Optimizing face detection machine (deep) learning tools for infant and child faces in video data for data de-identification (and face tracking purposes)

“BLUR THE BABY”

Harold Kim & Mary Zhuo Ke

What is the Problem?

Data Privacy

for Infants and Children of Parents who share their data with research institutions

- In 2016, Feinstein Institute for Medical Research was **fined $3.9 Million** for disclosing identifying data and violating HIPAA.
- https://www.hhs.gov/hipaa/for-professionals/compliance-enforcement/agreements/feinstein/index.html

Data Sharing

for Collaborative Research

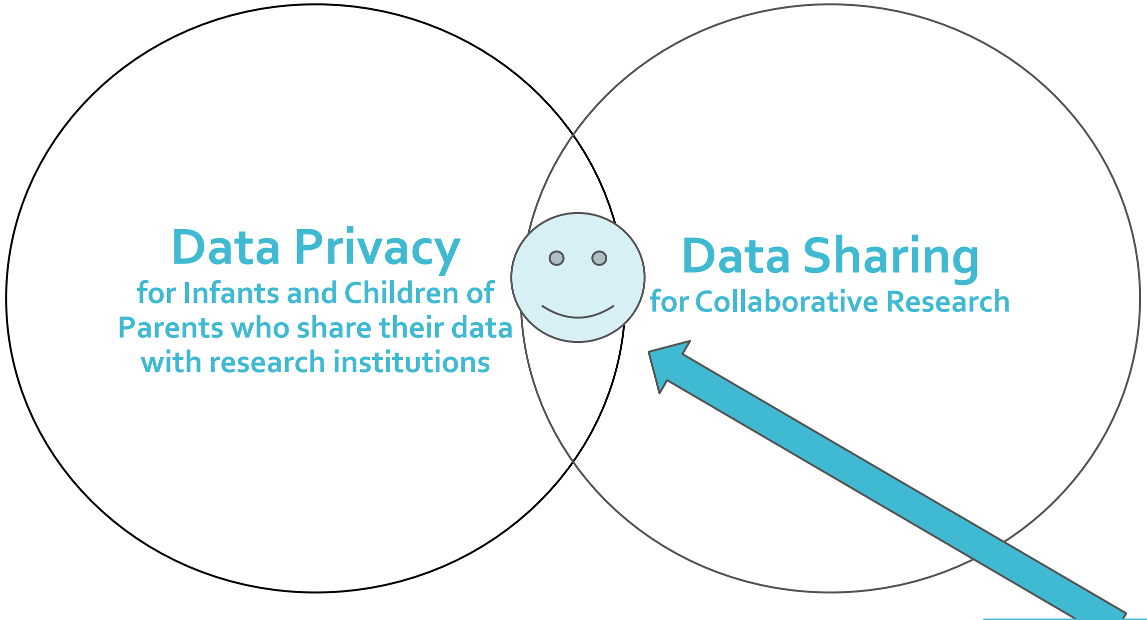
- **Identifiable data cannot be shared easily**, limiting further research on data types that can be better informed by video recordings
- **Consent forms** given to participants/patients **are not intuitive**, leading to sharing without full awareness of the risks.

Data Usability

for Innovative Research

- Video data is often difficult to load, store, and use.
- Measures to assess data quality can take a long time and be subjective.

How will we solve it?



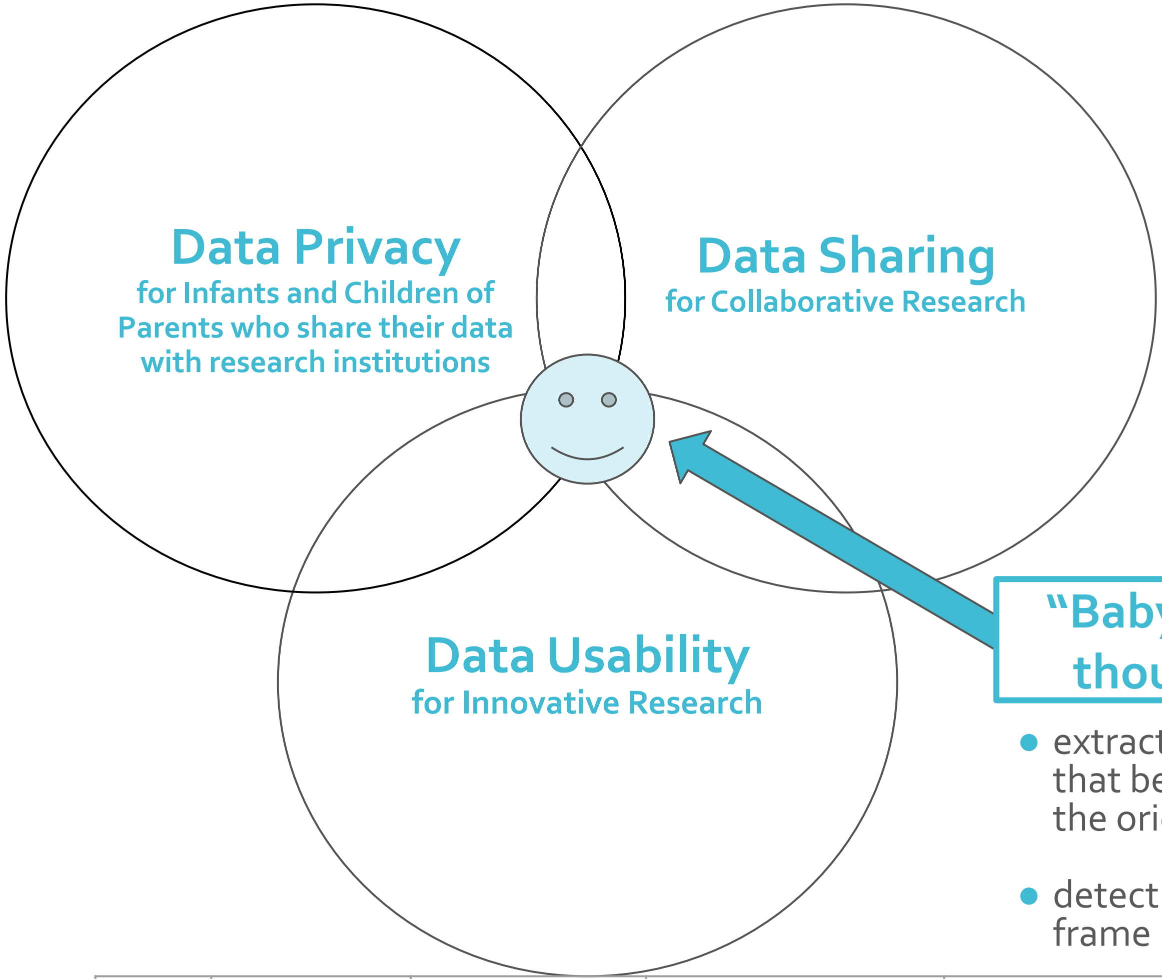
Data Privacy
for Infants and Children of Parents who share their data with research institutions

