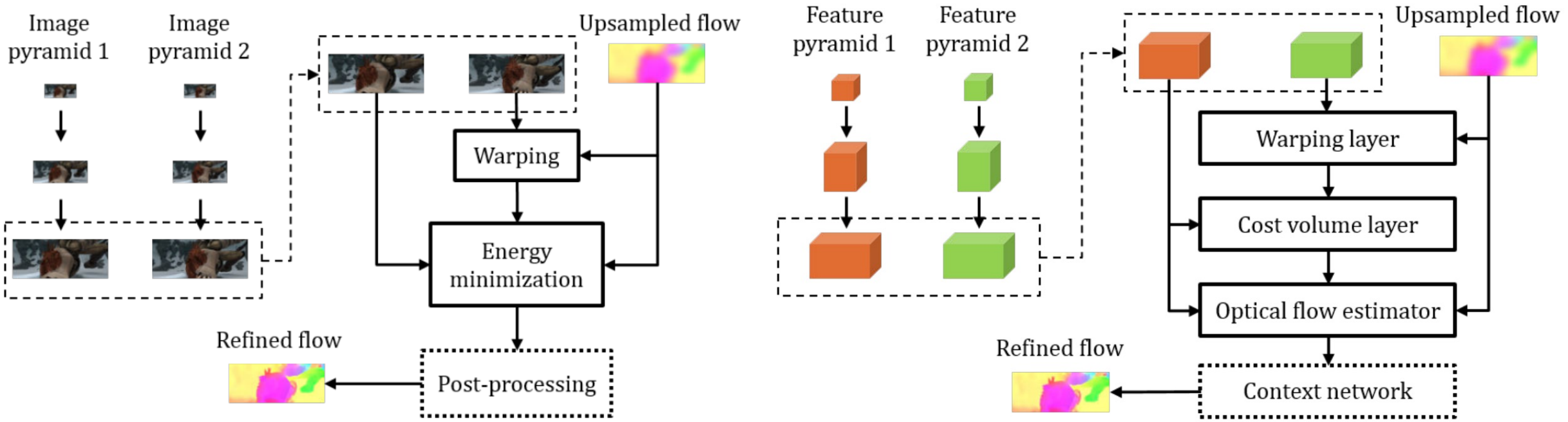
Flying chairs

	Frame pairs	Frames with ground truth	Ground truth density per frame
Middlebury	72	8	100%
KITTI	194	194	~50%
Sintel	1041	1041	100%
Flying Chairs	22872	22872	100%

- Не хватает данных для обучения
- Поэтому синтетический набор из летающих стульев

A. Dosovitskiy et. al. FlowNet: Learning Optical Flow with Convolutional Networks. 2015

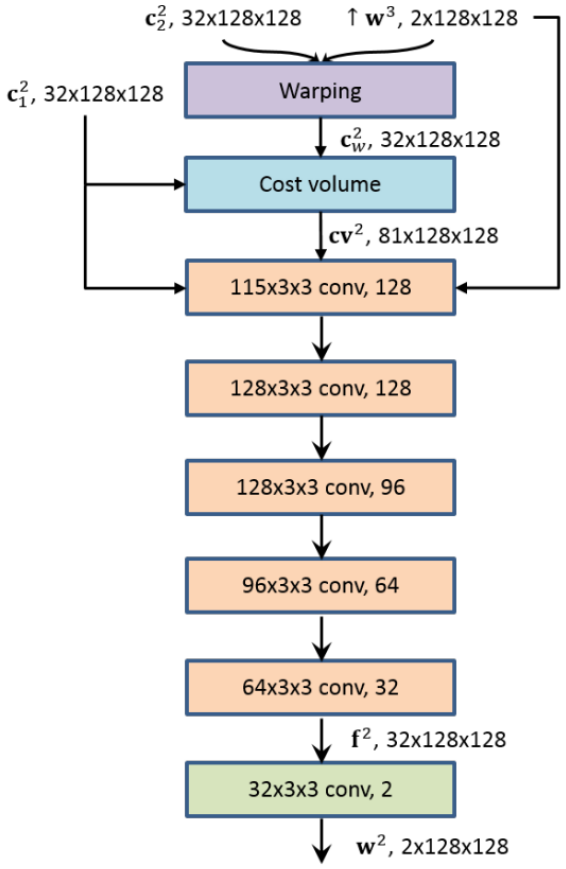
PWC-Net (Pyramid, Warping, Cost volume)



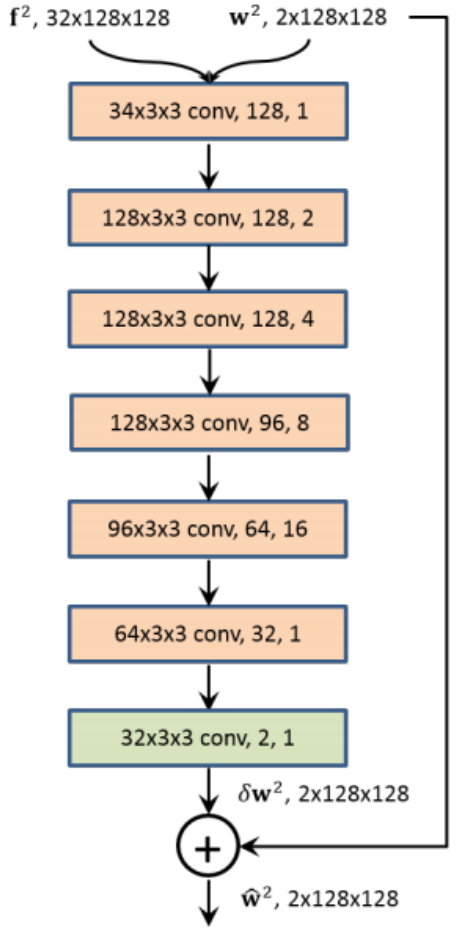
- Реализуем классический подход, но через нейросетевые признаки и нейросетевой вывод
- Вычисляем “Cost Volume” через корелляцию пикселей со сдвигом не более d
- Размер CV = $d^2 * Width * Height$
-

Sun et. al. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. CVPR2018

Некоторые детали

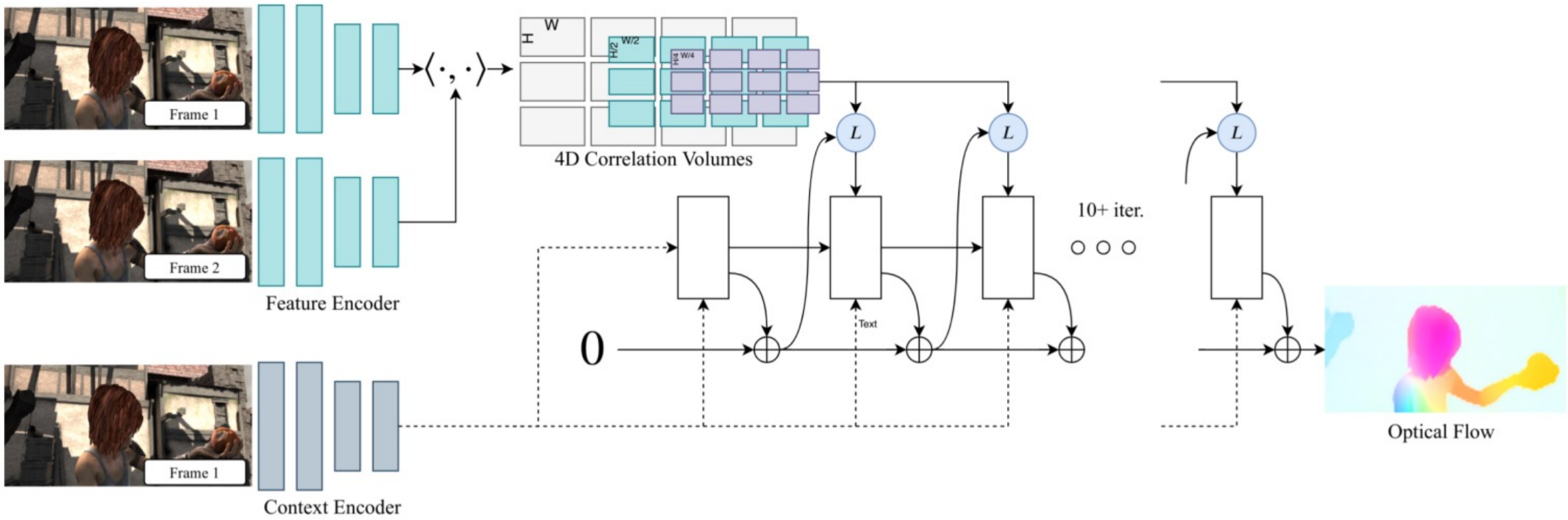


OF Estimator берёт на вход Cost Volume, Optical Flow, и Features



- Сеть Context делает пост-обработку
- На вход признаки и OF

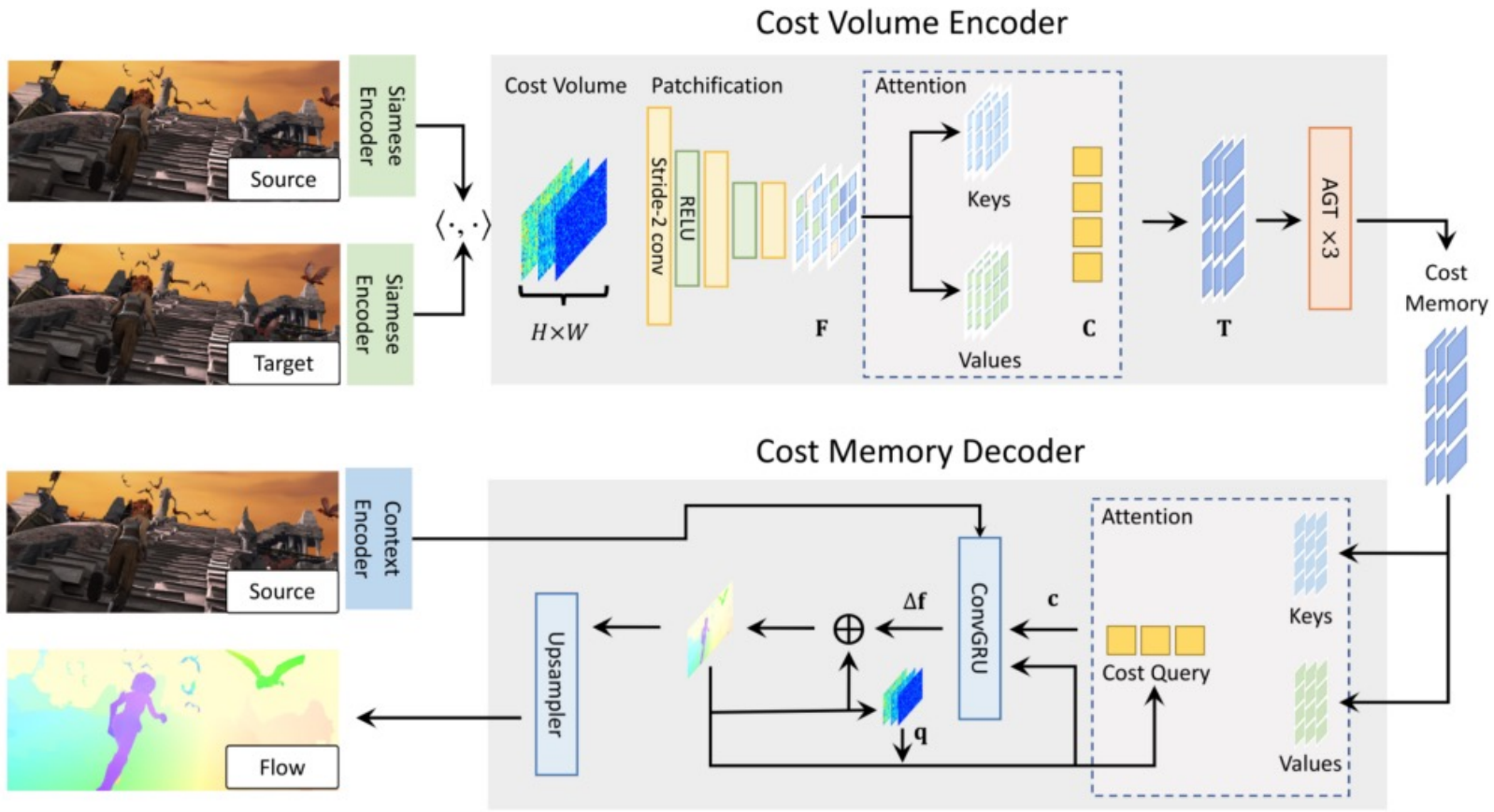
RAFT



- Поддерживаем оценку OF в высоком разрешении без пирамиды
- Используем рекурентный элемент для уточнения OF

