AlphaDog: No-Box Camouflage Attacks via Alpha Channel Oversight

Dr. Qi Xia
Dr. Qian Chen

UTSA
The University of Texas at San Antonio™
Department of Electrical and Computer Engineering
NDSS-2025

Background

Traditional black-box adversarial attacks on computer vision models:

- Gradient-based attack: Fast Gradient Sign Method (GoodFellow et al)
- Image Scaling Attacks (Xiao et al)

They face significant limitations, including:

- Intensive querying requirements
- Time-consuming iterative processes
- A lack of universality,
- Low attack success rates (ASR) and confidence levels (CL)

To overcome these limitations, we propose **AlphaDog:**

- No-box attack (Zero Query)
- Universal adaptability
- 100% confidence level and ASR

What is AlphaDog?

AlphaDog: an Alpha channel attack, the first universally efficient **targeted no-box** attack, exploiting the often overlooked ***Alpha channel*** in RGBA images to create visual disparities between human perception and machine interpretation, efficiently deceiving both.https://sites.google.com/view/alphachannelattack/home

Targeted Images (can be any images with same size)

Human Eyes

(Normal Targeted Image)

AI Models

(Malicious Targeted Image)

What is AlphaDog?

AlphaDog: an Alpha channel attack, the first universally efficient **targeted no-box** attack, exploiting the often overlooked ***Alpha channel*** in RGBA images to create visual disparities between human perception and machine interpretation, efficiently deceiving both.https://sites.google.com/view/alphachannelattack/home

Targeted Images (can be any images with same size)

Human Eyes

(Normal Targeted Image)

AI Models

(Malicious Targeted Image)

What is AlphaDog?

AlphaDog: an Alpha channel attack, the first universally efficient **targeted no-box** attack, exploiting the often overlooked ***Alpha channel*** in RGBA images to create visual disparities between human perception and machine interpretation, efficiently deceiving both.https://sites.google.com/view/alphachannelattack/home

Targeted Images (can be any images with same size)

Human Eyes

(Normal Targeted Image)

AI Models

(Malicious Targeted Image)

What is AlphaDog?

AlphaDog: an Alpha channel attack, the first universally efficient **targeted no-box** attack, exploiting the often overlooked ***Alpha channel*** in RGBA images to create visual disparities between human perception and machine interpretation, efficiently deceiving both.https://sites.google.com/view/alphachannelattack/home

Targeted Images (can be any images with same size)

Human Eyes

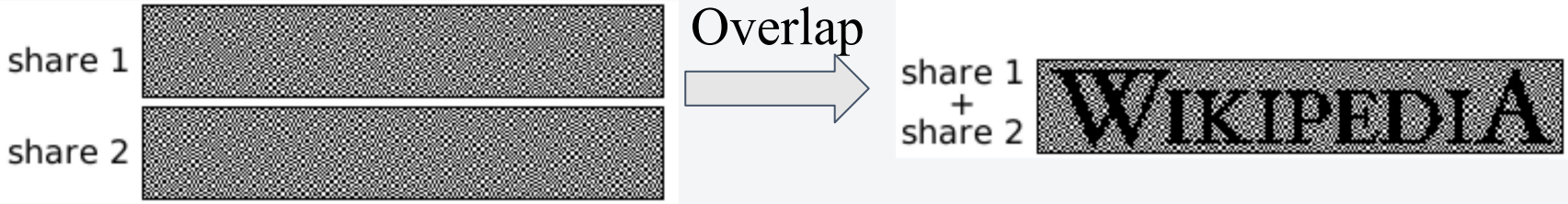
(Normal Targeted Image)

AI Models

(Malicious Targeted Image)

AlphaDog foundation

Visual Cryptography: When overlapping two seeming nonsense images, we get a new image



AlphaDog is inspired by Visual Cryptography!

When overlapping an image with a background, human eyes and AI models can see the same image differently!

Alpha Channel & RGBA Image Format

Old images like JPEG, BMP have only RGB channels,
But modern images like PNG, HIFF, GIF, WEBP, have four channels, RGBA