Data Sharing
for Collaborative Research

Blur the Baby!

- Takes video recordings of infants/children
- Detects identifiable faces with assessment of overall accuracy
- Blurs faces frame by frame
- Compiles a de-identified video that is safer to share in the research community.

How will we solve it?



Data Privacy
for Infants and Children of Parents who share their data with research institutions

Data Sharing
for Collaborative Research

Data Usability
for Innovative Research

"BabyFace"? (haven't thought of a name)

- extract features into tabular format that become deidentified derivatives of the originally identifiable data
- detect faces, movements by tracking frame by frame

SID	Frame # in Video	Coordinate left eye	Coordinate right eye	Coordinates mouth	Difference in x, y movement from last frame
123	1	143.123, 98.86	145.123, 98.76	….	….
123	2	140.123, 95.86	142.123, 95.76	….	….
123	3	….	….	….	….

Why videos?

- They are a **large data type** that otherwise is often archived, despite the richness in the features possible with video data.
- They take up a lot of space, so extracting features in tabular form would **save a lot of cost and storage** (secondary goal to de-identification).
- They have **more complexities** that pose a challenge to evaluate compared to photos (e.g., animations, transitions from frame to frame, blurred movements), which **makes it a great deep learning problem.**

PART 1

I’ll present the initial inspiration and use of the “Blur the Baby” tool envisioned for data de-identification.

What will Blur the Baby improve?

- Protect privacy of **minors who are not able to give consent yet, thus reducing their risk** of privacy concerns when they are older.
- Motivate ways to improve privacy consent **without completely discarding valuable data.**
- **Enable privacy-informed data sharing** by removing the identifiable piece of video recording data so that researchers can use other elements of the data with more assurance.
- **Stress-test** existing tools, find essential areas of improvement
  - Understand where tools fail, why they fail, and what can be learned about it when setting up data collection

What kind of solution do we need?

Affordably Efficient

There shouldn't be a high cost to data privacy. If we can propose a solution that d**oesn't take too much time and heavy resources**, that would be ideal.

The **solution should scale**, so if you have 1M videos, you don't want it to take forever.

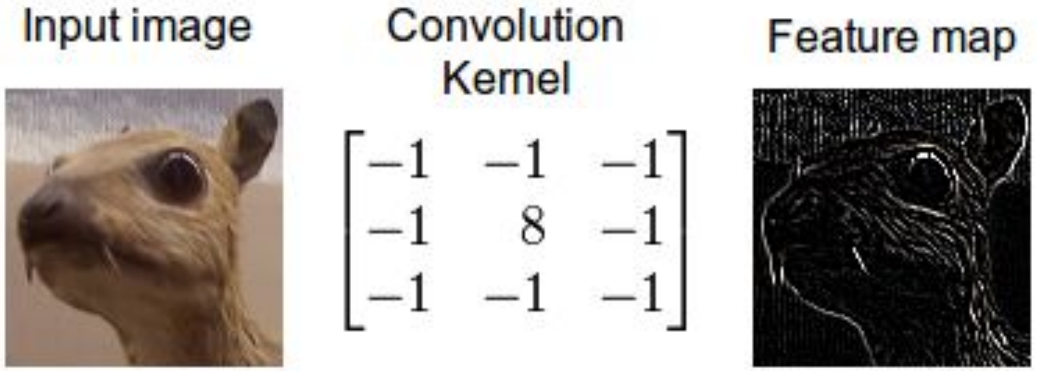
100% Accuracy

We **can't** have partially blurred videos, so we need to aim for 100% accuracy.

What's the point of de-identification if one of the frames has an exposed face?

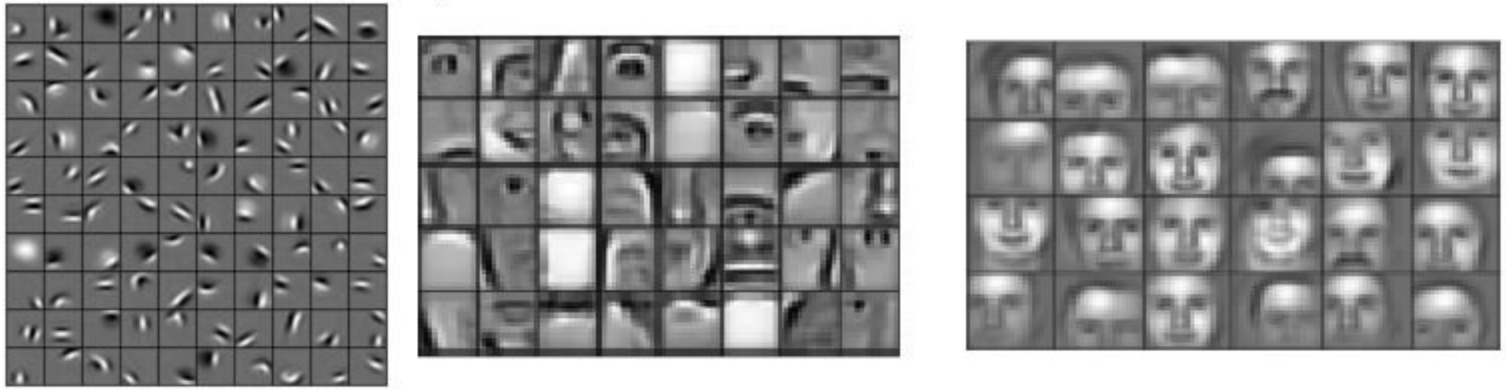
Visual of a Convolutional Neural Network

