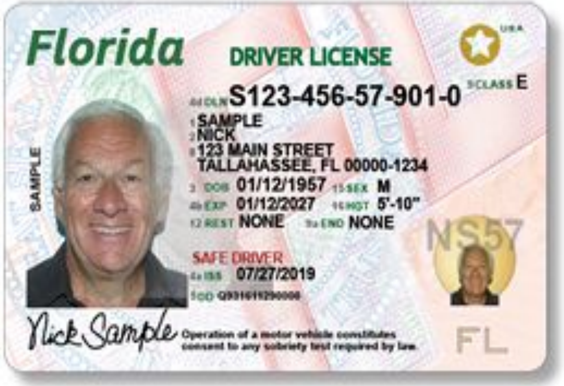
Reducing Geographic Performance Differentials for Face Recognition

Martins Bruveris

onfido

The Problem

- 1:1 Face Recognition between selfies and photos of documents



- Part of Onfido’s remote identity verification solution
- The document proves your identity
- The selfie proves the document belongs to you

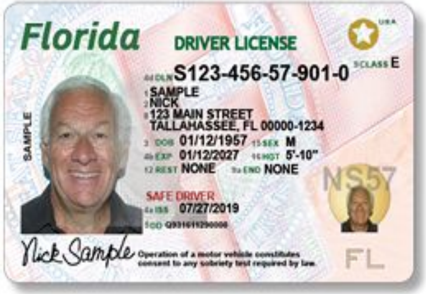
The Problem

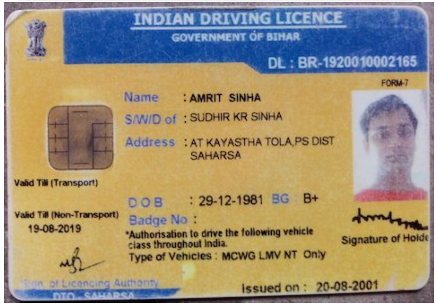
- 1:1 Face Recognition between selfies and photos of documents

- Part of Onfido’s remote identity verification solution
- The document proves your identity
- The selfie proves the document belongs to you

Challenges of Selfie-Doc Face Recognition

- User-controlled image capture: wide range of devices, light conditions
- Document images are photos of physical documents, *not* high-res images stored on chip
- Bi-sample data: only 2 images per identity
- Large number of document types

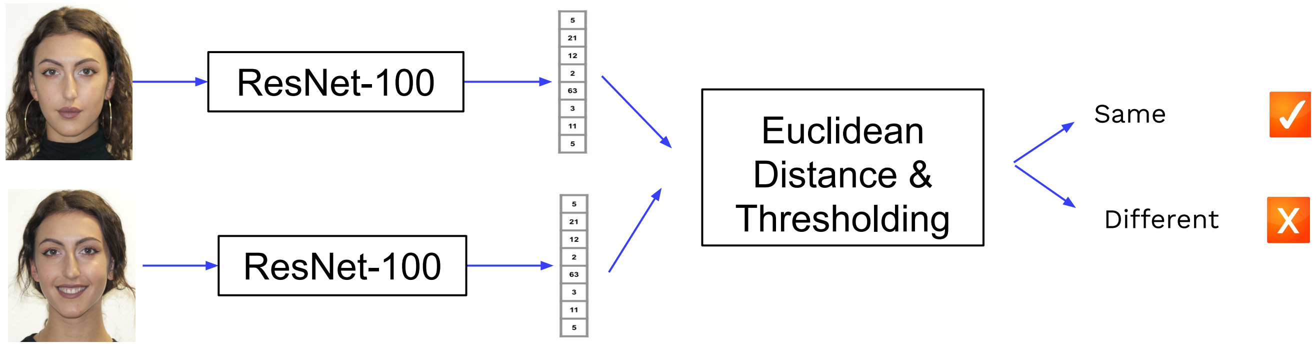




Previous Work

- Selfie-Doc matching
  - Chinese resident cards, using chip photo (Shi and Jain '18, '19, Zhu et al. '19)
  - Chilean ID cards (Albiero et al. '19)
- Geographic and Racial performance differentials
  - Race-based evaluation (Krishnapriya et al. '19, Cavazos et al. '19)
  - NIST FRVT Report Part 3 (Grother et al. '19)
- Mitigation strategies
  - Racial Faces in the Wild (Wang et al. '19)
- Bi-sample or shallow face learning
  - Semi-siamese networks (Du et al. '20)

Contribution



- Face Recognition model trained on selfie-doc data.
- Evaluation of performance differentials across geographies.
- Evaluate sampling methods to reduce performance differentials.
- Speculation about nature of bias

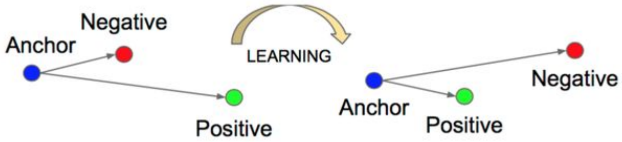
Selfie-Doc Dataset

- In-house dataset of 6.8M image pairs
- Available metadata
  - *Document issuing country*
  - Gender
- Test set of 100K image pairs.

	Male	Female	Unknown	All
Europe (EU)	29.0%	16.5%	15.5%	61.0%
America (AM)	9.2%	5.6%	0.3%	15.1%
Africa (AF)	0.3%	0.1%	0.1%	0.5%
Asia (AS)	2.4%	0.7%	1.6%	4.7%
Oceania (OC)	0.1%	0.1%	0.2%	0.3%
Unknown (UN)	0.0%	0.0%	18.3%	18.3%
All	41.0%	23.0%	36.1%	100.0%

