Анализ изображений лиц с помощью сверточных нейронных сетей

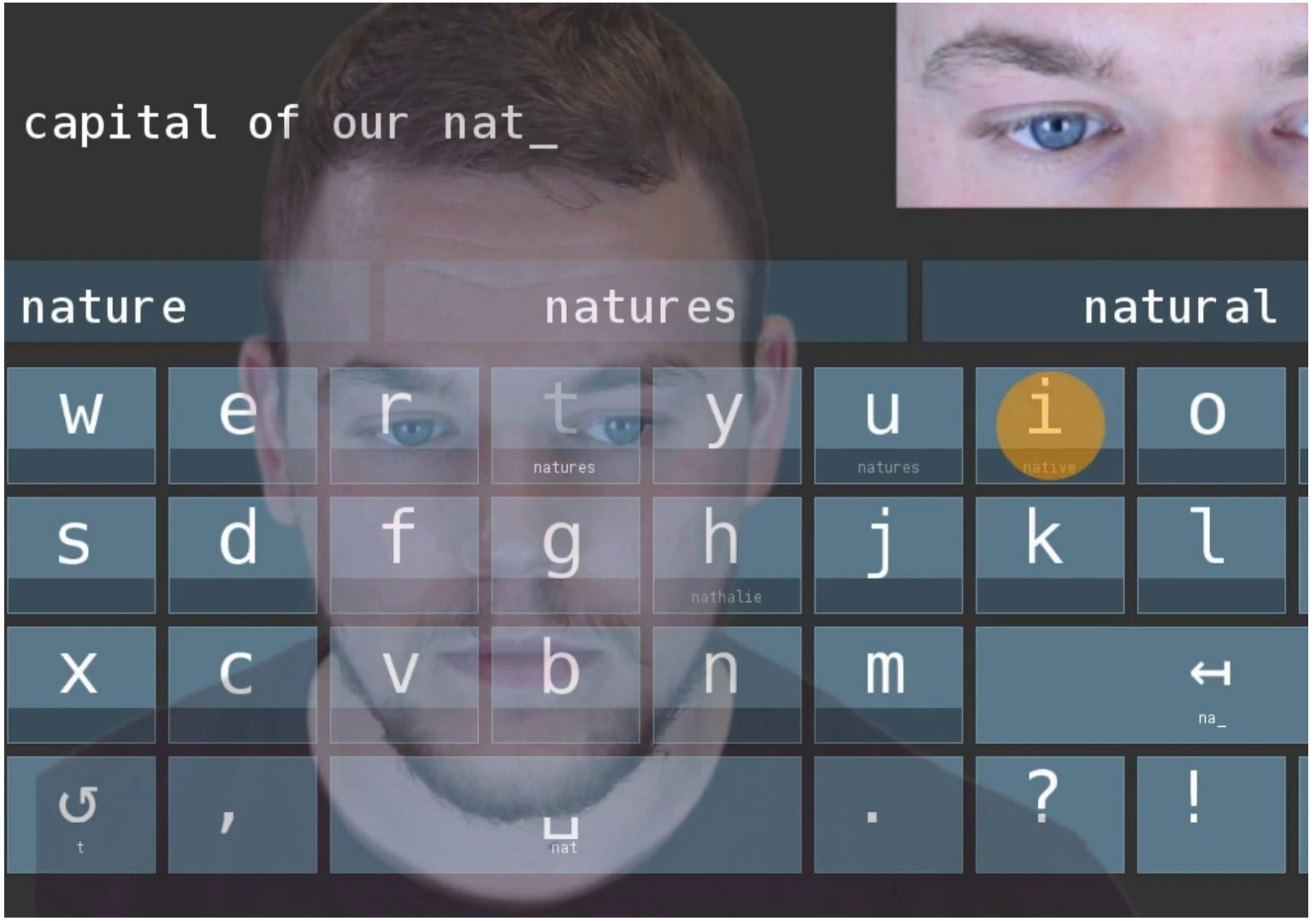
Harman
Дмитрий Яшунин, Роман Власов

Agenda

- Introduction
- Face and landmark detection
- Eye gaze direction estimation

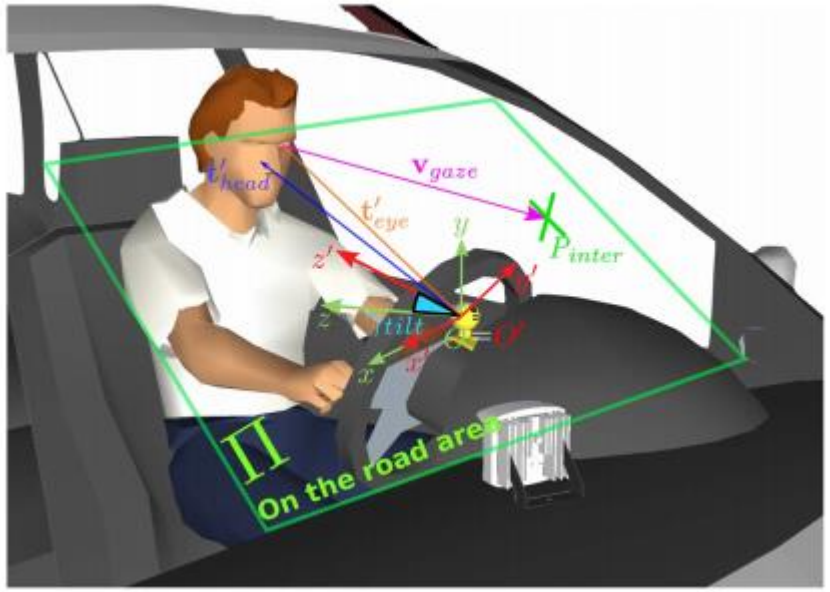
Applications of gaze estimation

- Support systems
  - Gaming and device controlling by eyes



Applications of gaze estimation

- Analysis systems
  - Automotive: driver attention analyzing through gaze estimation



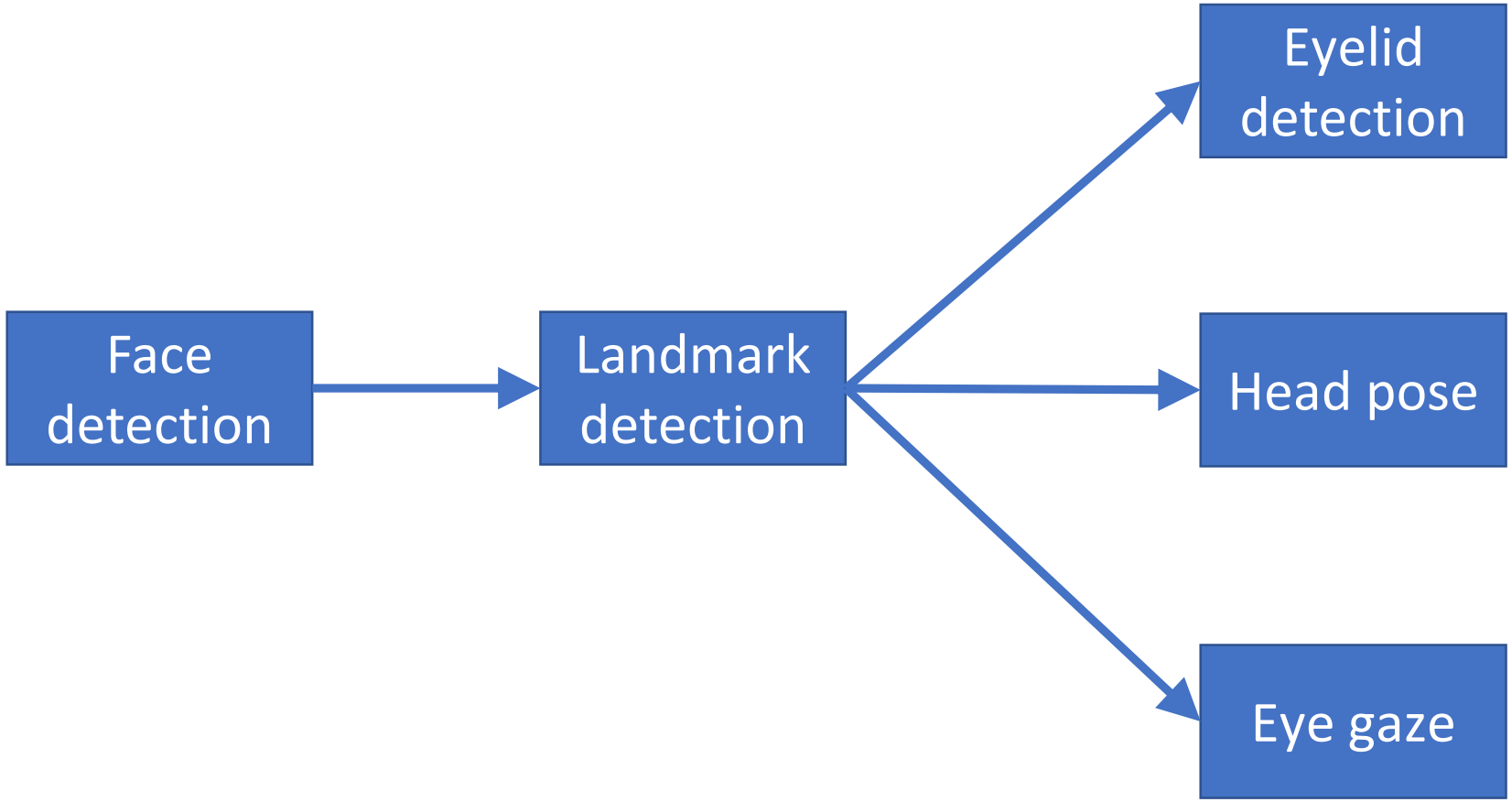
Applications of gaze estimation

- Analysis systems
  - VR/AR: detection of user attention, using comfortable glasses instead of smartphone

Applications of gaze estimation

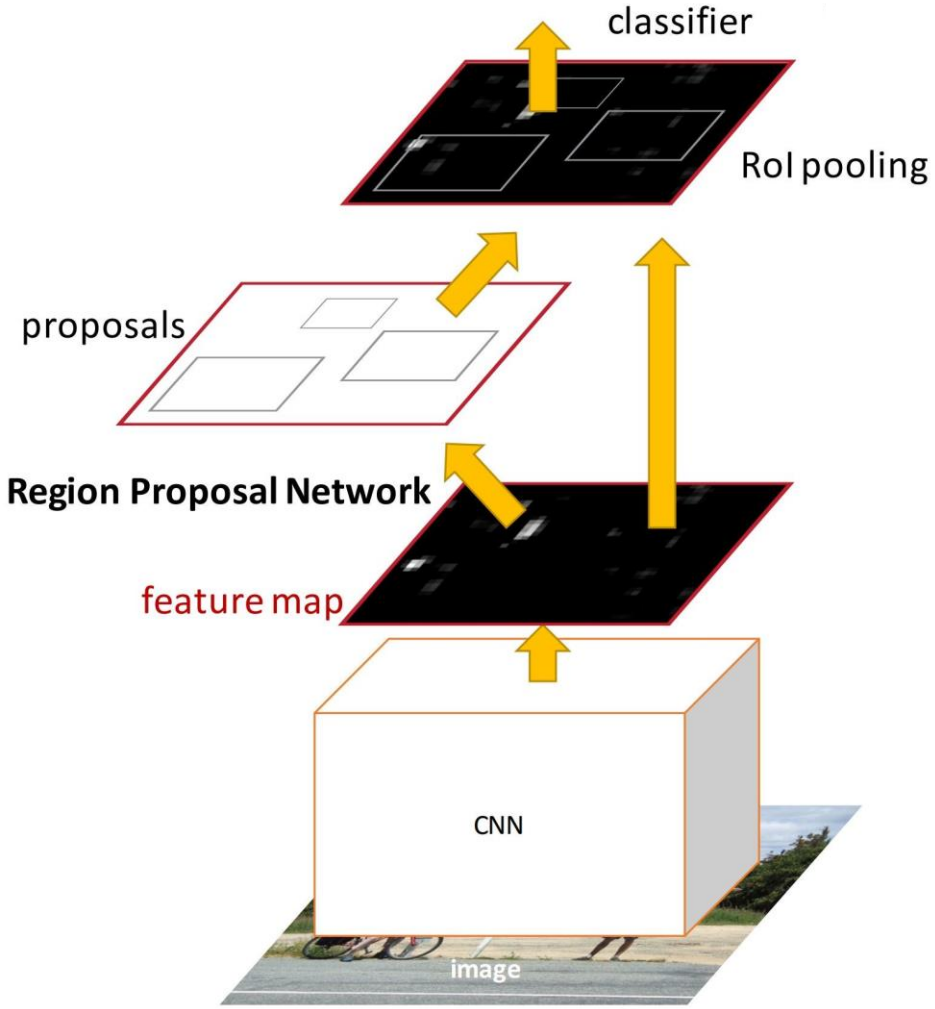
- Analysis systems
  - Marketing, development

Common pipeline for face analytics



Face and landmark detection

Single task face detectors: **Faster RCNN**



Two stage detector:

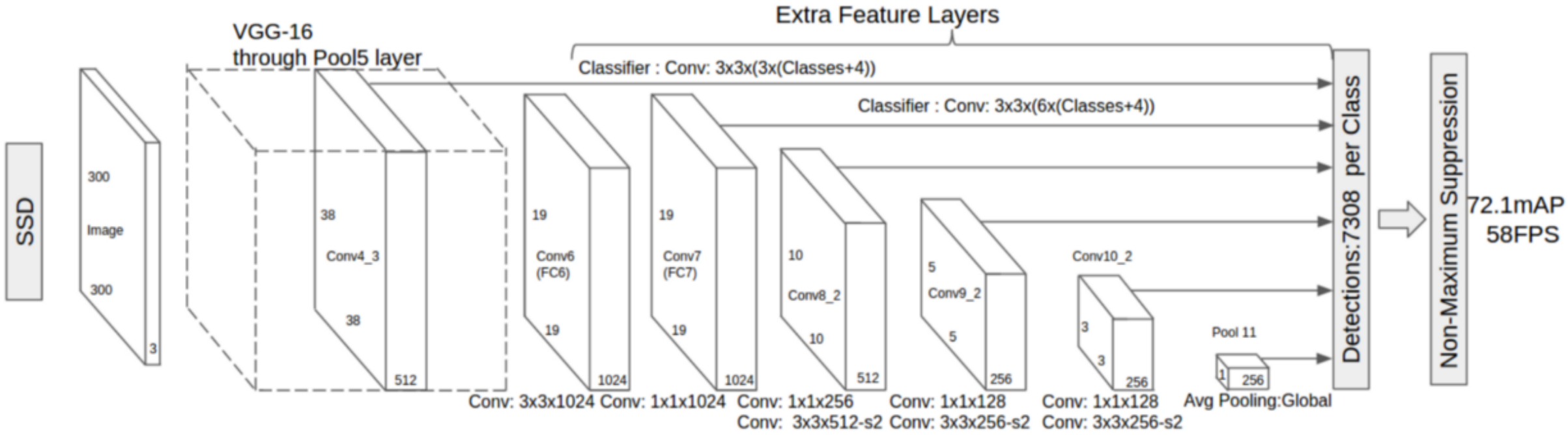
1. Region Proposal Network (RPN) generates proposals
2. Detection head classifiers and refines proposals

Shaoqing R., et al, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, 2015

Single task face detectors: **Single Shot MultiBox Detector (SSD)**

For each scale in feature map output detections (based on default boxes)

Shallow layers detect small object, deep layers – big objects

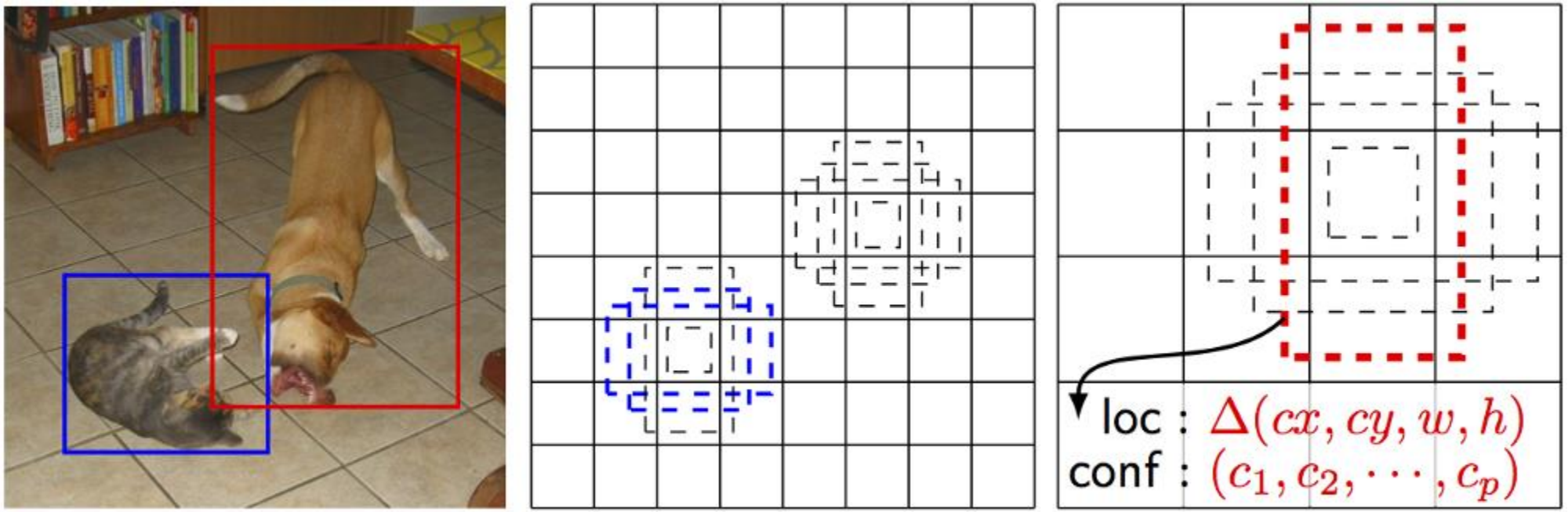


Wei L., et al, SSD: Single Shot MultiBox Detector, 2015

Single task face detectors: **Single Shot MultiBox Detector (SSD)**

Anchor bounding boxes

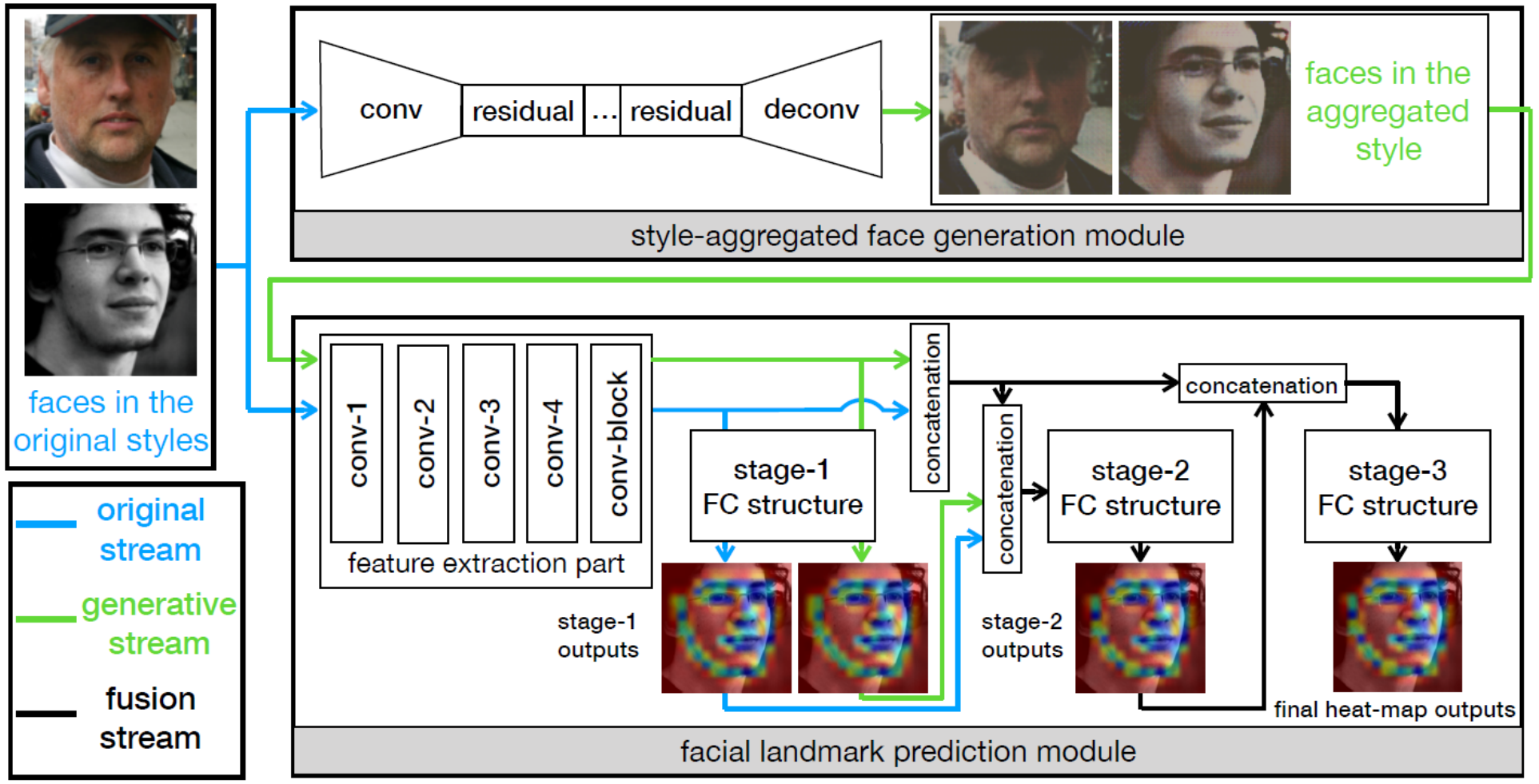
Outputs offsets to default bounding boxes and class probabilities



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

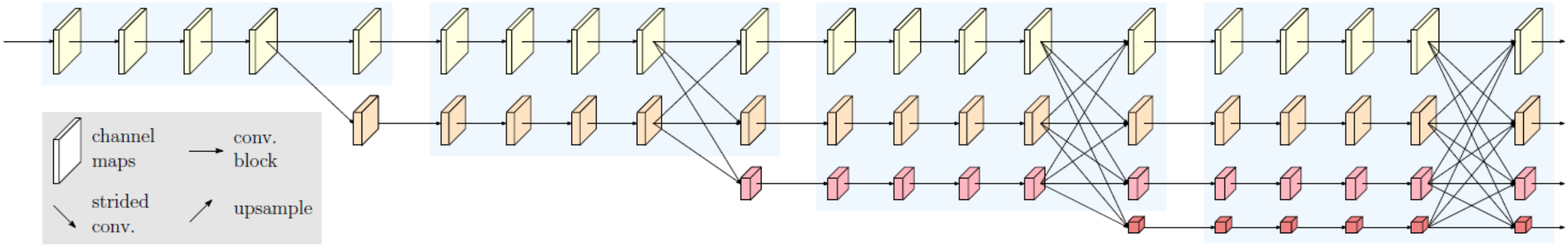
Wei L., et al, SSD: Single Shot MultiBox Detector, 2015

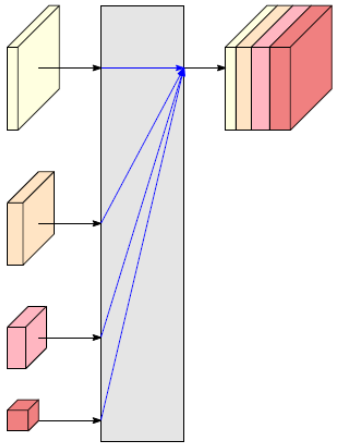
Single task landmark detectors: **Style Aggregated Network (SAN)**