Convolution - smoothing out of some features, and depending on the size, filters larger or more refined features

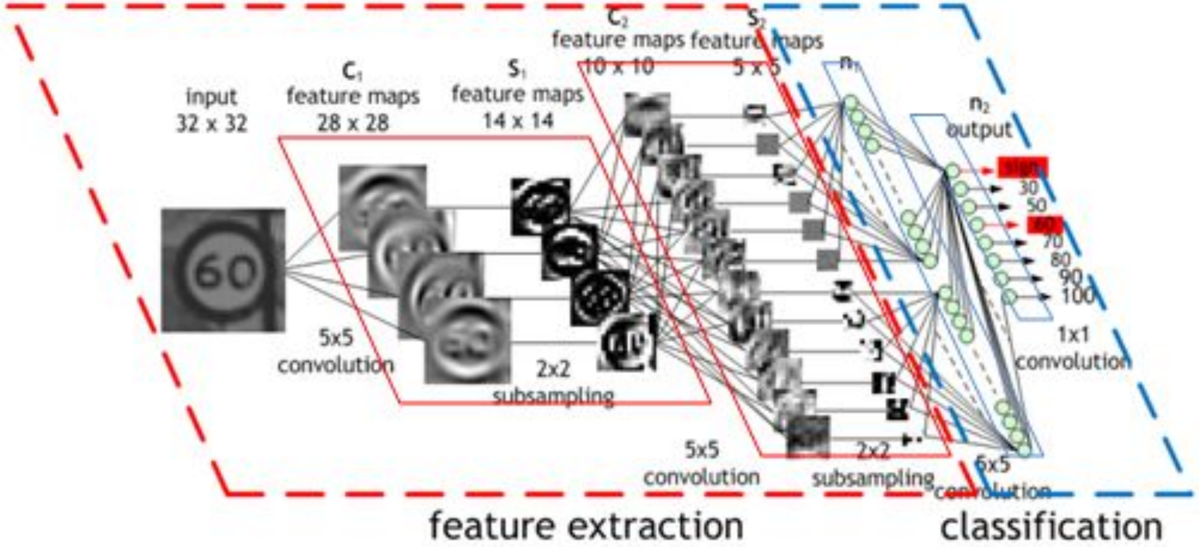


$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Learned features that indicate presence of a particular class



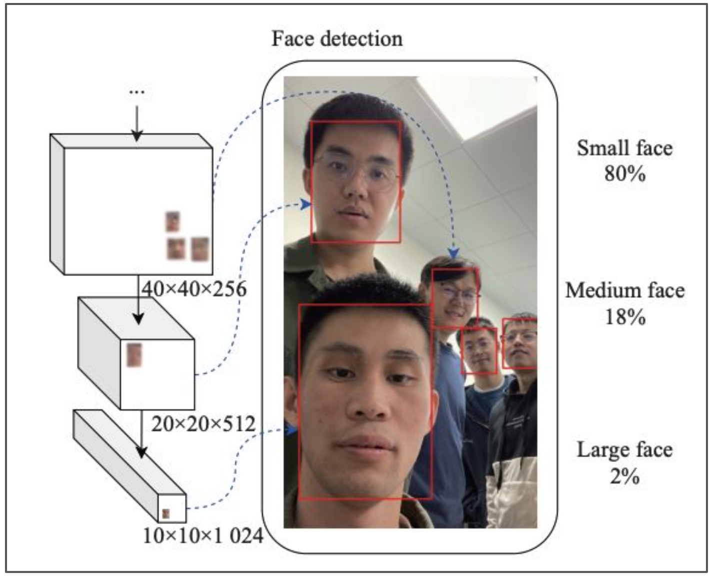
Layers that can stack together to piece together info about complex features to better distinguish the object



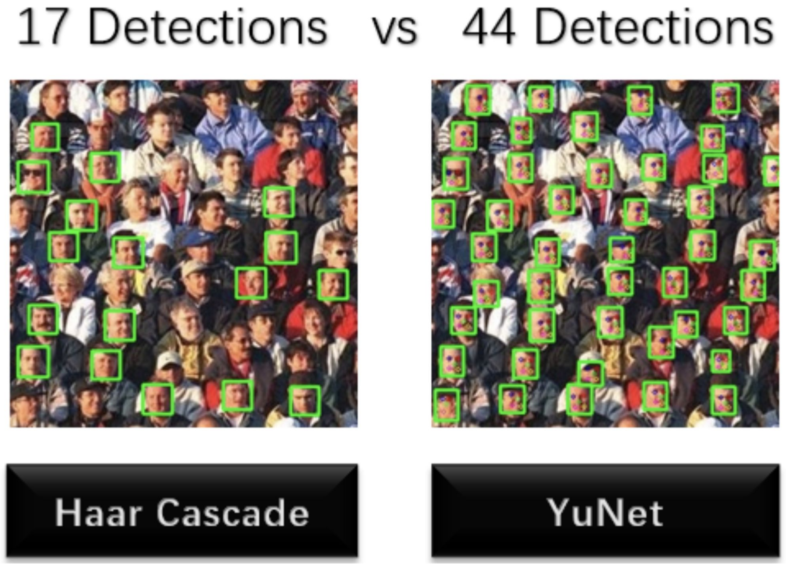
https://developer.nvidia.com/discover/convolutional-neural-network
https://developer.nvidia.com/blog/deep-learning-nutshell-core-concepts/