Loss Function and Training

- Image *x*, feature embedding *z=f(x)*
- Training with triplet loss



$$\mathcal{L} = \max\left(D_{ap}^2 - D_{an}^2 + \alpha, 0\right)$$

where (x_a, x_p, x_n) are triplets consisting on an *anchor* a *positive* and a *negative* image

$$D_{ap}^2 = \|f(x_a) - f(x_p)\|^2$$

- Online semi-hard triplet selection: for each pair x_a, x_p consider candidates x_c that violate the margin

$$\|f(x_a) - f(x_p)\|^2 + \alpha > \|f(x_a) - f(x_c)\|^2$$

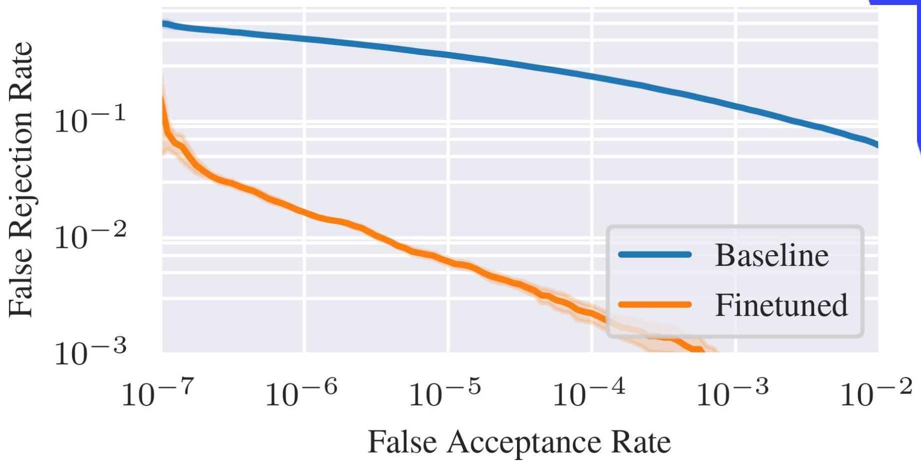
Algorithm 1: Training loop

Input : Batch of selfie-doc pairs (X^s, X^d)
$X^s = [x_1^s, \ldots, x_N^s]$
$X^d = [x_1^d, \ldots, x_N^d]$
Output: Updated network $f(\cdot)$

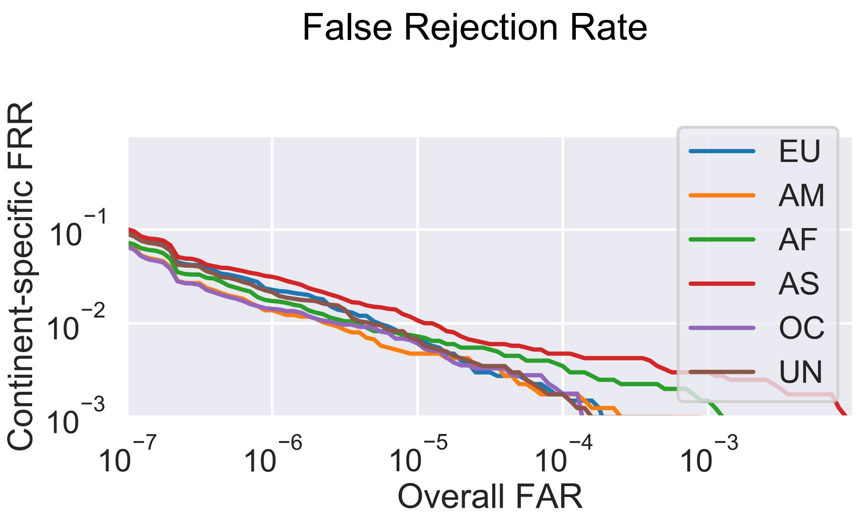
Compute embeddings for the whole batch
for $i = 1 \ldots N$ **do**
  $z_i^s,\ z_i^d = f(x_i^s),\ f(x_i^d)$
end
Use the embeddings for triplet selection
for $i = 1 \ldots N$ **do**
  select $j(i)$ s.t. $(x_i^s, x_i^d, x_{j(i)}^d)$ is a hard triplet
  select $k(i)$ s.t. $(x_i^d, x_i^s, x_{k(i)}^s)$ is a hard triplet
end
Train with triplets in minibatches of size N_{train}
for $i = 1 \ldots N$ **do**
  update network weights using triplets $(x_i^s, x_i^d, x_{j(i)}^d)$ and $(x_i^d, x_i^s, x_{k(i)}^s)$
end

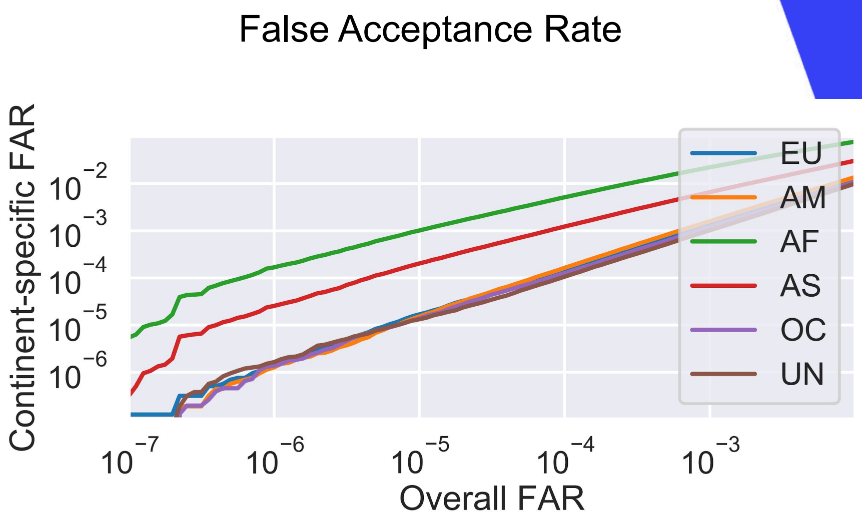
Baseline Model Performance

- **Baseline**
  - ResNet-100 model trained on MS-Celeb-1M.
  - Performance: 99.77% on LFW, 98.47% on MegaFace.
- **Fine-tuning**
  - Triplet selection batch size 10,240.
  - Optimization batch size 32.
  - Learning rate 1e-5, decaying to 1e-7.
  - Trained for 2.7M steps.

