

It Doesn't Look Like Anything to Me: Using Diffusion Model to Subvert Visual Phishing Detectors

Paper #684 (Major Revision)

A Diffusion Models

Basic Diffusion Model Denoising diffusion probabilistic models (DDPM) (referred to as “diffusion models” for short) are generative models that generate high-quality images while offering fine-grained controls over image fidelity and diversity [7, 11, 13]. The goal of a diffusion model is to learn to generate data similar to the training data. To do so, diffusion models “destroy” the training data by progressively adding Gaussian noise to the image, and then learn to recover the original image by reversing the noising process. As shown in Figure ??, it contains a forward (diffusion) process and a reverse (diffusion) process. The forward process is to add Gaussian noise to a clean image step by step until reaching a version of (close to) pure Gaussian noise. The reverse diffusion process is defined by a Markov chain to transform Gaussian noise back to a clean image. The goal is to learn the reverse process with a neural network p_θ .

More specifically, during the forward process, we take x_0 (i.e., a clean image) and apply q to add Gaussian noises at each timestamp t to produce a series of latent vectors from x_1 to x_T , with variance $\beta_t \in (0,1)$. This process is modeled by a Markov chain to obtain the approximate posterior:

$$q(x_1, \dots, x_T | x_0) := \prod_{i=1}^T q(x_i | x_{i-1}) \quad (\text{Eqn. 1})$$

$$q(x_t | x_{t-1}) := N(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I}) \quad (\text{Eqn. 2})$$

Given a large T and well-scheduled variance β_t , x_T is nearly an isotropic Gaussian distribution N [13]. A recent work [7] shows that we do not need to repeatedly sample from $x_t \sim q(x_t | x_0)$. Instead, with $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=0}^t \alpha_s$, we can calculate x_t as:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \epsilon \sim N(0, \mathbf{I}) \quad (\text{Eqn. 3})$$

Here, $1 - \bar{\alpha}_t$ defines the variance of the noise ϵ at an arbitrary timestamp t . Using Bayes theorem, the posterior $q(x_{t-1} | x_t, x_0)$ can also be presented as a Gaussian and calculated with mean

$\mu_q(x_t, x_0)$ and variance $\beta_{q(t)}$ [13] which:

$$\tilde{\mu}(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_t} - 1}{1 - \bar{\alpha}_t} x_0 + \frac{\sqrt{\bar{\alpha}_t} (1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t \quad (\text{Eqn. 4})$$

$$\tilde{\beta}_t := \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t \quad (\text{Eqn. 5})$$

$$q(x_{t-1} | x_t, x_0) = N(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I}) \quad (\text{Eqn. 6})$$

For the inverse process, to recover $q(x_0)$, we can sample from $q(x_T) \sim \text{Gaussian } N(0, \mathbf{I})$ and run the reverse distribution $q(x_t | x_{t-1})$ until we reach x_0 . The reverse process could be approximated using a neural network to predict a mean μ_θ and variance Σ_θ as follows:

$$p_\theta(x_{t-1} | x_t) := N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (\text{Eqn. 7})$$

In DDPM [7], the variance is fixed at each time t as a function of the α coefficient defined by Eqn. 5. The variance can also be predicted as Σ_θ using the neural network [13]. To train the model p_θ , we can optimize the variational lower-bound L_{vlb} to get $p_\theta(x_0)$ as follows:

$$L_{vlb} = L_0 + L_1 + \dots + L_{T-1} + L_T \quad (\text{Eqn. 8})$$

$$L_0 = -\mathbb{E}_{q(x_1 | x_0)} [\log p_\theta(x_0 | x_1)] \quad (\text{Eqn. 9})$$

$$L_T = D_{KL}(q(x_T | x_0) || p(x_T)) \quad (\text{Eqn. 10})$$

$$L_{t-1} = \sum_{i=2}^T \mathbb{E}_{q(x_i | x_0)} [D_{KL}(q(x_{t-1} | x_i, x_0) || p_\theta(x_{t-1} | x_i))] \quad (\text{Eqn. 11})$$

Note that, only L_{t-1} has trainable parameters—it is simpler for training the diffusion model by focusing on this L_{t-1} term [7]. This is referred to as simplified objective function $L_{simple}(\theta)$. So the goal of training is to minimize the difference between the learned denoising transition step $p_\theta(x_{t-1} | x_t)$ and the forward process posteriors $q(x_{t-1} | x_t, x_0)$ as the ground truth defined by their KL Divergence.