Solutions available today

YuNet



Wu, W., Peng, H. & Yu, S. YuNet: A Tiny Millisecond-level Face Detector. Mach. Intell. Res. 20, 656–665 (2023). https://doi.org/10.1007/s11633-023-1423-y

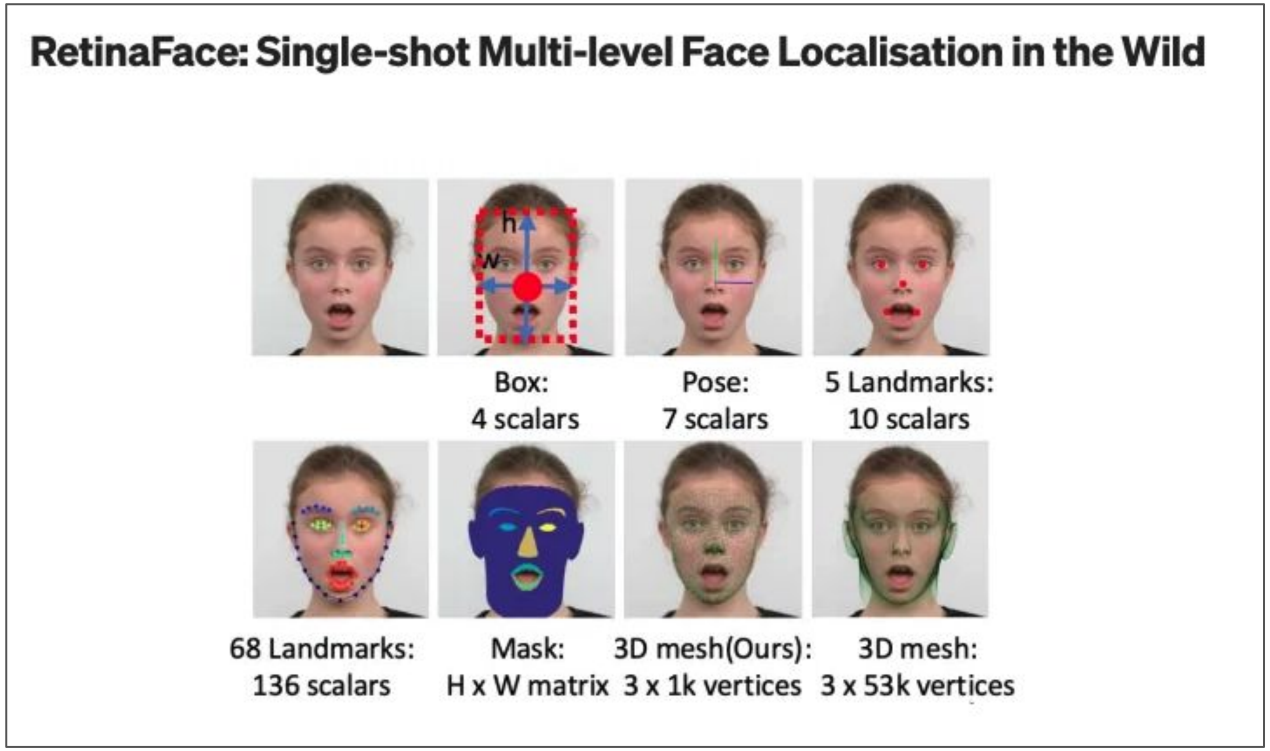


https://opencv.org/blog/opencv-face-detection-cascade-classifier-vs-yunet/

- CNN-based
- trained on WIDER FACE dataset
- lightweight, great for edge devices
- focuses on small faces that are hard to detect
- tiny feature pyramid network

Solutions available today

RetinaFace

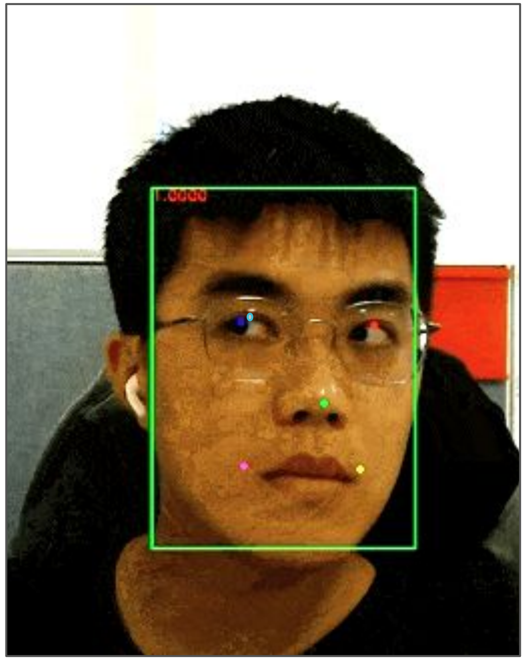


RetinaFace: Single-shot Multi-level Face Localisation in the Wild

https://medium.com/analytics-vidhya/exploring-other-face-detection-approaches-part-1-retinaface-9b00f453fd15