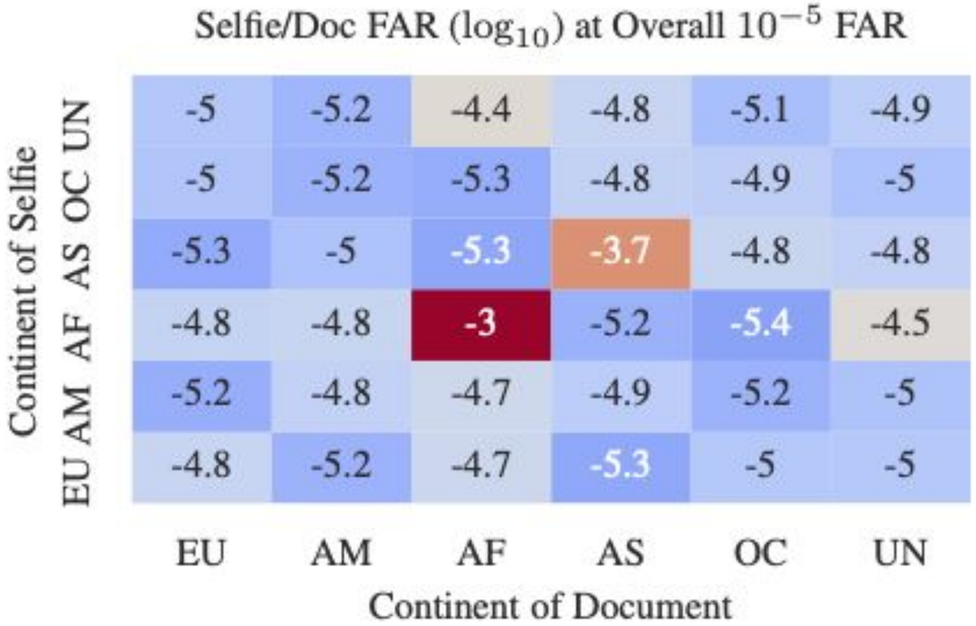


Fine-tuned Model Performance by Continent





Fine-tuned Model Performance by Continent



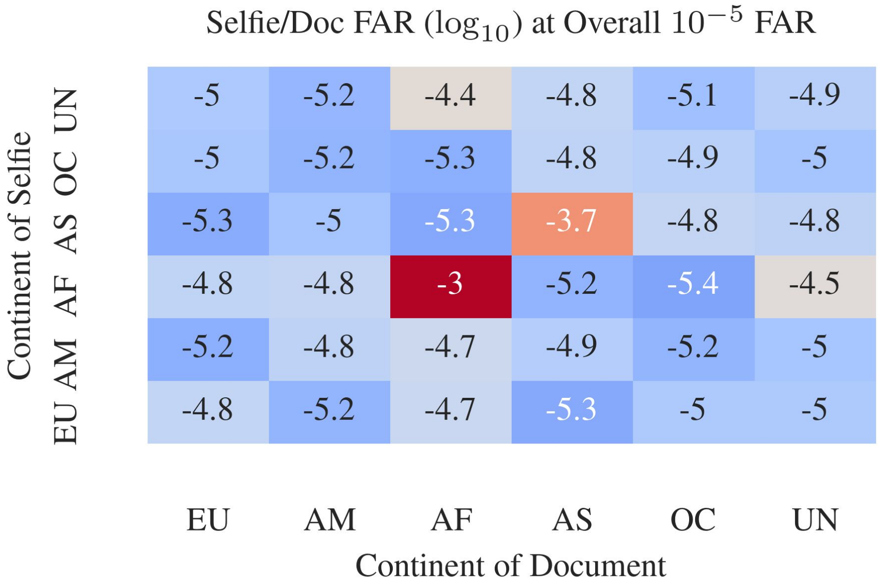
Mitigation Strategies

- **Dataset sampling**
  - *Equal Sampling* - Sampling equally from each continent
  - *Adjusted Sampling* - Weighted sampling as follows
    - EU, AM, OC an UN have weight 1
    - AF, AS have weight 3
  - *Dynamic Sampling* - Weighted sampling with weights dynamically adjusted during training based on within-class FAR.
    - 10-fold increase in FAR yields 4-fold increase in weights
    - Exponential averaging to avoid too sudden weight changes
- *Note:* We do not change the size of the dataset, only the frequency with which a sample from each continent is chosen.

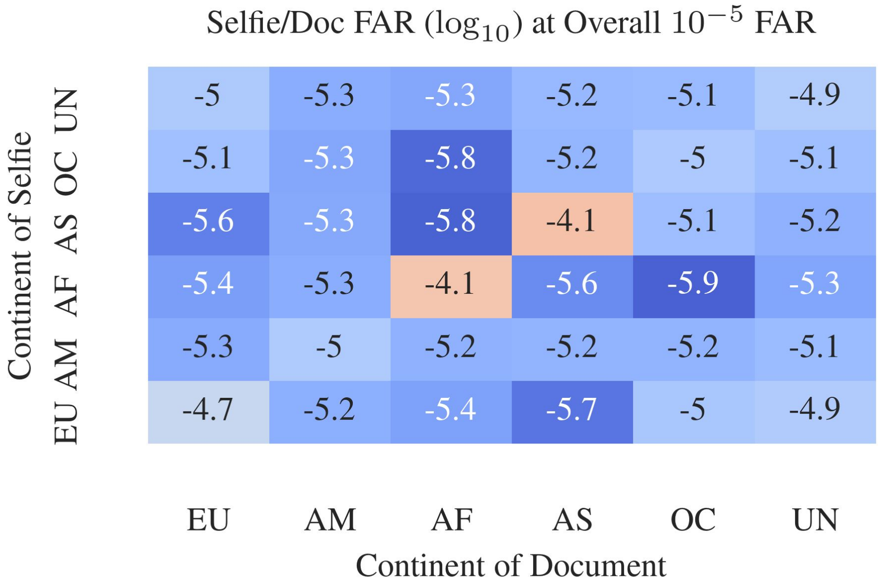
Mitigation Strategies

- **Training**
    - Training initialized with *fine-tuned model* weights.
    - Triplet loss with batch size 10,240 for triplet selection.
    - Optimization batch size 32, learning rate 1e-6, decaying to 1e-7.
    - Trained for 256,000 steps.