Teed, Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. ECCV 2020

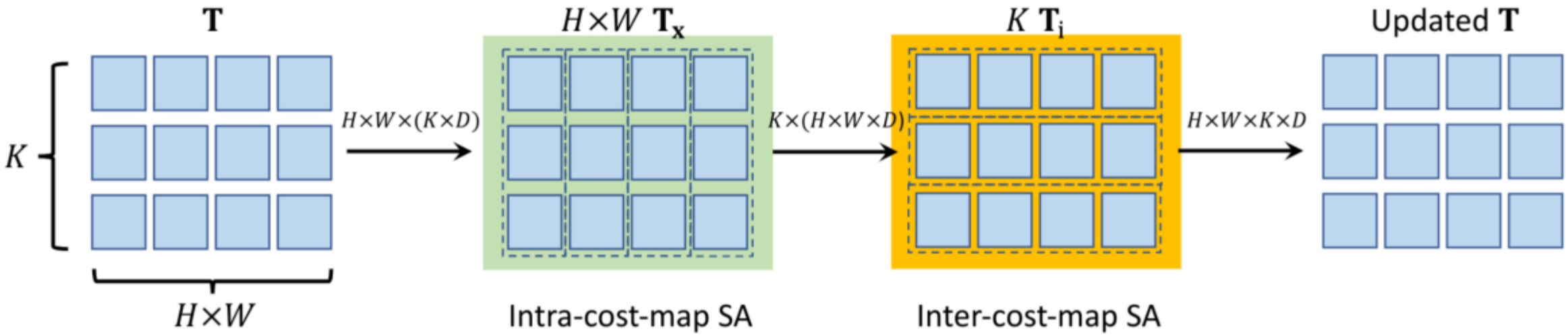
FlowFormer



- Сжимаем cost volume в токены ($H^2 \times W^2 \rightarrow H \times W \times K \times D$)
- 2-х стадийное внимание: вначале для токенов внутри cost map и затем между cost maps
- Декодер с cost queries

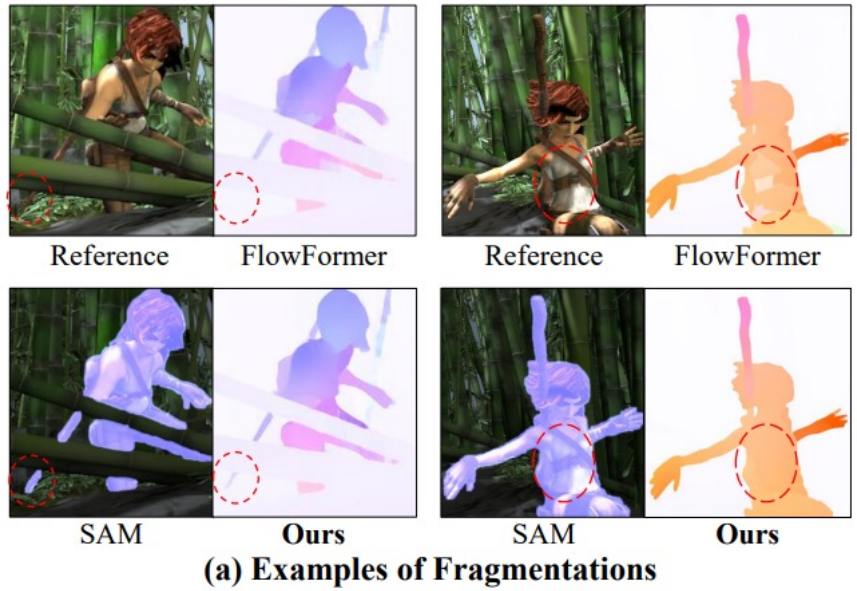
Huang et al. FlowFormer: A Transformer Architecture for Optical Flow. ECCV 2022

FlowFormer



Huang et al. FlowFormer: A Transformer Architecture for Optical Flow. ECCV 2022

SAMFlow



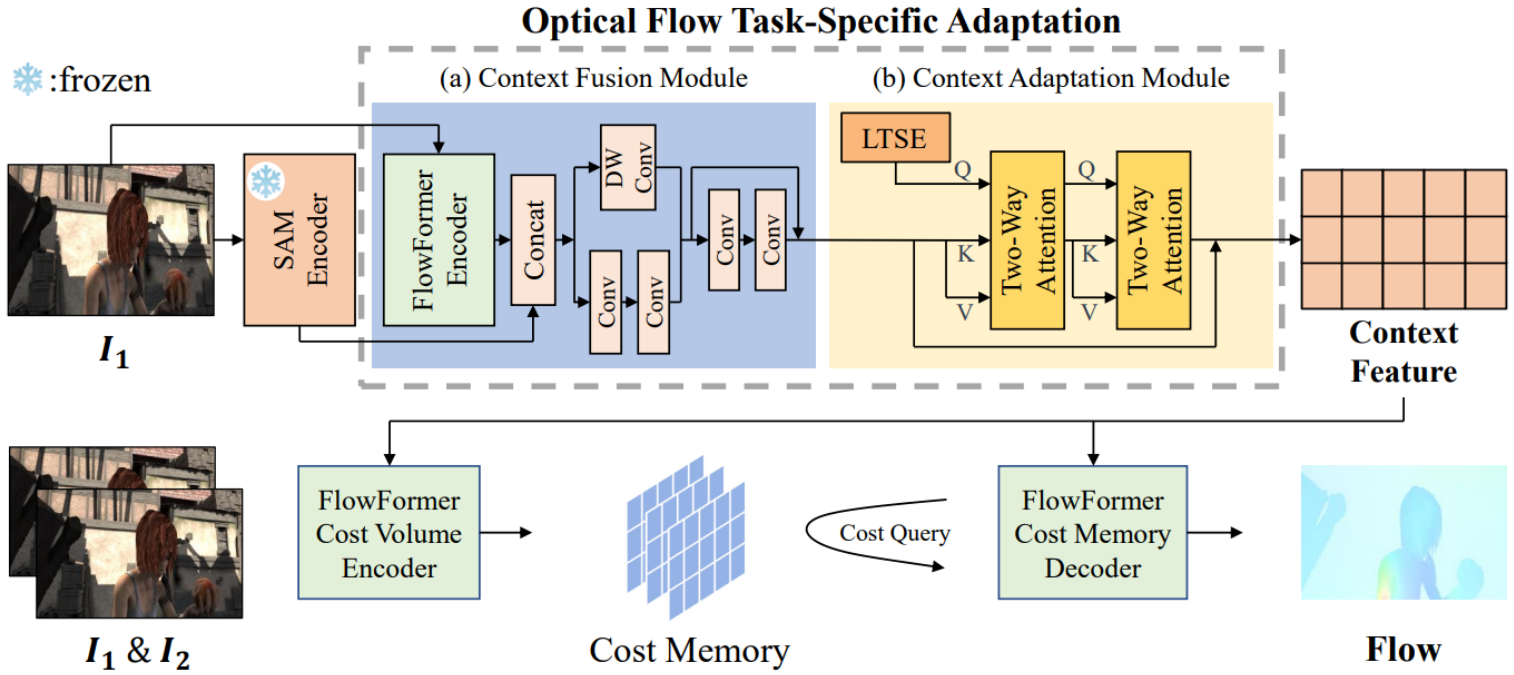


Figure 2: The overview of our SAMFlow, which utilizes the frozen SAM image encoder to boost the object perception of the optical flow model FlowFormer. We design two modules for in-depth utilizing SAM, including: (a) the CFM, which fuses SAM features with FlowFormer encoder, and (b) the CAM, which adapts the features with the Learned Task-Specific Embedding.

Всё становится лучше, если добавить признаки от фундаментальной модели

Zhou et. al. SAMFlow: Eliminating Any Fragmentation in Optical Flow with Segment Anything Model. AAAI2024

План лекции

1. Введение в обработку и анализ видео
2. Оптический поток и его оценка
3. Распознавание событий в видео
4. Отслеживание одного объекта
5. Отслеживание множества объектов

Action recognition

Human actions are the main content of movies, TV news and shows, home video and video surveillance

- Smart surveillance
  - *Abnormal situation detection*
- Video archive indexing and retrieval
  - *Search for a scene with Putin and Obama handshake*
- Content navigation
  - *Rewind to the next goal in the soccer match*

Source: http://www.di.ens.fr/~laptev/actions/hollywood2/

Human action

The most basic actions are simple body movements like walking or running or clapping hands.

Walking

Jogging

Running

Boxing

Waving

Clapping

Source: http://www.nada.kth.se/cvap/actions/

Actions

Short meaningful movements

Answer phone

Handshake

Actions and events

A set of small actions with a specific common goal can still be called an “action” but we can also call them “events”

Making sandwich

Doing homework

Source: http://trecvid.nist.gov/

Events

An event can include a lot of different actions of different people

Birthday party

Parade

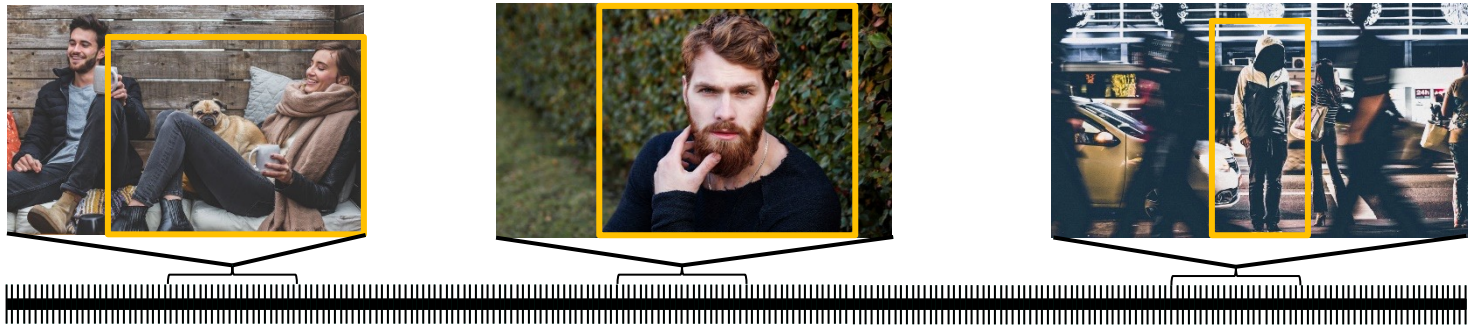
Source: http://trecvid.nist.gov/

Tasks

Action or event classification: assign a label of action or event to a video

Making sandwich: present
Feeding animal: not present
…

Action localization: search for a spatial region and time interval of a specific action in a video



Source: http://trecvid.nist.gov/

Actions and motions

- Actions are usually defined by motions
- But many visually similar motions can have very different meanings and correspond to different actions

Source: http://www.di.ens.fr/willow/teaching/recvis11/slides/lecture10_video.pdf

Datasets

It is much easier to collect and annotate datasets for action recognition compared to object tracking or optical flow estimation

Kiss

Drink

Source: http://www.di.ens.fr/~laptev/actions/hollywood2/

KTH Actions (2004)

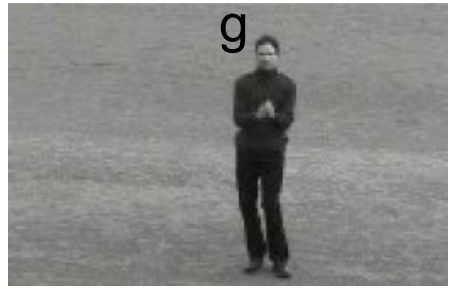
Walking

Jogging

Runnin
g

Boxing

Waving

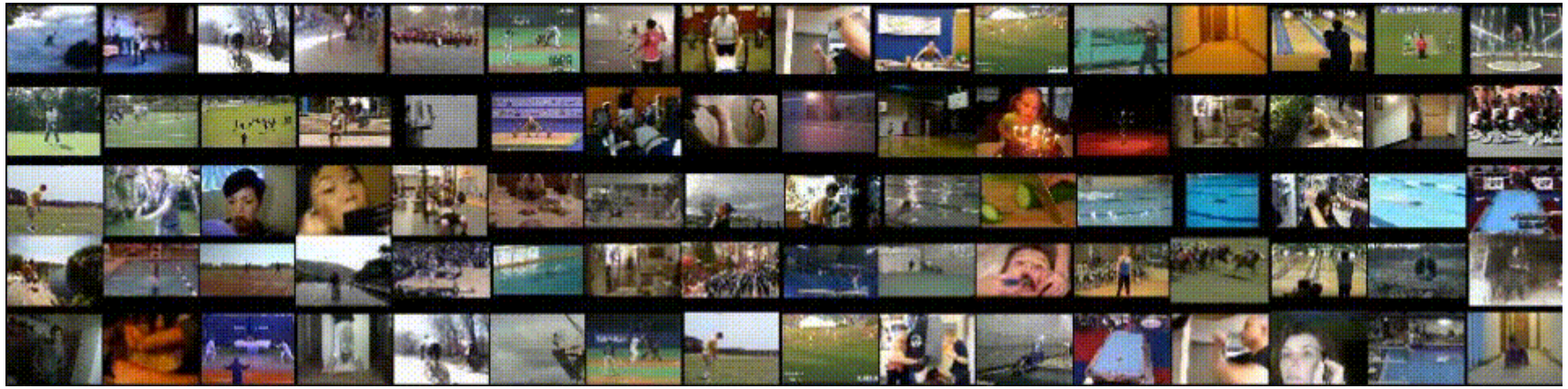


Clapping

- 25 people, 6 actions, 4 scenes (indoor, outdoor, outdoor with different scale, outdoor with different clothes)
- 2391 video

Source: C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. ICPR, 2004.

UCF101 (2012)



- 13320 videos from Youtube in 101 classes
- Five groups: (1)Human-Object Interaction; (2) Body-Motion Only; (3) Human-Human Interaction; (4) Playing Musical Instruments; (5) Sports.
- Examples: to dye a lips, to tint eyes
- Accuracy from 43.9 (baseline Oct 2013) to 87.9% (Oct 2014)

Source: Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild., CRCV-TR-12-01, November, 2012.