(a) Image-Text Logo



(b) Text-Only Logo

Figure 1: **Examples of Successful Attacks**— We show successful examples of (a) image-text logos and (b) text-only logos. These logos have a similarity range of 0.6 – 0.87. From left to right, the quality decreases.

According to the KL Divergence between two Gaussian distributions, the optimization of L_{simple} is presented by Eqn. 12.

$$\begin{aligned}
 & \arg \min_{\theta} D_{KL}(q(x_{t-1}|x_t, x_0) || p_{\theta}(x_{t-1}|x_t)), \\
 & = \arg \min_{\theta} D_{KL}(N(x_{t-1}; \tilde{\mu}, \Sigma_q(t)) || N(x_{t-1}; \mu_{\theta}, \Sigma_q(t))) \\
 & = \arg \min_{\theta} \frac{1}{2\sigma_q^2(t)} [\|\mu_{\theta} - \tilde{\mu}\|^2]
 \end{aligned}
 \tag{Eqn. 12}$$

In the basic DDPM [7], the variance for forward and reverse processes are fixed at each time t and can be calculated as a function of α coefficient. We can write both of the variances $\beta_t \mathbf{I}$ and $\tilde{\beta}_t \mathbf{I}$ as $\Sigma_q(t)$ for brevity [11]. The variance could be written as $\Sigma = \sigma_q^2(t) \mathbf{I}$, using Eqn. 5.

One way to optimize the loss function is to train a neural network to predict $\mu_{\theta}(x_t, t)$. Alternatively, the network can predict x_0 and use Eqn. 4 to get $\mu_{\theta}(x_t, t)$. In practice, training the network to predict the noise ϵ gives better image qual-



(a) Image-Only Logo

Figure 2: **Examples of Successful Attacks**— We show successful examples of (a) image-only logos. These logos have a similarity range of 0.6 – 0.87. From left to right, the quality decreases.

ity [7]. So the simplified objective is defined as:

$$L_{simple} = \mathbb{E}_{t, x_0, \epsilon} [\|\epsilon - \epsilon_{\theta}(x_t, t)\|^2] \tag{Eqn. 13}$$

$\mu_{\theta}(x_t, t)$ can be derived from $\epsilon_{\theta}(x_t, t)$ as:

$$\mu_{\theta}(x_t, t) = \frac{1}{\sqrt{\alpha_t}} (x_t - \frac{1 - \alpha_t}{1 - \sqrt{\alpha_t}} \epsilon_{\theta}(x_t, t)) \tag{Eqn. 14}$$

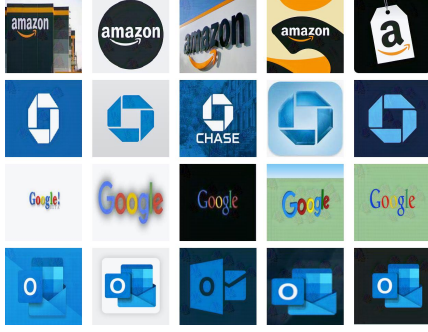
Improved Diffusion Model. The improved diffusion model [13] reduces the sampling by an order of magnitude during the forward passes without sacrificing the quality of generated images. Instead of fixing $\Sigma_{\theta}(x_t, t) = \sigma_t^2 \mathbf{I}$, the improved diffusion model also learns the variance for the reverse process to improve the log-likelihood (for $\log p(x)$). It defines the learning of variances as follows:

$$\Sigma_{\theta}(x_t, t) = \exp(v \log \beta_t + (1 - v) \log \tilde{\beta}_t) \tag{Eqn. 15}$$

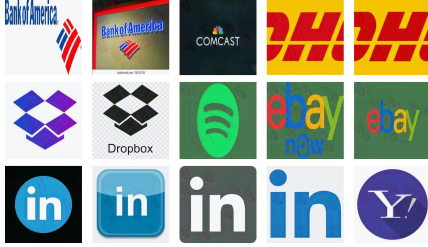
where β_t and $\tilde{\beta}_t$ represent the upper and lower bounds for the reverse process variances, and v is the predicted variance from the neural network. The improved diffusion model proposes a new hybrid loss which predicts for both mean $\mu_{\theta}(x_t, t)$ and variance $\Sigma_{\theta}(x_t, t)$:

$$L_{hybrid} = L_{simple} + \lambda L_{vib} \tag{Eqn. 16}$$

λ is the scaling factor (set to be small so the L_{vib} will not



(a) Example 1



(b) Example 2

Figure 3: Examples of Successful PhishGap Attacks— While these logo images succeed in bypassing PhishIntention, they are not necessarily ready to be added to webpages to deceive users end-to-end due to image cropping, disproportionate stretching, unusual background color, and the use of non-web logos.

overwhelm L_{simple}). It also applies a stop gradient to $\mu_\theta(x_t, t)$ for the L_{vib} term such that L_{vib} only influences $\Sigma_\theta(x_t, t)$.