Fine-tuned Model and Equal Sampling

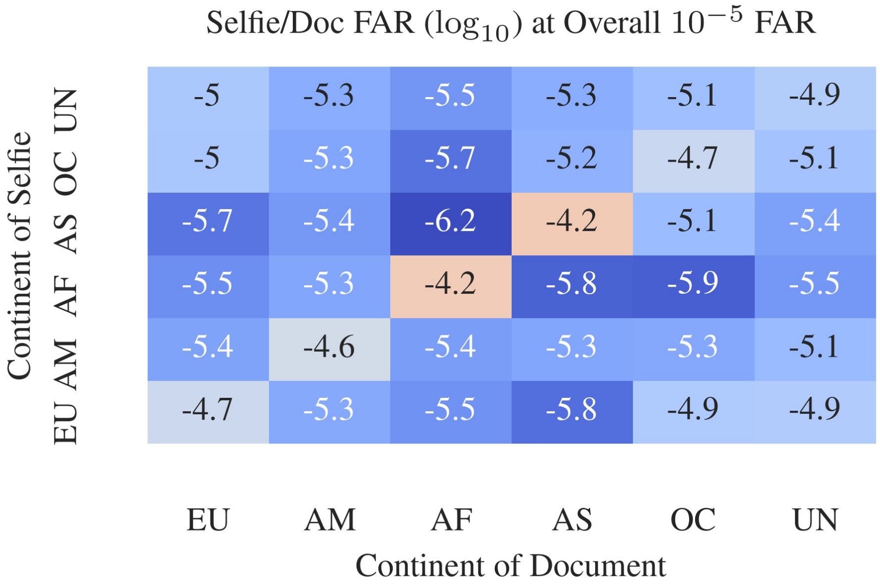


Fine-tuned Model

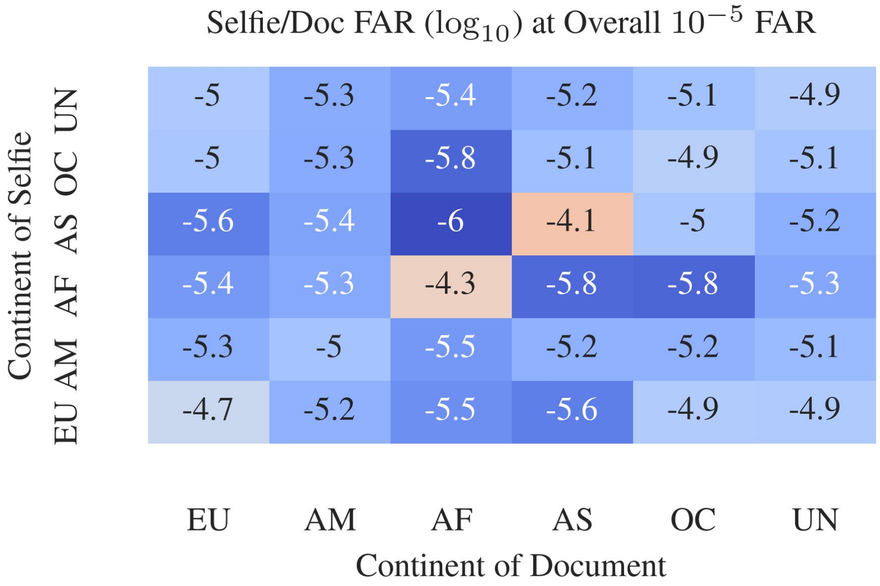


Equal Sampling

Adjusted and Dynamic Sampling



Adjusted Sampling



Dynamic Sampling

What Didn’t Work - Homogeneous Batches

- Why does adjusted sampling help?
- Having more similar samples in a batch increases chance of selecting a hard triplet.
- If more similar samples help, why not use batches that contain samples from one continent only?
- *Homogeneous Batch Sampling*
  - Each batch of 10,240 samples is chosen from a single continent
  - All continents are sampled equally

Selfie/Doc FAR ($\log_{10}$) at Overall 10^{-5} FAR

	EU	AM	AF	AS	OC	UN
Checkpoint 13	-4.8	-5.1	-3.6	-4	-4.9	-4.9
Checkpoint 14	-5	-5.5	-2	-6.1	-6.4	-4.7
Checkpoint 15	-4.8	-5	-2.5	-3.9	-5	-4.8
Checkpoint 16	-4.7	-5	-3.6	-4	-4.9	-4.9
Checkpoint 17	-5	-5.4	-2	-5.1	-6.1	-4.7

Continent of Selfie/Document

Sampling Methods Comparison

Selfie/Doc FAR ($\log_{10}$) at Overall 10^{-5} FAR

	EU	AM	AF	AS	OC	UN
Baseline	-5	-4.7	-3.9	-3.8	-4.9	-5
Finetuned	-4.8	-4.8	-3	-3.7	-4.9	-4.9
Equal Sampling	-4.7	-5	-4.1	-4.1	-5	-4.9
Adj. Sampling	-4.7	-4.6	-4.2	-4.2	-4.7	-4.9
Dyn. Sampling	-4.7	-5	-4.3	-4.1	-4.9	-4.9

