# It's Not What It Looks Like: Manipulating Perceptual Hashing based Applications

#### CCS 2021



Qingying Hao UIUC



Licheng Luo UIUC



Steve T.K. Jan UIUC Virginia Tech



Gang Wang UIUC



### **Perceptual Hashing**

• Generate fingerprints for images







(b) "e2a43f415352c3f5" Hamming Distance = 0



(c) "fc231cf82376b90c" Hamming Distance = 32

# Perceptual Hashing for Reverse Image Search Catfishing / Dating Scam Detection



Eli, 24 © 7 miles away

Sunday fundays > lazy Sundays Skiing > snowboarding Electric guitar < acoustic guitar (but I play both)

Peanut butter > jelly (Though, they still go together pretty well. Maybe we will too.)



Pages that include matching images

https://rearfront.com > fake-profile-using-models-pictur... ▼

#### Guy Made A Fake Profile On Tinder Using A Model's Picture ...

1200 × 627 · Apr 15, 2018 — Guy Made A Fake Profile On Tinder Using A Model's Picture To Show How Girls Respond To Creepy Messages From Hot Guys.

https://brainstudy.info > images > how-to-create-fake-ti... 💌

#### How To Create Fake Tinder Account - brainstudy.info

1200 × 627 — Tinder Without Phone Number: Can it Work and what are the ... So, I did a terrible thing and made a fake profile.: Bumble. Start · Woman ...

https://brainstudy.info > images > how-to-fake-a-tinder-... \*

#### How To Fake A Tinder Account Articles from 2021 - brainstudy ...

1200 × 627 — Tinder Line With Beth How To Find A Fake Tinder Profile. Perfect meme response for fake tinder accounts - 9GAG. How To Spot Fake Tinder Profiles, Bots, ...

https://cheezburger.com > guy-trolls-vain-tinderellas-by... 💌

#### Guy Trolls Vain Tinderellas By Using Model's Pics On His Profile

503 × 498 — Guy Trolls Vain **Tinderellas** By Using Model's **Pics** On His **Profile**. This is either **fake** AF or people are just always down to grab a free coffee, and disregard ...

https://pxs.hphmmelas.site > fake-tinder-profile-maker-... ▼

#### Category: Fake tinder profile maker online - Pxs

1200 × 627 · Jan 7, 2021 — The guy in the right however, will score 10x more matches because of his picture quality. Tinder FAKE Profiles - Is Tinder Dead? ♡ ...







#### www.tinderseduction.com

male-model-tinder-case-study/ - First found on Aug 11, 2018
male-model-tinder-case-study/ - First found on Jun 7, 2016

Filename: Model-Profile-Pic-2.png (280 x 280, 38 KB)

### Perceptual Hashing for Misinformation Detection





**Original image**  $\longrightarrow$  Semantic  $\longrightarrow$  **Edited image** Changes Xo

Xm

Pages that include matching images

www.snopes.com > Fact Checks > Fauxtography \*

#### Does This Photograph Show Stacey Abrams Touting the ...



797 × 980 · Nov 6, 2018 — A photograph of Georgia gubernatorial candidate Stacey Abrams posing with Linda Sarsour (a Palestinian-American Muslim political activist) ...

#### www.washingtonexaminer.com > weekly-standard > fac... \*

#### Fact Check: Did the Muslim Brotherhood Show Support for ...



998 × 954 · Oct 17, 2018 — A photo of activist Linda Sarsour and Georgia candidate for governor Stacey Abrams d over the ....

twitter.com > staceyabrams > status \*

#### Stacey Abrams على تويتر: "Pr



1591 × 2048 · Jan 20, 2018 @womensmarch national co women speak up, ...

mel Vhen

#### What if the Attacker Can Do More



Attack image



Pages that include matching images

www.ajc.com > blog > politics > out-state-political-forces-...