B SpacePhish Experiment

We follow an existing method (SpacePhish [2]) to generate adversarial phishing webpages that were originally designed to bypass HTML-based detectors. Then we test these adversarial phishing pages against a logo-based visual phishing detector, PhishIntention [10]. The goal is to understand if PhishIntention can detect such adversarial phishing webpages.

First, we follow the same method described in [2] to generate adversarial webpages. This process starts by first training a target phishing detector with a Random Forests (RF) classifier using HTML features. The classifier is trained with a dataset of 30K websites (15K benign and 15K phishing) [5]. The classifier uses 80% of the data for training and the rest 20% for testing. After training, the classifier is very accurate, with a true positive rate (TPR) of 0.98 and a false positive rate (FPR) of 0.04. Then, we randomly selected 1k phishing websites from the testing set for adversarial manipulation. Utilizing the methods proposed in [2] (i.e., adding links or wrapping all links with “onclick”), we generate 160 adversarial webpages. These adversarial webpages can bypass the RF classifier with 100% evasion success rate.

Brand	# Bypass VisualPhishNet	Rate
AT&T	81 (81)	1.00
Instagram	199 (199)	1.00
LinkedIn	6229 (6,249)	1.00
Yahoo	35 (39)	0.90
eBay	156 (183)	0.85
CIBC	102 (121)	0.84
DHL	108 (194)	0.56
Amazon	10,401 (37,970)	0.27
Dropbox	6,787 (29,773)	0.23
Chase	4,157 (18,601)	0.22
BOA	1,593 (13,479)	0.12
Google	14 (121)	0.12
Outlook	107 (11,387)	0.01
PayPal	72 (6,383)	0.01
Netflix	0 (80)	0.00

Table 1: Transferability to VisualPhishNet (All Logos)— Number of adversarial phishing pages (generated with a local PhishIntention model) that successfully transfer to bypass another phishing detector VisualPhishNet. Three brands from Table ?? (Comcast, Spotify, DocuSign) are omitted because they are not included in [1].

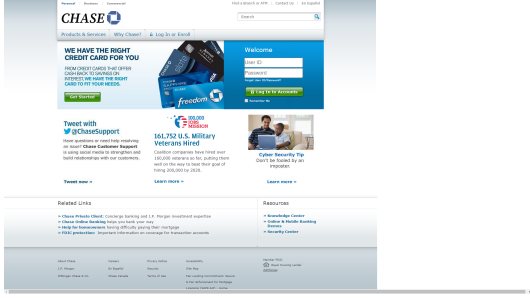
Brand	# Bypass (# tested)	Rate
Amazon	40 (40)	1.00
AT&T	18 (18)	1.00
BOA	40 (40)	1.00
Chase	40 (40)	1.00
CIBC	26 (26)	1.00
DHL	41 (41)	1.00
DocuSign	36 (36)	1.00
Dropbox	40 (40)	1.00
eBay	37 (37)	1.00
Google	25 (25)	1.00
Netflix	16 (16)	1.00
Outlook	40 (40)	1.00
PayPal	40 (40)	1.00
Spotify	40 (40)	1.00
Yahoo	8 (8)	1.00
LinkedIn	39 (40)	0.98
Comcast	27 (29)	0.93

Table 2: Gradient-Masking Defense—Number and rate of adversarial pages that bypass PhishIntention after gradient masking.

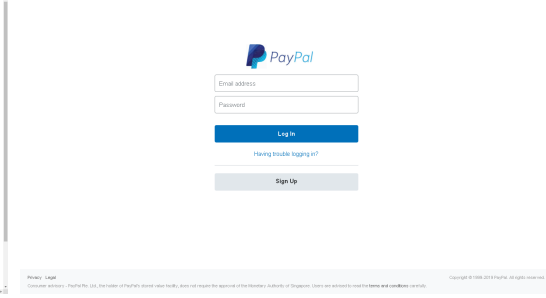
Then, we run PhishIntention [10] on these 160 adversarial phishing webpages, which are mimicking 34 brands in PhishIntention’s reference list (i.e., protected brands). Only 27% of these samples managed to bypass the detection of PhishIntention. The result indicates that PhishIntention has good resistance against such adversarial phishing pages that focus on manipulating HTML (provided that the corresponding webpages are included in the reference list!).

