SAPIENZA
UNIVERSITÀ DI ROMA

Protecting Users from Face Recognition using Adversarial Machine Learning

Zeeshan Hussain Khand

Prof. Iacopo Masi

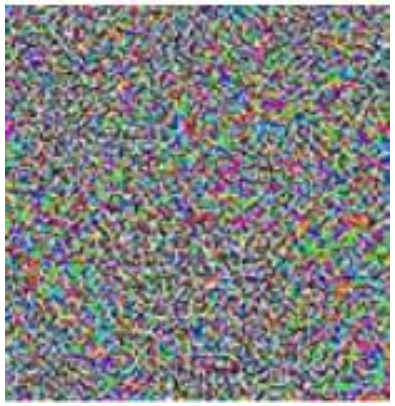
Supervisor

Adversarial Attacks against Deep Learning Models

Original Input image

$+ \ 0.003 \ x$

Adversarial perturbation



$=$

Adversarial Image

Adam Brody

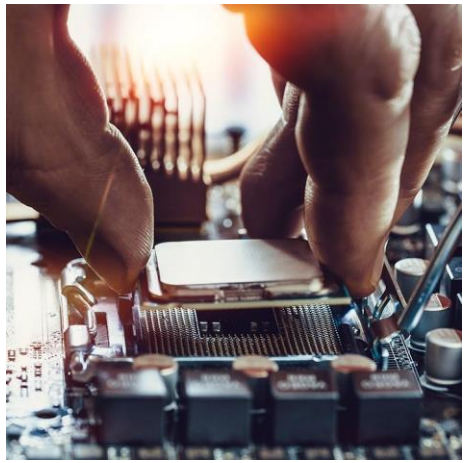
Penn Badgley

No difference in visual perception

Goal : Design an Adversarial Attack that protects user images from unauthorized Face Recognition

Why ?

Takes less time now to train larger and advanced FR systems



Cheaper and Hardware/GPU’s

Labelled Training data anywhere

But what if people with bad intentions take use of this technology in unethical ways ?

Why ?

Clearview.ai

Could lead to

- Racial Discrimination
- Prosecution

Our Work

Phase 1 : Re-design adversarial attack proposed by Lowkey in their paper.

Phase 2 : Evaluate lowkey against blur transformations from Kornia library.

Phase 3 : Design a new improved attack that is robust to blur transformations

Original Image

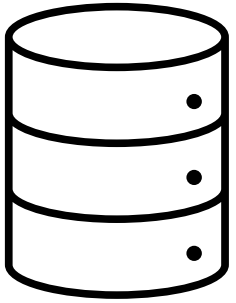
Face processing & alignment

Attack

Protected image

FR database (gallery)

Probe Image



Face Recognition System

Johnny Messner

Goal : Apply adversarial attack on potential gallery images so that they do not match probe images of the same user.

Lowkey Adversarial Attack

Proposed by team of students from University of Maryland in collaboration with United States Naval Academy.

Lowkey applies an adversarial filter to images that impedes their use in face recognition systems.

Shown to significantly reduce accuracy against Amazon Rekognition and Microsoft Azure Face to below 1 %.

Experimental Setup

Model Ensemble : IR-152 and ResNet-152 with ArcFace and CosFace loss functions.

Trained on : MS-Celeb 1M dataset.

Used as : Ensemble to generate attacks and feature extractors for recognition.

Experimental Setup

- Dataset : FaceScrub dataset developed by NUS Singapore.
- Subset : 12000 images belonging to 140 subjects.
- Randomly select 24 identities (12 male ,12 female) and apply lowkey and our attack to their gallery images and insert them back into gallery.
- To evaluate, we take each probe image and compare it against gallery images to find closest match.

Rank k accuracy

We consider model successful in rank k setting if correct identity appears among k closest gallery images in the models feature space. (Top 1 and Top 5)

Similarity Metric

Cosine similarity for matching.

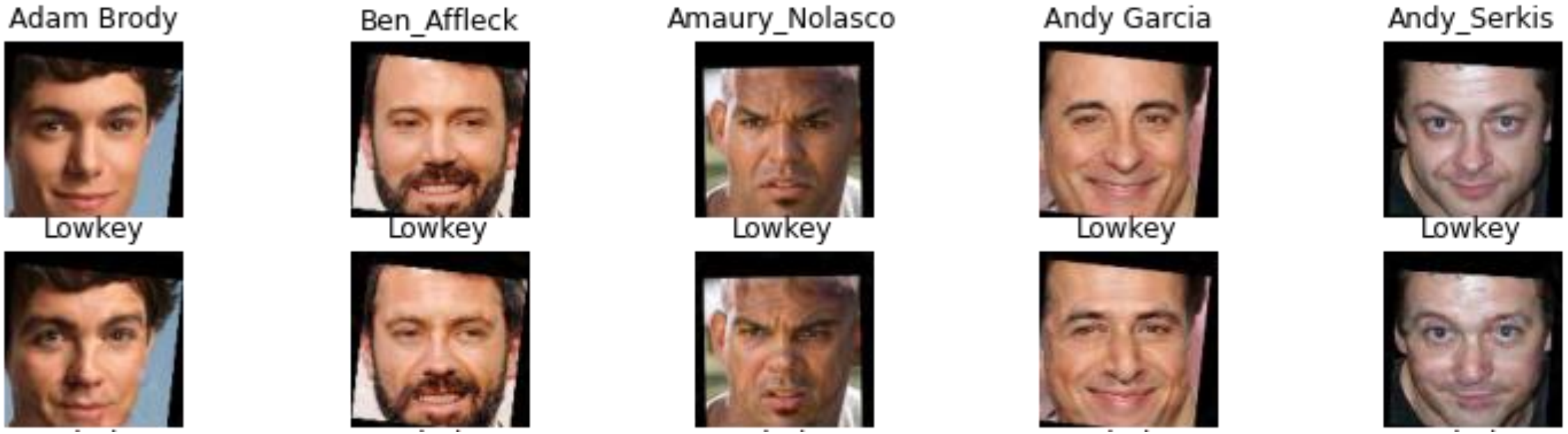
Lowkey Adversarial Attack

$$\max_{x'} \frac{1}{2n} \sum_{i=1}^{n} \frac{\overbrace{\|f_i(A(x)) - f_i(A(x'))\|_2^2}^{\text{non-smoothed}} + \overbrace{\|f_i(A(x)) - f_i(A(G(x')))\|_2^2}^{\text{smoothed}}}{\|f_i(A(x))\|_2} - \alpha \underbrace{\text{LPIPS}(x, x')}_{\text{perceptual loss}},$$

- x is the original image
- x' is the adversarial image.
- A(x) is the face alignment and extraction pipeline
- G(x) is the gaussian smoothing term
- LPIPS is used for perceptual similarity loss

Evaluation Results Top 1 Accuracy				
Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
clean	99.36	99.36	99.57	99.47
Lowkey	86.97	87.07	87.29	87.18

Evaluation Results Top 5 Accuracy				
Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
Clean	99.47	99.36	99.57	99.47
Lowkey	87.18	87.07	87.29	87.18