Xuanyi D., et al, Style Aggregated Network for Facial Landmark Detection, 2018

Single task landmark detectors: **High-Resolution Net (HRNet)**

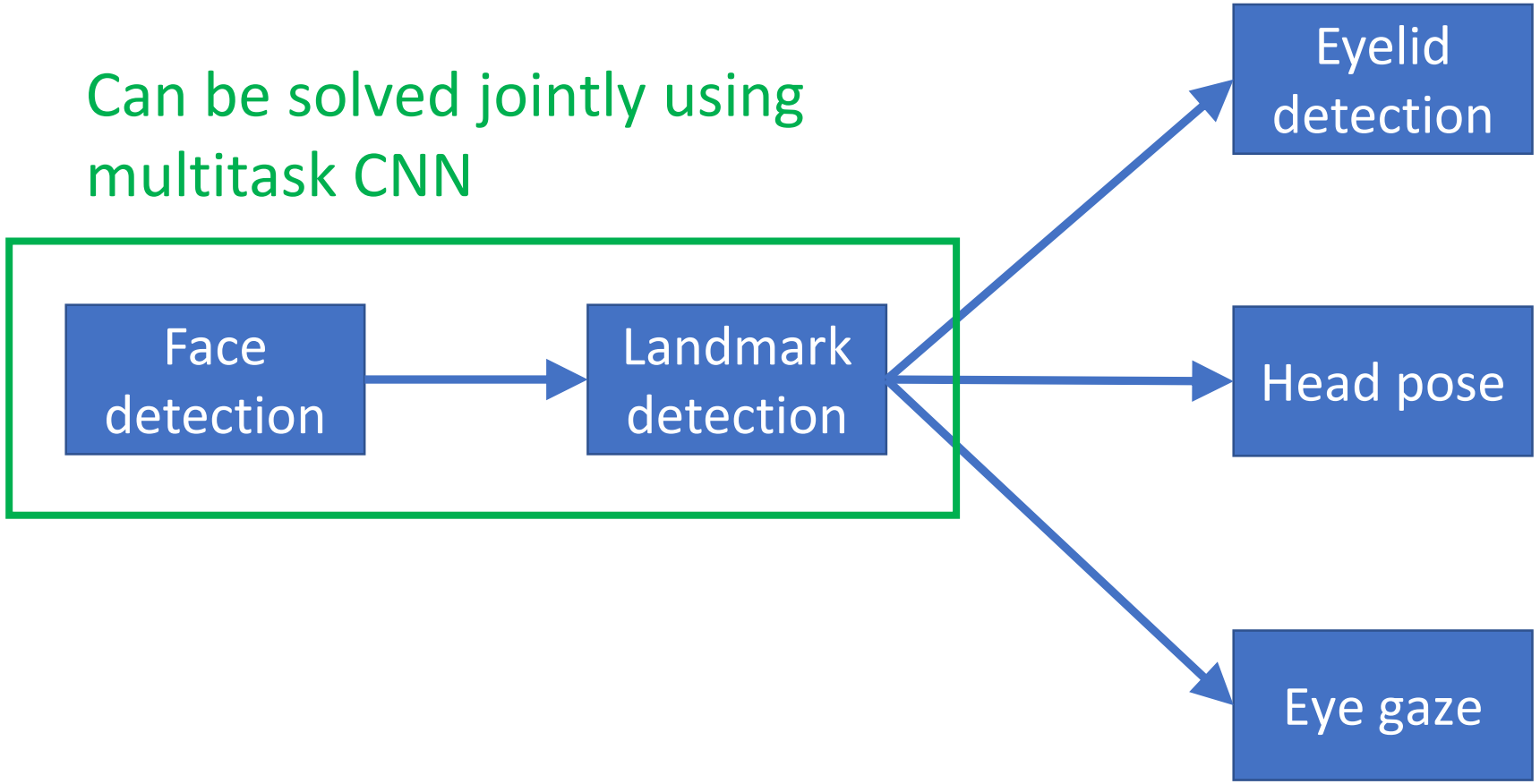




-> Landmark detection

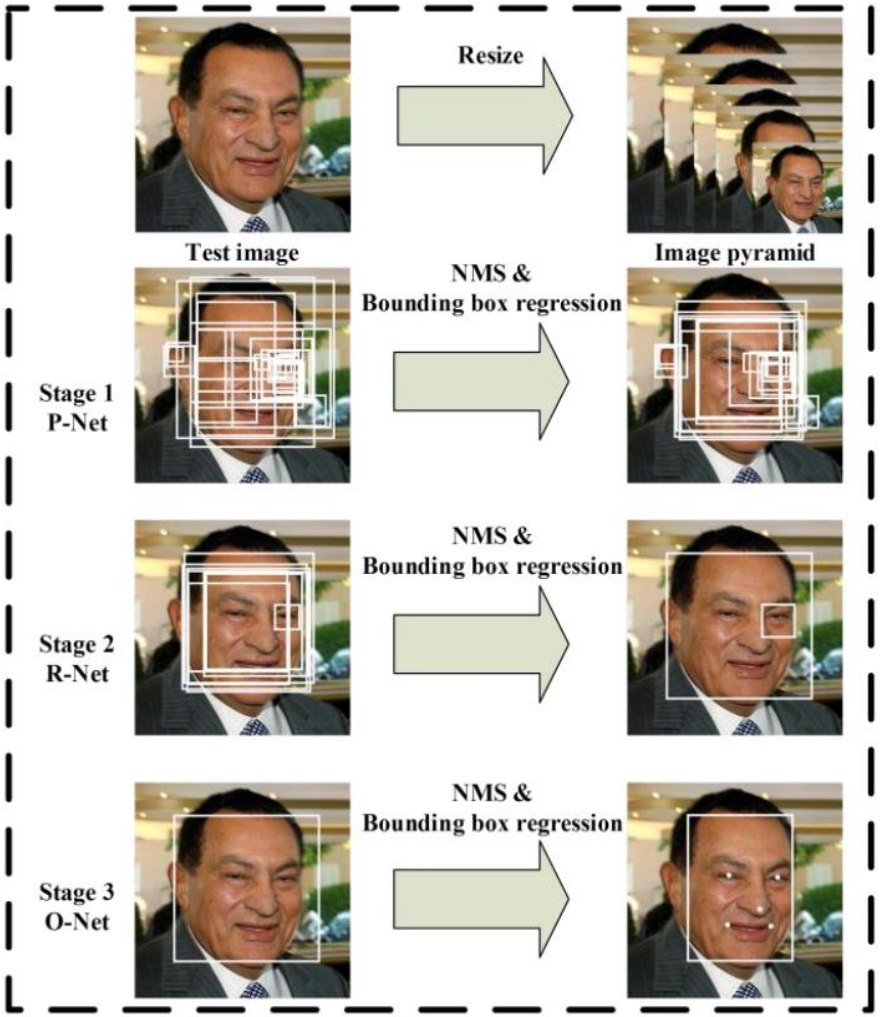
Ke S., et al, High-Resolution Representations for Labeling Pixels and Regions, 2019

Common pipeline for face analytics



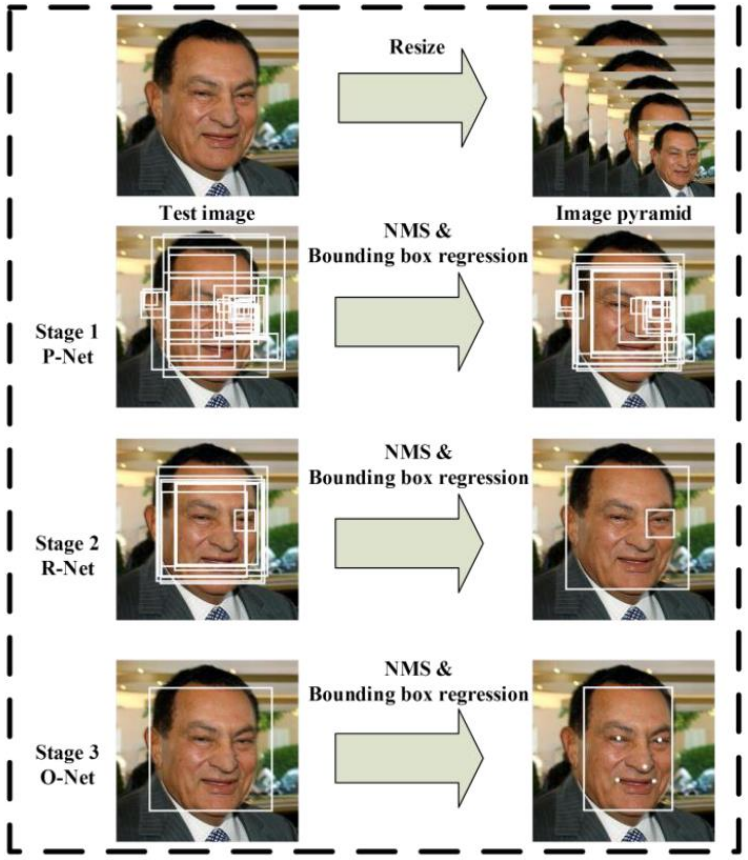
Multitask face detectors (face + landmarks): Multi-task Cascaded Convolutional Networks (**MTCNN**)

Cascade from 3 lightweight CNNs

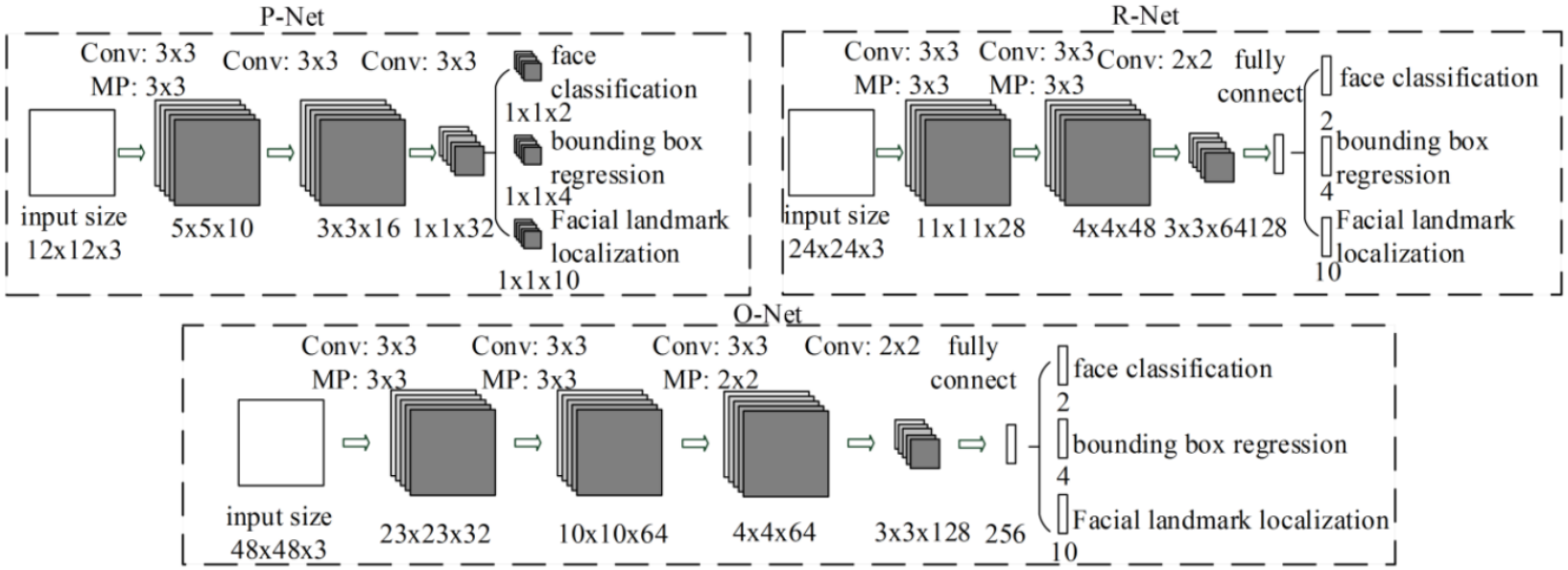


Kaipeng Z., et al, Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks, 2016

Multitask face detectors (face + landmarks): Multi-task Cascaded Convolutional Networks (**MTCNN**)

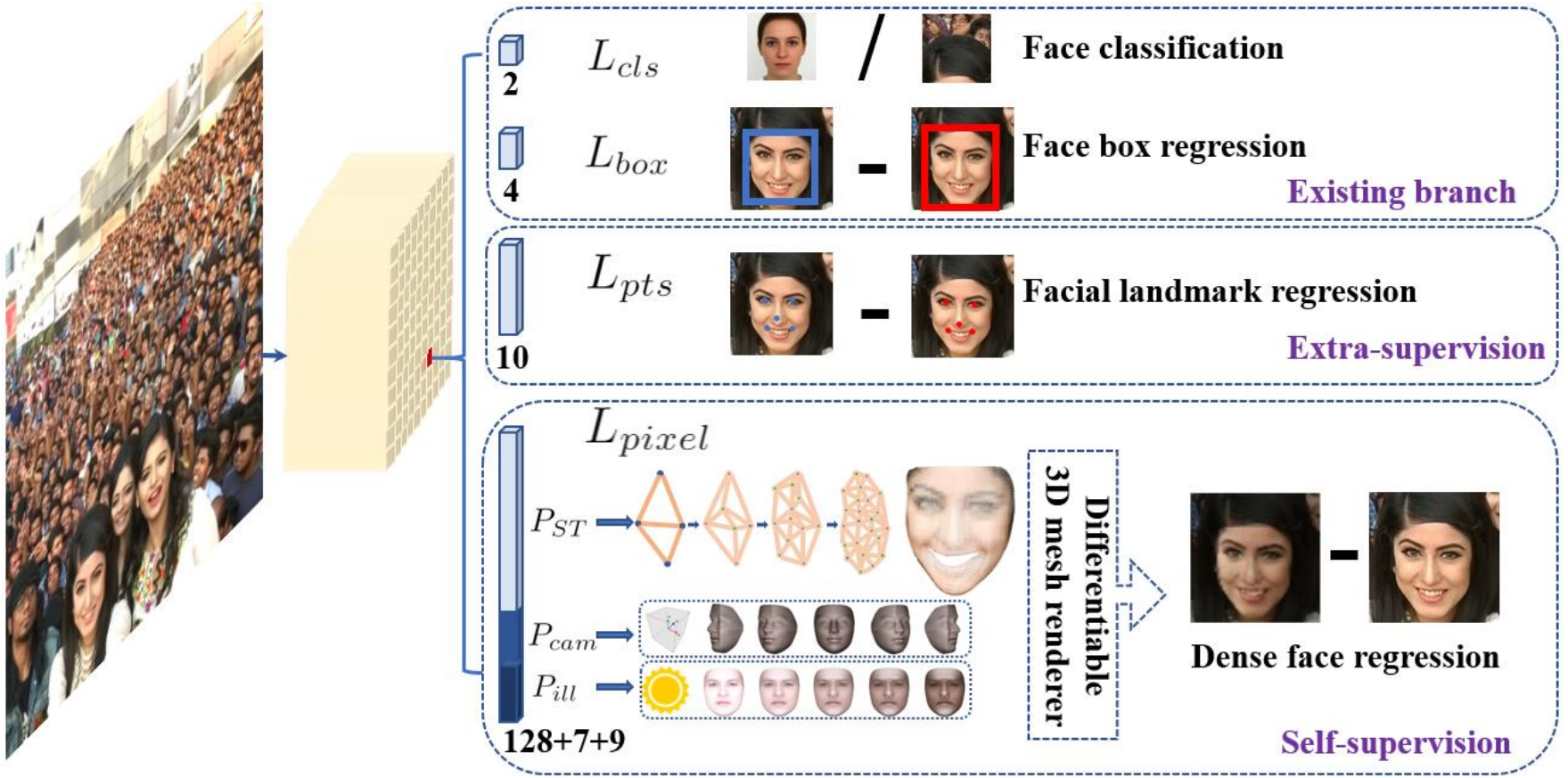


Architectures of CNNs



Kaipeng Z., et al, Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks, 2016

Multitask face detectors (face + landmarks): **RetinaFace**: Single-stage Dense Face Localisation in in the Wild