C Gradient-Masking

Another defense method proposed by prior work [9, 10] is gradient masking. The goal is to modify the phishing detection model such that it becomes more difficult for optimization-based algorithms to generate adversarial examples. More specifically, by changing the activation function of the phish-

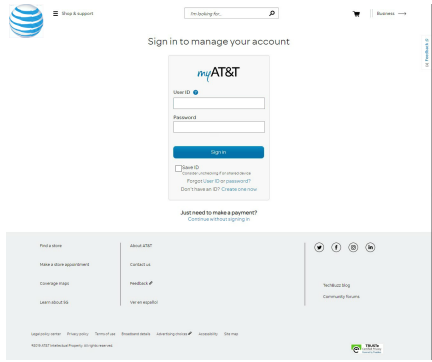


(a) Chase Bank

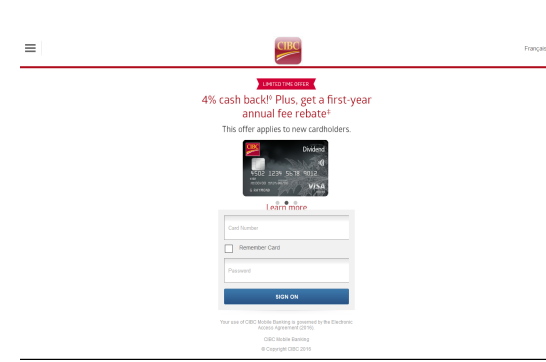


(b) PayPal

Figure 4: Phishing Page with Image-Text Logo Attack—Successful phishing pages that bypass PhishIntention, with the combined adversarial image-text logos inserted into the target webpages of CHASE Bank and PayPal.



(a) AT&T



(b) CIBC Bank

Figure 5: Phishing Page with Image Logo Attack—Successful phishing pages that bypass PhishIntention, with adversarial image logos inserted into the target webpages of AT&T and CIBC Bank.

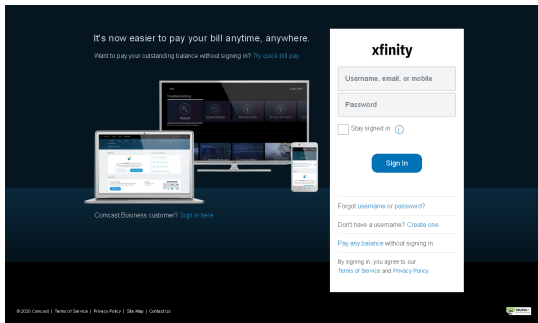


Figure 6: Phishing Page with Text Logo Attack—Successful phishing pages that bypass PhishIntention, with attack fonts inserted into the target screenshots of Comcast page.

ing detector to a “smooth” version (similar to defensive distillation [4]), we expect to increase the difficulty of searching the gradients for computing adversarial examples. We test the gradient-masking method used in [9], using a *Step ReLu* function: $f(x) = \max(0, \alpha \cdot \lceil \frac{x}{\alpha} \rceil)$. We replace all the ReLu functions in the OCR-Siamese model with the Step ReLu function. We then test this improved version of PhishInten-

tion end-to-end on webpage levels using the same 20% of testing data used in §???. The results are shown in Table 2.

We find that gradient-masking is not effective in defending against LogoMorph. Most of our adversarial webpages can still bypass PhishIntention. The reason is that our adversarial logos “attack” the similarity function to compare the input and reference logos. It relies on font manipulation and a diffusion model to introduce *large* and yet semantic-preserving perturbations. Gradient masking, on the other hand, is designed to defend against classification-gradient-based adversarial examples (that introduce small noises), which is not well suited to defend against our attack. Noteworthy is that also Phish-GAP [8] defeated detectors using *Step ReLu*.