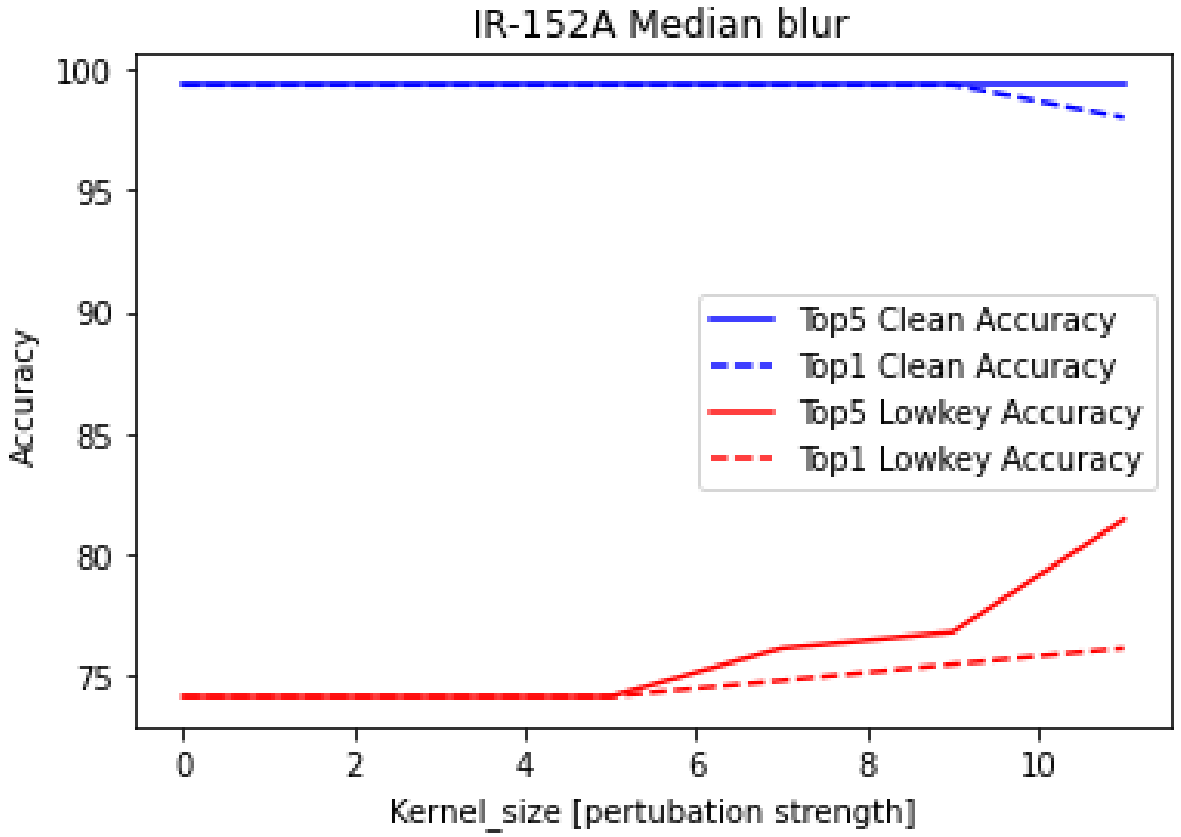


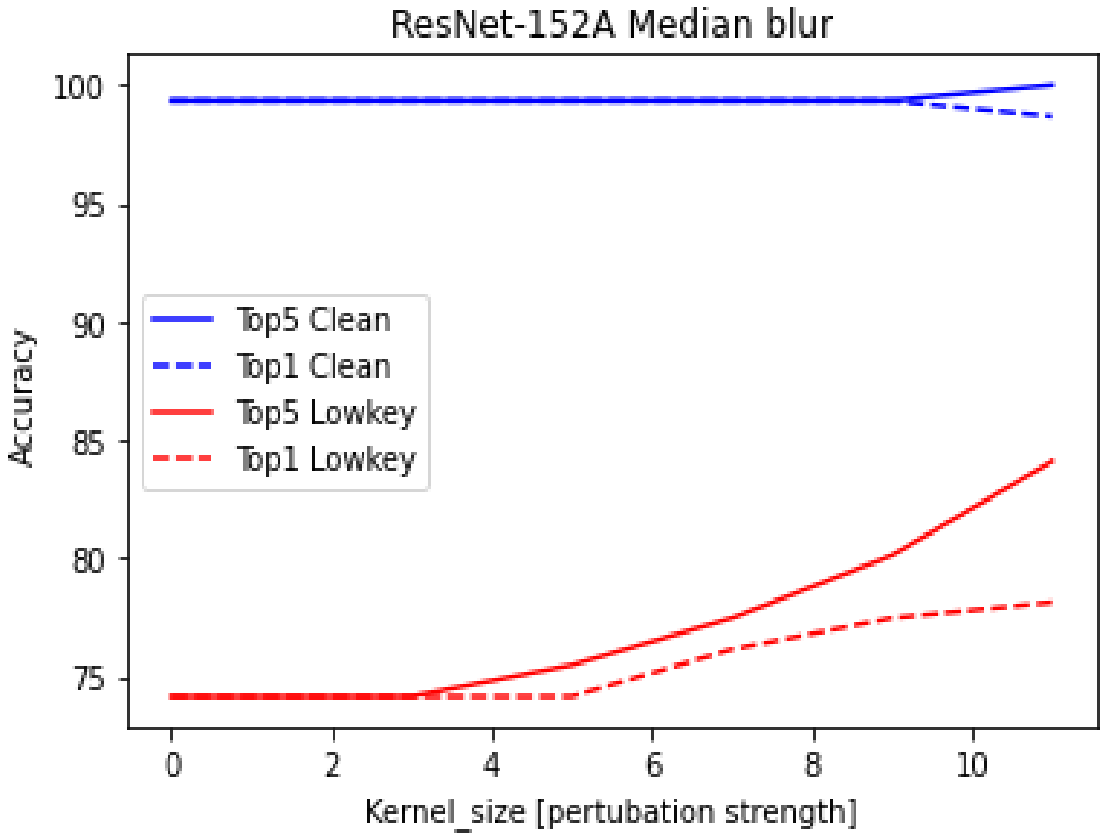
Phase 2 : Evaluating Robustness of Lowkey with Kornia Blur Transformations