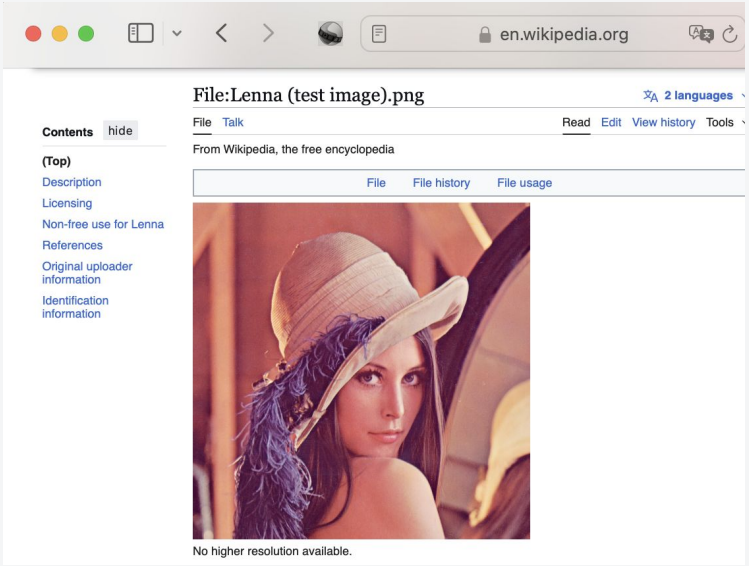
Red (R), Green (G), and Blue (B) color channels and an Alpha channel (A).

- Alpha channel values range from 0 (fully transparent) to 1 (fully opaque) for each pixel.
- Alpha channel controls the level of transparency.

AI model developers don't care about Alpha Channel !
They usually simply remove Alpha Channel and read only RGB channels!

Background Colors

Digital Image Media & Their Background Colors



Chrome

Safari

Paint

Preview

Background Colors

Digital Image Media & Their Background Colors

TABLE I: Background Colors of Image Media Apps. * indicates default Apps for the Operating System.

Composition Strategy	Thumbnail (Reduced-Size Image Display)	Viewer (Full-Size Image Display)
White Background	Google Chrome, Safari, Mozilla Firefox , Microsoft Edge, Microsoft IE, macOS Finder*, iPhone Photos*, Win10 file explorer* Ubuntu file explorer*	Google Chrome, Safari, Mozilla Firefox, Microsoft Edge, Microsoft IE, Adobe PDF viewer, Adobe Photoshop Ubuntu Image viewer* iPhone Photos*
Gray Background	N/A	Win10 Photos* (R:64 G:64 B:64), Mac Preview* (R:150,G:150,B:150)

Chrome

Safari

Paint

Preview

How AI Models Treat Alpha Channel?

- **80 open-source models** and **20 COTS models**
- **Consistency:** Most models remove the input images' Alpha Channel.
- **Outliers:** Google Bard and GoogleCloudVisionApi add black and white background colors.

No Box AlphaDog Attack

- No query. The attacker assumes that all victim AI models remove the input image Alpha channel.
- No response from the victim AI models.
- Human visualizes Normal target images.

Root Cause: Ignoring Alpha Channel

TABLE II: Computer vision models for Alpha channel treatment in input images. "Open" denotes open-source models, and "Cloud" signifies commercial cloud-based image recognition systems. Only the underscored cloud-based systems with ^{w} (Google Cloud Vision Api) and ^{b} (Bard) add a white background to the input image; others remove the input image's Alpha channel.

Training Dataset (Model Category)	100 Computer Vision Models 80 (Open) and 20 (Cloud)
COCO [33] (Open)	Mask RCNN [14]–[17], YOLOv3 [18]–[32] YOLOv4 [34], YOLOv5 [35], [36] RetinaNet [37]–[40], CenterNet [41]–[44] EfficientDet [45]–[49], Cascade RCNN [50], [51]
Pascal VOC [52] (Open)	Faster RCNN VGG [53]–[58] YOLOv1 [59],YOLOv2 [60]
ImageNet [72] (Open)	AlexNet [61]–[65], ResNet [66]–[71] EfficientNet [73]–[76], InceptionV3 [77] InceptionV4 [78], GoogLeNet [79]
MNIST [80] (Open)	LeNet-5 [81], [82]
FDDB [83] (Open)	Facedet [84] Cascade CNN [85]
KITTI [86] (Open)	MonoDepth [87]–[90]
BDD100k [91] (Open)	YOLOv3 [92], [93], YOLOv5 [94]
Wider Face [95] (Open)	Facenet [96], YOLOv3 [97], [98], YOLOv2 [99], [100]
CIFAR-10 [101] (Open)	ResNet [102]
Dataset Unknown (Cloud)	$\underline{Bard^{b}}$ [12] $\underline{GoogleCloudVisionApi^{w}}$ [13] Amazon Rekognition [11] GeminiProVisionAPI [103] Baidu Image [104], Baidu API [105] IMAGERecognize [10], TeachableMachine [106], Nyckel [107], Labelbox [108], ChatGPT4 [9], Wolfram [109], Vue.ai [110] Microsoft Azure [111], AliYunVision [112] Tencent Vision [113], Landing Lens [114] Clarifai [115], Imagga [116], ANYLINE [117]

How To Create an AlphaDog Attack Image?

