

# Towards a More Trustworthy Facial Analysis System using Self-Auditing

UNIVERSITY  
OF MIAMI



**Rahul Dass**

Ph.D. Student, U-LINK Predoctoral Fellow

April 5, 2021

Dr. Ubbo Visser (Computer Science)  
Dr. Nick Petersen (Sociology and Law)  
Dr. Odelia Schwartz (Computer Science)



## Overview

1. FPT performance – what’s the deal?
2. Highlights from 2020
3. Research questions
4. “Self-auditing” method
5. Results
6. On-going / future work
7. One more thing...

# FPT performance – what's the deal? (1/2)

## SUCSESSES: controlled development and evaluation

- Rigorous evaluation by NIST every 6-months for over 200 algorithms in facial processing technology or FPT (Grother et al., 2018)
  - Only 0.2% error rate
  - Since 2010 and 2014: 25% and 20% gains
- Rapid improvements across different components in FPT pipelines
  - Technical solutions (Ryu et al, 2017; Kärkkaäinen & Joo, 2019; Dass et al., 2020)
  - Societal repercussions (Buolamwini & Gebru, 2018; Hanna et al., 2020; Raji & Fried, 2021)
  - Force industrial accountability (Raji & Buolamwini, 2019; Heilweil, 2020)
  - Tech policies for stakeholders (Garvie, 2016; Moy, 2019; Learned-Miller et al., 2020)

# FPT performance – what's the deal? (2/2)

## **FAILURES: real-world deployments and repercussions**

- Continued failed pilots by Western law enforcement - UK and many states in the U.S. (Raji and Fried, 2021)
- FPT bans and moratoria across the U.S. (Raji and Fried, 2021)
- NIST reporting systemic classification disparities across facial characteristics based on recent varied tests assessing FPT robustness and generalizability (Raji and Fried, 2021)

# Highlights from 2020

*It's Not Just Black and White: Classifying Defendant Mugshots Based on the Multidimensionality of Race and Ethnicity* [Dass et al., 2020]

- **Task:** predict contemporary notions of racial identification for mugshots by considering race and ethnicity as multidimensional
- Aimed to tackle potential sources of bias within a FPT pipeline:
  - **Labeling bias:** using 2 sources of ground-truth (court text-based and human's visual-based)
  - **Dataset bias:** balancing training sample sizes and two data augmentation methods
  - **Algorithmic bias:** 7 vision architectures with ImageNet weights and fine-tune hyperparameters

# From last year...Future Work

- Difference between highest accuracies (Race is 0.37%) and (Race-Ethnicity is 3.31%), models' architecture less contributory when using transfer learning - investigate if training DLMs from scratch makes a difference?
- Inference learning via “Balanced Student Race-Ethnicity” SE-ResNet-50 model:
  - Predict race-ethnicity for remaining student annotated mugshots (14K stratified sample)
  - Generate new DLM-based race-ethnicity labels for remaining 180K mugshots and compare performance with Imbalanced Court trained SE-ResNet-50 (81.05%)
- Evaluate how biased each DLM is w.r.t. each race-ethnicity subgroup and assess if the new methodology fosters DLMs to be more demographically inclusive

# Research Questions\*

*\*building from last year's talk*

- Domain adaptation / transfer learning:
  - To what extent does (1) model architecture and, (2) pretrained weights for *out-of* or *in*-domain task classification
  - Generate (missing) race and race-ethnic labels and compare with ground-truths
- Improve DLM engineering:
  - Other face preprocessing methods, vision architectures, pretrained paradigms?
- Create DLM inference and interpretability pipeline:
  - *Experimentation*-based evaluation on unseen mugshots subject to varying data augmentations (Muthukumar et al., 2018; Balakrishnan et al., 2020)