- Median Blur
- Box Blur
- Gaussian Blur
- Motion Blur
- Max Blur
- Pool Blur

1.Evaluating Robustness against Median-blur

Applying Median blur on Lowkey protected images



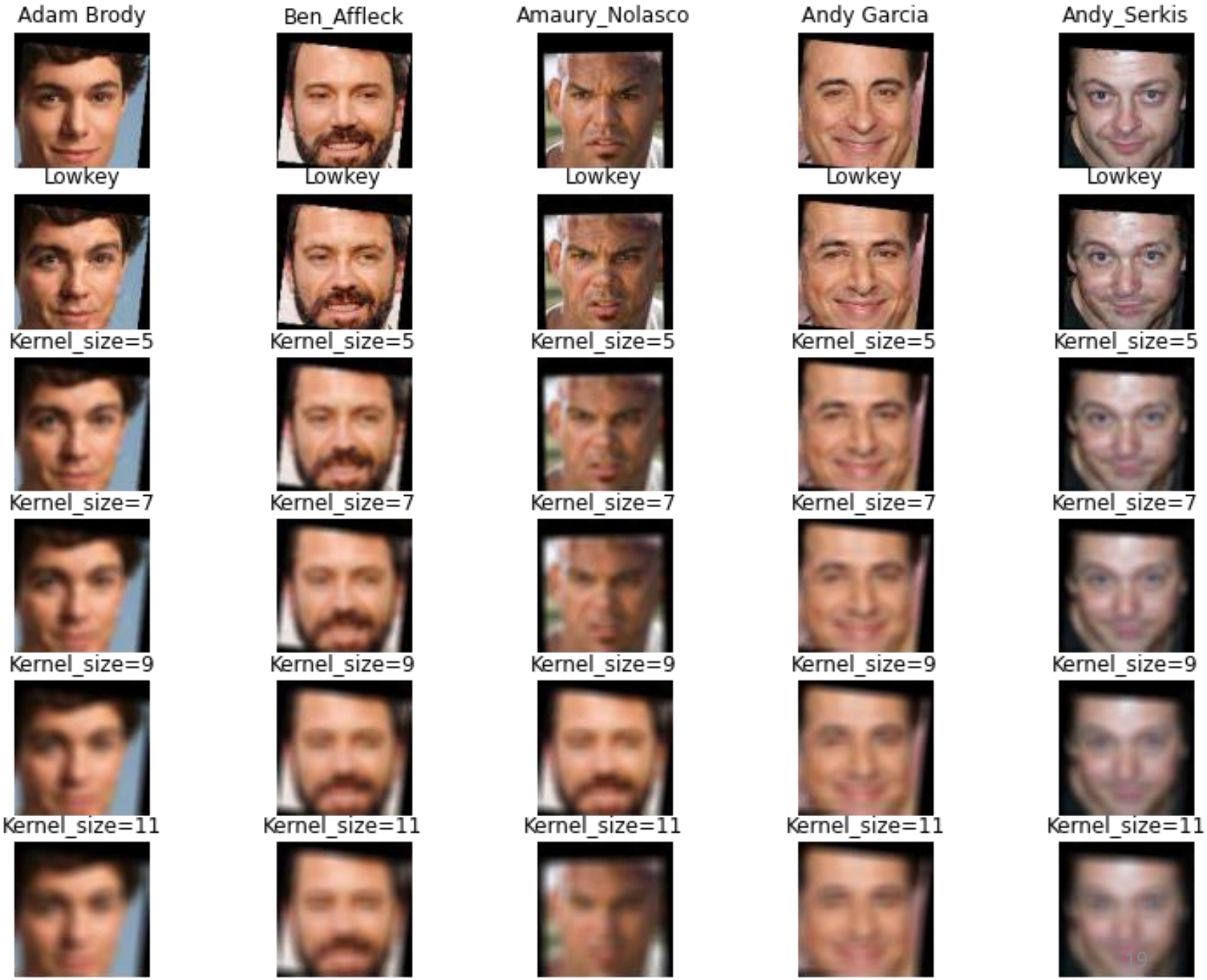


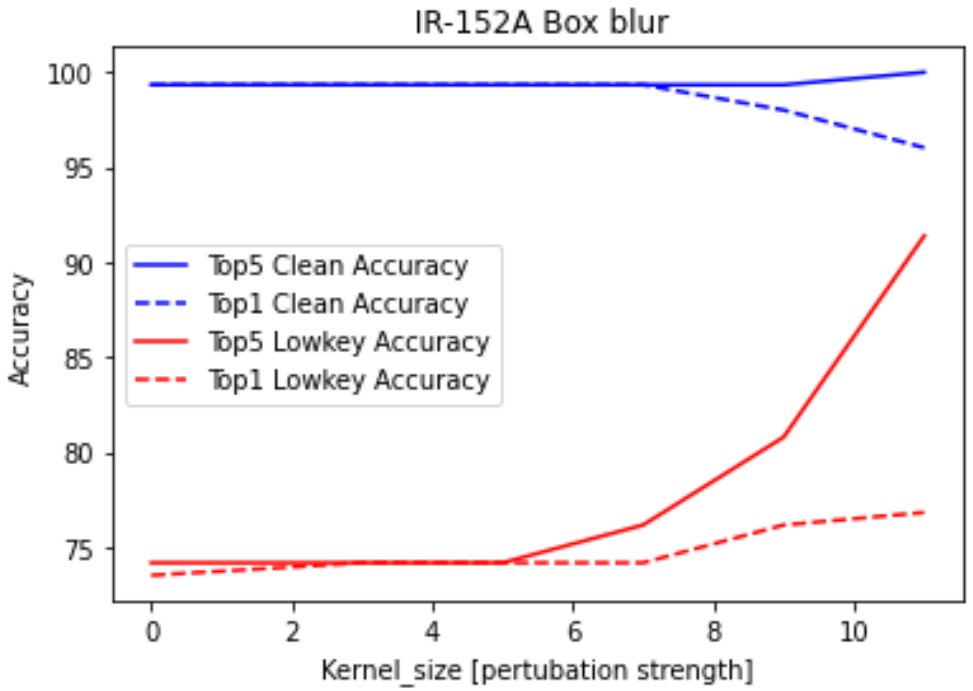
For lowkey protected images , accuracy increases as we increase kernel size k.

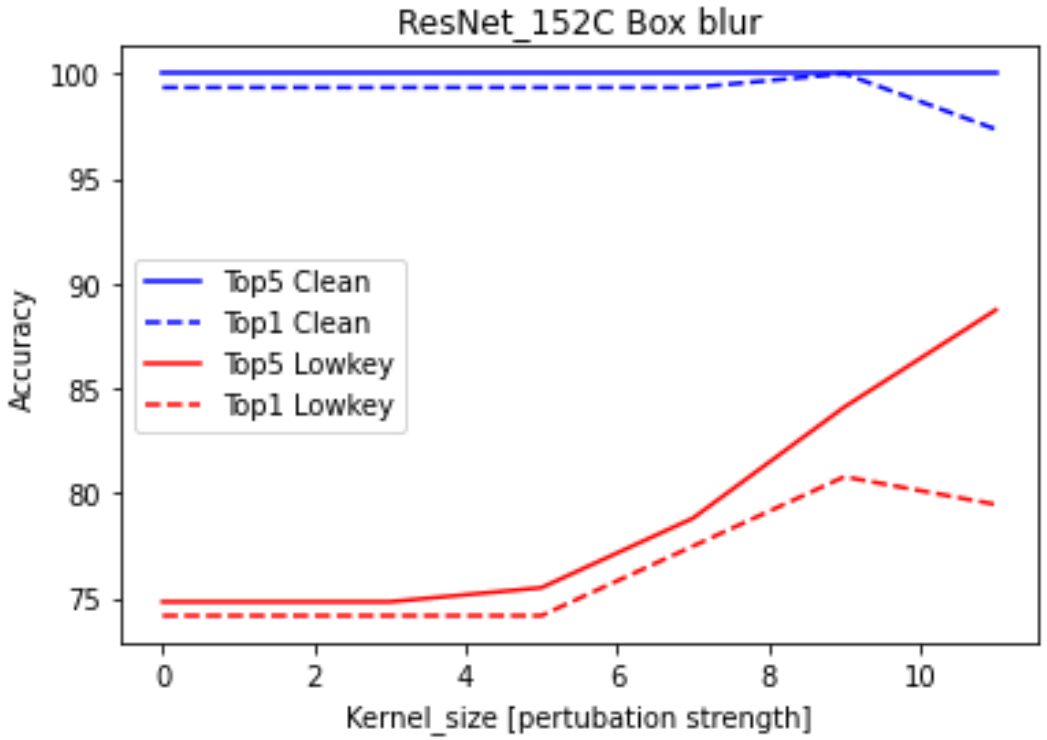
Conclusion : Lowkey is not robust to Median blur with kernel size 9 or above