TrecVid MED'13



- 100 positive video clips per event category, 5000 negatives
- Testing on 98000 videos clips, i.e. 4000 hours
- 20 known events, 10 adhoc events
- Videos from publicly available, user-generated content on various Internet sites
- Various descriptors for video, audio, text & speech recognition

Source: http://www-nlpir.nist.gov/projects/tv2013/tv2013.html

Kinetics

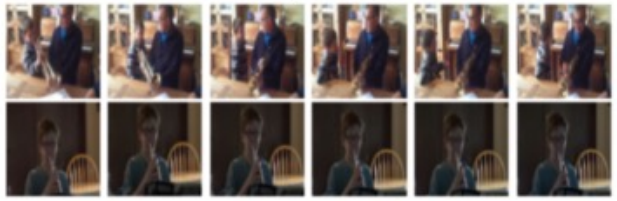
riding a bike

playing violin

braiding hair

dribbling basketball

riding unicycle



playing trumpet

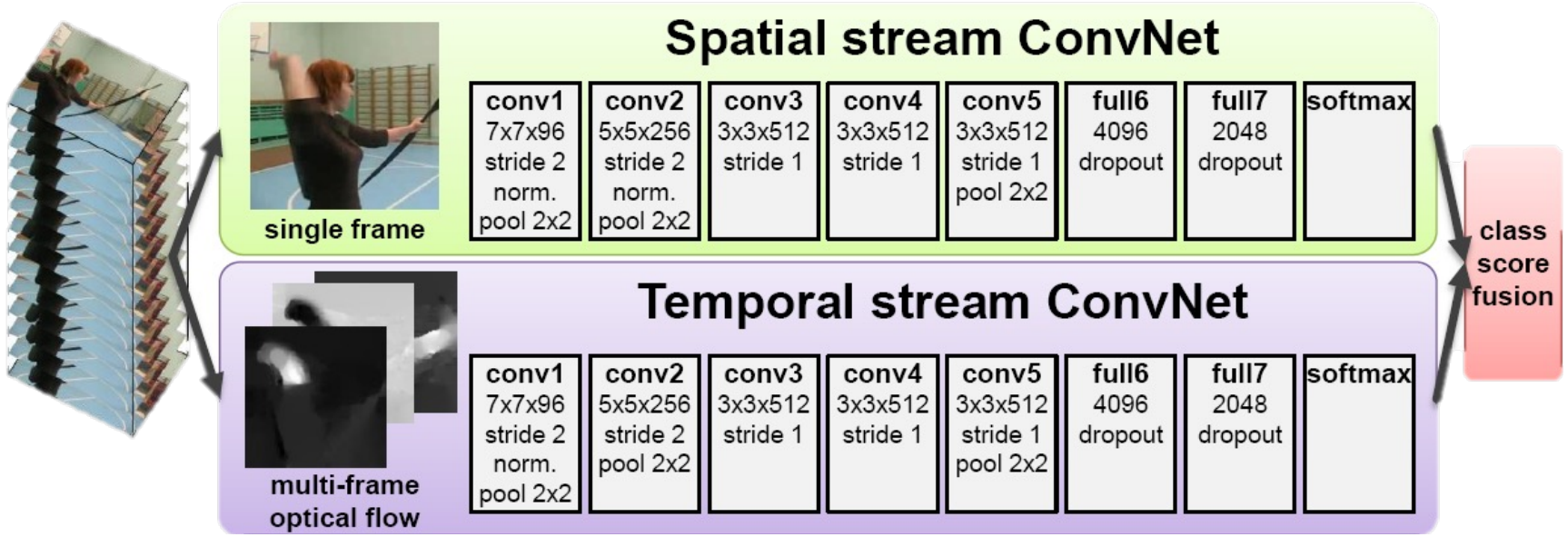
brushing hair

dunking basketball

- large video dataset scraped from YouTube
- 400, 600, 700 classes
- 300k, 650k, 700k videos

Kay et al. The Kinetics Human Action Video Dataset. arXiv:1705.06950

Two-stream CNN model



Temporal stream:

- Vertical and horizontal optical flow components
- 5-10 frames (input tensor $w \cdot h \cdot 2L$)
- Mean flow can be subtracted from each frame

Source: Karen Simonyan, Andrew Zisserman Two-Stream Convolutional Networks for Action Recognition in Videos, NIPS 2014

Самообучение

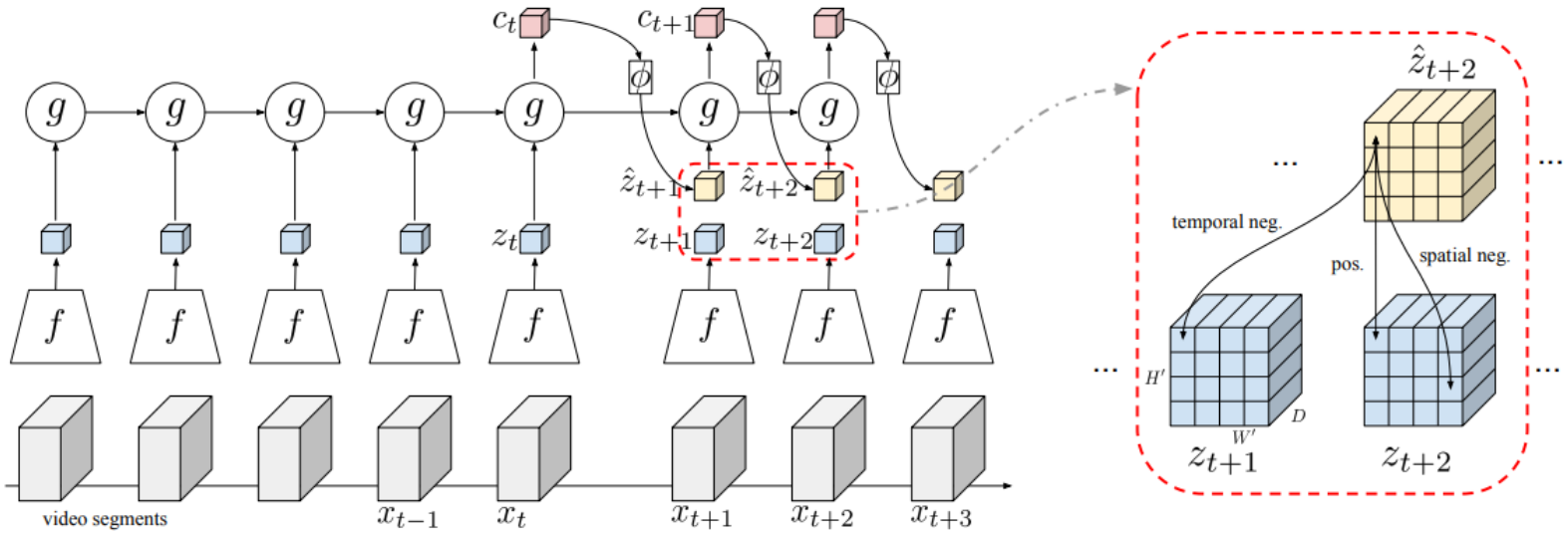


Figure 2: A diagram of **Dense Predictive Coding** method. The left part is the pipeline of the DPC, which is explained in Sec. 3.1. The right part (in the dashed rectangle) is an illustration of the Pred-GT pair construction for contrastive loss, which is explained in Sec. 3.2.

- Развитие self-supervised методов
- Учимся на задаче без разметки – предсказания признаков следующих видеофрагментов по предыдущим
- Нужно строить «похожие» на реальность

https://arxiv.org/pdf/1909.04656.pdf

План лекции

1. Введение в обработку и анализ видео
2. Оптический поток и его оценка
3. Распознавание событий в видео
4. Отслеживание одного объекта
5. Отслеживание множества объектов

Visual object tracking (VOT)

- Произвольный объект выделен на первом кадре
  - Про него ничего не знаем, нет «детектора»
- Мы хотим отследить движение объекта во всех последующих кадрах
- Выход – «следы» (track) объектов, последовательность локаций в каждом кадре видео

Сложность задачи

- Вычислительная нагрузка
  - Нужно обрабатывать N кадров в секунду
- Изменение по времени
  - Вид объекта меняется от кадра к кадру из-за ракурса, изменения освещения, внутренних изменений (скейтбордист)
- Взаимодействие объектов
  - Перекрытия объектов
  - Визуальное сходство объектов
  - И т.д.

VOT Challenge

Основной конкурс для оценки методов визуального трекинга
http://votchallenge.net/

- Открытые реализации
- Оценка качества и скорости
- Небольшой, но разнообразный набор роликов
- Короткие ролики (100 frames)

Развитие:

- Добавление других модальностей (RGB + Depth, RGB + IR)
- Появление «долгого» отслеживания, с пропаданием и появлением объектов

Примеры последовательностей

Most challenging:

Matrix

Rabbit

Butterfly

Least challenging:

Singer

Octopus

Sheep

Source: http://votchallenge.net/

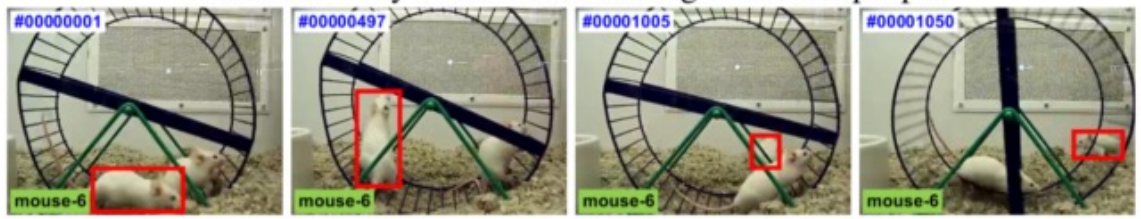
LaSOT

Bear-12: "white bear walking on grass around the river bank"

Bus-19: "red bus running on the highway"

Horse-1: "brown horse running on the ground"

Person-14: "boy in black suit dancing in front of people"

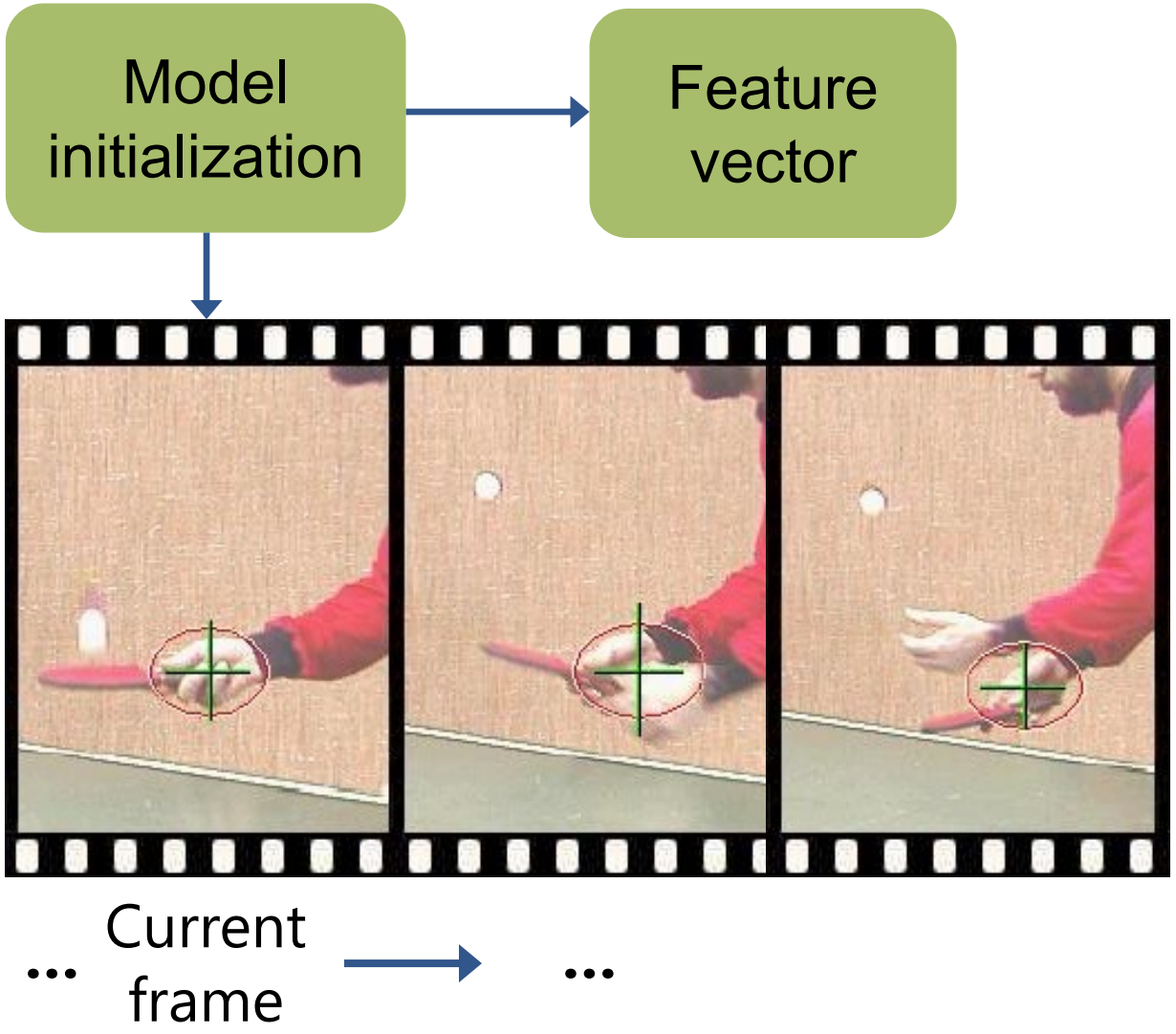


Mouse-6: "white mouse moving on the ground around another white mouse"