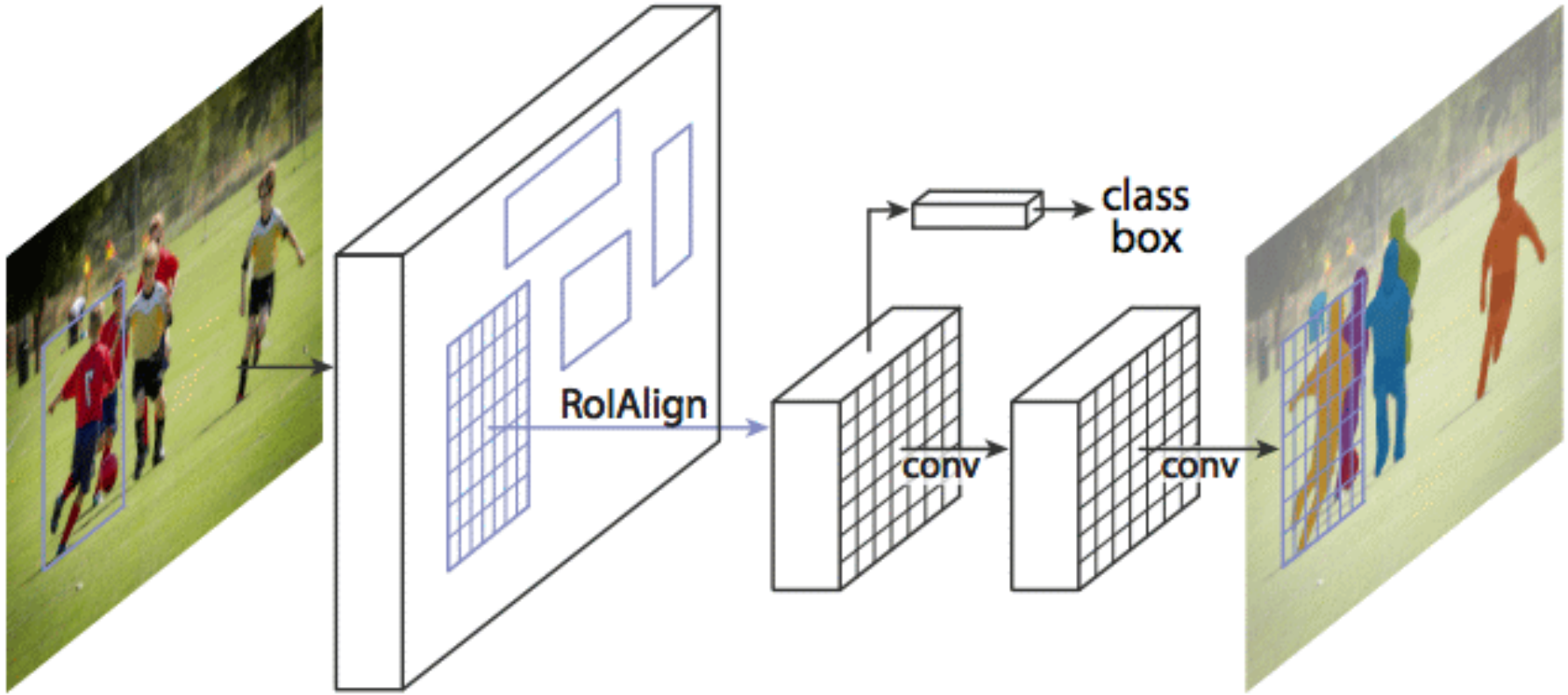


Jiankang D., et al, RetinaFace: Single-stage Dense Face Localisation in the Wild, 2019

Multitask face detectors (face + landmarks): **VisionLabs face detector**

- MobileNetv2 backbone
- Detection head – SSD
- Keypoint head – direct regression

Multitask face detectors (face + landmarks): **Mask R-CNN**

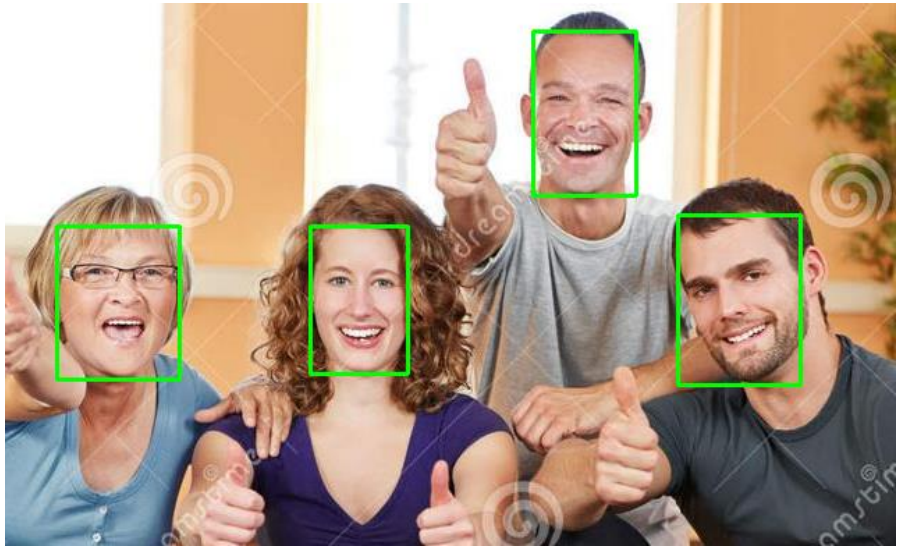


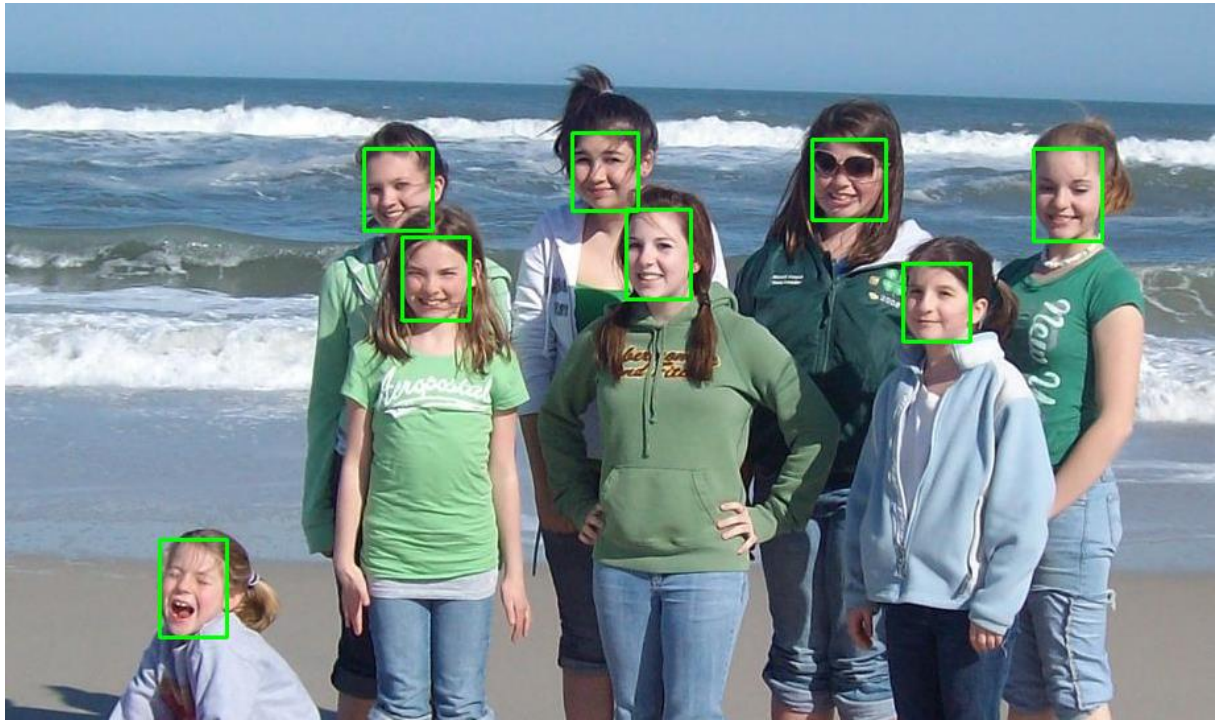
Kaiming H., et al, Mask R-CNN, 2017

Datasets for face and landmark detection

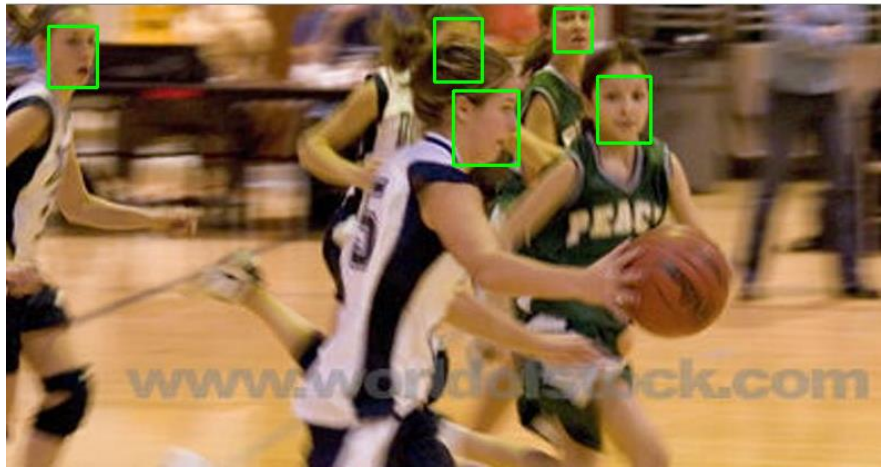
- WIDER Face: Used for training and validation
- AFLW: Used for validation