# Results (1/4) - Generate mugshot labels

- Generated approx. 194K mugshots
- High degree of correspondence with generated court labels,  $r = 0.8143$
- Suggests a viable method for generating missing race-ethnicity labels in court databases
- Expand to investigate disparities in criminal justice

```
[ $ cat mtCt2s_vgg19.csv | head -10
Jail_ID,Black_Prob,White_Prob,Prediction
110068060,0.99995,5e-05,Black
120000006,0.24984,0.75016,White
120000034,0.99997,3e-05,Black
120000104,0.9838,0.0162,Black
120000109,0.00037,0.99963,White
120000164,0.99998,2e-05,Black
120000171,0.99997,3e-05,Black
120000182,0.99999,1e-05,Black
120000195,1.0,0.0,Black
```

```
rdass@sickles ~/Research/ULINK/2020_Image_Vision_Computing/cat4_full_dataset
$ nvidia-smi
Mon Jun 1 00:01:40 2020

+-----+
| NVIDIA-SMI 384.81                Driver Version: 384.81                |
+-----+-----+
| GPU   Name                               Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf    Pwr:Usage/Cap|      Memory-Usage | GPU-Util  Compute M. |
+-----+-----+
| 0     Tesla P100-PCIE...            Off          | 00000000:04:00:0 | Off |
| N/A   44C    P0     43W / 250W | 14074MiB / 16276MiB | 12%      Default |
+-----+-----+
| 1     Tesla P100-PCIE...            Off          | 00000000:82:00:0 | Off |
| N/A   40C    P0     37W / 250W | 14145MiB / 16276MiB | 0%       Default |
+-----+-----+

+-----+
| Processes:                         GPU Memory |
| GPU       PID    Type   Process name                               Usage      |
+-----+-----+
| 0         23987   C     ...rdass/Research/tensorflowEnv/bin/python 723MiB |
| 0         23995   C     ...rdass/Research/tensorflowEnv/bin/python 1957MiB |
| 0         26953   C     ...rdass/Research/tensorflowEnv/bin/python 723MiB |
| 0         26973   C     ...rdass/Research/tensorflowEnv/bin/python 1947MiB |
| 0         27275   C     ...rdass/Research/tensorflowEnv/bin/python 723MiB |
| 0         27315   C     ...rdass/Research/tensorflowEnv/bin/python 723MiB |
| 0         31480   C     ...rdass/Research/tensorflowEnv/bin/python 723MiB |
| 0         31799   C     python                                     1279MiB |
| 0         36250   C     ...rdass/Research/tensorflowEnv/bin/python 1945MiB |
| 0         36519   C     ...rdass/Research/tensorflowEnv/bin/python 3299MiB |
| 1         23987   C     ...rdass/Research/tensorflowEnv/bin/python 2145MiB |
| 1         26953   C     ...rdass/Research/tensorflowEnv/bin/python 2161MiB |
| 1         27275   C     ...rdass/Research/tensorflowEnv/bin/python 3259MiB |
| 1         27315   C     ...rdass/Research/tensorflowEnv/bin/python 3299MiB |
| 1         31480   C     ...rdass/Research/tensorflowEnv/bin/python 3259MiB |
+-----+-----+
```



# Results (2/4) – Extent of Face preprocessing

Original  
[ varying resolution ]



Original resized  
[ 299 x 299 ]



OpenFace  
[ 299 x 299 ]



MTCNN  
[ 299 x 299 ]



[ N = 195,174 ]

[ N = 194,957 ]

[ N = 195,162 ]

**-217**

**-12**

[ Source: Dass et al., 2021a – working paper ]

# Results (3/4) - Improve DLM engineering

Vision architectures: DenseNet161

Pre-trained paradigms: ImageNet vs. Random vs. Face

Table 1: Comparing the validation accuracies of 7 DLMs subject to varying pretrained paradigms using MTCNN preprocessed mugshots for binary (Black vs. White) classification

Model	ImageNet		Random		VGGFace2 (Scratch)		MS-Celeb-1M (Scratch) VGGFace2 (Fine-Tune)	
	Courts	Students	Courts	Students	Courts	Students	Courts	Students
AlexNet	92.75%	92.75%	94.50%	<b>93.50%</b>	-	-	-	-
DenseNet161	96.50%	96.50%	96.00%	92.25%	-	-	-	-
InceptionV4	96.75%	96.75%	94.25%	91.50%	-	-	-	-
ResNet50	95.25%	95.25%	94.00%	92.25%	<b>96.00%</b>	<b>94.75%</b>	<b>97.00%</b>	93.25%
SE-ResNet50	96.75%	96.75%	96.00%	92.00%	96.00%	94.25%	94.75%	<b>93.50%</b>
SE-ResNeXt-50	96.75%	96.75%	<b>96.00%</b>	90.50%	-	-	-	-
VGG19_bn	<b>96.75%</b>	<b>96.75%</b>	92.50%	90.00%	-	-	-	-

[ Source: Dass et al., 2021b – working paper ]

# “Self-auditing” method (1/2)

672 model inference interpretability scenarios

= 28 DLMs x 24 data augmentations (unseen mugshots from same dataset)

- **Inference:** predict binary (Black vs. White) racial categories
  - **Extent of face preprocessing:** Original resized vs. OpenFace vs. MTCNN
  - **Facial orientation:** Natural vs. Inverse (upside-down)
  - **Ground-truth source:** Courts vs. Students
  - **Racial category:** Black vs. White
- **Interpretability:** visualize DLM top-layer for “best” and “worst” mugshots
  - **Saliency maps:** Grad-CAM and guided backpropagation
  - **Greatest model confidence:** correctly (best) and incorrectly (worst) mugshots

# “Self-auditing” method (2/2)

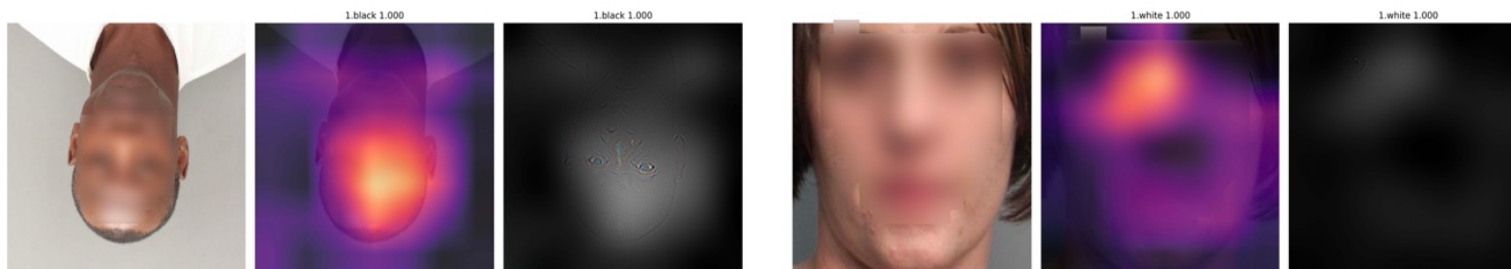
Table 2: Breakdown of 10 test time datasets out of 24 in total based on four data augmentation parameters for model evaluation