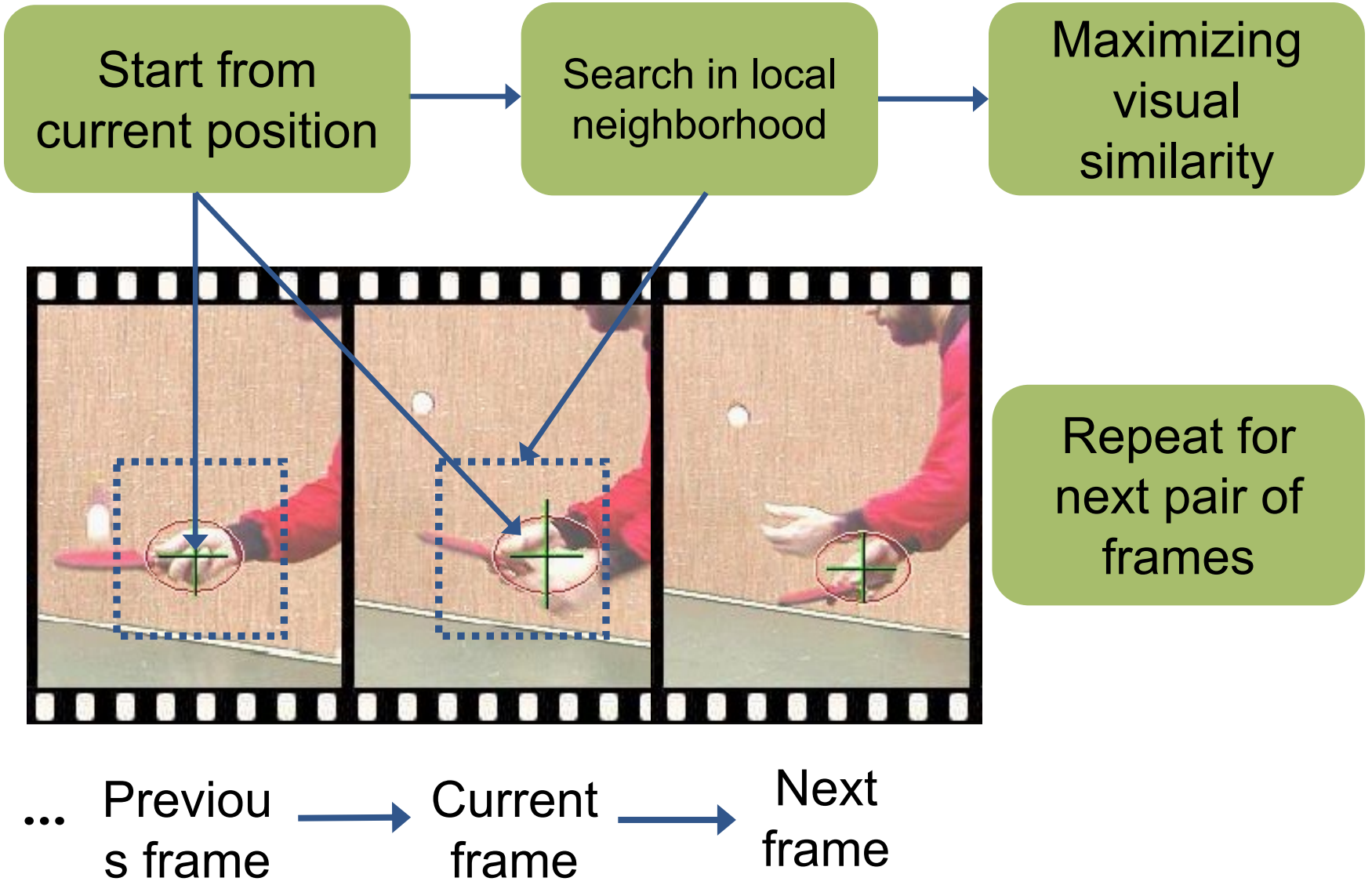
- 1400 videos (YouTube CC license)
- 84s duration on average
- 3.5M frames in total
- 70 classes chosen for popular applications
- labelled by 10 volunteers and PhD students

Han et al. LaSOT: A High-quality Benchmark for Large-scale Single Object Tracking. CVPR 2019

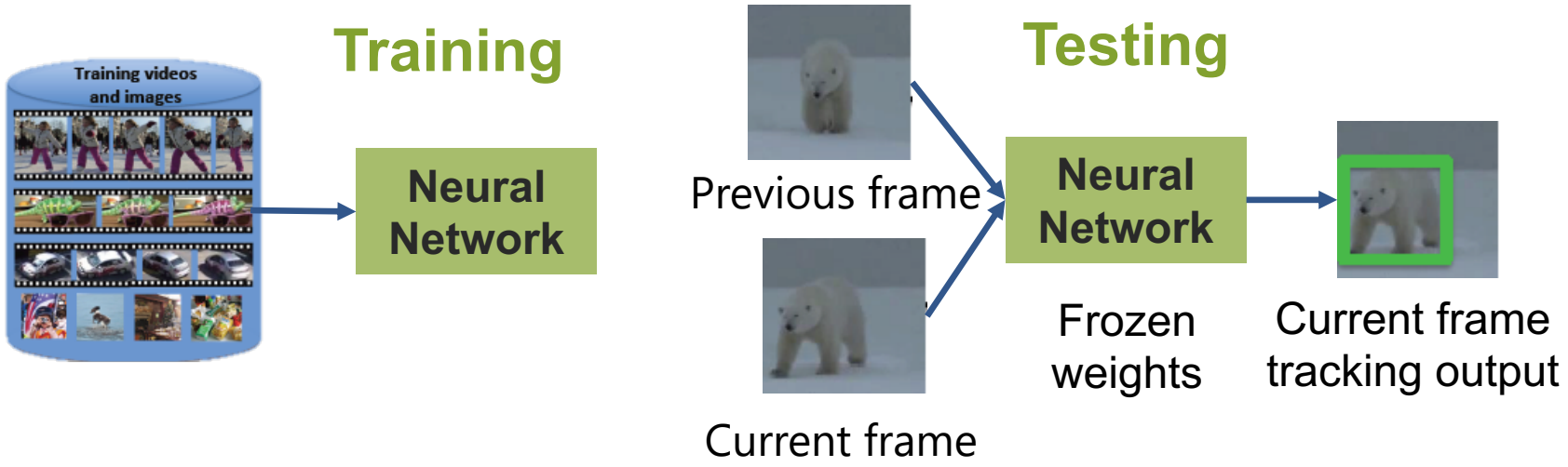
Стандартная схема итеративного трекинга



Стандартная схема итеративного трекинга



GOTURN

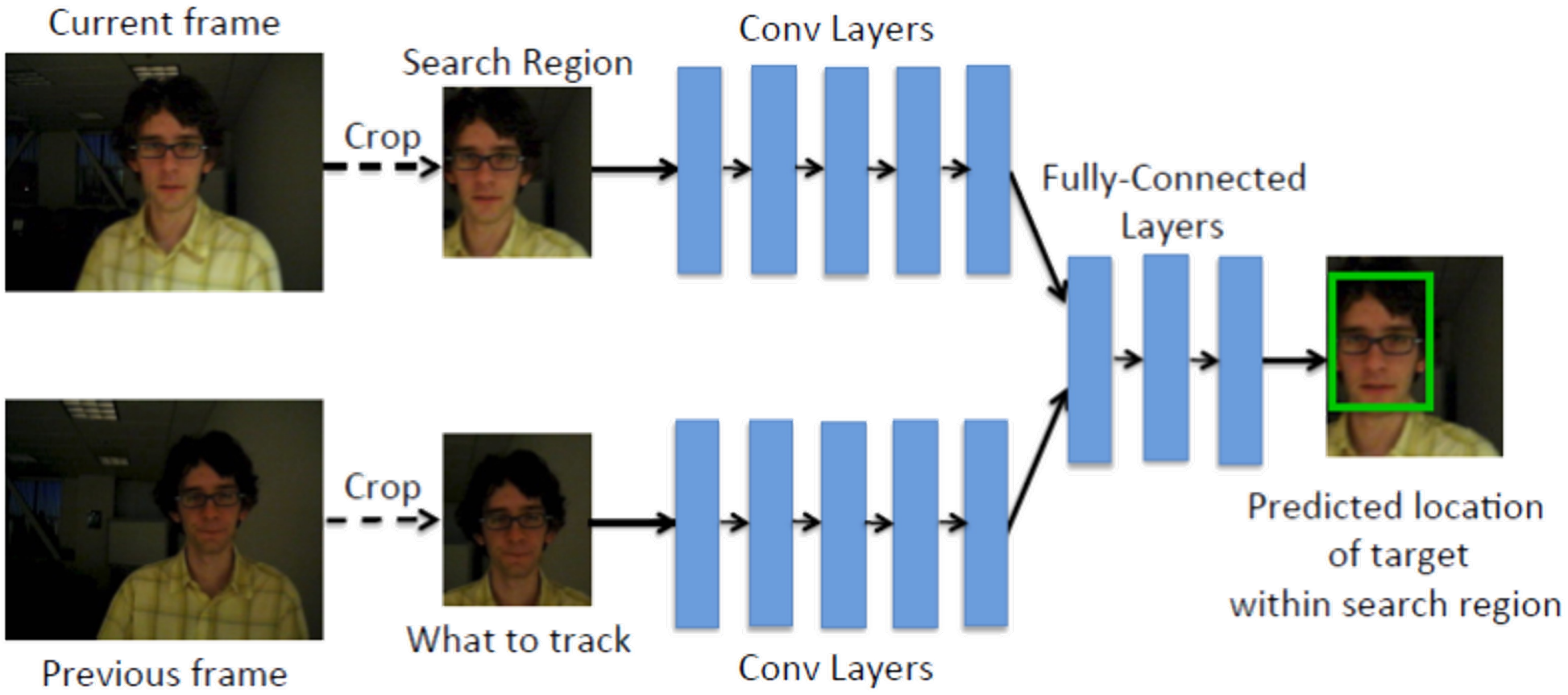


Generic Object Tracking Using Regression Networks (GOTURN):

- Обучаем нейросеть на коллекции видео и изображений с bounding box
- Просто применяем её к паре кадров и за счёт этого получаем скорость 100 кадров/сек

Source: Held et.al. Learning to Track at 100 FPS with Deep Regression Networks, ECCV 2016

GOTURN network



Source: Held et.al. Learning to Track at 100 FPS with Deep Regression Networks, ECCV 2016

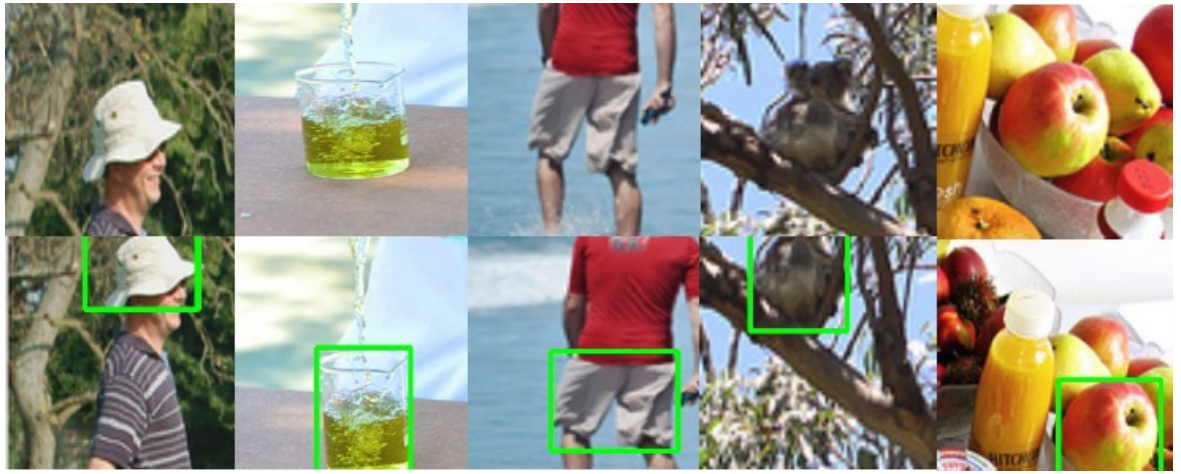
Training GOTURN

Previous video frame centered on object

Current video frame, shifted, with ground-truth bounding box

Image centered on object

Shifted image with ground-truth bounding box



Small shifts are sampled more often than large shifts

Source: Held et.al. Learning to Track at 100 FPS with Deep Regression Networks, ECCV 2016