D Evasion Rate vs. Similarity Threshold

In this section, we further explore the trade-off between the similarity threshold and evasion success rate. For this analysis, we still focus on logos having $\text{Sim} \geq 0.6$ (to make sure the logos are of reasonable visual quality). More specifically, we collect all the candidate image- and text-logos (from §?? and

User Study	Adv. Attack	Baseline	Total
Gender			
Male	52	24	76
Female	47	23	70
Non-binary / third gender	1	3	4
Age			
18-29	23	15	38
30-39	32	14	46
40-49	18	5	23
50-59	15	9	24
60-69	8	5	13
70 or above	4	2	6
Education			
Some high school or less	0	1	1
High school diploma or GED	11	1	12
Some college, but no degree	22	11	33
Associates or technical degree	15	8	23
Bachelor's degree	35	22	57
Graduate or professional degree	17	7	24
Phish knowledge			
Yes	68	36	104
No	29	13	42
Prefer not to say	3	1	4
Technical Background in Computer Science / Engineering			
Yes	14	10	24
No	84	39	123
Prefer not to say	2	1	3
Technical Background in Computer Security			
Yes	3	2	5
No	96	46	142
Prefer not to say	1	2	3
Total	100	50	150

Table 3: Demographics of our main user studies (in §??).

§?? in Table ?? and Table ??)¹, sort them based on their similarity, and then group them into three buckets: $0.6 \leq \text{Sim} < 0.7$, $0.7 \leq \text{Sim} < 0.8$, and $0.8 \leq \text{Sim} < 0.87$. Then for each bucket, we add the adversarial logos to their corresponding webpages and test the webpages against PhishIntention end-to-end pipeline. The success rate of these adversarial logos is shown in Figure 7. Overall, the result confirms the expected trade-off: choosing logos with a lower similarity improves the evasion success rate. However, in the ancillary user study (Appendix F), we show that the cost of using lower-similarity logos is on users: from users’ perspectives, such logos tend to also have a lower resemblance to the target brands. In practice, attackers should prioritize higher-quality logos (high similarity) as long as they can bypass the detector. However, given that attackers may not have perfect knowledge of the target system (and hence do not know the exact threshold), attackers may consider using logos that are slightly below the “anticipated” threshold (we use this intuition for choosing some of the logos of our blackbox experiment in §??).

¹We did not combine the text and image parts at the logo level; as discussed in §??, they are directly combined at the webpage level.

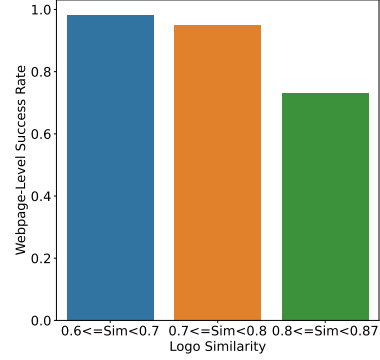


Figure 7: **Evasion Rate vs Logo Similarity**—We group the adversarial logos into 3 buckets based on their similarity, and report the aggregated webpage-level evasion rate against PhishIntention.

E Additional Analysis of Main User Study

We perform an additional analysis of the user study data collected in §??. Table 3 shows the demographics and background information of participants. We want to examine whether users’ phishing knowledge, technical background in Computer Science and Engineering (CSE), and their brand familiarity would affect their detection performance on *adversarial phishing pages*. We did not perform the analysis for “technical background in computer security” because the vast majority of users (95%) do not have such background.

Does Expertise with IT Affect LogoMorph? We report the overall accuracy, true positive rate (TPR), and true negative rate (TNR) for users in different groups in the *adversarial study* in Table 4. Users are divided into different groups based on whether they have phishing knowledge and a technical background in CSE. We find that the technical background in CSE does not affect users’ detection rate of adversarial phishing pages (TPR=0.59 for users with CSE background, and TPR=0.60 for users without CSE background). Although the TNR of users without CSE background (TNR=0.79) is slightly higher than that of users with CSE background (TNR=0.73), this difference is not statistically significant ($p=0.1$, $\chi^2=2.70$). The same conclusion holds for users with phishing knowledge: whether or not users have phishing knowledge does not statistically affect their performance on adversarial phishing pages (TPR). However, the result of our Chi-squared statistic test indicates that people with phishing knowledge are more likely to correctly identify benign webpages ($p=0.02$, $\chi^2=5.29$).

Impact of Brand Familiarity. We also analyze the correlation between users’ detection performance and their brand familiarity. In our adversarial study (§??), we captured 1,800 responses, and 85.7% (1,542) indicated that users know the brands. Users knowing the brand get TPR=0.60 and TNR=0.81; in contrast, users who do not know the brand get

Users technical background	Accuracy	TPR	TNR
With tech background of CSE	0.66	0.59	0.73
Without tech background of CSE	0.70	0.60	0.79
With phish knowledge	0.70	0.59	0.81
Without phish knowledge	0.67	0.60	0.74

Table 4: **Results of Users with Different Technical Background**—We report the accuracy, TPR and TNR for people with/without phishing knowledge and technical background in Computer Science and Engineering in the *adversarial study*.