2. Evaluating Robustness of Lowkey against Box-Blur Transformation

Applying Box blur on Lowkey protected images







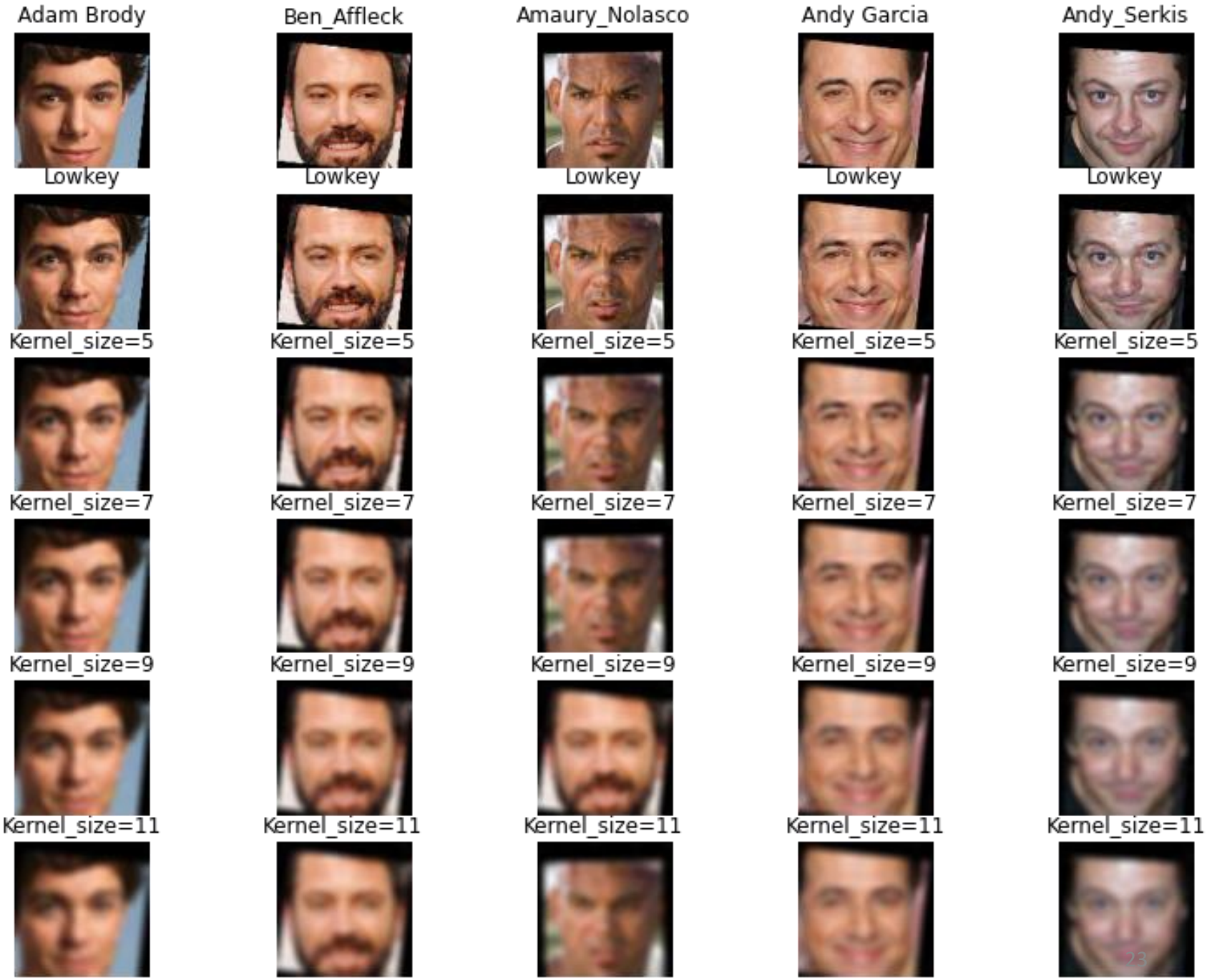
For lowkey protected images , accuracy increases as we increase kernel size k.

Conclusion : Lowkey is also not robust to Box blur with kernel size 9 or above

Evaluating Robustness against Gaussian-blur

Note : Lowkey uses Gaussian-Blur with kernel size 7 and $\sigma = 3$ as part of its pipeline to generate attacks.

Applying Gaussian blur on Lowkey protected images



Evaluation Top 1 and Top 5 for Gaussian-Blur

Top 1

Top-1 Evaluation pipeline Gaussian Blur - 24 probe,24 gallery				
Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
Clean	99.34	99.34	99.34	99.34
Clean + gaussian_blur with k=3	99.34	99.34	99.34	99.34
Clean + gaussian_blur with k=5	99.34	99.34	99.34	99.34
Clean + gaussian_blur with k=7	99.34	99.34	99.34	99.34
Clean + gaussian_blur with k=9	99.34	99.34	99.34	99.34
Clean + gaussian_blur with k=11	99.34	99.34	100.0	100.0
Clean + gaussian_blur with k=17	99.34	99.34	100.0	100.0
Lowkey	73.51	74.17	74.17	74.17
Lowkey + gaussian_blur with k=3	73.51	74.17	73.51	74.17
Lowkey + gaussian_blur with k=5	72.85	73.51	74.17	72.19
Lowkey + gaussian_blur with k=7	72.85	70.86	74.83	72.85
Lowkey + gaussian_blur with k=9	72.19	73.51	76.82	75.5
Lowkey + gaussian_blur with k=11	72.19	74.83	76.82	76.16

Top 5

Top-5 Evaluation pipeline Gaussian Blur - 24 probe,24 gallery				
Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
Clean	99.34	99.34	99.34	100.0
Clean + gaussian_blur with k=3	99.34	99.34	99.34	100.0
Clean + gaussian_blur with k=5	99.34	100.0	99.34	100.0
Clean + gaussian_blur with k=7	99.34	99.34	99.34	99.34
Clean + gaussian_blur with k=9	99.34	99.34	99.34	100.0
Clean + gaussian_blur with k=11	99.34	99.34	100.0	100.0
Clean + gaussian_blur with k=17	99.34	99.34	100.0	100.0
Lowkey	74.17	74.17	74.17	74.83
Lowkey + gaussian_blur with k=3	74.17	74.17	74.83	74.83
Lowkey + gaussian_blur with k=5	73.51	74.17	75.5	74.83
Lowkey + gaussian_blur with k=7	74.17	72.85	77.48	74.83
Lowkey + gaussian_blur with k=9	75.5	74.83	81.46	76.82
Lowkey + gaussian_blur with k=11	75.5	74.83	83.44	78.15

Conclusion : Lowkey is robust to Gaussian Blur. This is because Lowkey uses it as part of its pipeline to generate attacks.