STARK

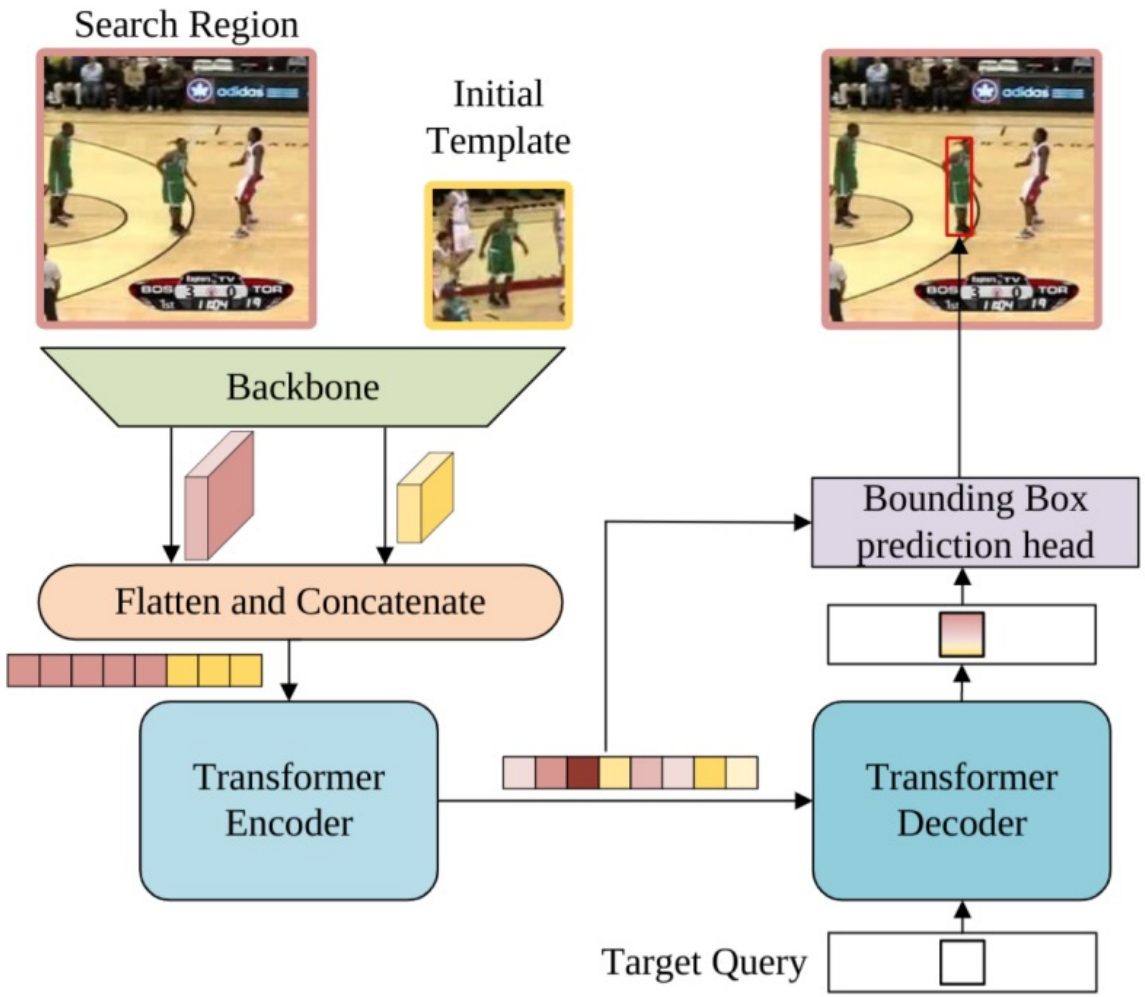


Figure 2: Framework for spatial-only tracking.

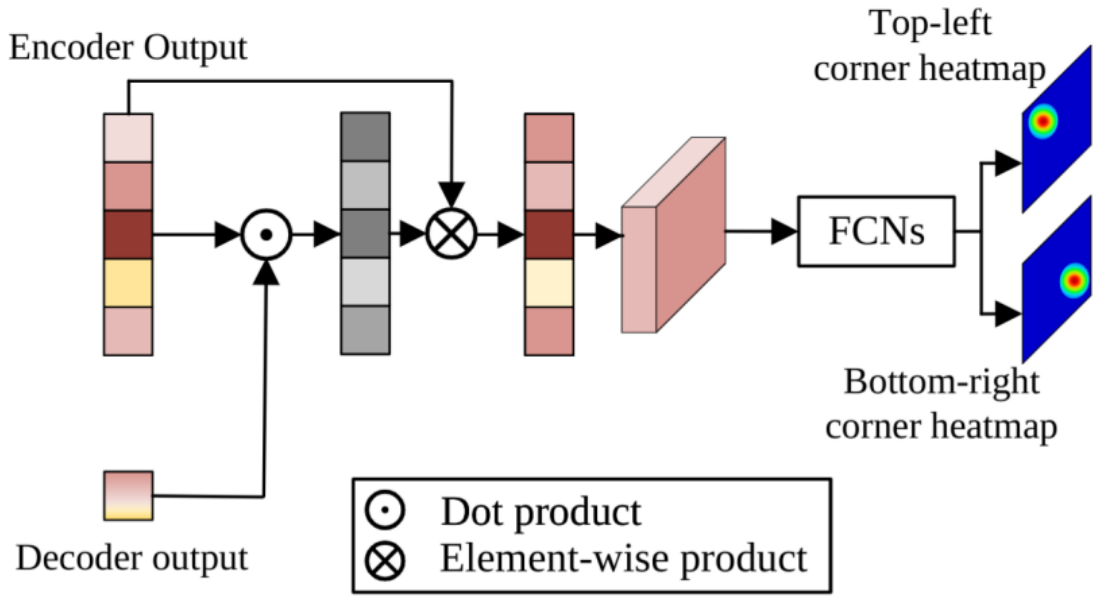


Figure 3: Architecture of the box prediction head.

Yan et al. Learning Spatio-Temporal Transformer for Visual Tracking. ICCV 2021

STARK

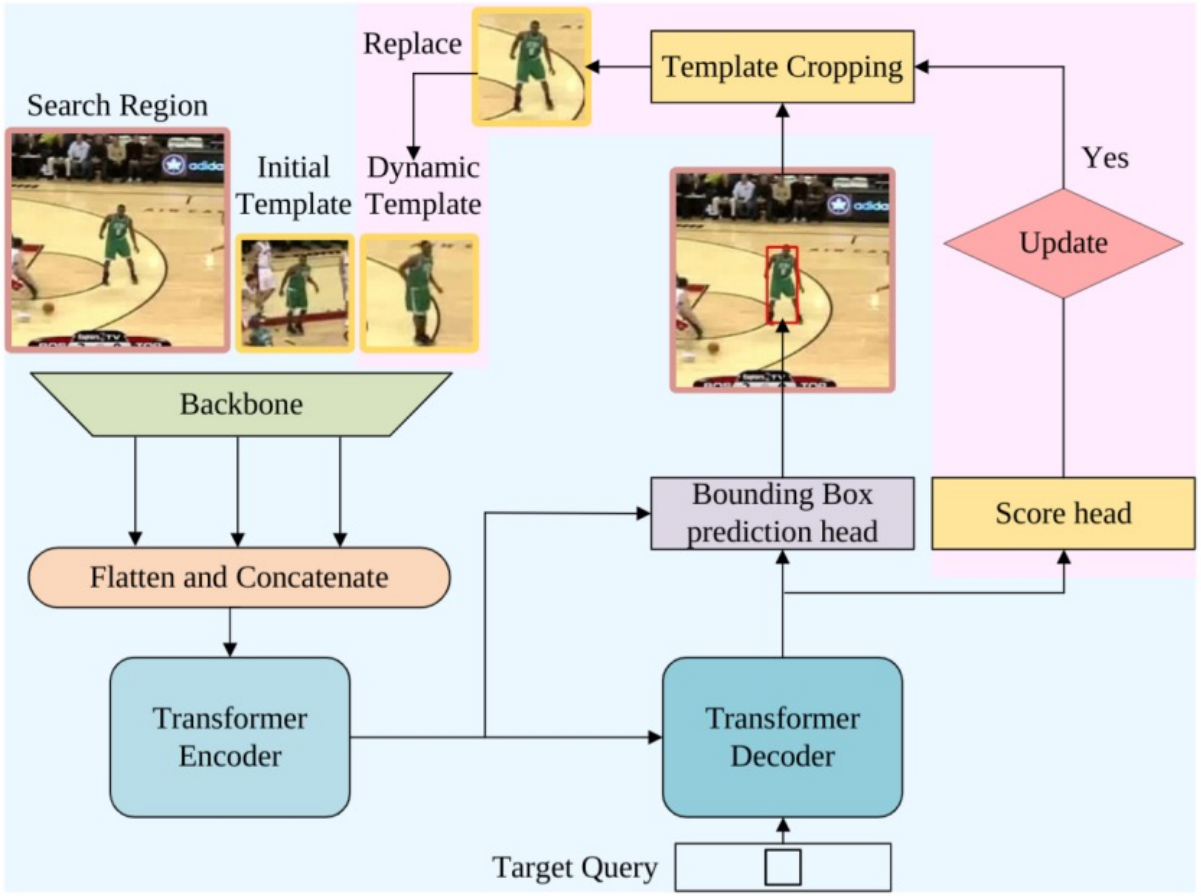


Figure 4: Framework for spatio-temporal tracking. The differences with the spatial-only architecture are highlighted in pink.

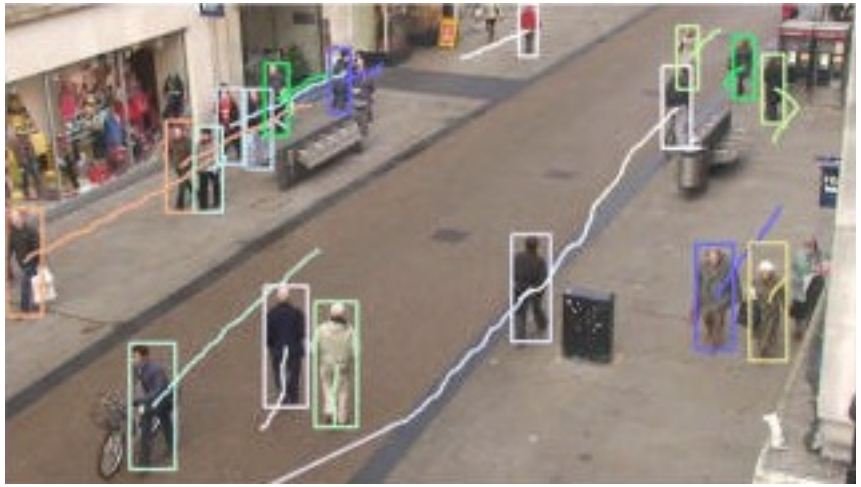
Two stage-training: first localization, then classification

Yan et al. Learning Spatio-Temporal Transformer for Visual Tracking. ICCV 2021

План лекции

1. Введение в обработку и анализ видео
2. Оптический поток и его оценка
3. Распознавание событий в видео
4. Отслеживание одного объекта
5. Отслеживание множества объектов

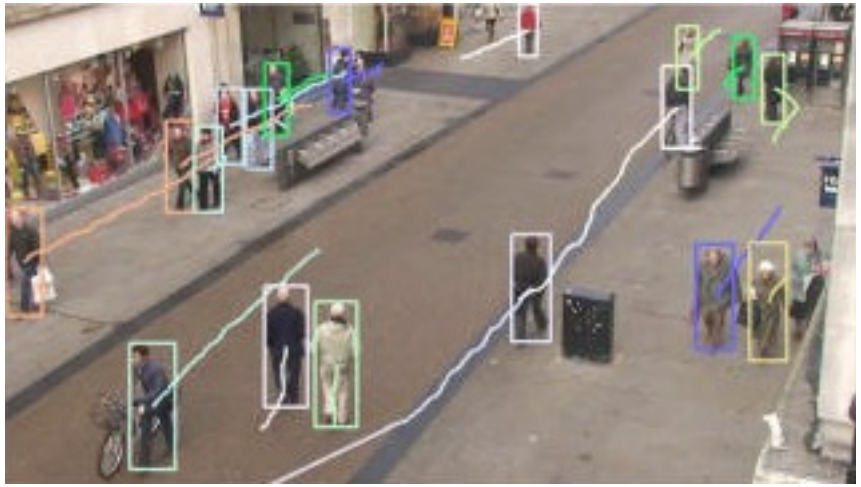
Multiple object tracking (MOT)



- Работаем со множеством объектов
- На длительных промежутках времени
- Варианты :
  - Detection Based Tracking (DBT)
  - Detection Free Tracking (DFT)

Source: http://www.robots.ox.ac.uk/~lav/Papers/benfold_reid_cvpr2011/benfold_reid_cvpr2011.html