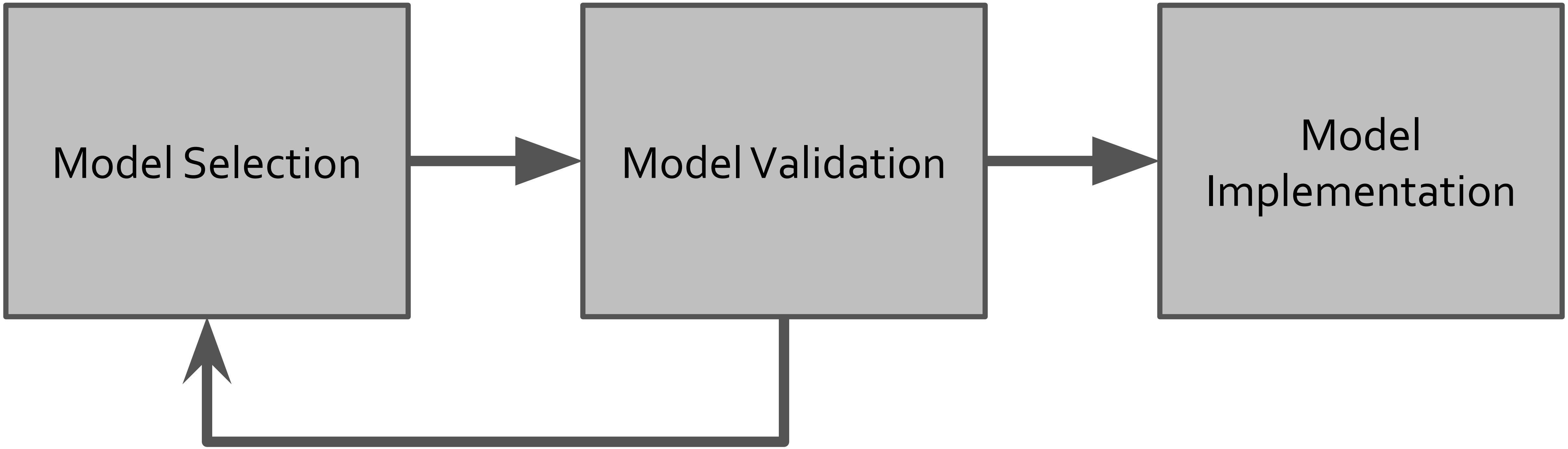
- trained on WIDER FACE dataset
- ResNet, MobileNet using pretrained ImageNet
- computationally intense
- uses comprehensive techniques that help estimate 3D attributes too
- focuses on accuracy at multiple scales

Face Detection Output Data



https://github.com/opencv/opencv_zoo/tree/main/models/face_detection_yunet

- Face features:
  - left eye
  - right eye
  - nose
  - left edge of mouth
  - right edge of mouth
- bounding box of the face
- confidence score in face detection



Model Selection

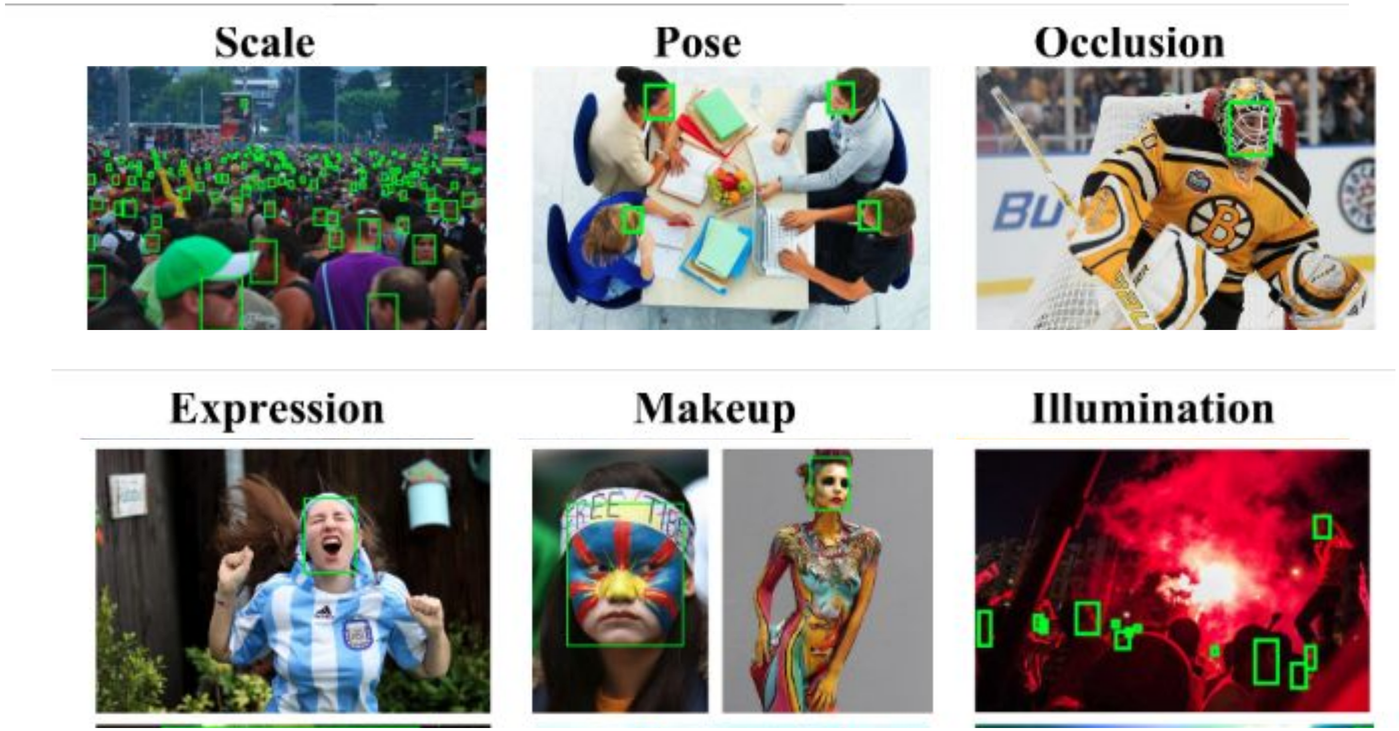
Model Validation

Model Implementation

Refine model based on validation

Datasets

WIDER FACE validation data



- Focused on validation datasets of the **Family Group** photos
- Ground truth labels of number of faces per photo
- http://shuoyang1213.me/WIDERFACE/

Assessment of models using WIDER FACE validation dataset

Quick sanity check:
WIDER FACE validation data

Metric	YuNet	RetinaFace
Number of Images	58	58
Time to Run	1.71 seconds	27.85 seconds
Accuracy	36.21%	77.59%

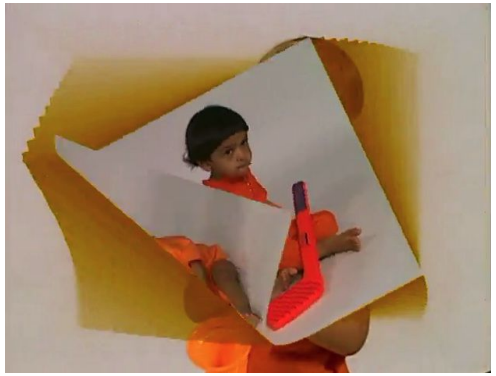
If an image has 10 faces, it would be considered a fail if it detected 9/10 faces.