Evaluating against Motion blur, Max Blur and Pool Blur Transformations.

Lowkey is highly robust to these transformations !

Top-5 Evaluation pipeline Max Blur - 24 probe,24 gallery

Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
Clean	99.34	99.34	99.34	100.0
Clean + max_blur with k=3	99.34	99.34	99.34	99.34
Clean + max_blur with k=5	99.34	99.34	99.34	100.0
Clean + max_blur with k=7	99.34	99.34	99.34	100.0
Clean + max_blur with k=9	99.34	99.34	99.34	100.0
Clean + max_blur with k=11	99.34	99.34	99.34	99.34
Clean + max_blur with k=17	99.34	99.34	99.34	100.0
Lowkey	74.17	74.17	74.17	74.83
Lowkey + max_blur with k=3	72.85	74.17	74.17	74.83
Lowkey + max_blur with k=5	74.17	74.17	75.5	74.83
Lowkey + max_blur with k=7	74.83	73.51	76.16	75.5
Lowkey + max_blur with k=9	74.17	70.86	76.16	74.83
Lowkey + max_blur with k=11	74.17	71.52	76.16	72.85

Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
Clean	99.34	99.34	99.34	100.0
Clean + pool_blur with k=3	99.34	99.34	99.34	100.0
Clean + pool_blur with k=5	99.34	99.34	99.34	100.0
Clean + pool_blur with k=7	99.34	99.34	99.34	100.0
Clean + pool_blur with k=9	99.34	99.34	99.34	99.34
Clean + pool_blur with k=11	99.34	99.34	99.34	99.34
Clean + pool_blur with k=17	99.34	99.34	99.34	100.0
Lowkey	74.17	74.17	74.17	74.83
Lowkey + pool_blur with k=3	74.17	74.17	74.83	75.5
Lowkey + pool_blur with k=5	74.17	74.17	76.16	74.83
Lowkey + pool_blur with k=7	74.17	74.17	76.16	75.5
Lowkey + pool_blur with k=9	73.51	74.17	75.5	75.5
Lowkey + pool_blur with k=11	73.51	72.85	76.16	74.83

Top-5 Evaluation pipeline Motion_Blur with intensity=3,5,7,9,11 &17

Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
Clean	99.34	99.34	99.34	100.0
Clean + motion_blur=3	99.34	99.34	99.34	100.0
Clean + motion_blur=5	99.34	99.34	99.34	99.34
Clean + motion_blur=7	99.34	99.34	99.34	99.34
Clean + motion_blur=9	99.34	100.0	99.34	100.0
Clean + motion_blur=11	99.34	100.0	99.34	100.0
Clean + motion_blur=17	99.34	99.34	99.34	100.0
Lowkey	74.17	74.17	74.17	74.83
Lowkey + motion_blur=3	74.17	74.17	74.83	74.83
Lowkey + motion_blur=5	74.17	74.17	74.17	74.83
Lowkey + motion_blur=7	74.17	74.17	74.17	74.83
Lowkey + motion_blur=9	74.17	74.17	74.83	75.5
Lowkey + motion_blur=11	74.17	74.83	74.83	78.81

Conclusions

Lowkey is robust to Motion-blur , Pool-blur and Max-blur Transformations.

Lowkey incorporates Gaussian-Blur as part of its attack pipeline thus it is also robust to it.

However, when we apply Median-blur and Box-blur to lowkey images , the blur acts as remover of noise and perturbation and images get recognized. Hence, lowkey is not robust to them.

Our attack for Protection against Face Recognition

Goal : Design an adversarial attack that is robust to blur transformations based on learning from Lowkey and controlled experiments with Kornia.

Our Attack

$$\max_{x'} \frac{1}{2n} \sum_{i=1}^{n} \frac{\overbrace{\|f_i(A(x)) - f_i(A(x'))\|_2^2}^{\text{non-smoothed}} + \overbrace{\|f_i(A(x)) - f_i(A(G(x')))\|_2^2}^{\text{smoothed}} + \overbrace{\|f_i(A(x)) - f_i(A(M(x')))\|_2^2}^{\text{transferability}}}{\|f_i(A(x))\|_2}$$

$$-\alpha \underbrace{\text{LPIPS}(x, x')}_{\text{perceptual loss}}$$

Results

Here we probe only against protected subjects in gallery

Apply Median Blur to Lowkey and Our Attack

Evaluation Results of Top 1 Accuracy for Median-blur transformation				
Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
clean	99.36	99.36	99.57	99.47
Lowkey	86.97	87.07	87.29	87.18
Our Attack	87.07	87.07	87.39	87.29
Lowkey + median_blur with k = 9	86.54	86.86	85.36	86.43
Our Attack + median_blur with k = 9	86.75	86.43	85.15	86.32
Lowkey + median_blur with k = 11	82.91	82.8	76.28	80.66
Our Attack + median_blur with k = 11	80.88	80.98	75.21	79.27

Evaluation Results of Top 1 Accuracy for Median-blur transformation				
Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
clean	100.0	100.0	100.0	100.0
Lowkey	0.0	0.0	0.0	0.0
Our Attack	4.35	0.0	6.96	1.74
Lowkey + median_blur with k = 9	4.35	8.7	13.04	13.04
Our Attack + median_blur with k = 9	6.09	4.35	7.83	6.96
Lowkey + median_blur with k = 11	19.13	18.26	17.39	25.22
Our Attack + median_blur with k = 11	2.61	0.0	3.48	1.74

Apply Median Blur to Lowkey and Our Attack

Evaluation Results of Top 5 Accuracy for Median Blur Transformation				
Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
Clean	99.47	99.36	99.57	99.47
Lowkey	87.18	87.07	87.29	87.18
Our Attack	87.18	87.07	87.82	87.29
Lowkey + median_blur with k = 9	87.07	87.82	87.39	87.07
Our Attack + median_blur with k = 9	87.29	87.5	86.97	86.75
Lowkey + median_blur with k = 11	87.93	86.65	83.76	85.79
Our Attack + median_blur with k = 11	85.79	85.47	83.01	84.08

Evaluation Results of Top 5 Accuracy for Median Blur Transformation				
Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
Clean	100.0	100.0	100.0	100.0
Lowkey	0.87	0.0	0.87	0.87
Our Attack	10.43	5.22	22.61	6.09
Lowkey + median_blur with k = 9	20.87	16.52	25.22	25.22
Our Attack + median_blur with k = 9	10.43	6.09	14.78	14.78
Lowkey + median_blur with k = 11	40.0	38.26	40.87	43.48
Our Attack + median_blur with k = 11	7.83	5.22	13.91	6.96

Conclusion and Discussion

Generating Adversarial attack on large datasets is computationally expensive and time-consuming task.

No guarantee of complete protection.

A step towards privacy preservation and ethical use of Face Recognition Technology.

Giving Control back to user.

Future Work

01

Improve Run time of the attack

02

Incorporate Box-blur in the attack pipeline

03

Evaluate in white box setting against state of Art FR systems

04

Evaluate against commercial APIs of Microsoft Azure and Amazon Rekognition.