Test scenario	Test dataset size	Test augmentation parameters			
		Ground-truth	Face preprocessing	Face orientation	Racial category
1	5,931	court	original	natural	Black
2	6,244	court	original	natural	White
3	5,924	court	OpenFace	natural	Black
4	6,242	court	OpenFace	natural	White
5	5,931	court	MTCNN	natural	Black
6	6,244	court	MTCNN	natural	White
7	6,931	court	original	inverted	Black
8	7,244	court	original	inverted	White
9	6,924	court	OpenFace	inverted	Black
10	7,242	court	OpenFace	inverted	White

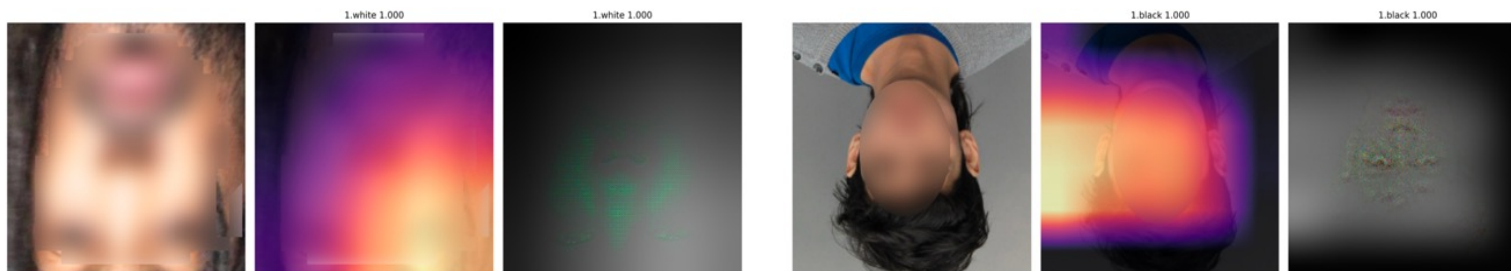
[ Source: Dass et al., 2021a – working paper ]

# Result (4/4) - “Self-auditing” pilot results

Ground-truth source (training)	Initialization paradigm	Model architecture	Test scenario	Racial Category	Model accuracy
student	ImageNet	SE-ResNeXt-50	7	Black	99.94%
court	ImageNet	VGG19_bn	16	White	99.76%
court	Random	InceptionV4	21	Black	28.29%
student	Random	AlexNet	8	White	4.14%



(a) Correctly predicted Black mugshot with highest model confidence. (b) Correctly predicted White mugshot with highest model confidence.



(c) Incorrectly predicted Black mugshot with highest model confidence. (d) Incorrectly predicted White mugshot with highest model confidence.




# On-going / future Work

- *Detecting Racial Inequalities in Criminal Justice: An Ethical Deep Learning Approach for Generating and Interpreting Racial Identification using Mugshots*  
Rahul K. Dass, Nick Petersen, Marisa Omori, Tamara Lave, Ubbo Visser (2021a) – Working paper
- *From ImageNet to Facial Analysis Classification: Rethinking CNN Initialization Paradigms for Out of Domain Adaptation using Self-Auditing*  
Rahul K. Dass, Odelia Schwartz, Ubbo Visser (2021b) – Working paper
- Expanding the *trustworthy* vision methodology to other domains:
  - 2D/3D multi-model object detection (RoboCanes)
  - Retinal fundus images (Bascom Palmer)

# One more thing...

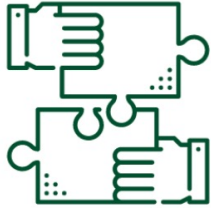
UNIVERSITY OF MIAMI

 GRADUATE SCHOOL



This event covers a topic in the Responsible Conduct of Research.

**A Dialogue with U-LINK's Graduate Fellows on  
Interdisciplinary Research & Team Science**

**Tuesday, April 6, 2021, from 5:30 to 6:30 pm**  
Via Zoom



UM's Laboratory for Integrative Knowledge (U-LINK) teams are addressing some of today's most pressing problems through interdisciplinary research. Graduate students are playing key roles in this work, while building skills in team science. Please join us for an interactive discussion about their experiences, ideas, and advice to other graduate students and early career researchers.

  GRADUATE SCHOOL

April 6, 2021 – **Tomorrow @ 5:30 pm!**

<https://miami.zoom.us/meeting/register/tJUtce6hrDMpGdy4kEpJzOEH4EWdSwaRE1r4>