WIDER FACE examples: easy

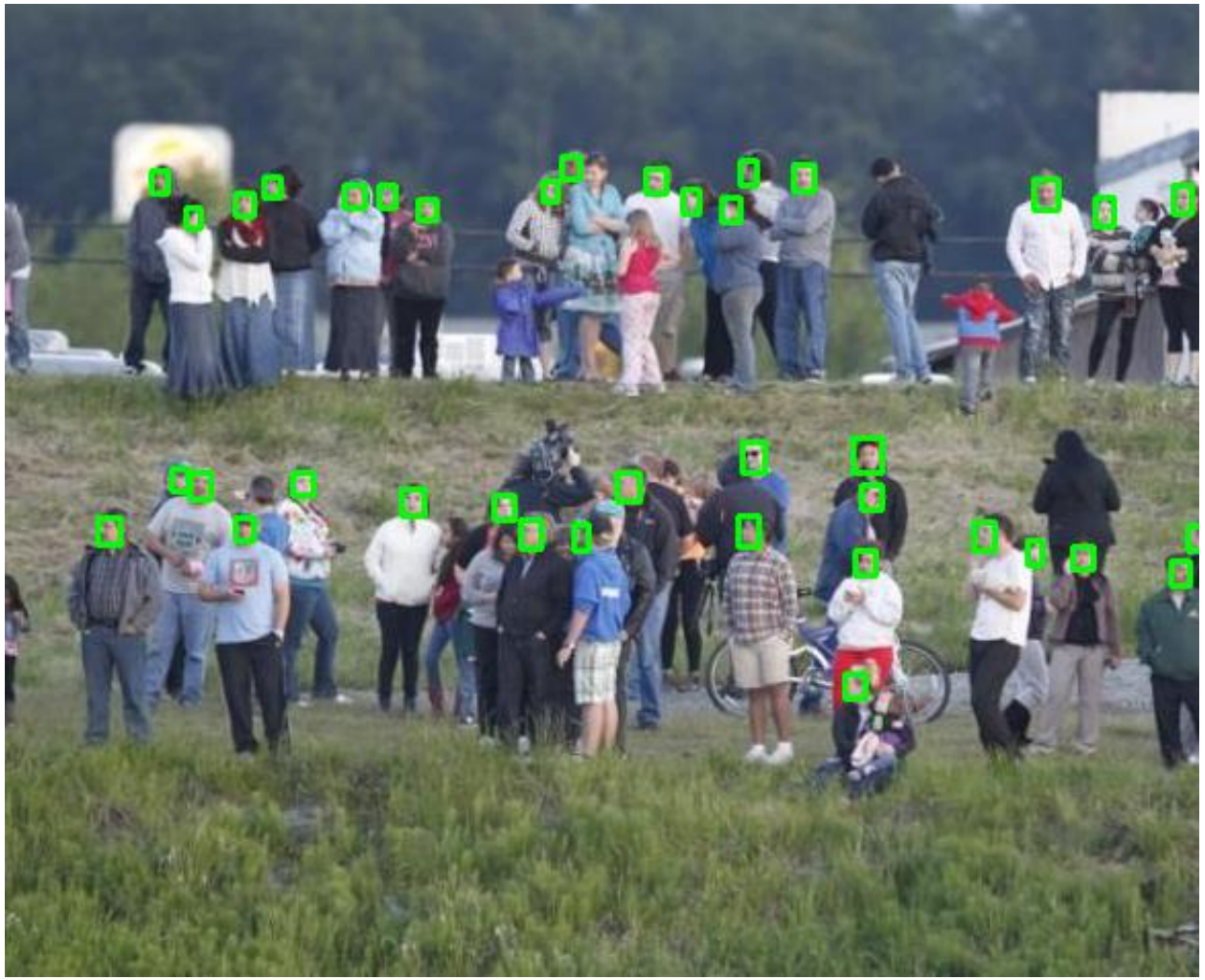




WIDER FACE examples: medium

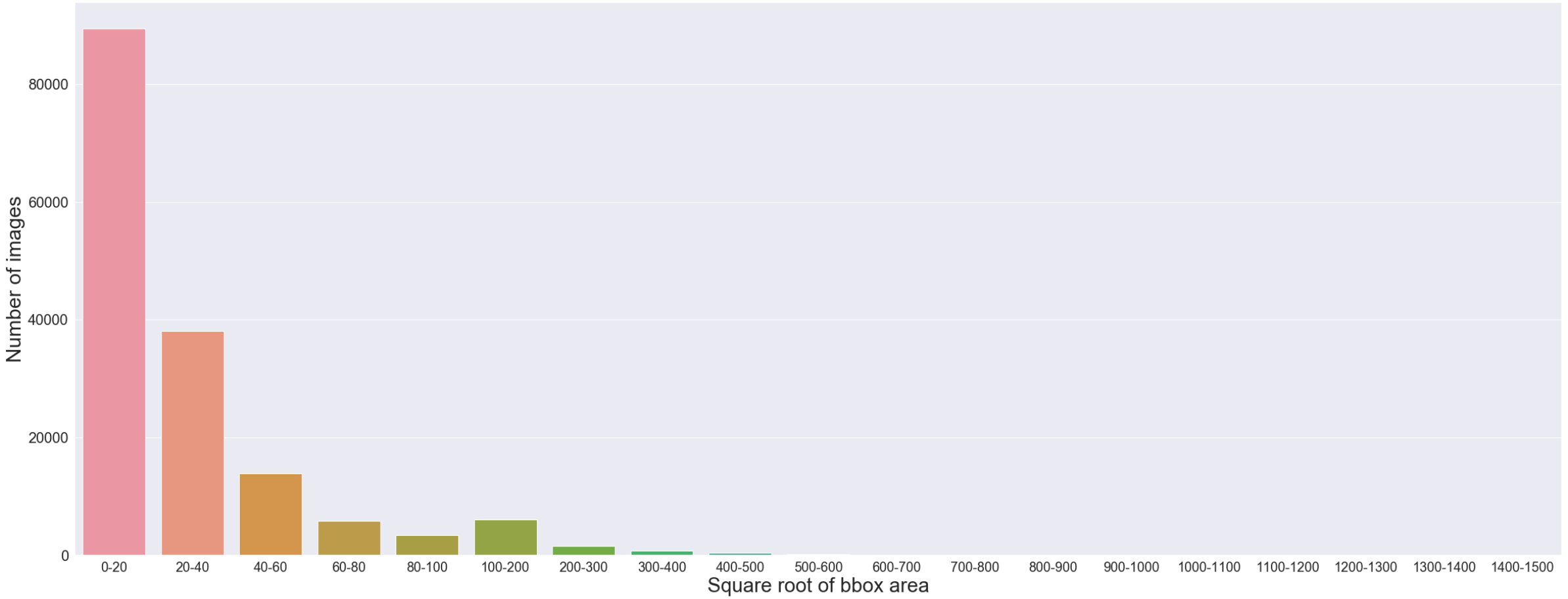


WIDER FACE examples: hard

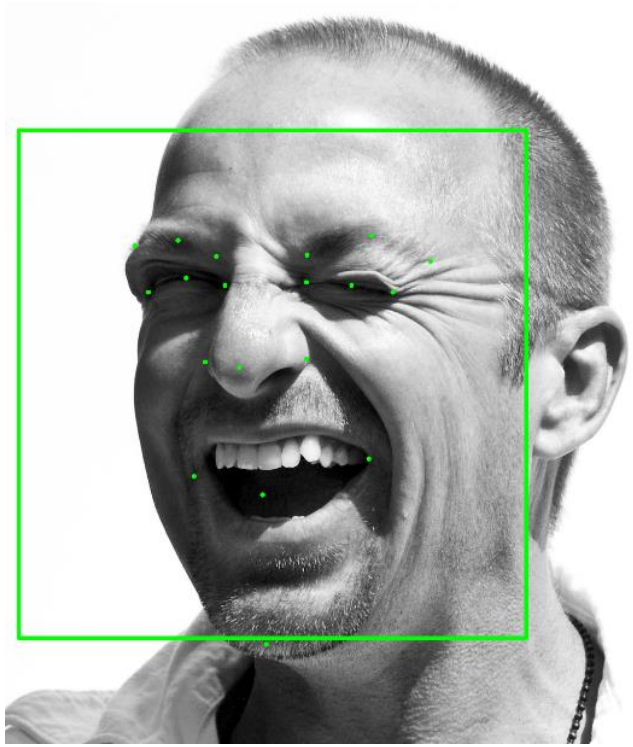


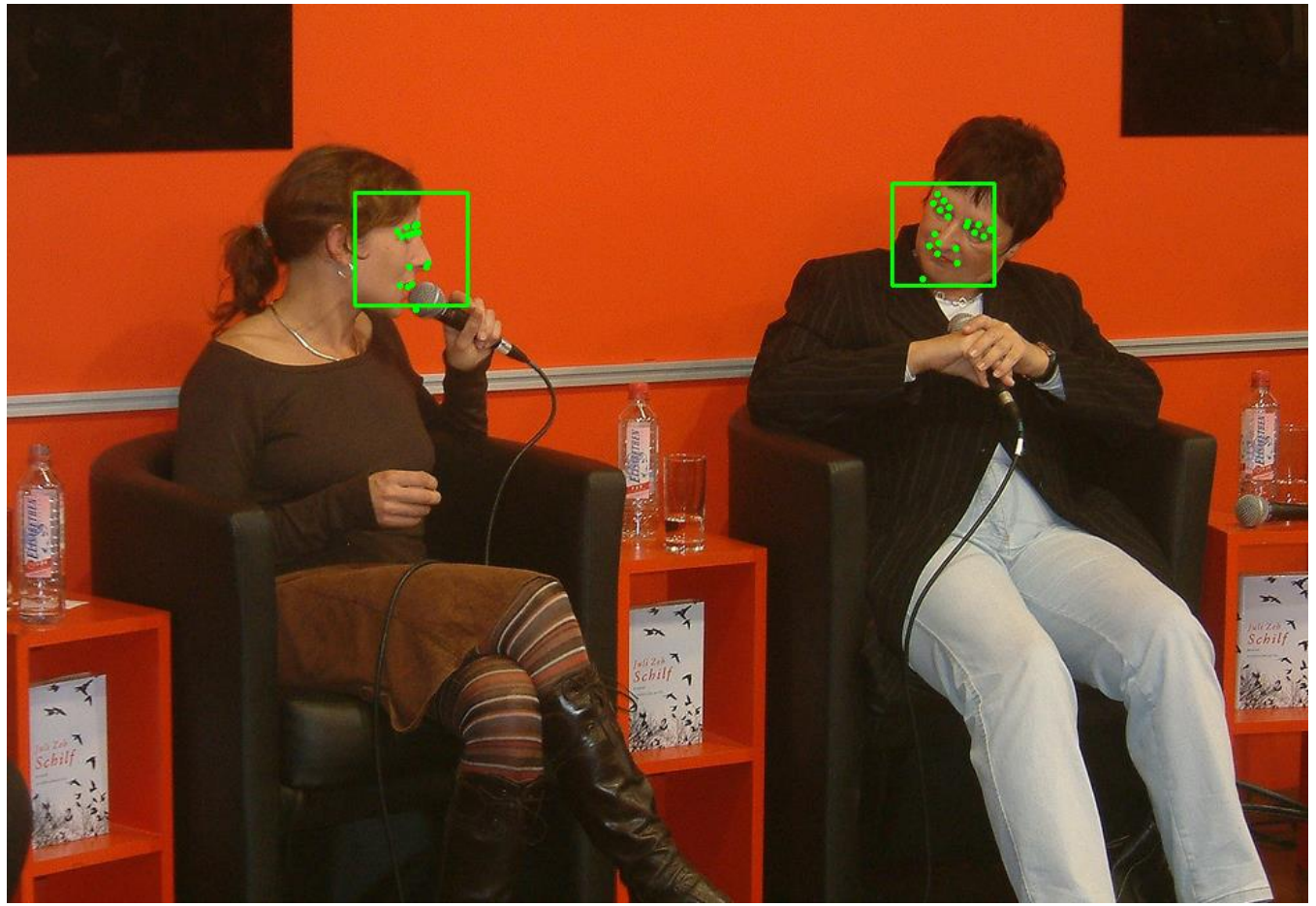
WIDER FACE examples

Datasets: WIDER FACE

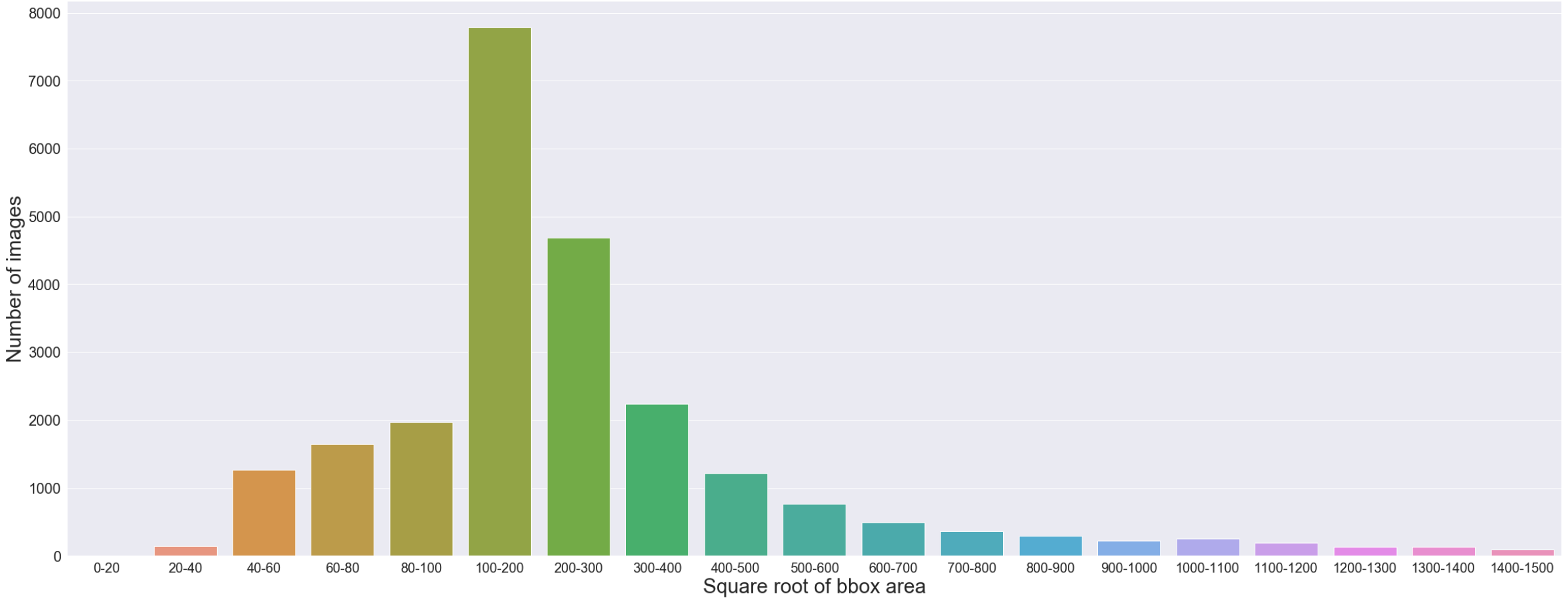


AFLW examples

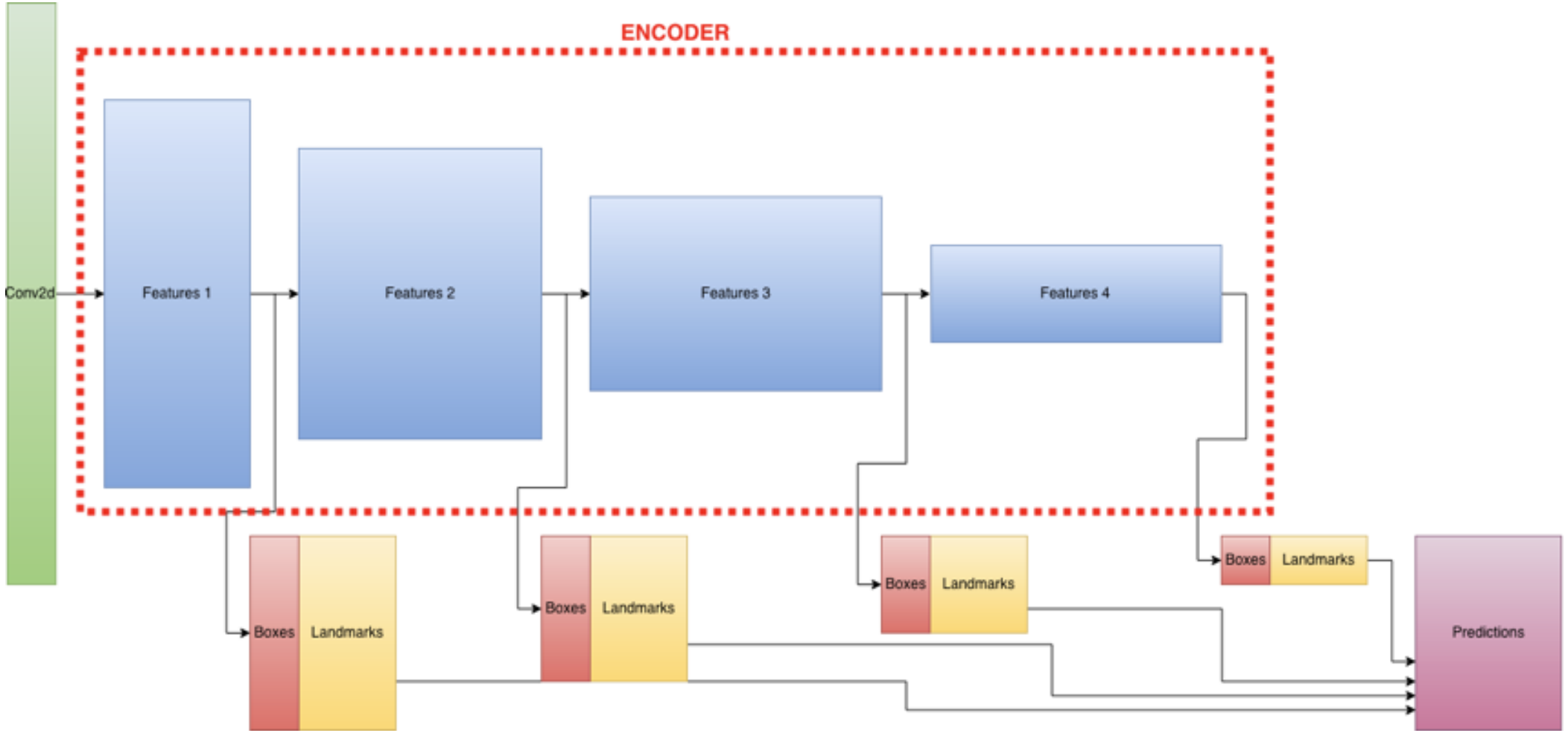




Datasets: AFLW



Face detection with direct regression of landmarks

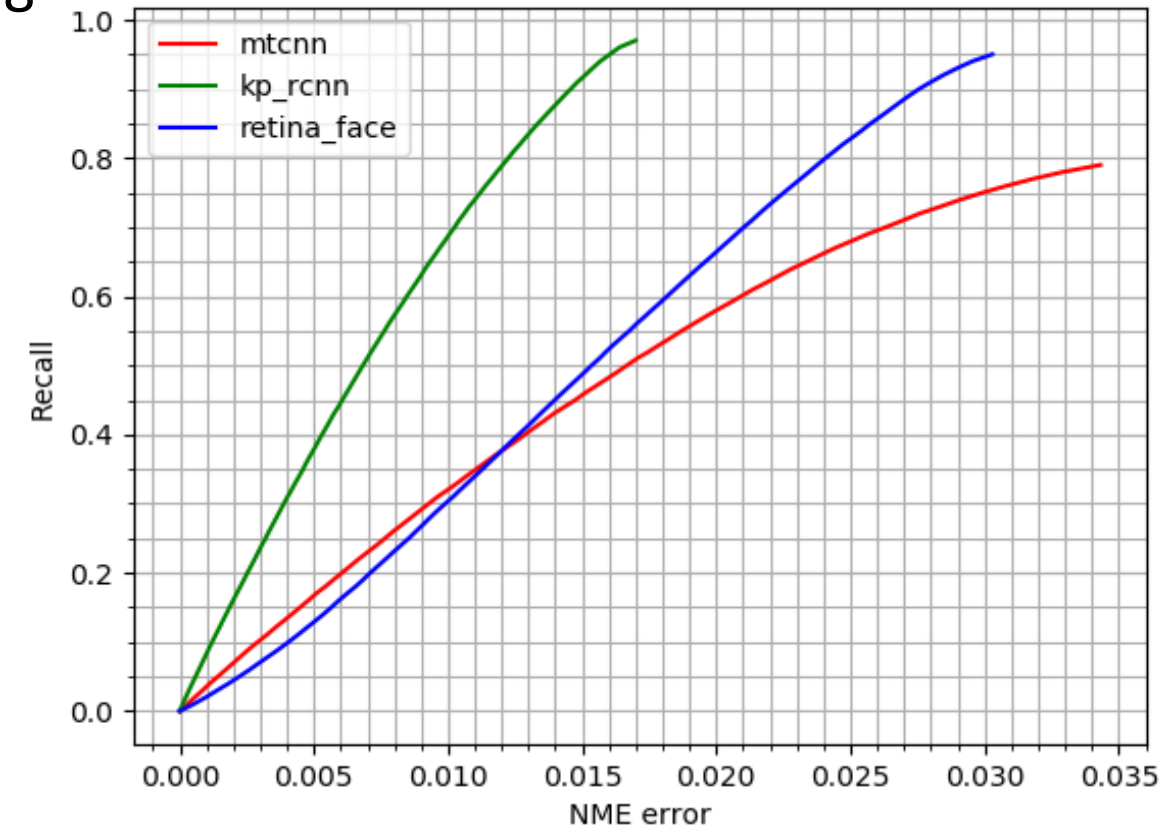


Our results for direct landmark regression

Mask R-CNN: results

- Accuracy:
  - WIDER Face AP:
    - MobileNetv2: 0.958 / 0.949 / 0.882 (easy, medium, hard)