Thank you

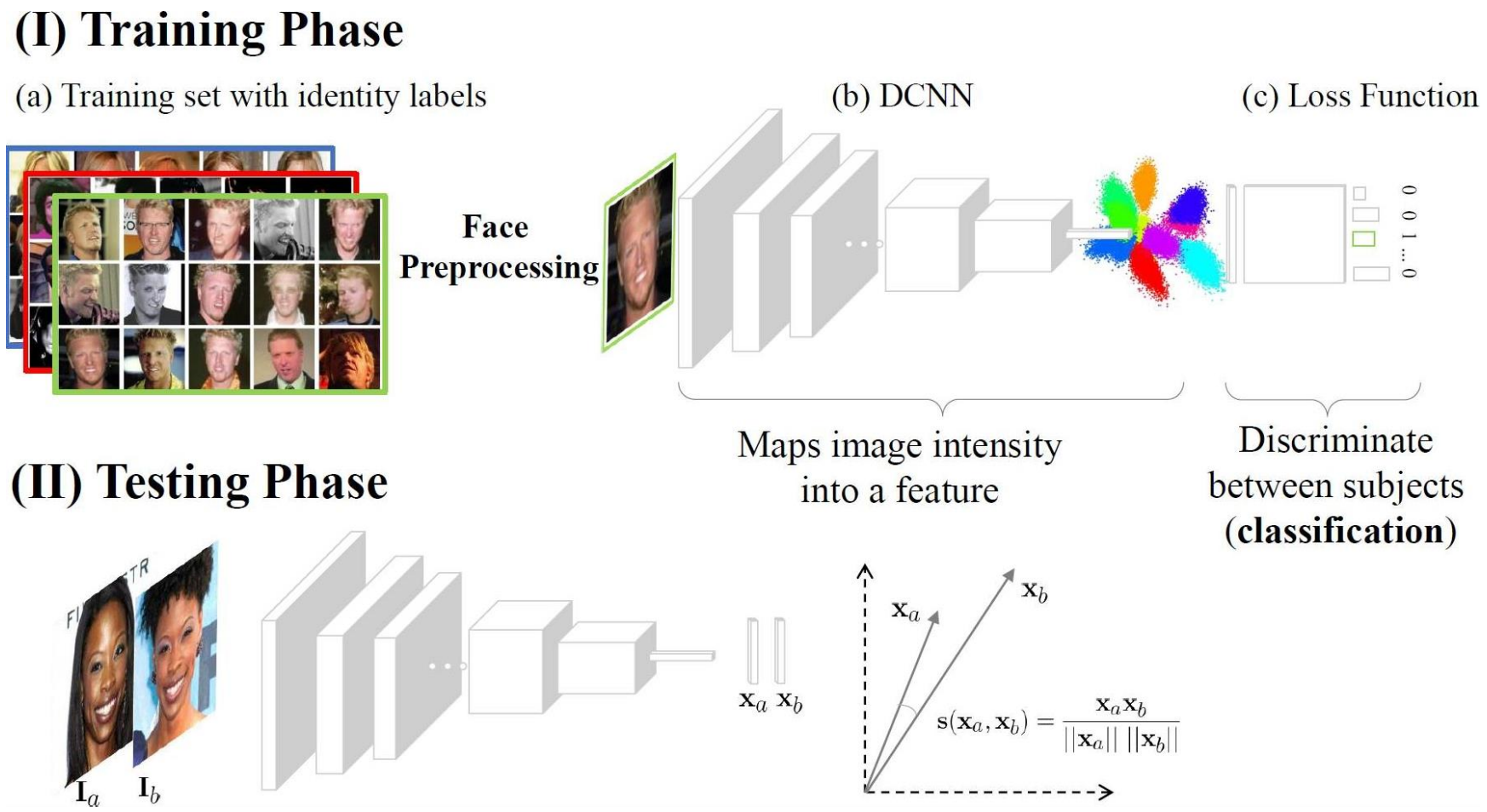
Top 5 rank accuracy for Box Blur

Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
Clean	99.34	99.34	99.34	100.0
Clean + box_blur with k=3	99.34	99.34	99.34	100.0
Clean + box_blur with k=5	99.34	100.0	99.34	100.0
Clean + box_blur with k=7	99.34	99.34	99.34	100.0
Clean + box_blur with k=9	99.34	99.34	100.0	100.0
Clean + box_blur with k=11	100.0	99.34	100.0	100.0
Clean + box_blur with k=17	43.71	70.86	61.59	51.66
Lowkey	74.17	74.17	74.17	74.83
Lowkey + box_blur with k=3	74.17	74.17	75.5	74.83
Lowkey + box_blur with k=5	74.17	74.17	75.5	75.5
Lowkey + box_blur with k=7	76.16	78.15	78.15	78.81
Lowkey + box_blur with k=9	80.79	79.47	86.75	84.11
Lowkey + box_blur with k=11	91.39	84.11	85.43	88.74

Top 5 accuracy for Median-Blur

Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
Clean	99.34	99.34	99.34	100.0
Clean + median_blur=3	99.34	99.34	99.34	100.0
Clean + median_blur=5	99.34	99.34	99.34	99.34
Clean + median_blur=7	99.34	99.34	99.34	99.34
Clean + median_blur=9	96.03	96.03	86.09	83.44
Clean + median_blur=11	99.34	99.34	100.0	99.34
Clean + median_blur=17	77.48	92.05	77.48	75.5
Lowkey	74.17	74.17	74.17	74.83
Lowkey + median_blur=3	74.17	74.17	74.17	74.83
Lowkey + median_blur=5	74.17	74.17	75.5	74.83
Lowkey + median_blur=7	76.16	74.17	77.48	76.82
Lowkey + median_blur=9	76.82	78.81	80.13	78.15
Lowkey + median_blur=11	81.46	82.78	84.11	84.11

How Face Recognition Systems work ?



What is LPIPS ?

- LPIPS (Learned Perceptual Image Patch Similarity) is a metric that quantifies the similarity between two images based on their perceptual features.
- LPIPS uses a pre-trained deep neural network, typically a VGG or AlexNet model, to extract high-level features from image patches. It then computes the distance between these features to obtain a similarity score between the two images.
- Unlike traditional image similarity metrics like SSIM and MSE, LPIPS is designed to be more robust to changes in image content and style, and to better reflect human perception of image similarity.

What is Arc Face and CosFace ?