Alpha Compositing in Computer Graphics

$$I_{Atk} = \text{Concat}(I_{IN}, A).$$

$$I_{Eye} = A \circ I_{IN} + (1 - A) \cdot BKG.$$

RGB channel intensity matrix: I_{IN} (AI Models Remove the Alpha channel, $I_{IN} = I_{AI}$)
Alpha Channel Matrix : A
BKG: digital image media (app)'s background color (Human Eye)

I_{Eye}, is targeted for human eye; I_{IN} is targeted for AI models
I_{Eye}, and I_{IN} are pre-selective. We want to calculate A to create I_{Atk}
Example (I_{Atk} creation):

$$I_{AI} = I_{IN} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$I_{Eye} = A \circ I_{IN} + 1 - A = \begin{bmatrix} 1 & 0.5 & 1 & 1 & 0.5 \\ 1 & 0.5 & 1 & 0.5 & 1 \\ 1 & 0.5 & 0.5 & 1 & 1 \\ 1 & 0.5 & 1 & 0.5 & 1 \\ 1 & 0.5 & 1 & 1 & 0.5 \end{bmatrix}$$

white BKG

, and therefore, we obtain A as:

$$A = \begin{bmatrix} 0 & 0.5 & 0 & 0 & 0.5 \\ 0 & 1 & 0 & 0.5 & 0 \\ 0 & 1 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 1 & 0 \\ 0 & 0.5 & 0 & 0 & 0.5 \end{bmatrix}.$$

Then we have I_{Atk}:

$$I_{Atk} = \text{Concat}(I_{IN}, A).$$

How To Create an AlphaDog Attack Image?

Alpha Compositing in Computer Graphics

Attack process:

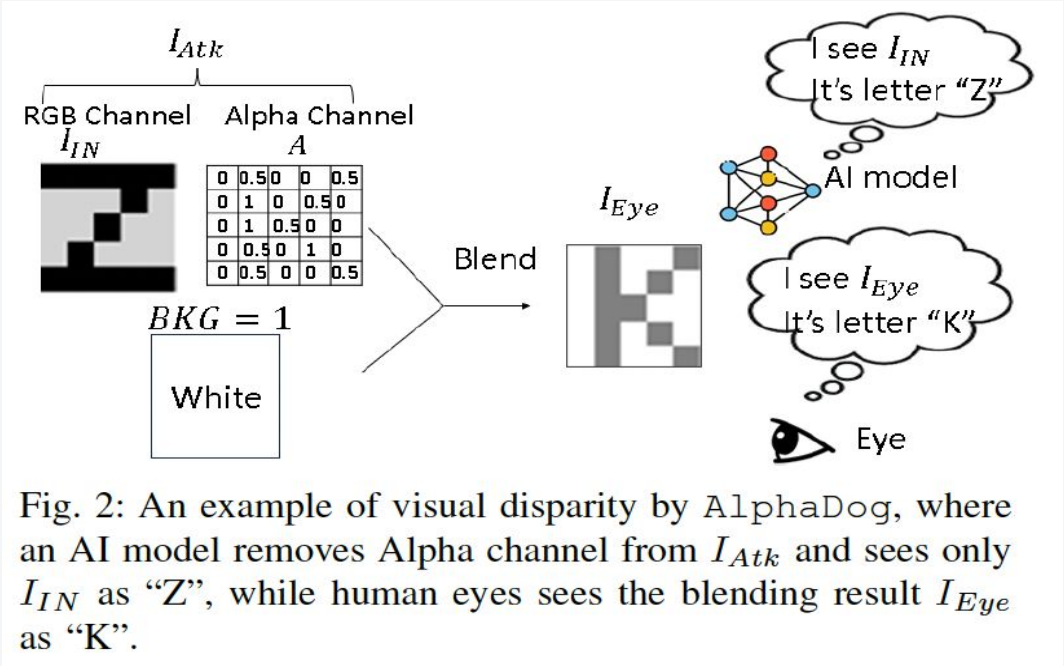


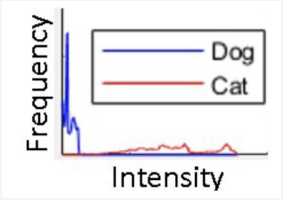
Fig. 2: An example of visual disparity by `AlphaDog`, where an AI model removes Alpha channel from I_{Atk} and sees only I_{IN} as "Z", while human eyes sees the blending result I_{Eye} as "K".