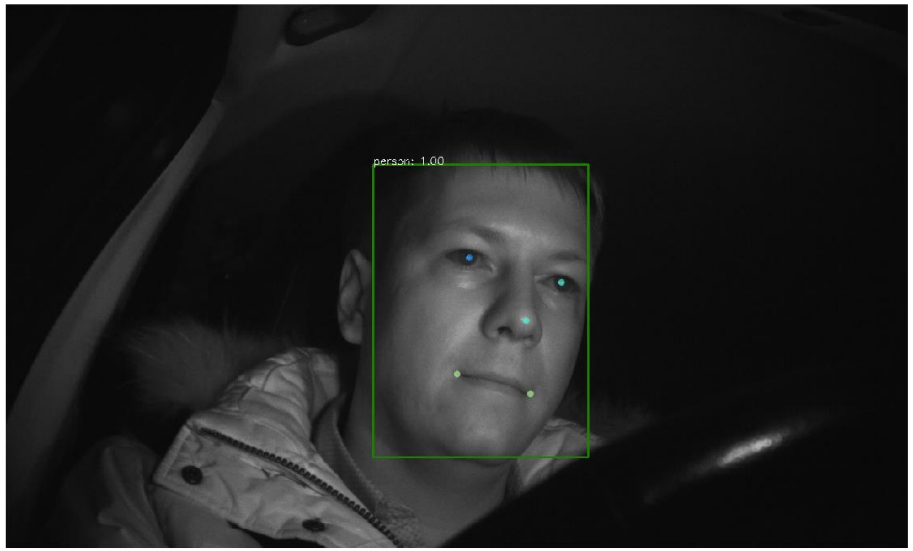
      SOTA: 0.969 / 0.961 / 0.918
  - AFLW NME: 0.0168



Mask R-CNN: AFLW examples

Mask R-CNN: WIDER FACE examples

Mask R-CNN examples



Eye Gaze direction estimation

Gaze estimation problem

- Approaches
  - Pupil Center Corneal Reflection (PCCR) methods
    - 2D regression based
    - 3D model based*
    - Cross-ratio based
  - Appearance based methods*
    - ML/DL
  - Combined methods
  - Auxiliary methods
    - Shape based methods
      (eye region, landmarks detection, etc)*

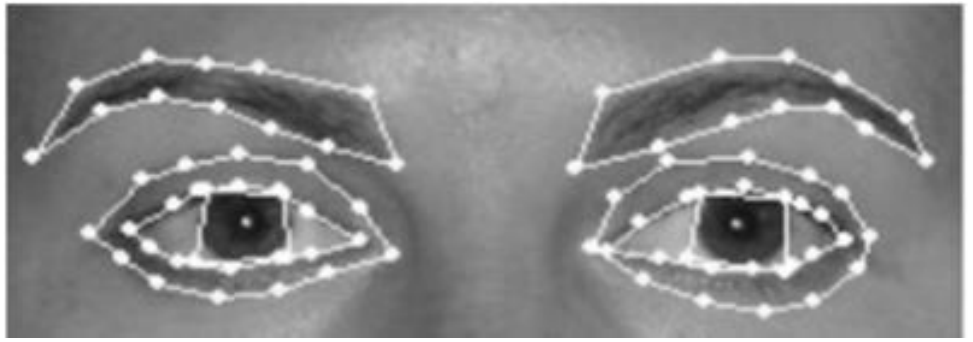
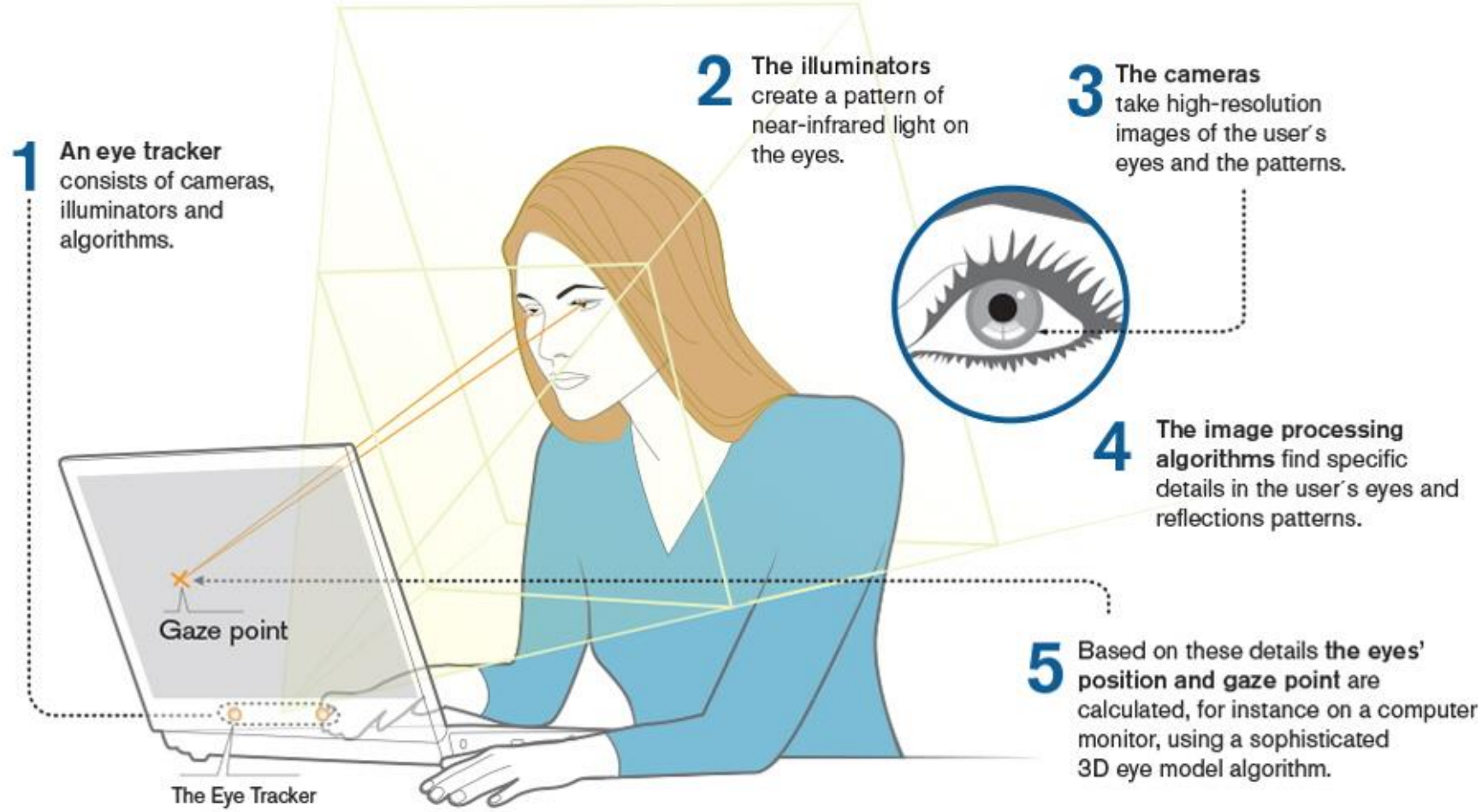


Image fitted with an Active Appearance model of the eye region[1]