Continent of Selfie/Document

False Acceptance Rate

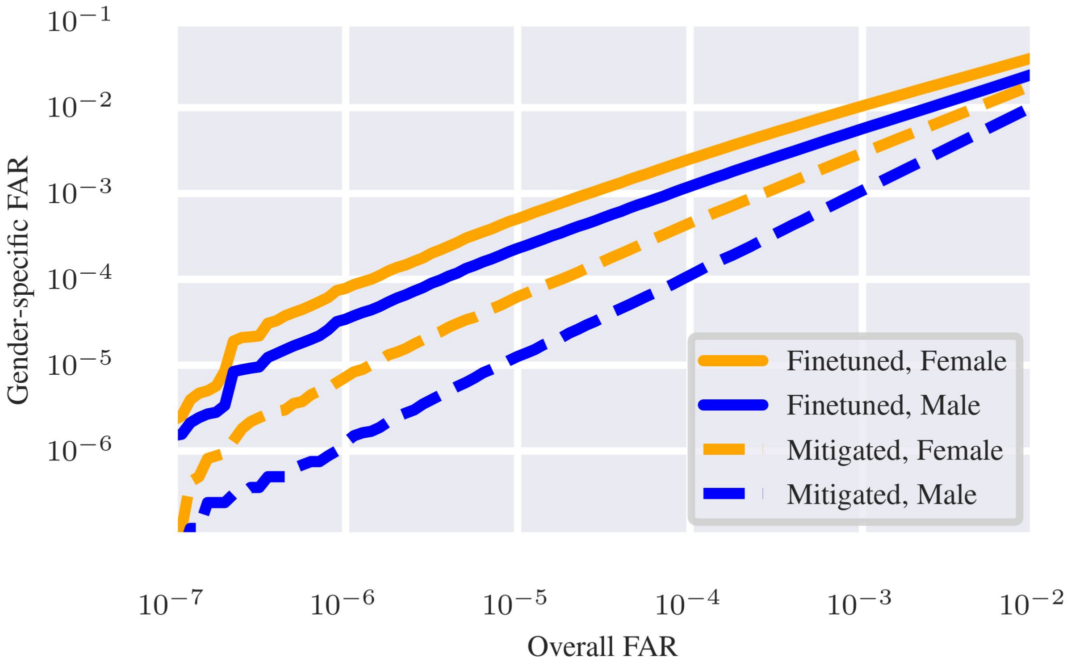
Selfie/Doc FRR at Overall 10^{-5} FAR

	EU	AM	AF	AS	OC	UN
Finetuned	0.68%	0.48%	0.72%	1.1%	0.61%	0.68%
Equal Sampling	0.75%	0.55%	1.5%	1.7%	0.69%	0.92%
Adj. Sampling	1.2%	0.82%	2.5%	2.3%	0.9%	1.4%
Dyn. Sampling	0.8%	0.68%	1.8%	1.9%	0.76%	0.95%

Continent of Selfie/Document

False Rejection Rate

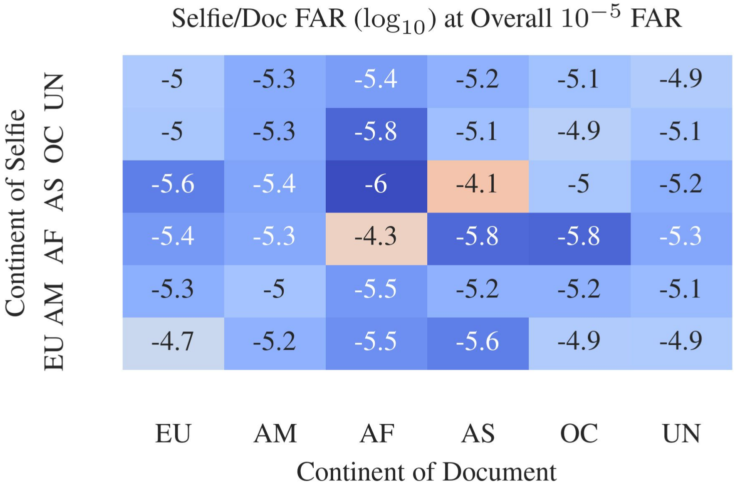
Open Questions

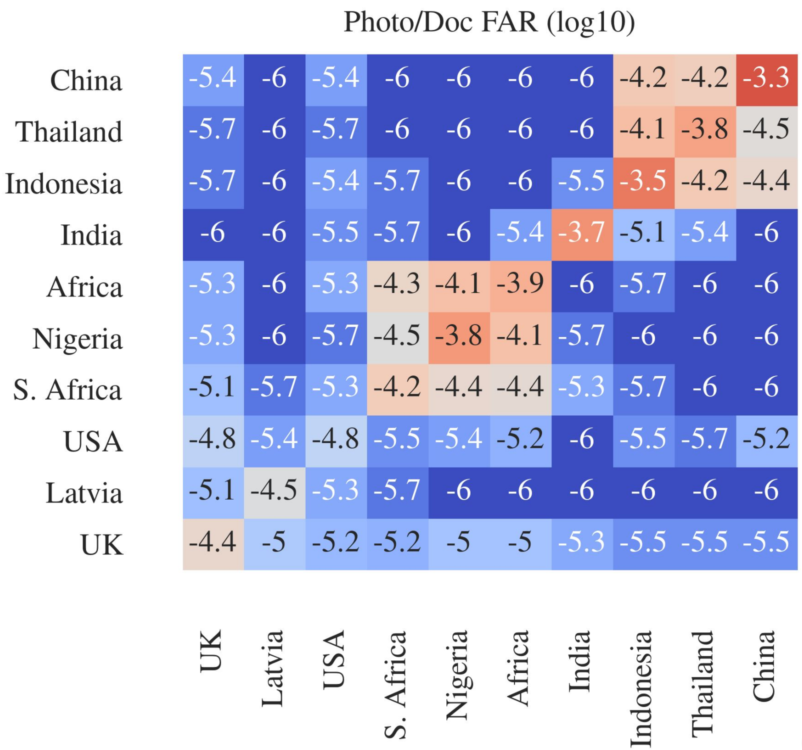


- Continent-based mitigation improves male and female performance
- Male performance improves more than female
- The gender-differential is larger for the mitigated model

Continents or Countries?