#### National forces help shape Georgia governor's race



1080 × 1080 · May 14, 2018 — Kamala Harris to about Stacey Abrams: "Make sure she wins." The appearance ...

twitter.com > staceyabrams > status \*

#### Stacey Abrams on Twitter: "Humbled to ha



 $1080 \times 1080 \cdot$  May 11, 2018 — Humbled to have as we fight to build a Georgia where everyone ha thrive.



www.facebook.com > stacey.abrams.77 > posts > it-was... \*

#### Stacey Abrams - It was a pleasure to host Sen. Kamala ...



898 × 960 — **Kamala Harris** yesterday in Atlanta during our "Drive to History." Sen. Harris' commitment to public education, criminal justice reform and social justice makes her a ...

# This Paper.

- Primary goal:
  - Generate attack images to subvert perceptual hashing based image search





Original image  $X_o$  Attack image  $X_a$ 



- Secondary goal: theme here . Textbox a little bit left
  - Blackbox model
  - Robust against practical defenses

# Background: pHash Algorithms



- Challenges:
  - pHash is not diffentiable
  - Design a function to approximate pHash process and estimate the gradient

#### Our Attack Algorithm



For a 64-bit hash string, if the attacker aims to cause 20-bit difference between  $x_a$  and  $x_{o_i}$ targeted  $d_t=20/64=0.3125$ 

#### Our Attack Framework: Basic Attack

Grayscale initialization



- Attacking grayscale image will reduce the attack search space
- Standard perceptual hashing algorithms calculate hash codes based on grayscale images

### **Our Attack Framework: Advanced Attacks**

Idea	Detailed methods
Attack over Input Ensemble (AoE) Motivation: search engines may have processed the original images	$\begin{split} \min_{X_a} D(x_a, x_o) + c \cdot (f(x_a, x_o) + \sum_{t \in T} f(x_a, t(x_o))) \\ \textbf{X}_a \text{ differs from } \textbf{X}_o \qquad \qquad$
	Julicu

### **Experiments Setups**

- Datasets
  - ImageNet
    - 1.28 million images
  - Face Images Catfishing / social network impersonation detection
    - 1,081 images
  - Image Manipulation Dataset (IMD) Misinformation detection
    - ~ 2400 images
- Target perceptual hashing algorithms
  - Standard pHash
  - Blockhash (Block Mean Value Based Hash)
    - Slice the image into blocks
    - Each bit is calculated for one block by comparing it with neighboring blocks

### **Emulate Image Reverse Search**



#### Backend Database (pHash)

Search Evaluation Metrics	Explanation
Average false positives (FP)	Average number of returned irrelevant images
Top-K Hit Rate ( <b>Top-K HR</b> )	True match images are among top-k returned images
Failed Query Rate (FQR)	True match images are not in returned images

### Impact of Grayscale Initialization

Efficiency of grayscale initialization VS. direct RGB attack

Method	Target hash distance d <sub>t</sub> = 0.15								
	Hash Distance	Perceptual Diff. (LPIPS)	Num of Iterations*						
Direct RGB (Baseline)	0.153	0.039	198.52						
Grayscale + RGB (Our)	0.156	0.005	22.98 + 1.52 = 24.5						

Basic attacks against pHash using 50 images from ImageNet

Num of Iterations\*: number of iterations to find the (first) targeted attack image

### **Basic Attack Standard pHash**

	Highly successfu	ıl attack			
Dataset		1			
	Hash Dist	Perceptual Diff.	#FP	Top-5 HR	FQR
ImageNet	0.310	0.034	3.38	0%	100%
Face	0.293	0.050	3.32	6%	94%

 $\tau$  = 0.2 for phHash in order to control FP of the search engine < 10





Clean LPIPS=0.015 LPIPS=0.097 LPIPS=0.185

### **Basic Attack Blockhash**

	Highly s									
Dataset	Target hash distance d <sub>t</sub> = 0.31									
	Hash Dist	Perceptual Diff.	#FP	Top-5 HR	FQR					
ImageNet	0.314	0.049	0	0%	100%					
Face	0.313	0.119	0	0%	100%					

 $\tau$  = 0.14 for Blockhash in order to control FP of the search engine < 10



**Clean**  $d_t = 0.15$   $d_t = 0.31$ 

### **Advanced Attack Results**



Cropping (t1) Remove 2.5% edge of the image





#### Disproportionate Scaling (t3)

Slightly stretch the image by increasing the width to **w** (scaling factor, e.g., 1.1) times the original width. Crop the image.