We see that RetinaFace outperformed for WIDER FACE validation photos.

This serves as a benchmark for us to compare the performance of each tool with data that we have the ground truth for, so we can know what to expect for data we do not have the ground truth for (a.k.a. NEW data) in terms of relative performance.

Datasets

Brainy Baby Video

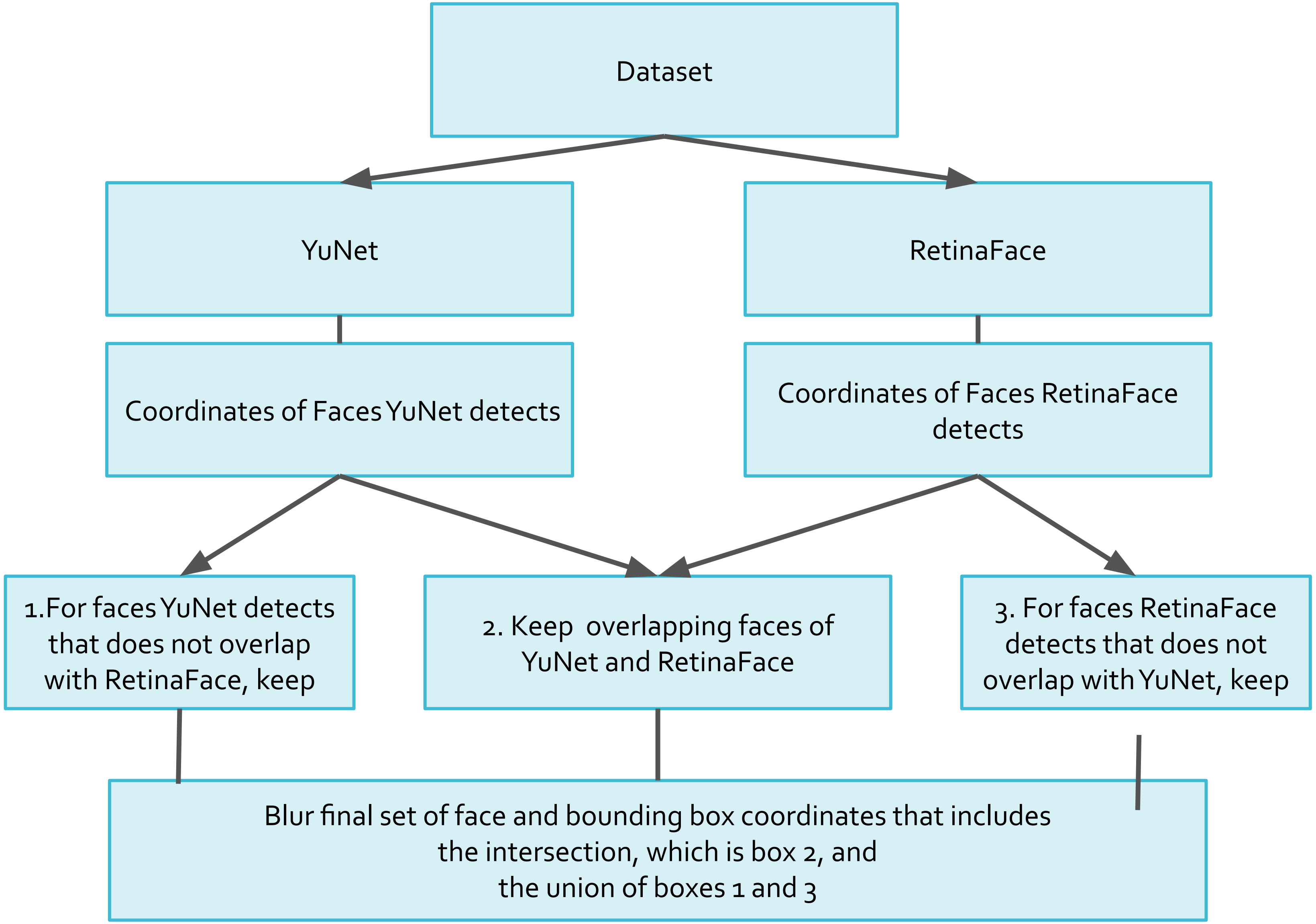


- Videos of babies and toddlers that are for public use
- https://archive.org/about/
- Breakdown .mp4 video into individual frames
  - 30 frames per second video
  - 14,500+ frames used
- Over 15 different baby faces, 10 different toddler faces, and 5 different adult faces
- Multiple orientations and occluded faces

Model 1:

BTB (Blur The Baby)