- Evaluate Dynamic Sampling model by country





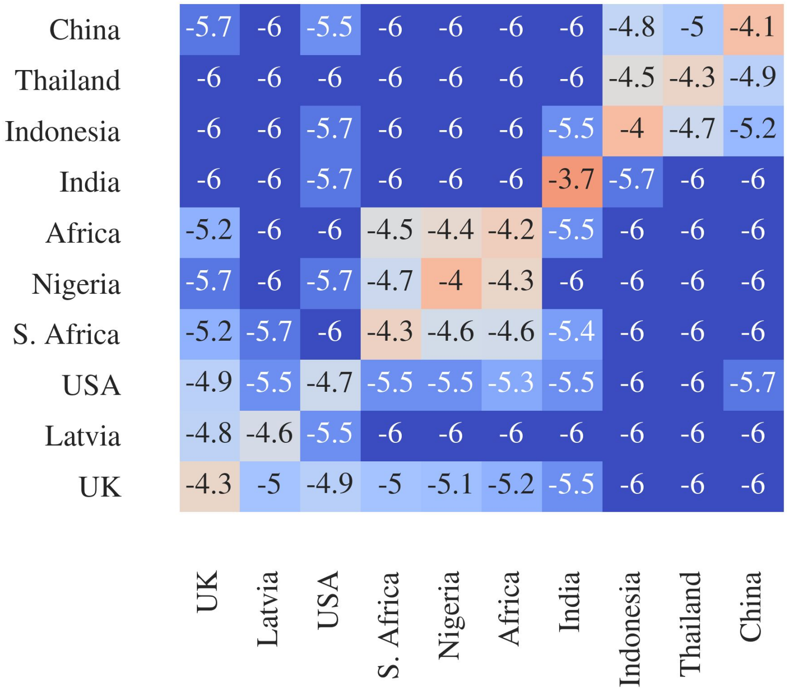
Country-based Sampling Strategies

- **Dataset sampling**
  - *Adjusted Sampling* - Weighted sampling as follows
    - Countries from Africa, Asia and America (except USA and Canada) have weight 4
    - All other countries have weight 1
  - *Dynamic Sampling* - Weighted sampling with weights dynamically adjusted during training based on within-class FAR.
- **Training**
  - Training initialized with *finetuned model* weights.
  - Triplet loss with batch size 10,240 for triplet selection.
  - Optimization batch size 32, learning rate 1e-6, decaying to 1e-7.
  - Trained for 256,000 steps.

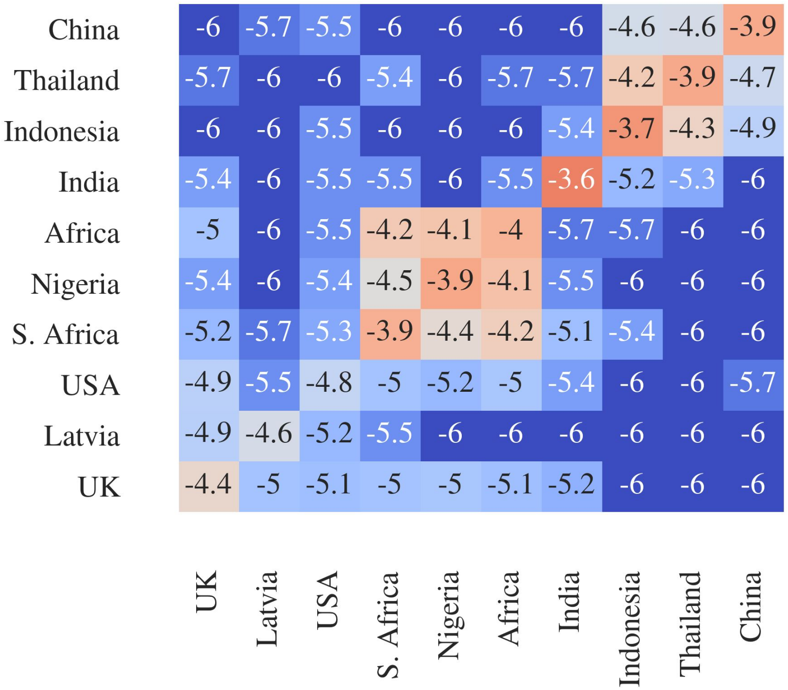
onfido

Adjusted and Dynamic Sampling

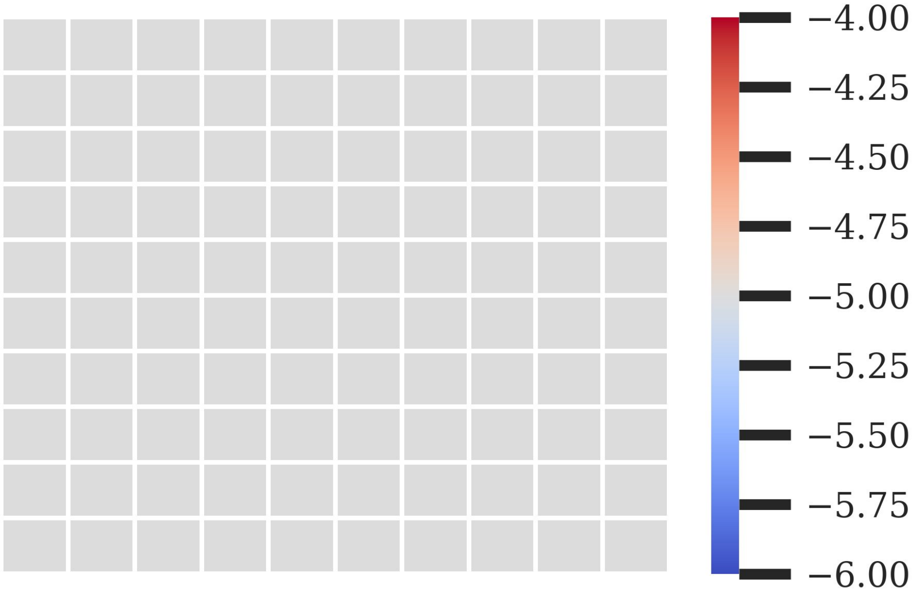
Photo/Doc FAR (log10)



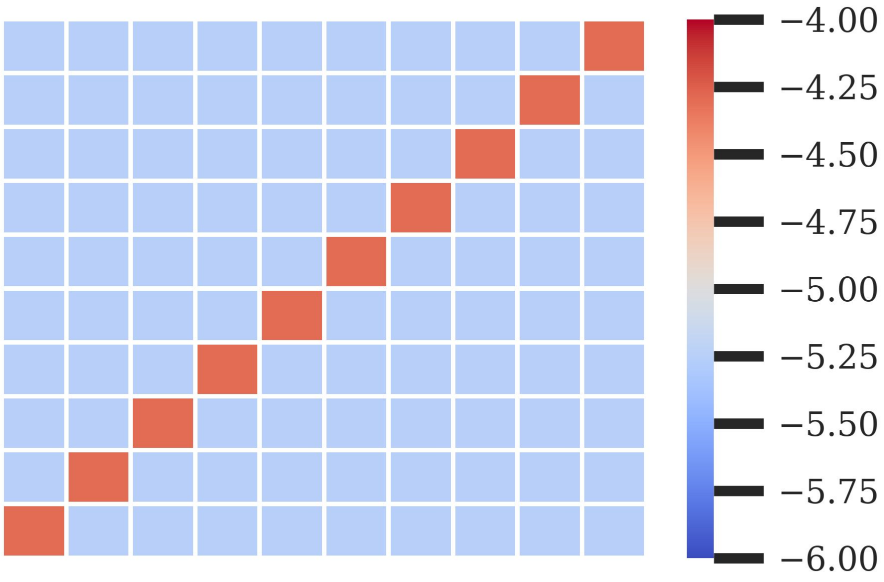
Photo/Doc FAR (log10)



The Ideal Scenario?