How to ensure *stealthiness*

I_{AI}

I_{Eye}

Must ensure histogram separation



Math derivative of histogram separation:

$$I_{Eye} = A \circ I_{IN} + 1 - A \Rightarrow A = (1 - I_{Eye}) \oslash (1 - I_{AI})$$

A

$$0 \leq (1 - I_{Eye}) \oslash (1 - I_{AI}) \leq 1. \quad (5)$$

$$0 \leq I_{AI} \leq I_{Eye} \leq 1$$

We empirically find that the value 0.2 serves as a universal threshold, independent of the specific target (normal or malicious) images.

$$0 \leq I_{AI} \leq 0.2 \leq I_{Eye} \leq 1 \quad (9)$$

AlphaDog Attack Images:

1. Any images could be selected as the targeted images
2. The image pairs must be the same size
3. Format: **PNG**, TIFF, HEIF, WebP, and GIF
4. Grayscale Images
5. Efficient generation O(n^2)

Algorithm 1 Attack Image Generation

fun ACAIMAGEGENERATION(I_{AI}, I_{Eye}, m, n)
  $I_{Atk} \leftarrow$ empty 3D array with dimensions $m \times n \times 2$
  $A \leftarrow$ empty 2D array with dimensions $m \times n$
  $I_{AI} \leftarrow \text{PREPROCESS}(I_{AI}, m, n) \times 0.8 + 0.2$
  $I_{Eye} \leftarrow \text{PREPROCESS}(I_{Eye} m, n) \times 0.2$ ▷ Eq. 9
  for $i \leftarrow 1$ **to** m **do**
    for $j \leftarrow 1$ **to** n **do**
      $A[i][j] \leftarrow \frac{1 - I_{Eye}[i][j]}{1 - I_{AI}[i][j]}$ ▷ Eq. 4
    end for
  end for
  $I_{Atk} = [I_{AI}, A]$ ▷ concatenate RGB and Alpha
  return $I_{Atk} \times 255$ ▷ return 8-bit attack image
end function
function PREPROCESS(I, m, n)
  $I \leftarrow$ remove I's alpha channel
  $I \leftarrow$ resize(I,m,n) ▷ Resize the image to $n \times m$
  $I \leftarrow I/255$ ▷ Normalize the 8-bit image
  return I
end function

AlphaDog Evaluation

Test Image Dataset: Our dataset encompasses the following, and is publicly accessible at [1].

1) 2,000 `AlphaDog` attack images in PNG format, each sized 256×256, under the assumption that AI models remove the input Alpha channel.
2) 4,000 attack images of identical dimensions but in alternative formats (TIFF, WebP, SVG, and GIF), with 1,000 examples per format.
3) 500 attack images each specifically tailored for testing the Bard and Gemini AI models, known for their deviations from the norm.

1. Stealthiness: IRB Approved Study.

20 participants: 18 and 45 years (with a mean age of 25.18 and a standard deviation of 6.8), comprised 50% females. Ethnic composition among participants was 50% Hispanic, 45% Caucasian, and 5% Asian, African American, and other ethnicities. All participants correctly identified images from.

Results: identity attack images as normal images

AlphaDog Evaluation

2. Image Formats: Attack Success Rate (ASR) depends on AI models

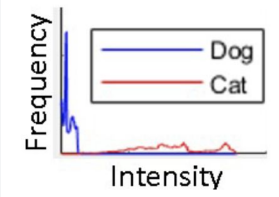
PNG format is the best choice now.

TABLE IV: Effects of RGBA image formats on `AlphaDog` success for open-source and cloud-based models. Checkmark (✓) indicates successful `AlphaDog` attacks with the corresponding format, while cross (×) indicates "format unsupported error".

Model	I/O Library or Model	RGBA Image Format				
		PNG	TIFF	Webp	SVG	GIF
80 open-sourced models shown in Table II	OpenCV	✓	✓	✓	×	✓
	TensorflowIO	✓	×	×	×	✓
	Pillow	✓	✓	✓	×	✓
	Imageio	✓	✓	✓	×	✓
	SimpleTik	✓	✓	×	×	×
20 Cloud-based systems	ChatGPT4	✓	✓	✓	✓	✓
	BaiduImage	✓	×	×	×	✓
	BaiduAPI	✓	×	×	×	×
	FreeImage	✓	×	×	×	×
	Wolfram	✓	✓	✓	✓	✓
	Google Api	✓	✓	✓	×	✓
	Gemini	✓	×	✓	×	×
	Teaching	✓	✓	×	×	✓
	LabelBox	✓	×	×	×	✓
	Nyckel	✓	✓	×	×	×
	MS Azure	✓	✓	✓	×	✓
	AliYun	✓	×	✓	×	✓
	Tencent	✓	×	×	×	×
	Amazon	✓	×	×	×	×
	Bard	✓	×	×	×	×
	GoogleCloud	✓	✓	✓	×	✓
	LandingLens	✓	×	×	×	×
	Clarifai	✓	✓	✓	×	✓
	Imagga	✓	×	×	×	×
	Anyline	✓	×	×	×	×
	Vue.ai	✓	×	×	×	×