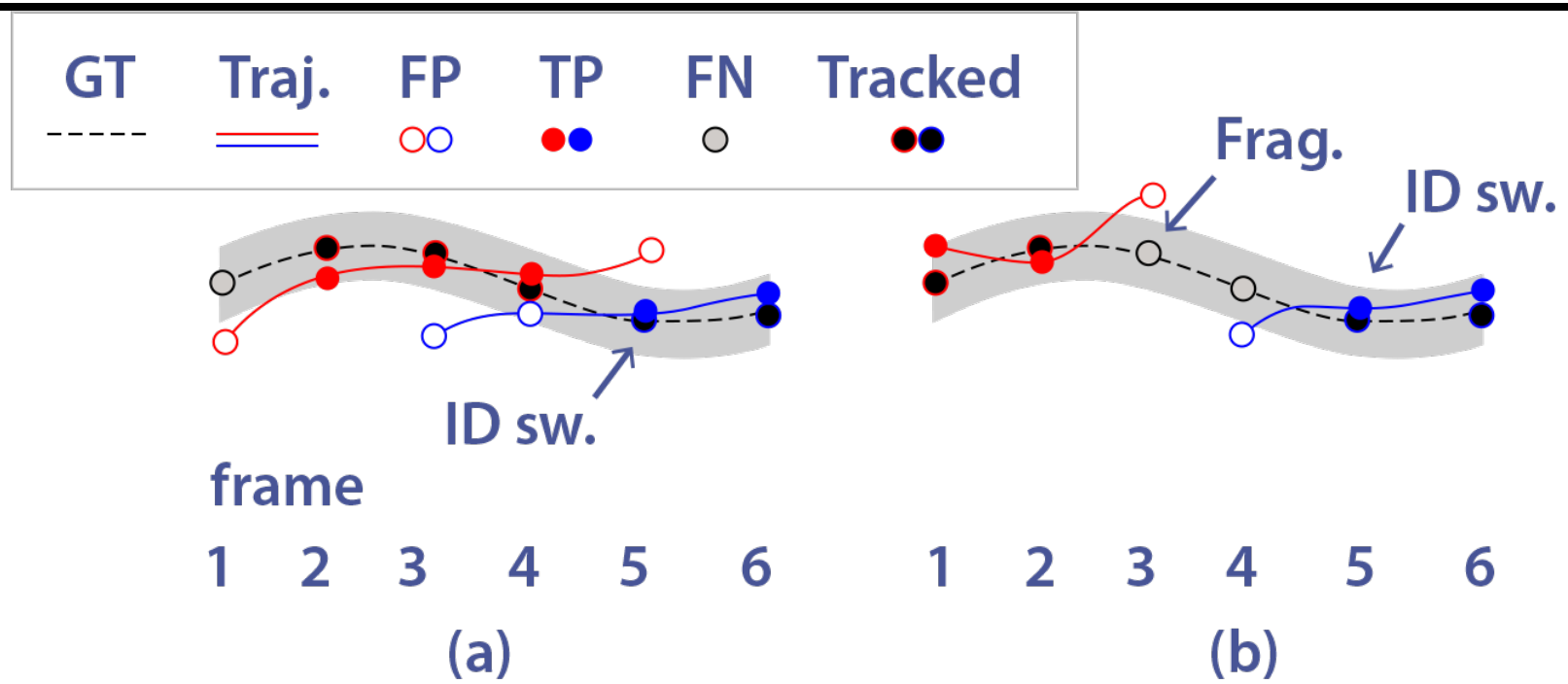
Выход MOT



- Набор траекторий объектов для всех объектов заданного типа, которые видны в видео
- Положение объекта обычно описывается с помощью bounding box

Source: http://www.robots.ox.ac.uk/~lav/Papers/benfold_reid_cvpr2011/benfold_reid_cvpr2011.html

Ошибки MOT



- ID switches – для одной эталонной траектории выдается две и более траектории
- Fragmentations – для одной эталонной траектории выдается две траектории с пропуском между ними, т.е. объект не виден на ряде кадров

Метрики качества

Multiple Object Tracking Accuracy (MOTP) –
надежность построения траектории

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t GT_t}$$

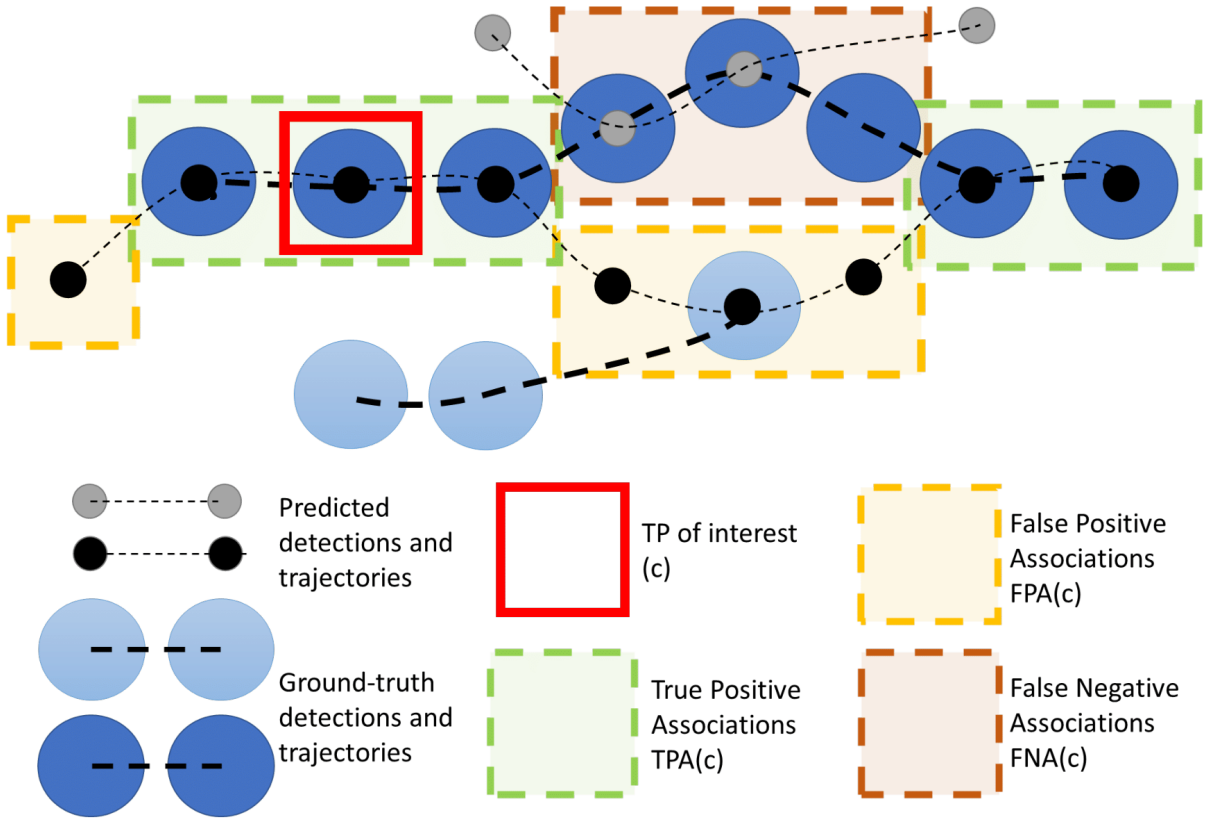
Multiple Object Tracking Precision (MOTP) –
точность локализации объектов

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t}$$

Метрики качества

- Mostly tracked (MT)
  - Объект успешно сопровождался > 80% длины эталонной траектории
- Mostly lost (ML)
  - Объект успешно сопровождался < 20% длины эталонной траектории
- Partially tracked (PT)
  - Все остальные

HOTA



$$\text{Ass-IoU} = \frac{|\text{TPA}|}{|\text{TPA}| + |\text{FNA}| + |\text{FPA}|}$$

$$\begin{aligned}\text{AssA} &= \frac{1}{|\text{TP}|} \sum_{c \in \text{TP}} \text{Ass-IoU}(c) \\ &= \frac{1}{|\text{TP}|} \sum_{c \in \text{TP}} \frac{|\text{TPA}(c)|}{|\text{TPA}(c)| + |\text{FNA}(c)| + |\text{FPA}(c)|}\end{aligned}$$

$$\begin{aligned}\text{HOTA}_\alpha &= \sqrt{\text{DetA}_\alpha \cdot \text{AssA}_\alpha} \\ &= \sqrt{\frac{\sum_{c \in \text{TP}_\alpha} \text{Ass-IoU}_\alpha(c)}{|\text{TP}_\alpha| + |\text{FN}_\alpha| + |\text{FP}_\alpha|}} \\ \text{HOTA} &= \int_{0<\alpha\leq 1} \text{HOTA}_\alpha \\ &\approx \frac{1}{19} \sum_{\substack{\alpha=0.05 \\ \alpha+=0.05}}^{0.95} \text{HOTA}_\alpha\end{aligned}$$

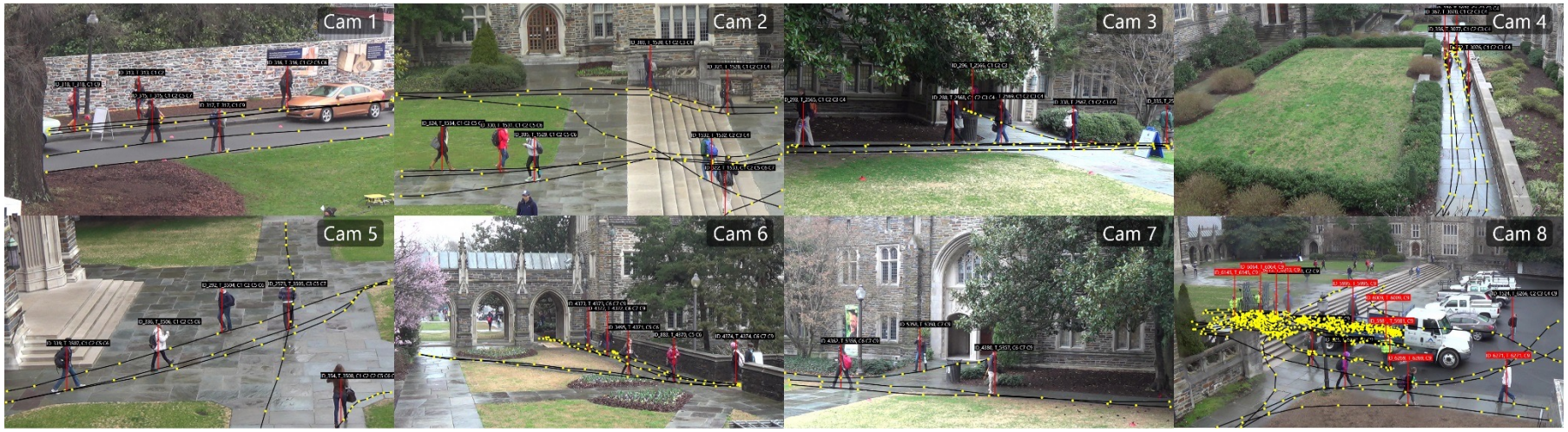
https://autonomousvision.github.io/hota-metrics/

MOT20 Challenge

- 4 train and 4 test videos with challenging crowded scenes
- 14k frames, 9 minutes
- 1.5M and 0.7M bboxes for training and testing pedestrian detector

Source: https://motchallenge.net

Duke Multi-Target, Multi-Camera



- 8 static cameras x 85 minutes of 1080p 60 fps video
- More than 2,000,000 manually annotated frames
- More than 2,000 identities
- Manual annotation by 5 people over 1 year
- More identities than all existing MTMC datasets combined
- Unconstrained paths, diverse appearance

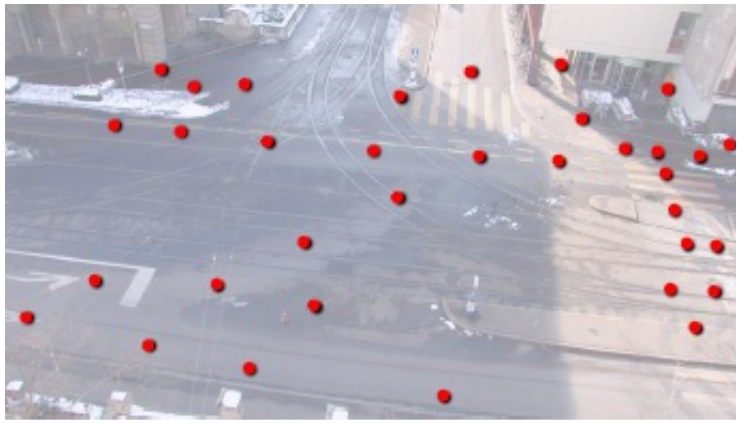
Source: http://vision.cs.duke.edu/DukeMTMC

UA-DETRAC Benchmark

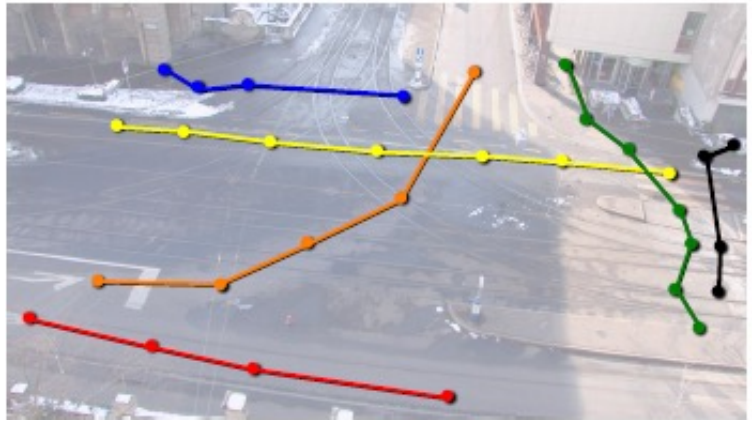
- 10 hours of videos, 25 fps, resolution 960×540 pixels
- 24 different locations at Beijing and Tianjin in China
- > 140000 frames
- 8250 vehicles
- 1.21M labeled boxes

Source: http://detrac-db.rit.albany.edu/

Detection by tracking



Детектирование объектов



Ассоциация обнаружений

Сопоставления обнаружения и траектории (треки)

Source: S.Manen Leveraging single for multi-target tracking using a novel trajectory overlap affinity measure. WACV 2016

Типичные проблемы

- Ошибки детектора – пропуски обнаружений (FN) и ложные срабатывания
- Хороший трекер должен устранять эти проблемы
  - Заполнение пропусков в детекциях
  - Фильтрация ложных срабатываний

Ограничения трекеров

- MOT Challenge findings:
  - 18% of tracks are not covered by detections at all
  - 37% of tracks are covered by low-confidence detections
  - Trackers reduce FP and raise FN
- Detector is a key!