Intersection and Union Ensemble Model of YuNet and RetinaFace



Dataset

YuNet

RetinaFace

Coordinates of Faces YuNet detects

Coordinates of Faces RetinaFace detects

1.For faces YuNet detects that does not overlap with RetinaFace, keep

2. Keep overlapping faces of YuNet and RetinaFace

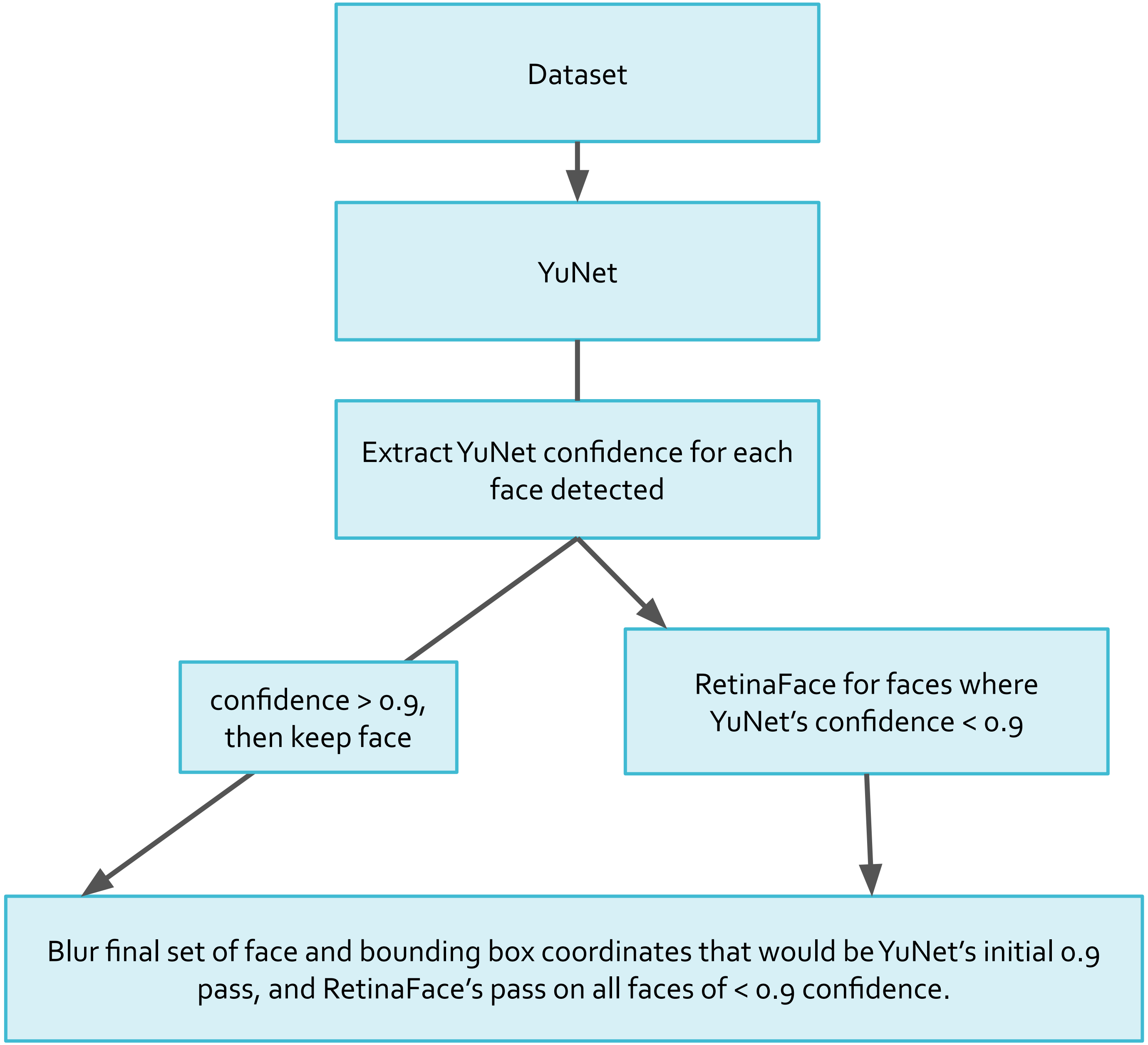
3. For faces RetinaFace detects that does not overlap with YuNet, keep

Blur final set of face and bounding box coordinates that includes
the intersection, which is box 2, and
the union of boxes 1 and 3

Model 2:

BTB (Blur The Baby) w/ YuNet as first pass and RetinaFace as second pass

that takes ⅓ of the time.

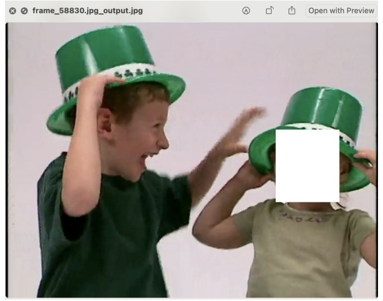

Dataset
YuNet
Extract YuNet confidence for each face detected
confidence > 0.9, then keep face
RetinaFace for faces where YuNet's confidence < 0.9
Blur final set of face and bounding box coordinates that would be YuNet's initial 0.9 pass, and RetinaFace's pass on all faces of < 0.9 confidence.

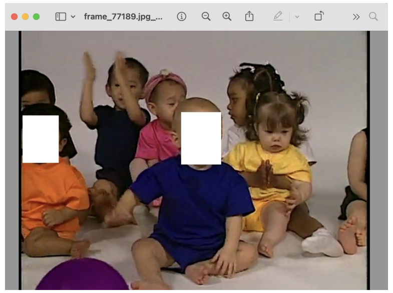
Assessment of BTB using Brainy Baby Video

CHALLENGE CASE: BABY VIDEO

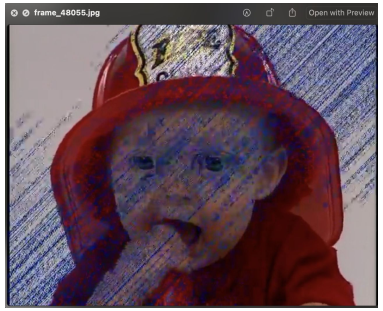
Metric	YuNet	RetinaFace	BTB w/ Intersection over Union	BTB w/ YuNet and RetinaFace ensemble
Number of Frames (with and without human faces)	14,434	14,434	14,434	14,434
Number of Faces Detected	11,066	15,456	15,479	15,400+ (forgot to print this info, but est based on rewatching the video)
Time to Run	5.95 minutes	2.72 hours	2.75 hours	1 hour