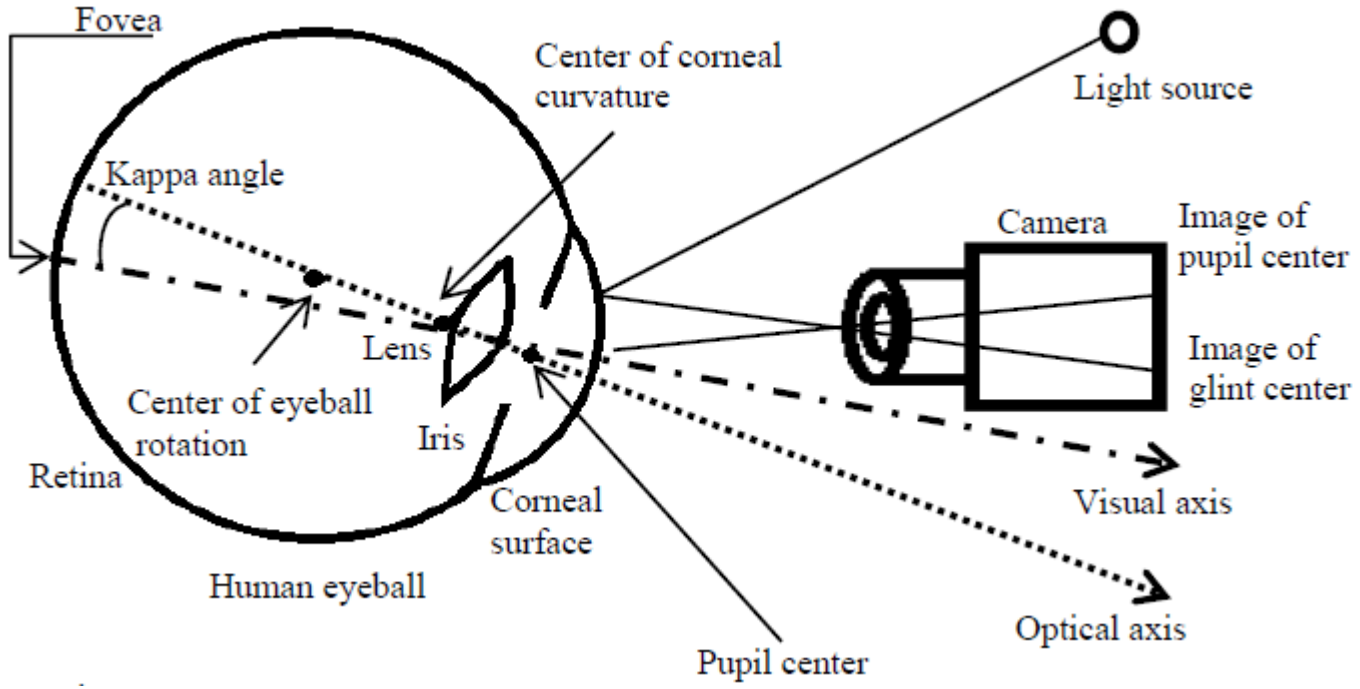
How does PCCR methods work?

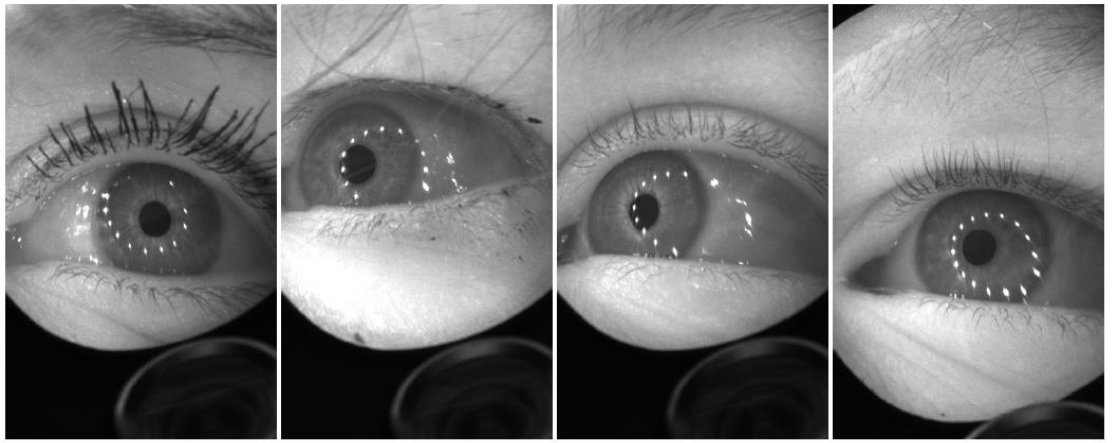


https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/how-do-tobii-eye-trackers-work/

PCCR 3D model based method

- These methods use a geometrical model of the human eye to estimate the center of the cornea, optical and visual axes of the eye and estimate the gaze coordinates as points of intersection where the visual axes meets the scene. 3D model based methods can be categorized on the basis of whether they use single or multiple cameras and type of user calibration required.



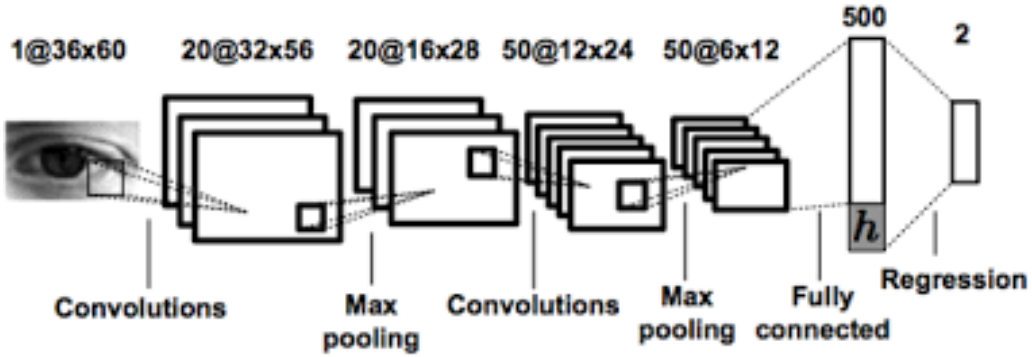


PCCR methods overview

- Advantages
  - High accuracy
  - Eye model based on
    - Physical eye features
    - Eye geometry
- Disadvantages
  - Requires lots of equipment (high hardware requirements)
  - Sensitive to illumination/light conditions
  - Requires calibration (except cross ratio)

Appearance based methods

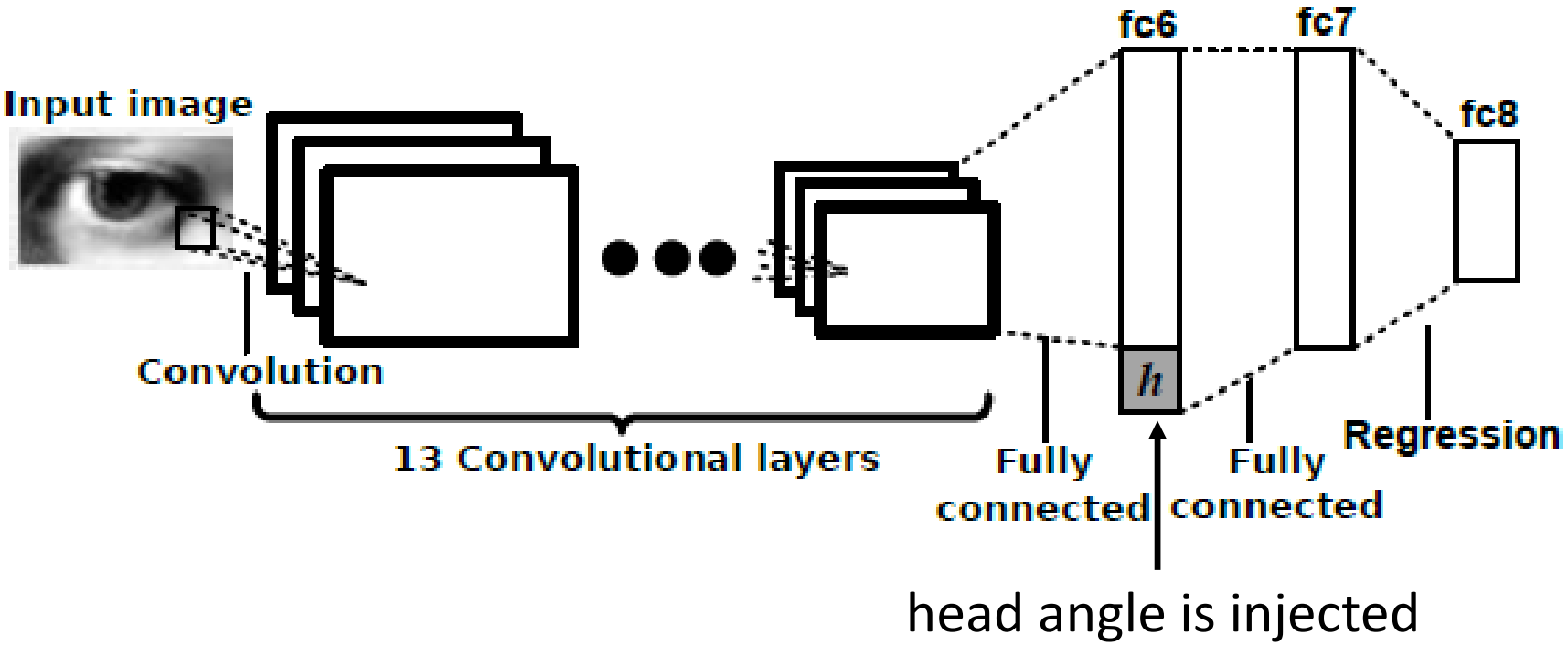
- In appearance based methods the information from the eye region is represented using a model trained with a set of features extracted from eye images.



- Main advantage: low hardware requirements, more robust in general
- Main disadvantage: data hungry, worse accuracy in comparison with PCCR methods.

Most famous architectures

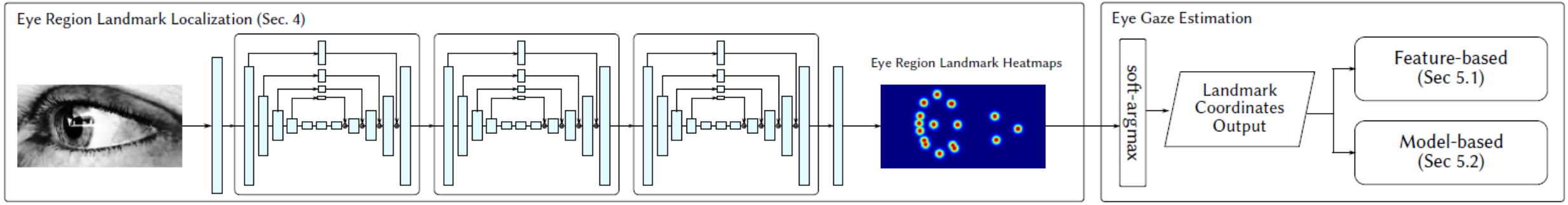
- GazeNet from MPIIGaze, Zhang et al



Most famous architectures

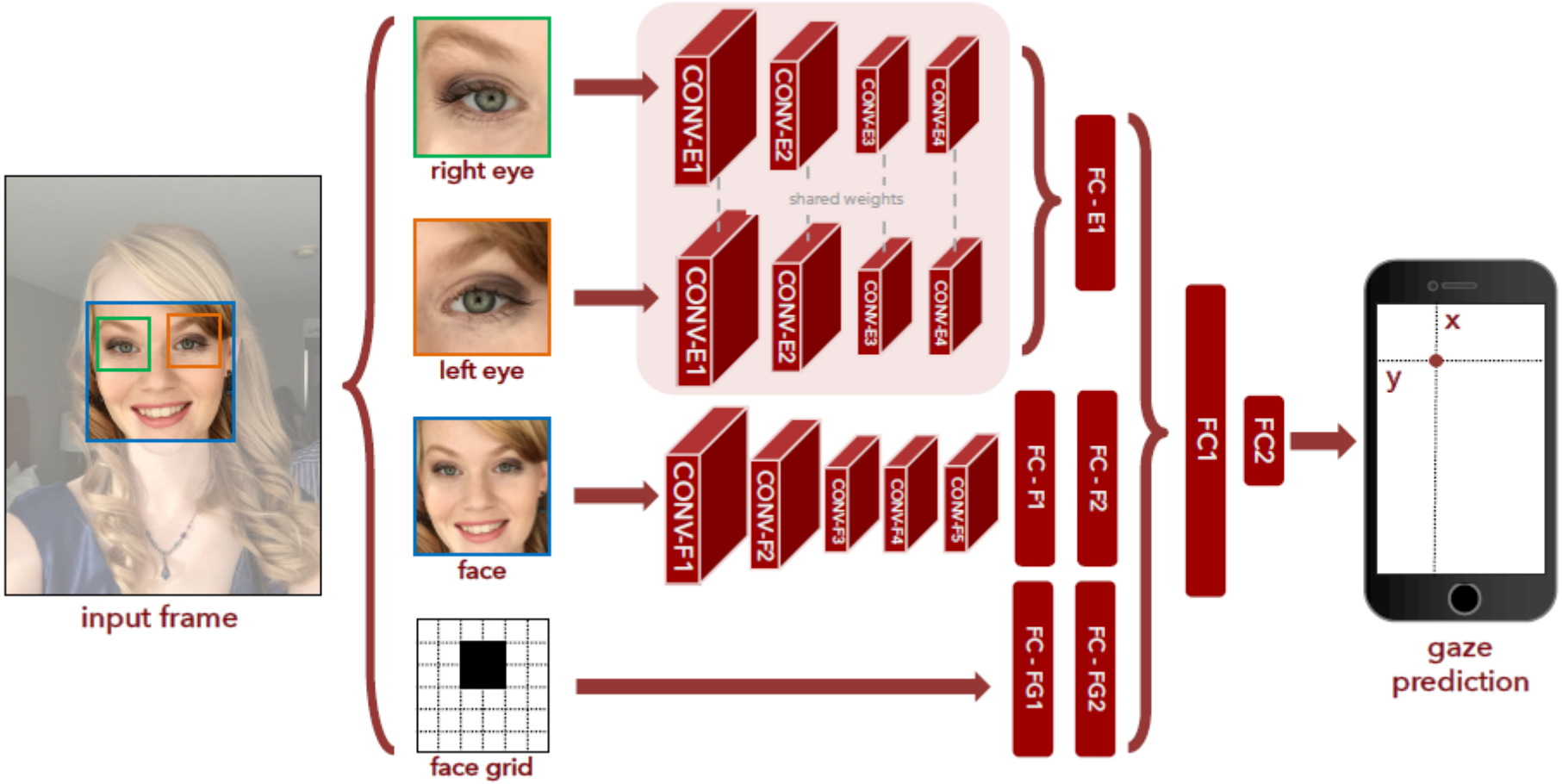
- Learning to Find Eye Region Landmarks for Remote Gaze Estimation in Unconstrained Settings