Attack Method	Blockhash, Target hash distance d <sub>t</sub> = 0.15						
	Hash Dist	Perceptual Diff.					
AoE	0.15	0.049					
AoT1	0.150	0.122					
AoT2	0.160	0.164					
AoT3	0.155	0.122					

Advanced attacks against Blockhash using 50 images from Face dataset



Example images for the advanced attacks against Blockhash with target hash distance  $d_t$  =0.15

### **Real World Experiments**



- Real-world search engines return multiple versions of the images
  - Need to manually examine the results to calculate true/false positives
- Procedure:
  - Small scale
    - Run multiple algorithms on 5 random images
    - Pick the strongest attack
  - Large scale evaluation

#### **Real World Experiments: Tineye Search**











TinEye searched over 50.0 billion images but didn't find any matches for your search image. That's probably because we have yet to crawl any pages where this image appears. TinEye is always crawling, so try your search again soon. See our FAQ for other reasons we may not have found your image.

Using TinEye is private. We do not save your search images.



Filter by website / collection



evecentre.com.my

index.php - First found on Nov 4, 2019

Filename: home-hero-about-eve-7a-a050e017.jpeg (2560 x 1140, 114 KB)



herstelgerichtluisteren.nl

inzichten/ - First found on Nov 7, 2019 Filename: nathan-dumlao-533483-unsplash-400x250.jpg (400 x 250, 18.9 KB)

### Real World Experiments: Large Scale

- Pick the best attack algorithm (AoT3 Disproportionate Scaling pHash)
- Test on 50 images from the Face dataset
  - Slightly tune up targeted hash distance **d**<sub>t</sub> from 0.31 to 0.34
  - Increase the disproportionate scaling factor from 1.1 to 1.24

<b>Evaluation Metrics</b>	Tineye	Google	Bing	Yandex
Avg. Reduction Rate*	100%	88%	100%	-1364%
Top1 Hit Rate	100% -> 0%	88% -> 36%	34% -> 0%	82% -> 58%
Top10 Hit Rate	100% -> 0%	96% -> 36%	34% -> 0%	86% -> 58%
False Postive Rate*	0% ->100%	11% -> 66%	66% -> 100%	21% -> 90%

Avg. (Average) Reduction Rate\*: The rate of returned results reduction False Positive Rate\*: The rate of irrelevant images among all returned results

# **Real-world Attack Case Study**



November 6, 2020 presidential election

(a) **Original**: a ballot counter was transcribing ballots that had been damaged

(b) **Cropped** image : the election worker is filling out the ballot

- Can still find the original / source image
- Show in fact-checking websites
- (c) **Attack** image:
  - AoT3-pHash
- (d) **Attack** image
  - AoT3-Blockhash •

### **Potential Defense**



Dataset	Hash Distance Target Dist d <sub>t</sub> =0.31 P						Searching	Results	
	pHash	Blockhash	aHash	dHash	wHash		#FP	Top-5 HR	FQR
ImageNet	0.368	0.327	0.366	0.438	0.386	0.109	2.88	0%	100%
Face	0.371	0.317	0.354	0.421	0.375	0.146	3.38	0%	100%

# **Potential Defense**



- Generate a set of search images using randomized transformations
  - Rotation, scaling, denoise and image filters
  - Parameters are randomized
- The search engine could fuse the results for all transformed images
  - Improve the top-1 and top-10 hit rates while reducing false positives

# **Conclusions and Future Directions**

- In Conclusion:
  - Design new attacks against perceptual hashing based applications in a black-box manner
  - Evaluate our attack against the standard and robust perceptual hashing algorithms
  - Test our attack using real-world image search engines
- Future:
  - To evaluate other security sensitive areas (e.g., child pornography content detection)
  - Experiment with more robust defense methods
  - Understand vulnerabilities of deep learning based hash functions

# Thank You

# https://qingyinghao.web.illinois.edu

# Question list:

# Backup slides Start here

# **Attack Effects on Online Abuse Detection Systems**

Harris' commitment to public education, criminal justice reform and social justice

nakes her a ..