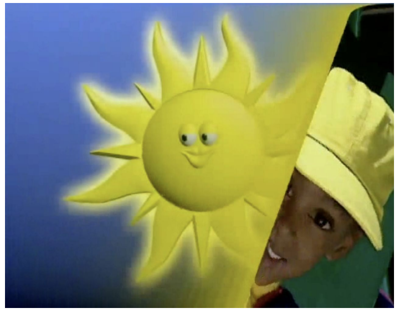
Cases YuNet misses





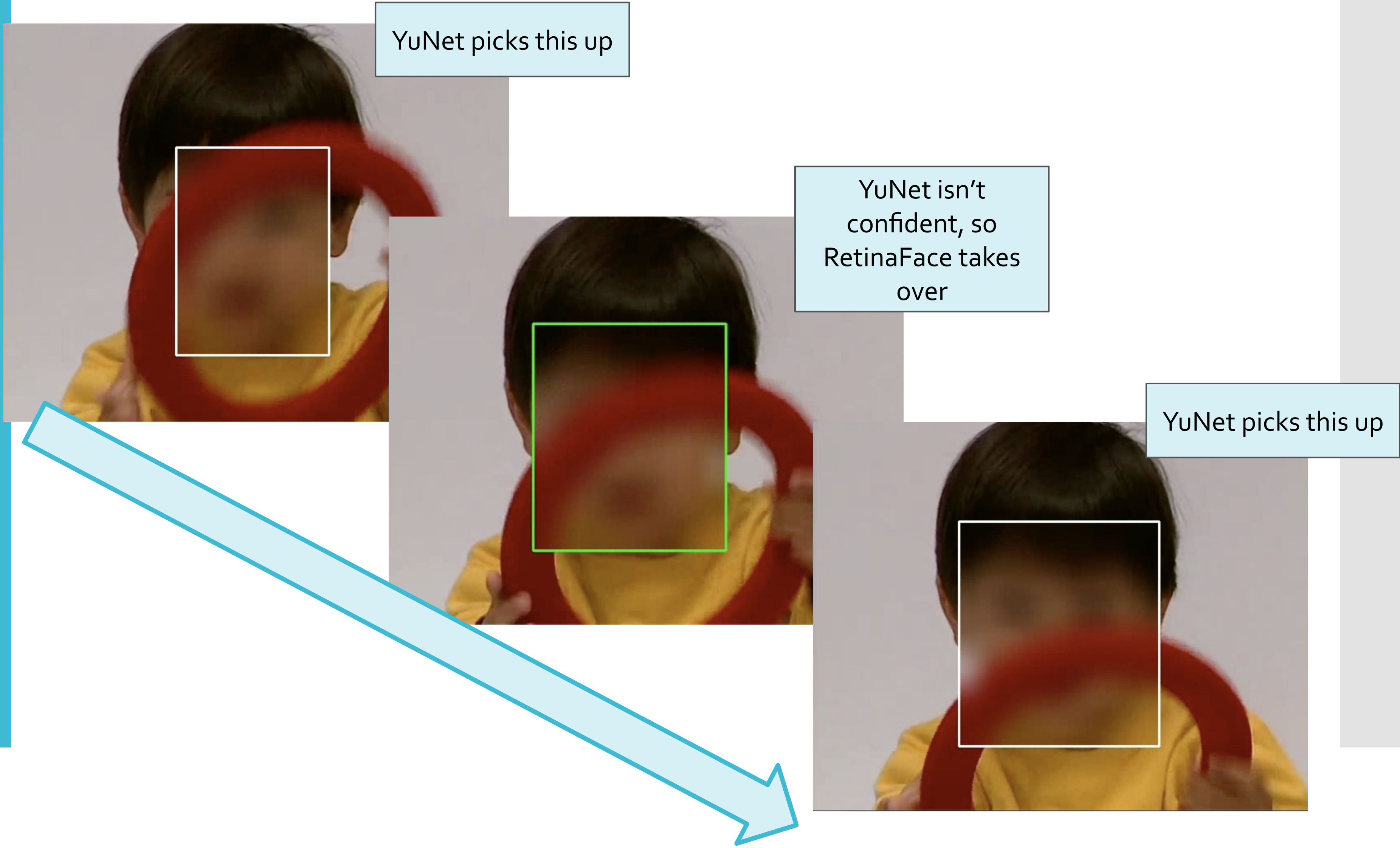
Cases RetinaFace misses





Video Demonstration of Final Output Video (30 seconds)

Final blurred video:
https://drive.google.com/file/d/115nT6ytIaGr_e2PAyakU_4yG_iZp4Jqn/view?usp=sharing



YuNet picks this up

YuNet isn't confident, so RetinaFace takes over

YuNet picks this up

Challenges

Setting up environment with conflicting packages (errors with tensorflow, keras, cv2)

Needing to have the correct format of the data and verifying the type of data being fed into the model when trying to train

Balancing performance with accuracy, deciding on that tradeoff based on level of risk

Ways I’ve tried to assess accuracy with an unlabeled video

- Suspicious if there are frames with zero detected faces.
  - manually check just those (there’s a dataframe you can filter that tells you how many were detected per frame)
- If you know there are 2 people in the frame (parent and child), then expect there to always be 2 faces per frame. Check the ones that aren’t.
- These tools capture most frames, so I found it to be very few frames in the end (99% detected, blurred)

Future work that could enable better solutions

- Having **more open source face datasets that are labeled** to enable better training and assessment of accuracy
- **Handle transitions/animations more robustly** as videos increase the variations of orientations/scales, especially if it were 60 fps.
- More **efficient ways** to **construct 3D poses** in order to differentiate faces of people from those of animals, objects, portraits

References

1. Yang, S., Luo, P., Loy, C. C., & Tang, X. (2016). WIDER FACE: A face detection benchmark. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5525–5533. http://shuoyang1213.me/WIDERFACE/

2. Wu, W., Peng, H. & Yu, S. YuNet: A Tiny Millisecond-level Face Detector. Mach. Intell. Res. 20, 656–665 (2023). https://doi.org/10.1007/s11633-023-1423-y
   1. https://github.com/opencv/opencv_zoo/blob/main/models/face_detection_yunet/face_detection_yunet_2023mar.onnx (need to download this model)
   2. Support from chatGPT to help understand YuNet.

3. J. Deng, J. Guo, E. Ververas, I. Kotsia and S. Zafeiriou, "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 5202-5211, doi: 10.1109/CVPR42600.2020.00525.
   1. Support from chatGPT to help understand RetinaFace.

4. Baby video dataset: https://archive.org/details/BrainyBabyShapesandColors