Learning to Find Eye Region Landmarks for Remote Gaze Estimation ETRA '18, June 14–17, 2018, Warsaw, Poland



Most famous architectures

- Eye Tracking for Everyone, Krafka et al.

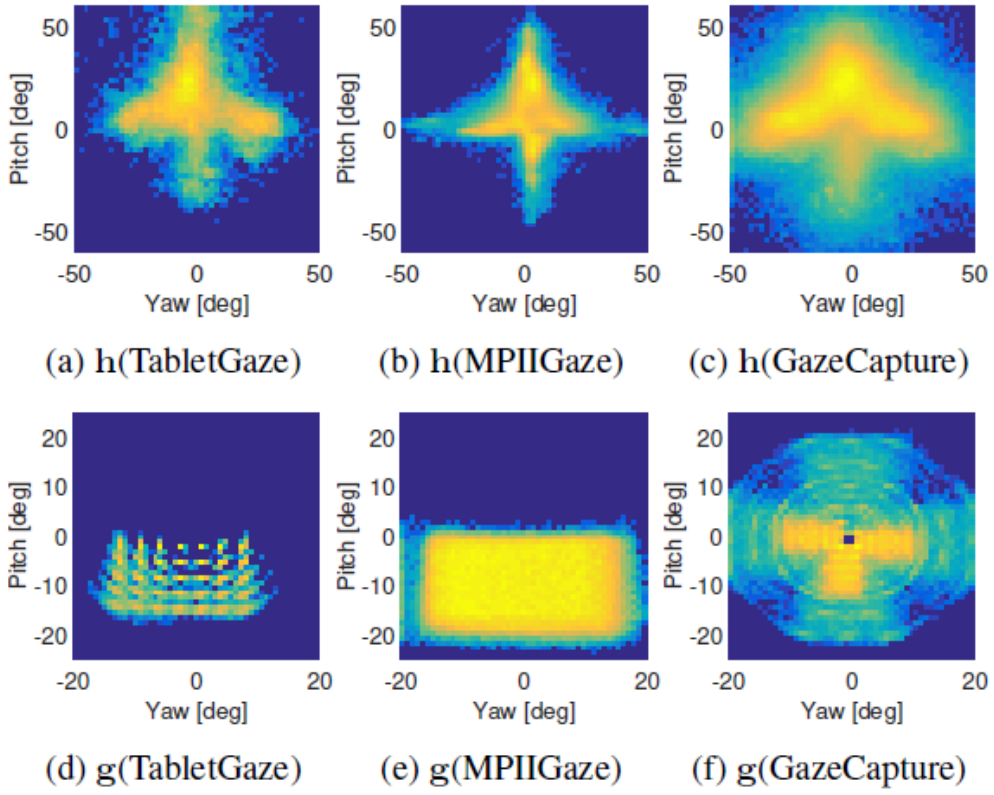


Datasets

Datasets overview.

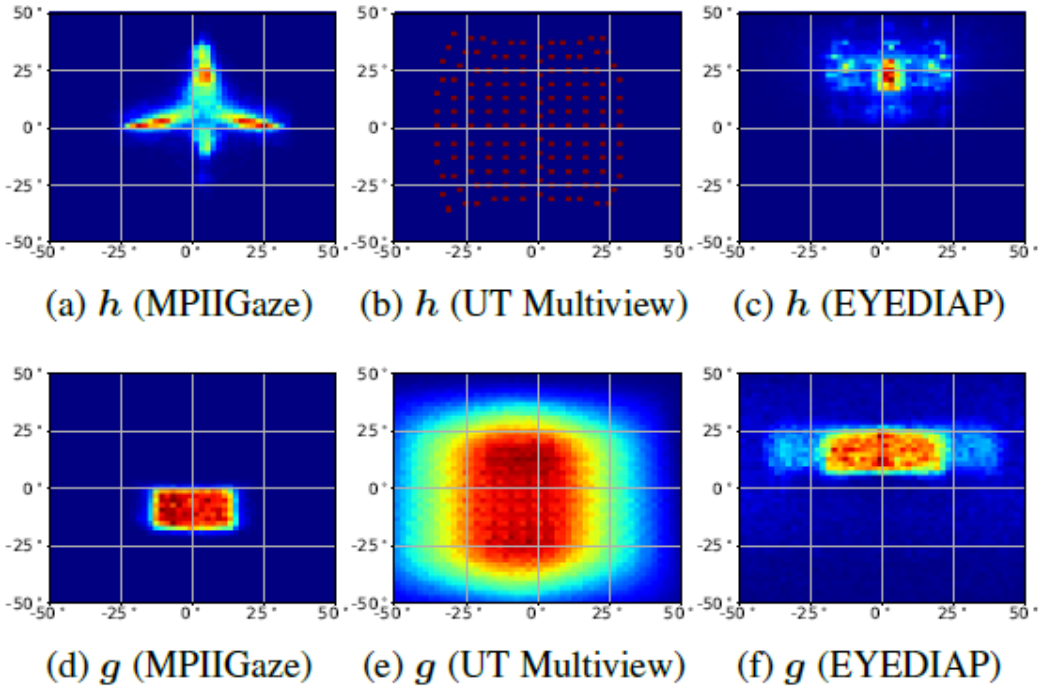
	Participants	Head poses	Gaze targets	Illumination conditions	Face annotations	Amount of data	Collection duration	3D anno.
Villaneuva et al.	103	1	12	1	1,236	1,236	1 day	No
TabletGaze	51	continuous	35	1	none	1,428 min	1 day	No
GazeCapture	1,474	continuous	continuous	daily life	none	2,445,504	1 day	No
Columbia	56	5	21	1	none	5,880	1 day	Yes
McMurrough et al.	20	1	16	1	none	97 min	1 day	Yes
Weidenbacher et al.	20	19	2-9	1	2,220	2,220	1 day	Yes
OMEG	50	3 + continuous	10	1	unknown	333 min	1 day	Yes
EYEDIAP	16	continuous	continuous	2	none	237 min	2 days	Yes
UT Multiview	50	8 + synthesised	160	1	64,000	64,000	1 day	Yes
MPIIGaze	**15**	**continuous**	**continuous**	**daily life**	**37,667**	**213,659**	**9 days ∼ 3 months**	**Yes**

Most popular datasets comparison



(a) h(TabletGaze) (b) h(MPIIGaze) (c) h(GazeCapture)

(d) g(TabletGaze) (e) g(MPIIGaze) (f) g(GazeCapture)



(a) h (MPIIGaze) (b) h (UT Multiview) (c) h (EYEDIAP)

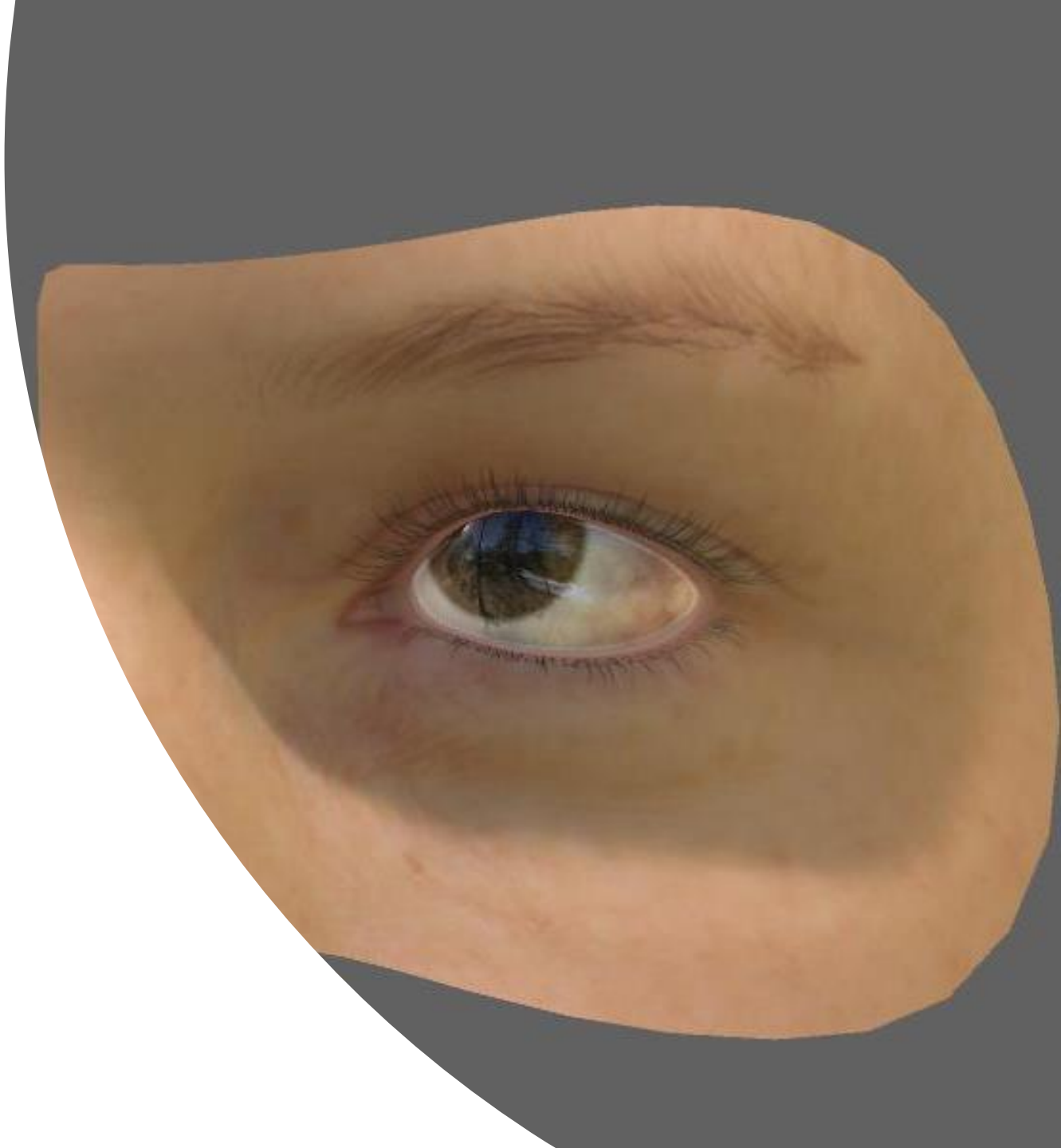
(d) g (MPIIGaze) (e) g (UT Multiview) (f) g (EYEDIAP)

Distribution of head pose h (1st row) and gaze direction g relative to the head pose (2nd row) for datasets TabletGaze, MPIIGaze, GazeCapture, UT-Multiview, and EYEDIAP. All intensities are logarithmic.

Synthetic dataset with Unity Eyes

UNIVERSITY OF CAMBRIDGE COMPUTER LABORATORY
RAINBOW RESEARCH GROUP

Unity Eyes tool makes possible rapidly synthesize large amounts of variable eye region images as training data.



Results. Gaze estimation mean error in degrees

Model	MPIIGaze(cross-person)	UT Multiview	EYEDIAP	Columbia	TabletGaze
GazeNet (Zhang et al)	5.4°	9.8°	9.6°	-	3.63°
ELG (Park et al)	4.6°	11.5°	7.5°	6.2°	-
i-Tracker (Krafka et al)	-	-	-	-	2.58°

Our approach results

- MPIIGaze (cross-person) mean degree error: 3.3° (SOTA ~4.5°)
- Predict examples on MPIIGaze. Top – ground truth, bottom – predicted .