+1 example, as Animation!



vomen speak up, ...

### **Existing Attacks and Limitations**

- Introduce image distortions to alter the semantic meaning of the image<sup>1 - old days.</sup>
  - Maintain same(similar) hash values
  - Target at image authentication applications
- Most existing attacks noises are manually crafted<sup>2</sup>





(a) original

(b) object change





(c) object removal

(d) object insertion

# Targeted Perceptual hashing algorithms (backup)

• Attack two types of perceptual hashing algorithms. Show examples. Real images



DCT\*: Discrete Cosine Transformation (Frequency calculation method)

# Compare Advanced and Basic Attacks: Transferability

Attack	<b>Transferred Distance (target hash dist</b> $d_t = 0$							
Method	pHash	Blockhash	aHash	dHash	wHash			
pHash Basic	0.313	0.122	0.073	0.164	0.081			
pHash AoE	0.324	0.158	0.105	0.215	0.109			
pHash AoT3	0.246	0.163	0.100	0.200	0.103			
Attack	Tran	sferred # FP	' (target l	nash dist	$d_t = 0.31$ )			
Method	pHash	Blockhash	aHash	dHash	wHash			
pHash Basic	3.38	2.6	74.12	15.34	22.66			
pHash AoE	2.64	1.82	80.74	9.28	12.44			
pHash AoT3	3.2	1.48	130.54	12.4	14.08			
Attack	Transfe	erred Top-5	HR (targ	et hash d	<b>ist</b> $d_t = 0.31$ )			
Method	pHash	Blockhash	aHash	dHash	wHash			
pHash Basic	0%	68%	36%	46%	30%			
pHash AoE	0%	36%	20%	24%	8%			
pHash AoT3	10%	28%	16%	24%	12%			
Attack	Tran	sferred FQR	(target ]	hash dist	$d_t = 0.31$ )			
Method	pHash	Blockhash	aHash	dHash	wHash			
pHash Basic	100%	32%	58%	54%	66%			
pHash AoE	100%	64%	78%	76%	90%			
pHash AoT3	88%	72%	78%	74%	84%			

- Advanced attacks (AoE, AoT3) are more transferable than the basic attack.
- AoE is three times slower than AoT3, so AoT3 is a better option

Transfer the attack image optimized for pHash to other hash functions using ImageNet dataset

# **Real-world Attack Case Study**

#### (a) Original Image

#### (b) Cropped Image to Spread Misinfo

"A Pennsylvania elections worker actually filling out ballots rather than counting them"

Google: 160 matched TinEye: 4 matched

Google: 0 matched

TinEye: 0 matched



Google: 11 matched TinEye: 0 matched

(c) Attack Image

(d) Attack Image





# Basic Attack Standard pHash: design choices

- Randomly select 50 images for ImageNet
- Impact of perceptual distance function:

Distance		Target h	ash dist	tance $d_t$ =0.15 Target hash distance $d_t$ =0.31				
Function	Hash Dist.	$L_2$	Pdist	Pdist # Iterations (Gray+RGB) H		$L_2$	Pdist	# Iterations (Gray+RGB)
L2 as $D()$	0.156	25.306	0.033	90.42+13.18	0.292	50.449	0.116	278.54+149.86
LPIPS as $D()$	0.156	6.788	0.005	22.98+1.52	0.310	20.828	0.034	87.56+1.40
	$\mathbf{A}$						40 (M)	

Comparison of different perceptual distance functions using the basic attack (ImageNet)