Важность хорошего детектора

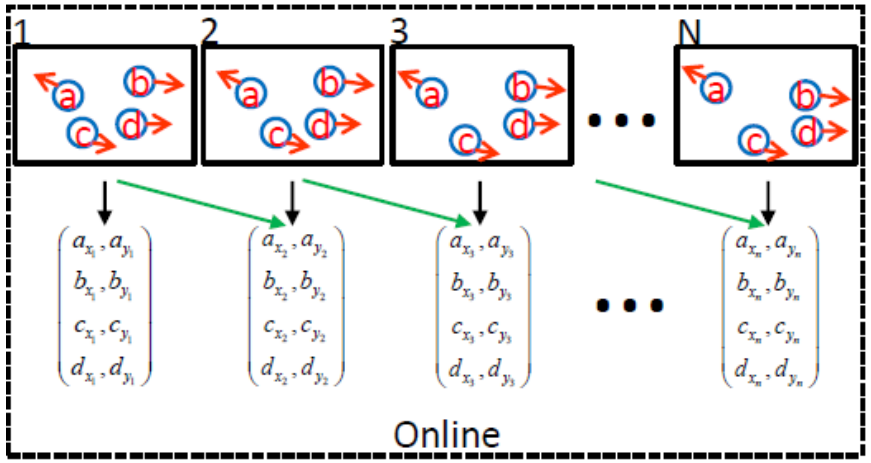
Tracker	Avg.Rank	↑MOTA		MOTP
HCC 1. ✓	9.6	**49.3**	±10.2	79.0
LMP 2. ✓	12.8	48.8	±9.8	**79.0**
FWT 3. ✓	21.7	47.8	±9.4	75.5
				R. Henschel, L. I
DeepSORT_2 13. ○	10.3	61.4	±10.6	79.1
TMO 14. ○	10.6	61.4	±10.1	79.3
SORTwHPD16 15. ○	10.2	59.8	±10.3	79.6

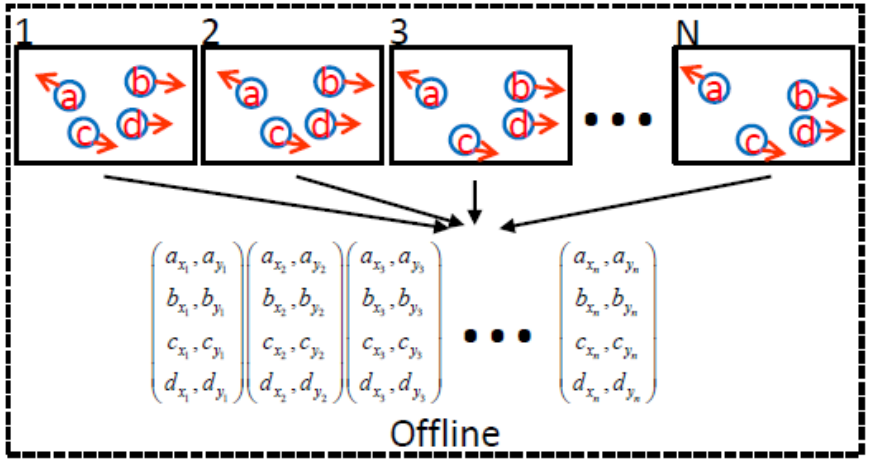
With DPM (Deformable part models) detector

With Faster R-CNN detector

Source: https://motchallenge.net

Online vs offline tracking





On-line tracking

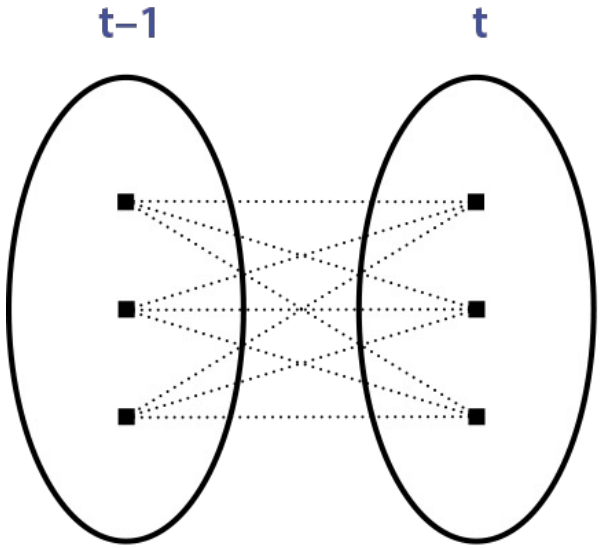
(Only current and previous frames are available)

Off-line (batch) tracking

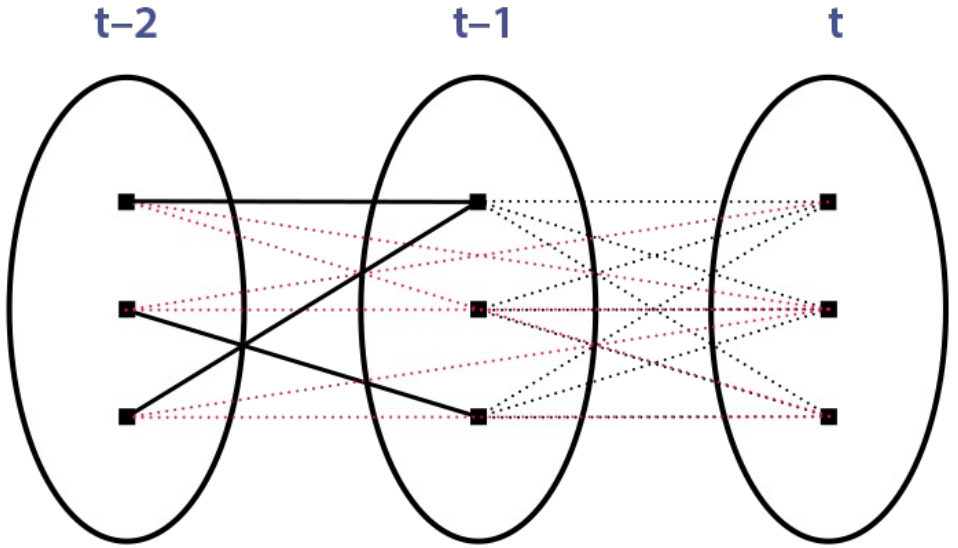
(All frames (including future) are available)

Source: https://arxiv.org/abs/1409.7618

Data association

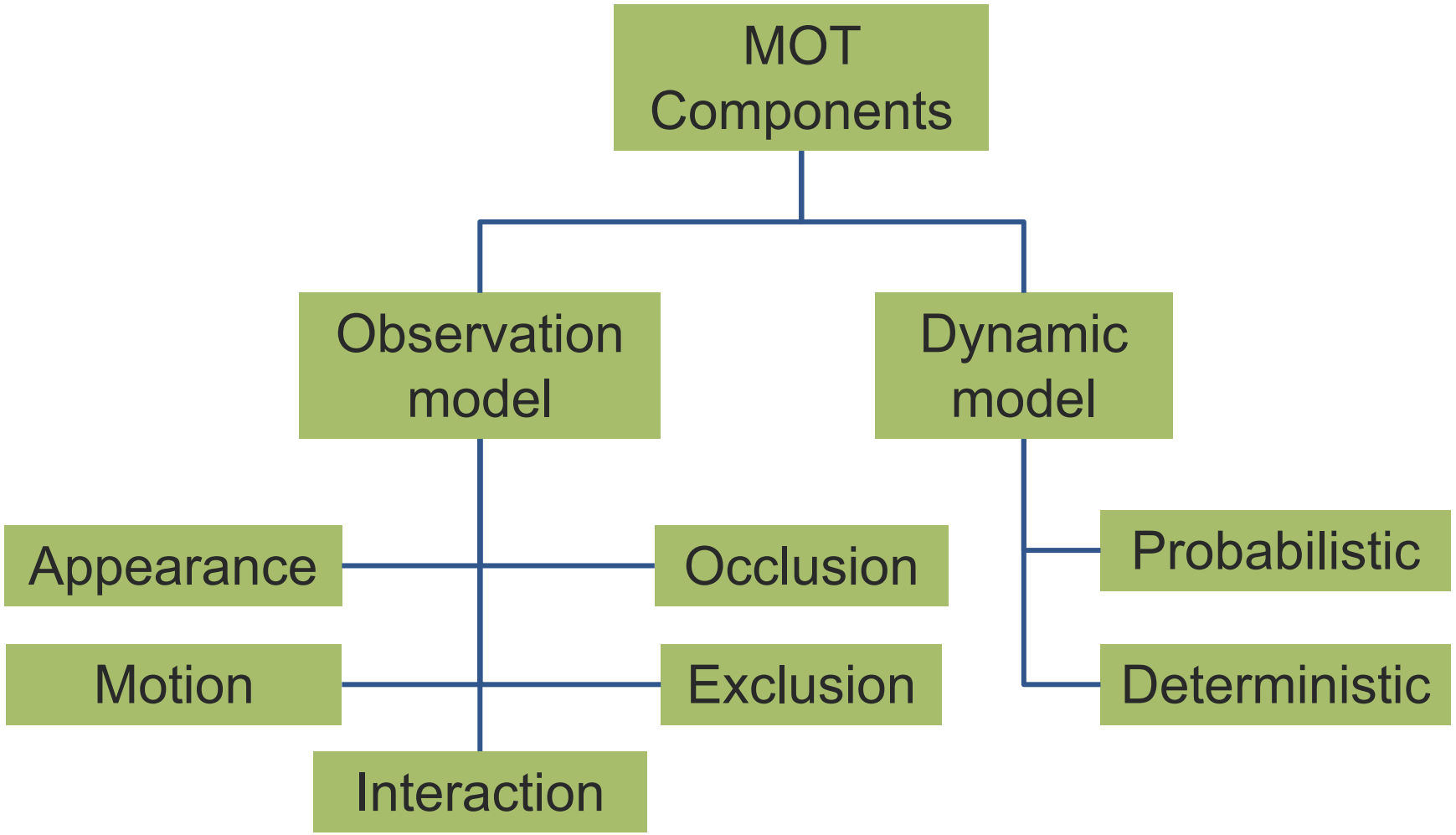


Two-frame methods

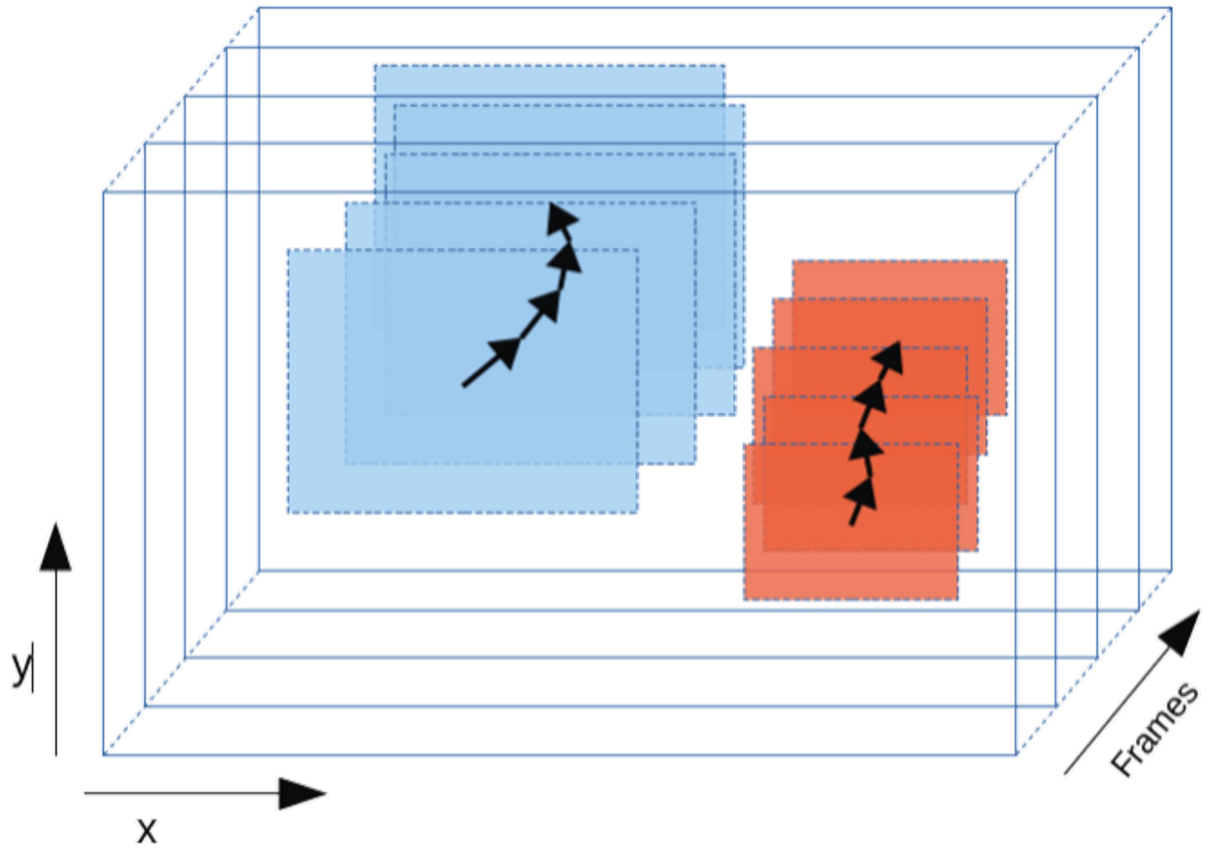


Multi-frame methods

Компоненты функции сходства (affinity)