Defense

Recall that an AlphaDog attack image has histogram separation. We can detect whether an image is an attack by checking if this image's histogram is separated



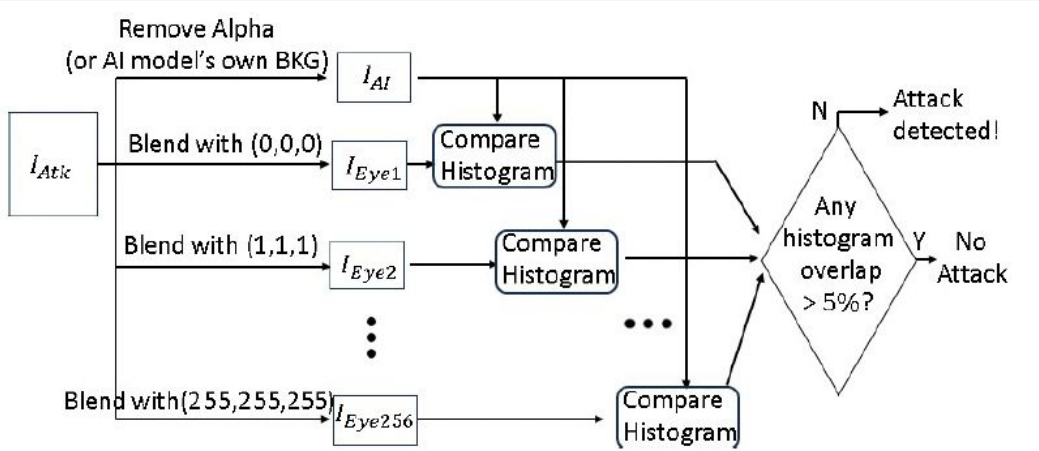


Fig. 4: Diagram illustrating intensity histogram-based detection.

Defense

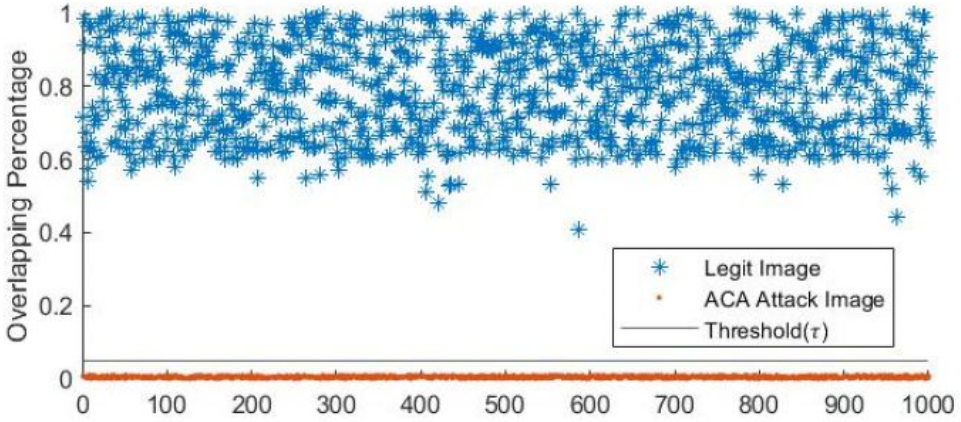


Fig. 5: Intensity histogram-based detection results for 1,000 `AlphaDog` attack images compared to 1,000 benign images.

Conclusion

AlphaDog represents a groundbreaking advancement in adversarial attacks by exploiting I/O flaw of Computer vision models. AlphaDog has such advantages over traditional adversarial attack:

- no-box (Zero query required)
- universal adaptability
- Can be generated effciently
- 100% confidence level and ASR

AlphaDog can be applied in :

- data poisoning,
- evasion attacks
- content moderation

AlphaDog can be potentially used to harm

- Autonomous driving,
- Medical
- Facial recognition

We create a dataset of 6500 attack images for researcher to test

Q&A

Email: Qi.Xia@utsa.edu