• Impact of Grayscale Initialization

Method Target hash distance $d_t=0$									
		Hash Dist	Pdist	# Iterations					
	Direct RGB (Baseline)	0.153	0.039	198.52					
	Grayscale+RGB (Our)	0.156	0.005	22.98+1.52					
-									

Comparison of direct optimizing RGB images vs. using grayscale initialization (ImageNet)

Dataset	Target hash distance $d_t$ =0.15					Tar	get hasl	n distar	nce $d_t = 0.31$		Tar	get hasl	n distai	nce $d_t = 0.47$	
	Hash Dist	Pdist	# FP	Top-5 HR	FQR	Hash Dist	Pdist	# FP	Top-5 HR	FQR	Hash Dist	Pdist	# FP	Top-5 HR	FQR
ImageNet	0.156	0.005	6.72	96%	0%	0.310	0.034	3.38	0%	100%	0.441	0.120	1.68	0%	100%
Face	0.156	0.009	4.76	100%	0%	0.293	0.050	3.32	6%	94%	0.383	0.145	2.90	4%	96%

- $\tau$  = 0.2 (Search engine distance threshold)
- When d<sub>t</sub>=0.15 and 0.31, most images can reach the targeted hash distance
- Near 100% FQR and top-5 HR is no more than 6%



Dataset	Target hash distance d <sub>t</sub> = 0.31				
	Hash Dist	Perceptual Dist.	#FP	Top-5 HR	FQR
ImageNet	0.310	0.034	3.38	0%	100%
Face	0.293	0.050	3.32	6%	94%

# Attack Evaluation using IMD Dataset

- IMD (Image Manipulation Dataset) dataset
  - Over 50% pairs have a distance below 0.2
    - Easy to trace the manipulated image
  - Create x<sub>a</sub> based on x<sub>m</sub> (modified version) to further increase hash distance

Attack Method	Target Hash Dist d <sub>t</sub> = 0.31		
	Hash Dist	Pdist	
pHash Basic	0.30	0.027	
pHash AoT3	0.244	0.146	



Clean Manipulated Basic AoT3

Basic attack and AoT3 Disproportionate Scaling attack using IMD dataset

Pages that include matching images

www.brylin.com > programs-services > mental-health-tr... 💌

#### Mental Health Treatment for Children and Adolescents ...



1920 × 1280 — Mental health treatment program for children and adolescents at BryLin offers assessment, crisis stabilization and treatment for kids 5-17 years of age.

www.fay3.com > ... 
Translate this page

#### صور لـ #شخص #مشاعر #مراهق #صورة #صبي #طفل - موقع فايع

375 × 250 — Adolescent #Portrait #Person #Child #Feelings # Boy





tarbiazakia.com > 2018/05 • هذا ما يع... < 7 Translate this page هذا ما يعلمه الأهل الناجحون لأو لادهم قبل عمر 18 سنة - التربية الذكية



الهم : المصروف يجب أن لا يكون فقط من أجل "المتعة". الأولوية يجب أن تكون لمراقبة — 1024 × 1024 كيفية إنفاق هذا المال. عَلَموا أولادكم أن لا يشتروا كيفما كان، وأن يوزعوا مصاريفهم ...

gramho.com > explore-hashtag > Ind... Translate this page

#### #IndonesianSubculture Instagram posts - Gramho.com



640 × 640 — new piercing : flat, helix and upperlobe Each body is unique and healing times vary ...

carine-llamas-psychanalyste.business.site 🔻 Translate this page

#### Carine Llamas Psychanalyste - Psychanalyste Enfants ...



1280 × 1280 — Le confinement a fait bouger les lignes dans une multitude de domaines. Le couple a pu être mis à rude épreuve par une proximité pas toujours bien vécue, une ...

1 2

menteor-organizzarea.com > hankou... Translate this page

#### 【子供反抗期】10歳からの対応の仕方!子供との関係性は…



800 × 533 · May 9, 2019 — 子供の反抗期。最近では特に低年齢化しているともいわれていますが、こればかりは親として見守ることなどして対応していくしかないです…

#### Previous

**Backup slide** 

#### Tineye