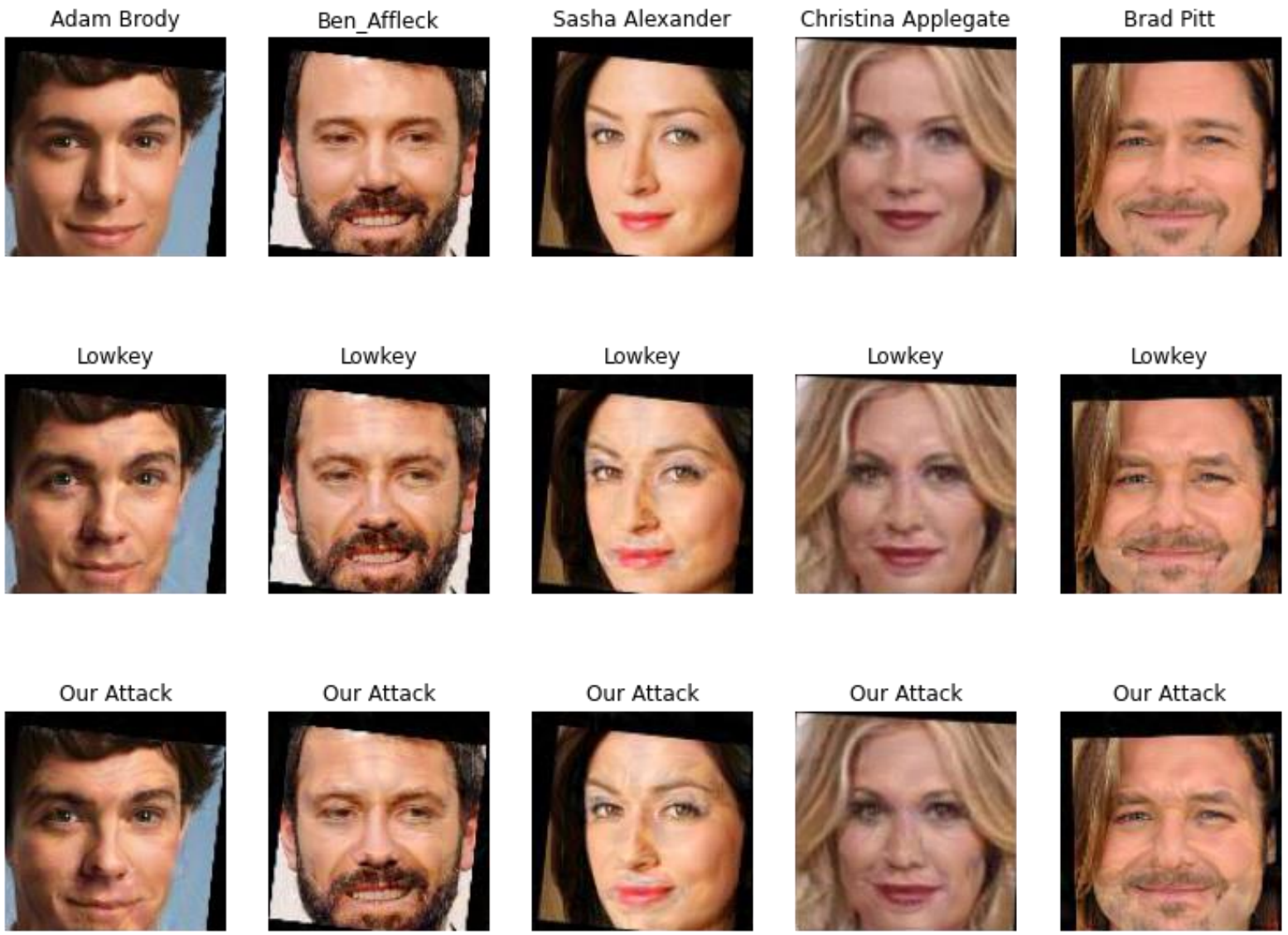
- ArcFace tries to maximize the difference between the correct identity's embedding and the embeddings of other identities. This is done by using a large margin between the embedding of the correct identity and those of the other identities.
- CosFace, on the other hand, applies a scaling factor to the embedding before computing the softmax probabilities. This scaling factor ensures that the cosine similarity between the embeddings is maximized for the correct identity and minimized for other identities.
- In summary, ArcFace and CosFace are two different techniques for training a neural network to recognize faces by minimizing intra-class variations and maximizing inter-class variations in the embedding space.

Evaluation Results

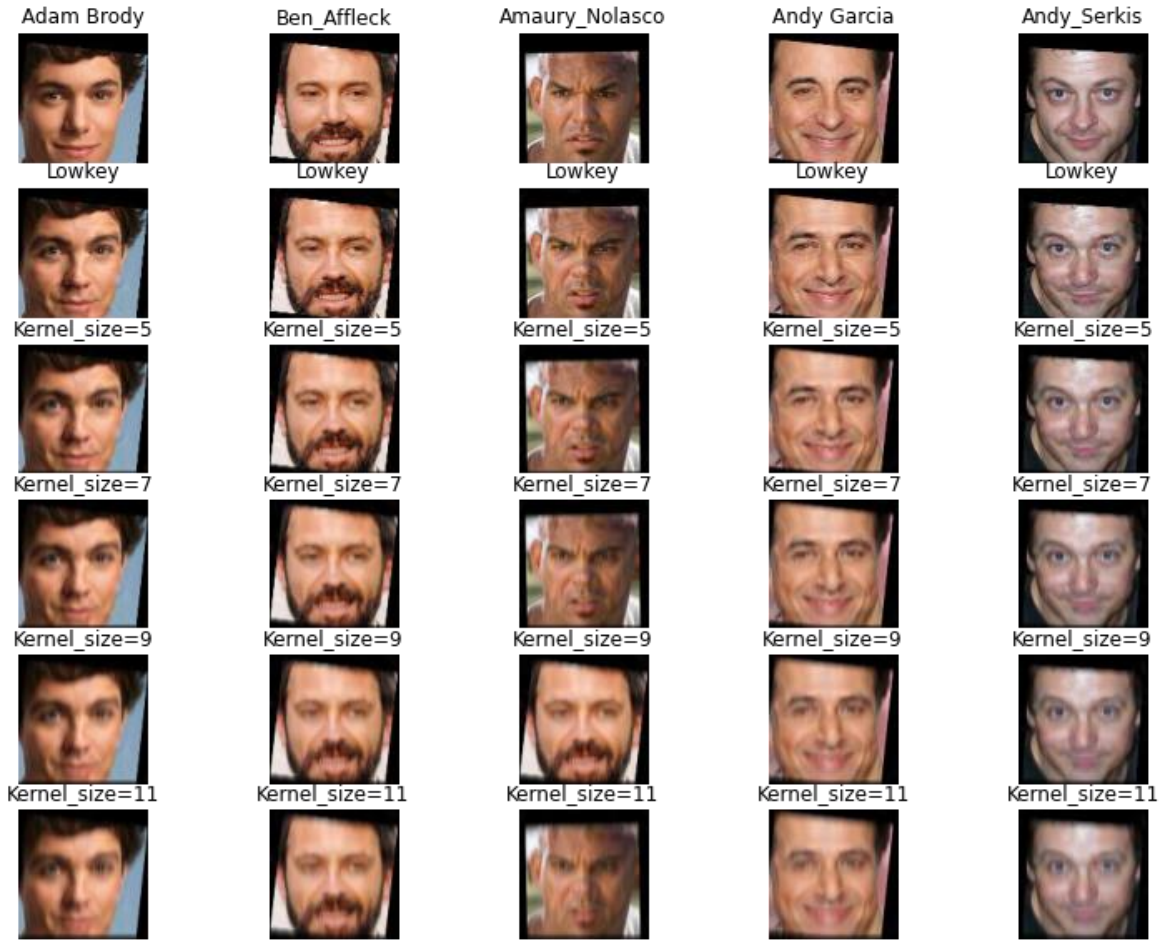
Evaluation Results Top 1 Accuracy				
Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
clean	99.36	99.36	99.57	99.47
Lowkey	86.97	87.07	87.29	87.18
Our Attack	87.07	87.07	87.39	87.29