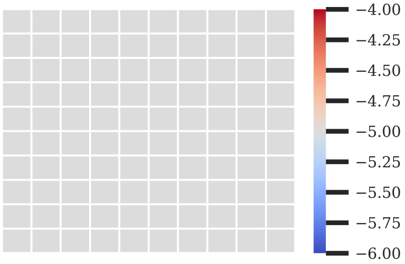


Uniform FAR across groups



Uniform FAR within groups

Thought Experiment



- Consider a perfectly unbiased model with
  - FAR = 10^{-5}
  - FRR = 10^{-2}
- Assume that we have a gender classifier with
  - Accuracy = 0.999
  - Error rate, $\varepsilon = 10^{-3}$
- Combine this into a *new model* as follows
  - Given two images, we determine the gender via classifier
  - If the genders are equal, we use original model for similarity
  - If the genders are different, the images don’t match
- What is the performance of the new model?

Thought Experiment

FAR	Male	Female
Male	10^{-5}	10^{-5}
Female	10^{-5}	10^{-5}
FRR	0.01	0.01

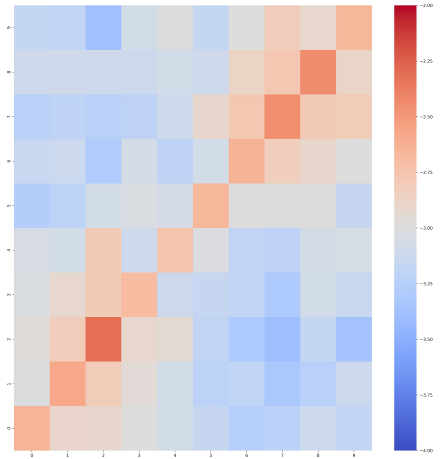
Original model

FAR	Male	Female
Male	$5 \cdot 10^{-6}$	$2 \cdot 10^{-8}$
Female	$2 \cdot 10^{-8}$	$5 \cdot 10^{-6}$
FRR	0.012	0.012

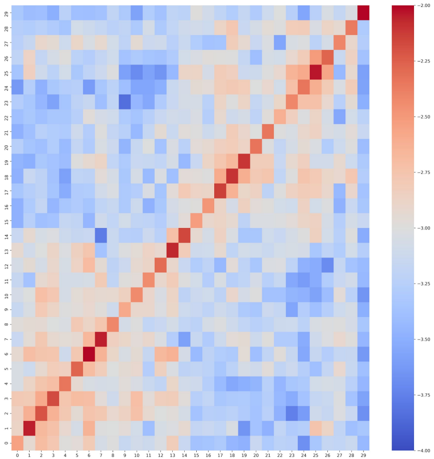
Model with gender classifier

- New model overall performance
  - FAR = $5 \cdot 10^{-6}$
  - FRR = $1.2 \cdot 10^{-2}$