Ассоциация прямо по IoU

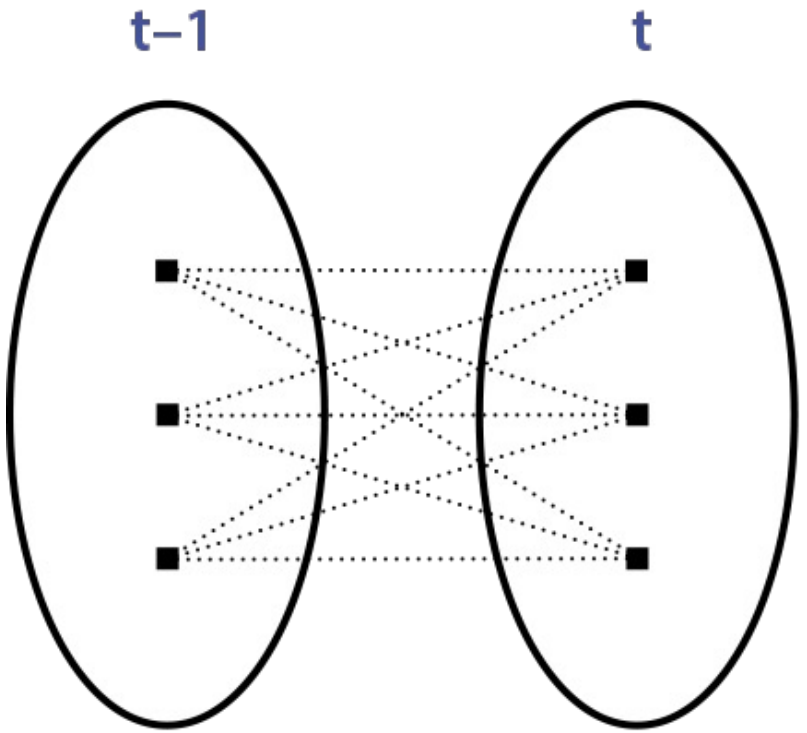


Matching object in neighouring frames by bounding box overlap

Bochinski et. al. High-Speed Tracking-by-Detection Without Using Image Information. AVSS 2017

Simple Online and Realtime Tracking (SORT)

- CNN-based object detector
- Kalman filter for prediction of object position in current frame based on positions in previous frames
- Hungarian algorithm for matching object detections in current frames with predicted positions
- IoU of detected and predicted bounding boxes as affinity measure for matching detection and track



A. Bewley, Z. Ge, L. Ott, F. Ramos, B. Upcroft. Simple Online and Realtime Tracking. In arXiv:1602.00763, 2016.

Influence of motion prediction

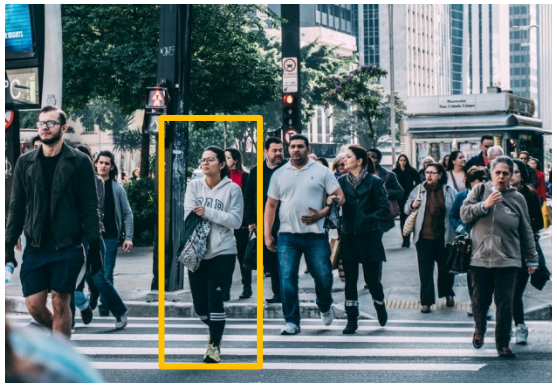
	MOTA	MOTP	MT	ML	FP	FN	ID sw	Frag	Hz
SORT	**59.8**	**79.6**	**25.4%**	**22.7%**	8698	**63245**	**1423**	**1835**	59.5
IOU	57.1	77.1	23.6%	32.9%	**5702**	70278	2167	3028	**3004**

Comparison of SORT to simple IoU tracker:

- Less False Negatives (FN), Fragmentations (Frag), ID Switches (ID Sw), Mostly Tracked (MT)
- Increase in MOTA and MOTP
- Increase in False Positive (bad)

Source: https://motchallenge.net

Re-identification



Detections in video

Probe

Gallery

Matches

SORT + DA

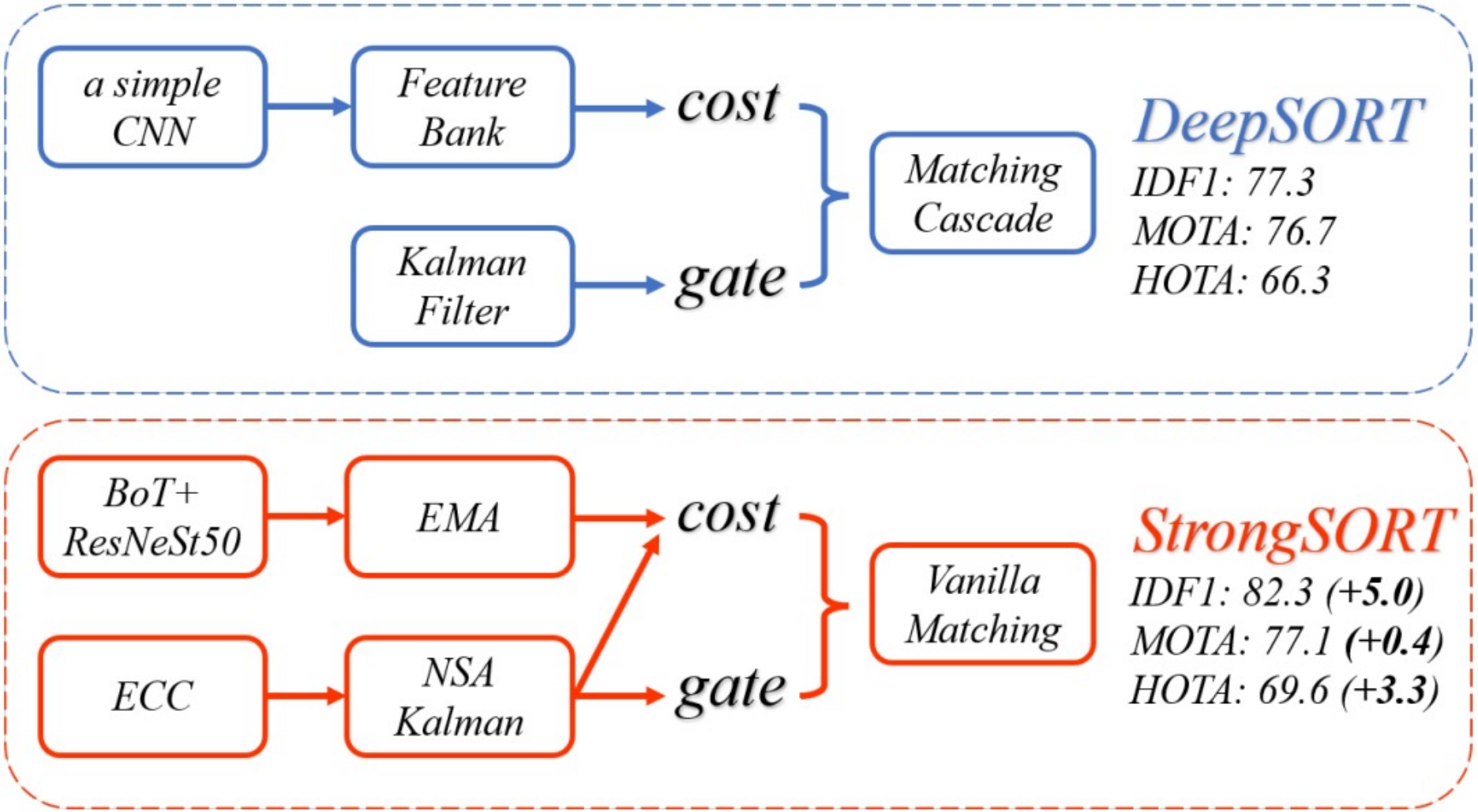
	MOTA	MOTP	MT	ML	FP	FN	ID sw	Frag	Hz
SORT	59.8	79.6	25.4%	22.7%	**8698**	63245	1423	**1835**	**59.5**
Deep SORT	**61.4**	79.1	**32.8%**	**18.2%**	12852	**56668**	**781**	2008	40

Addition of re-identification to affinity between detections and tracks:

- Reduces ID switches (ID sw), False negatives (FN), Mostly Lost (ML) and increases Mostly Tracked (MT)
- Somewhat raises False Positives (FP) and Fragmentations (Frag)

Source: https://arxiv.org/pdf/1703.07402.pdf

StrongSORT



Для начала улучшим базовые компоненты, включая детектор

Y.Du et. Al. StrongSORT: Make DeepSORT Great Again. IEEE TMM 2023

Дополнительные plug-in модули

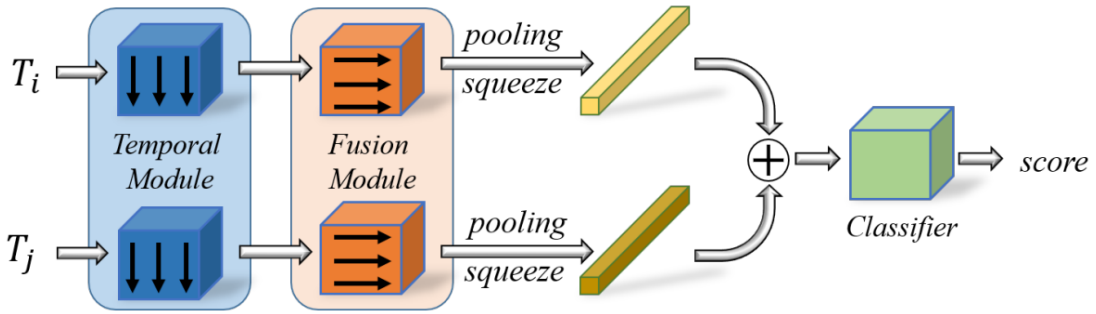


Fig. 3: Framework of the two-branch AFLink model. It adopts two tracklets T_i and T_j as input, where $T_* = \{f_k^*, x_k^*, y_k^*\}_{k=k^*}^{k^*+N-1}$ consists of the frame id f_k^* and positions (x_k^*, y_k^*) of the recent $N = 30$ frames. Then, the temporal module extracts features along the temporal dimension with 7×1 convolutions and the fusion module integrates information along the feature dimension with 1×3 convolutions. These two tracklet features are pooled, squeezed and concatenated, and then input into a classifier to predict the association score.

Ассоциация удалённых треклетов

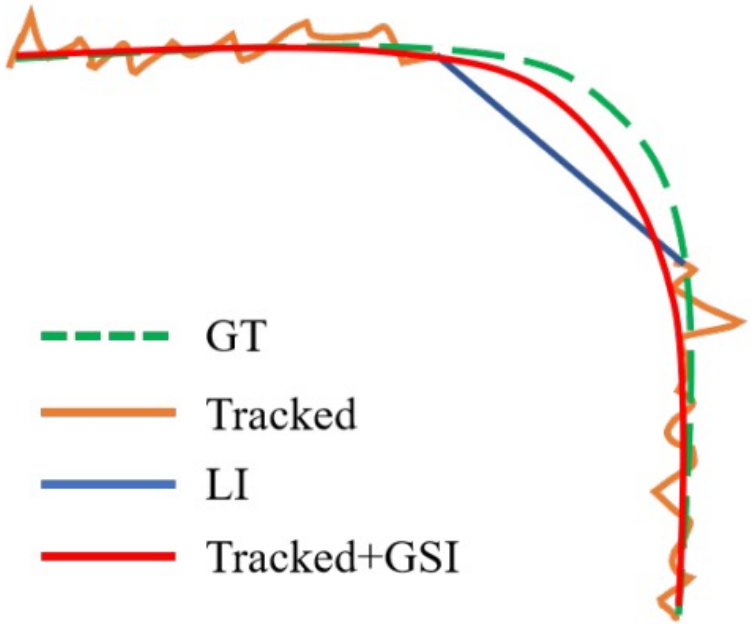


Fig. 4: Illustration of the difference between linear interpolation (LI) and the proposed Gaussian-smoothed interpolation (GSI).

«Грамотная» интерполяция траекторий между треклетами через сплайны

Y.Du et. Al. StrongSORT: Make DeepSORT Great Again. IEEE TMM 2023

Резюме

- 2 варианта видео – видеопоследовательность и видеопоток
- Движение – ключевой новый признак для распознавания по видео
- Мы рассмотрели 4 задачи анализа видео:
  - Оценку оптического потока
  - Распознавание действий и событий
  - Визуальное сопровождение объектов
  - Сопровождение множества объектов
- Все задачи сейчас успешно решаются нейросетевыми моделями
- Контекст (признаки изображения) и движение учитывают отдельно