Evaluation Results Top 5 Accuracy				
Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
Clean	99.47	99.36	99.57	99.47
Lowkey	87.18	87.07	87.29	87.18
Our Attack	87.18	87.07	87.82	87.29

Evaluation Results



Evaluation for Motion-Blur

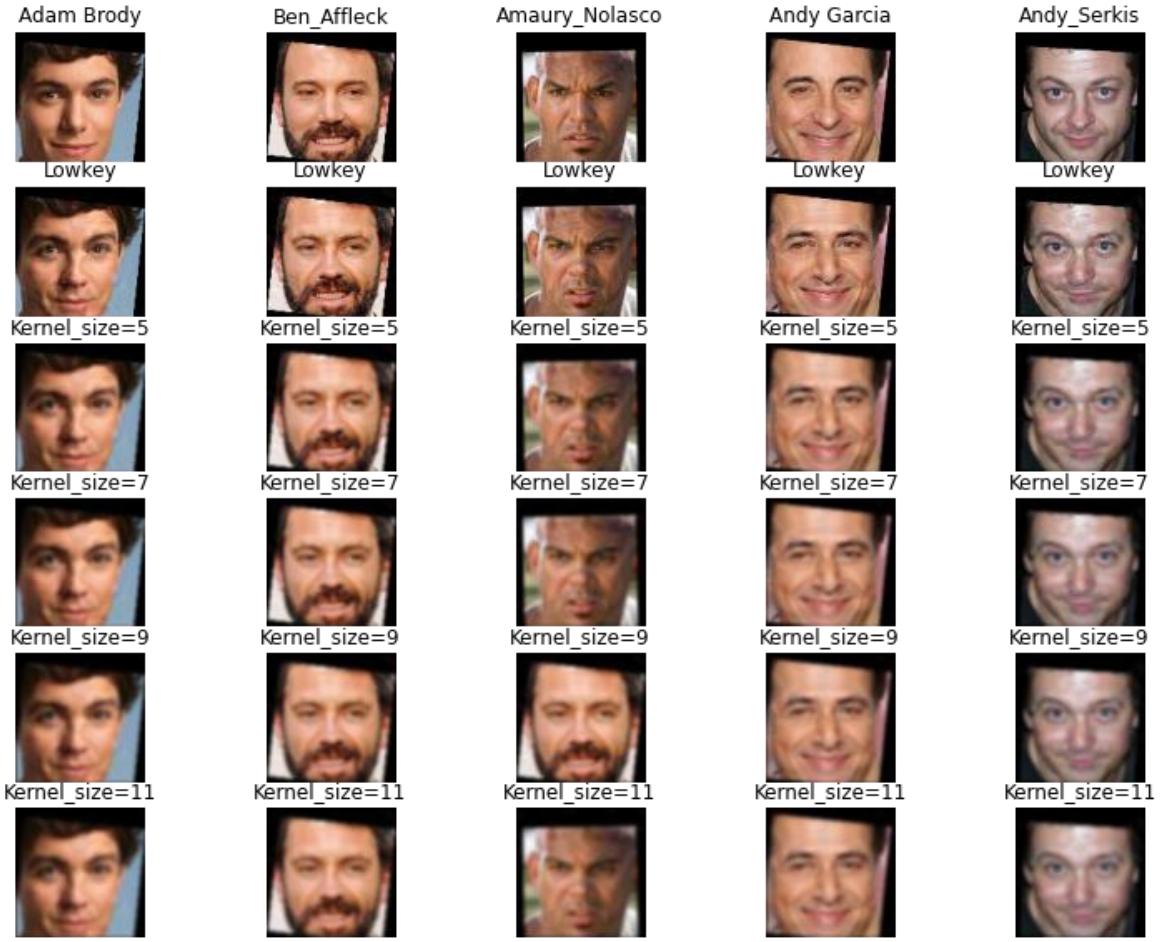


Top-5 Evaluation pipeline Motion_Blur with intensity=3,5,7,9,11 &17				
Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
Clean	99.34	99.34	99.34	100.0
Clean + motion_blur=3	99.34	99.34	99.34	100.0
Clean + motion_blur=5	99.34	99.34	99.34	99.34
Clean + motion_blur=7	99.34	99.34	99.34	99.34
Clean + motion_blur=9	99.34	100.0	99.34	100.0
Clean + motion_blur=11	99.34	100.0	99.34	100.0
Clean + motion_blur=17	99.34	99.34	99.34	100.0
Lowkey	74.17	74.17	74.17	74.83
Lowkey + motion_blur=3	74.17	74.17	74.83	74.83
Lowkey + motion_blur=5	74.17	74.17	74.17	74.83
Lowkey + motion_blur=7	74.17	74.17	74.17	74.83
Lowkey + motion_blur=9	74.17	74.17	74.83	75.5
Lowkey + motion_blur=11	74.17	74.83	74.83	78.81

Evaluation for Max-Blur

Top-5 Evaluation pipeline Max Blur - 24 probe,24 gallery				
Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
Clean	99.34	99.34	99.34	100.0
Clean + max_blur with k=3	99.34	99.34	99.34	99.34
Clean + max_blur with k=5	99.34	99.34	99.34	100.0
Clean + max_blur with k=7	99.34	99.34	99.34	100.0
Clean + max_blur with k=9	99.34	99.34	99.34	100.0
Clean + max_blur with k=11	99.34	99.34	99.34	99.34
Clean + max_blur with k=17	99.34	99.34	99.34	100.0
Lowkey	74.17	74.17	74.17	74.83
Lowkey + max_blur with k=3	72.85	74.17	74.17	74.83
Lowkey + max_blur with k=5	74.17	74.17	75.5	74.83
Lowkey + max_blur with k=7	74.83	73.51	76.16	75.5
Lowkey + max_blur with k=9	74.17	70.86	76.16	74.83
Lowkey + max_blur with k=11	74.17	71.52	76.16	72.85

Evaluation for Pool-Blur



Attacker	IR_152A	IR_152C	RESNET_152A	RESNET_152C
Clean	99.34	99.34	99.34	100.0
Clean + pool_blur with k=3	99.34	99.34	99.34	100.0
Clean + pool_blur with k=5	99.34	99.34	99.34	100.0
Clean + pool_blur with k=7	99.34	99.34	99.34	100.0
Clean + pool_blur with k=9	99.34	99.34	99.34	99.34
Clean + pool_blur with k=11	99.34	99.34	99.34	99.34
Clean + pool_blur with k=17	99.34	99.34	99.34	100.0
Lowkey	74.17	74.17	74.17	74.83
Lowkey + pool_blur with k=3	74.17	74.17	74.83	75.5
Lowkey + pool_blur with k=5	74.17	74.17	76.16	74.83
Lowkey + pool_blur with k=7	74.17	74.17	76.16	75.5
Lowkey + pool_blur with k=9	73.51	74.17	75.5	75.5
Lowkey + pool_blur with k=11	73.51	72.85	76.16	74.83