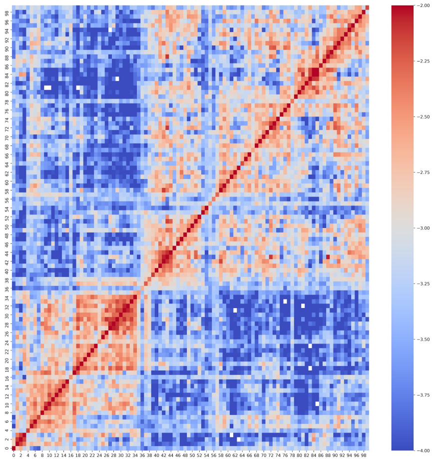
Algorithmic Grouping via Clustering



10



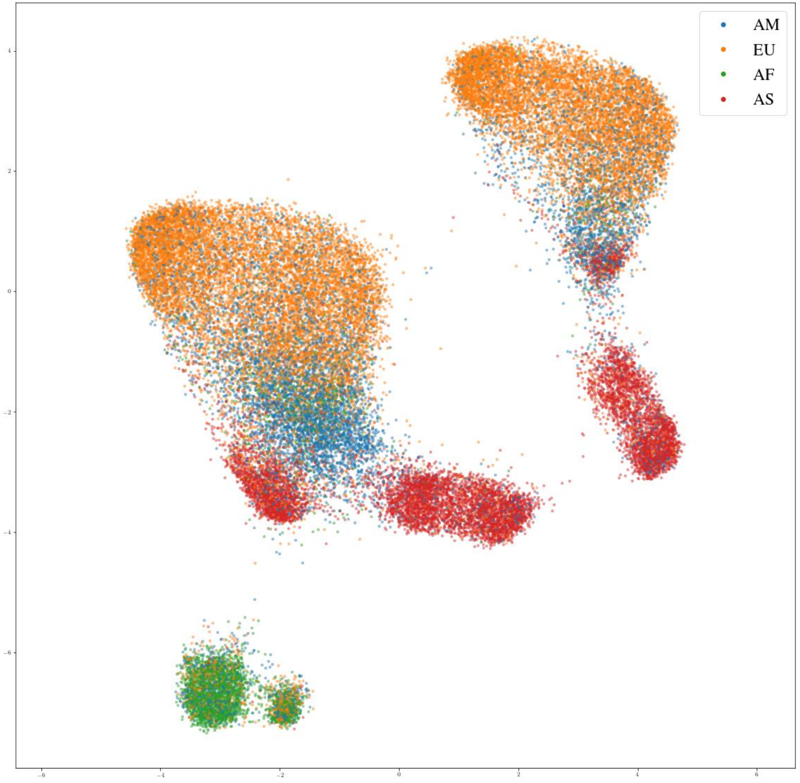
30



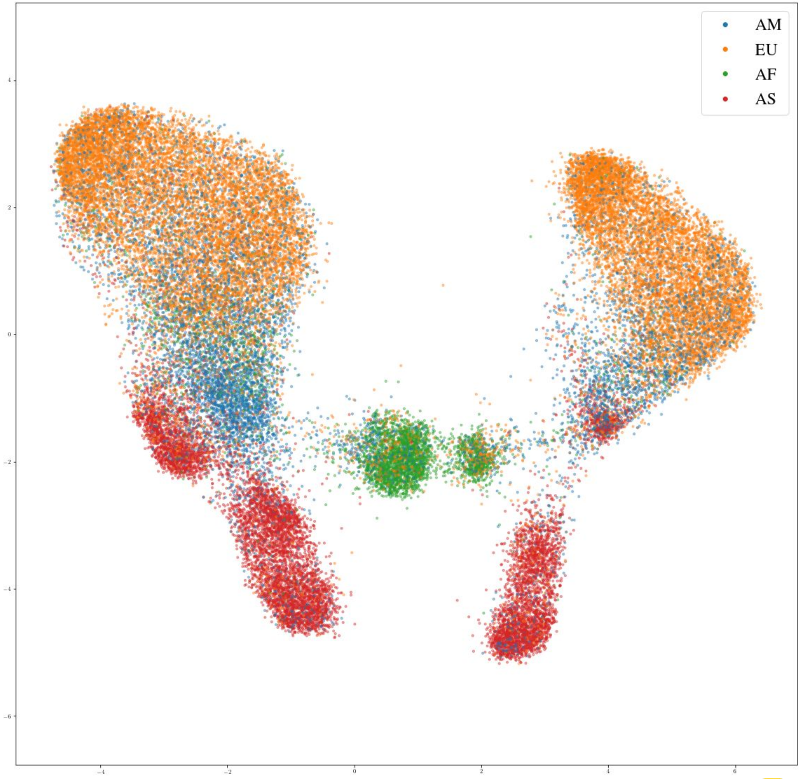
100

- Cluster a dataset of 1M face embeddings into 10, 30 or 100 clusters
- Compute the FAR between clusters at a fixed threshold
- Blue … lower FAR; Red … higher FAR

Visualization of Embedding Space

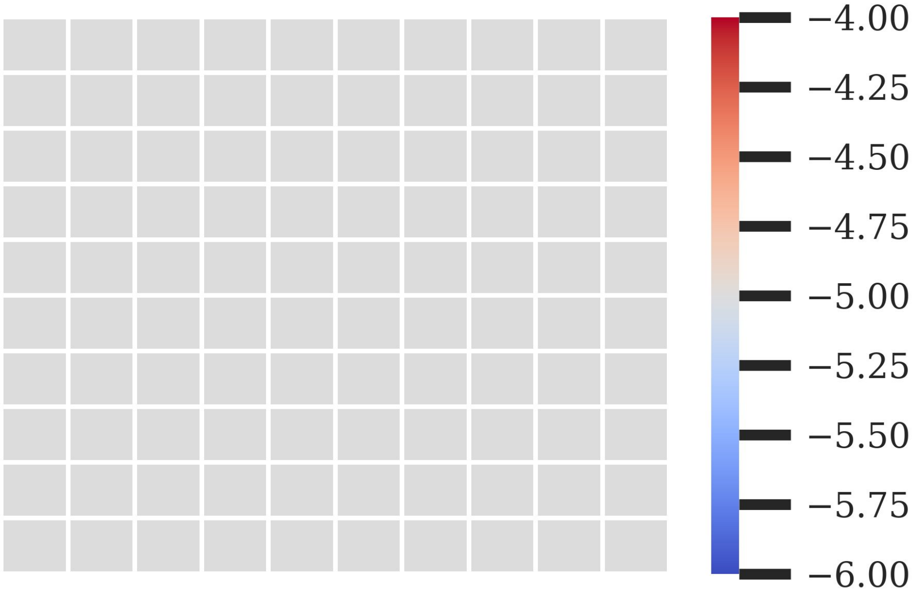


Baseline

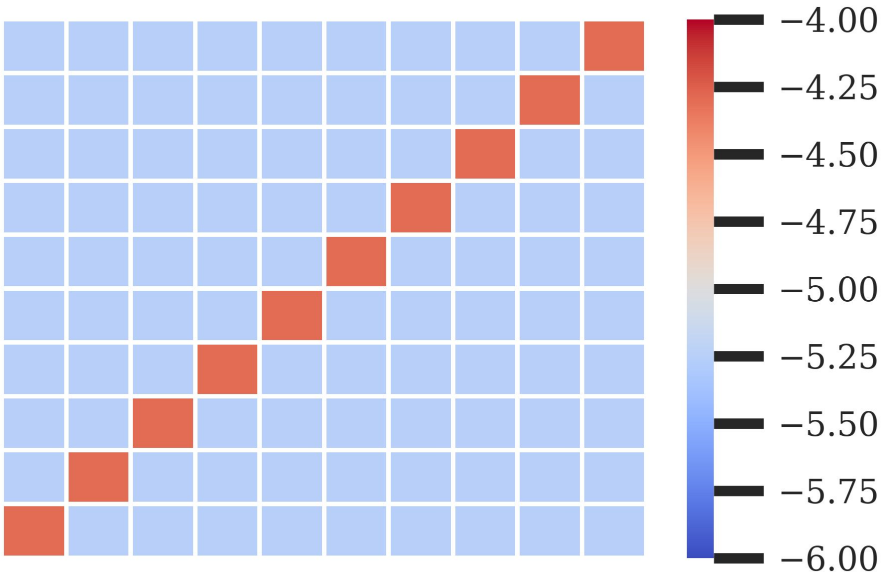


Continent-based dynamic sampling

The Ideal Scenario?



Uniform FAR across groups



Uniform FAR within groups

Discussion Points

- Performance differentials can be reduced without balanced data
  - Only 0.5% of images are from African documents
- Having fine-grained labels for the training set is an advantage
  - Future work to explore unsupervised clustering methods
- Dynamic sampling strategies require a clean validation set
  - Noise in the validation set will amplify errors in sampling weights
- Removing performance differentials is a multi-objective optimization problem.
  - Reducing FAR differential can lead to increased FRR differentials.
- What is the end-state of bias reduction?