TPR=0.54 and TNR=0.66. Based on the Chi-squared statistic test, no evidence suggests that brand familiarity would significantly affect users’ performance on detecting adversarial phishing webpages ($p=0.20$, $\chi^2=1.66$). However, users who know the brand are more likely to correctly identify legitimate webpages ($p<0.001$, $\chi^2=13.12$).

Discussion. The result indicates that the effectiveness of LogoMorph is not affected by users’ technical background in CSE, phishing knowledge, or brand familiarity. However, phishing knowledge and brand familiarity can help users recognize legitimate web pages. This finding is interesting in light of a recent work [6], wherein a user study ($n=126$) was carried out to assess the ability of individuals to recognize “adversarial webpages from the wild web” (i.e., without using LogoMorph). In this study, users with expertise in IT performed better than “amateurs”. It can hence be said that LogoMorph is a subtle threat, since it can fool users independently from their background. In contrast, other forms of adversarial webpages, such as those employed by real phishers [6], can be more easily spotted by IT experts.²

F Ancillary User Study (Logo-Level)

In our main paper, we carried out a user study (§??) whose goal was assessing if users were able to identify phishing webpages that included an adversarial logo crafted via LogoMorph. Here, we carry out an orthogonal user study, whose goal is to determine the extent to which the logos crafted via LogoMorph can retain the semantics of the targeted brand. Specifically, we show various logos crafted via LogoMorph which achieve a certain similarity (according to PhishIntention), and ask users to rate how much such logos “resemble” the targeted brand.

²Abundant prior work has carried out user studies on phishing (for a recent survey, see [3]), but the work by Draganovic et al. [6] is the only one that considers adversarial webpages and collects information on the IT expertise of their participants (the latter was not provided in PhishGAP [8]).

F.1 Questionnaire Design

This user study follows a similar setup (in terms of platform and structure) to the one in our main paper (§??). However, here, we only display the logo in isolation—whereas in our main paper we showed the entire webpage. **(I)** We begin by informing users that the study involves “understanding the human’s ability to identify website brands”.³ We then explain the participants’ rights and the instructions of our survey. **(II)** Each participant is then shown a total of 52 adversarial logos. All logos used in the study have bypassed the detection threshold of PhishIntention (i.e., $\text{Sim}<0.87$). Specifically, we show 3 logos for each of the 18 brands listed in Table ?? . These 3 logos pertain to three categories:

- *Very Similar*, i.e., adversarial logos crafted via LogoMorph with $0.8 \leq \text{Sim} < 0.87$ according to PhishIntention;
- *Somewhat Similar*, i.e., adversarial logos crafted via LogoMorph having $0.7 \leq \text{Sim} < 0.8$;
- *Barely Similar*, i.e., adversarial logos crafted via LogoMorph having $0.6 \leq \text{Sim} < 0.7$.⁴

We randomize the order of the logos for each participant to reduce bias. **(III)** Each logo (and corresponding questions) is shown in a dedicated section of our questionnaire. Upon reaching any new section, the user is first asked “Do you know the brand b ”, where b is the brand targeted by the logo, and the answers are “Yes, I know and visit it often” / “Yes, I know it but I do not use it often” / “No, I have never heard of it.” Then, we show the logo, and ask “Does this logo resemble the brand?” and the answer is picked from a five-point Likert scale: 1=not resemble; 2=probably not resemble; 3=undecided; 4=probably resemble; 5=resemble. Also, we specifically write “Please do not look up the logo” to prevent users from checking the Internet and biasing their responses. **(IV)** At the end of the questionnaire, we ask an attention check question, showing the image of a social network and inquiring whether it represents a bank. Our study does not collect any personally identifiable information, and our participants’ identities are anonymous and participation was voluntary and could withdraw their data at any time (during and after the study). On average, each participant spends ≈ 7 minutes.

F.2 Results (do our logos resemble the brand?)

We collected 5,200 valid responses from 100 participants.

Main Findings. To provide an accurate representation, we only consider the responses of users who “know” the brand

³We note that, differently from what we did in the user study in our main paper, we are not priming the users here: we *never* mention the term “phishing” in our questionnaire. This is to ensure that the users are not biased to think that a given logo may be related to a malicious purpose, thereby increasing the quality of user responses (i.e., this study is *not* about phishing).

⁴We do not have any logo for PayPal and LinkedIn within this range, which is why we have 52 logos (given by: $3 \times 18 - 2$).

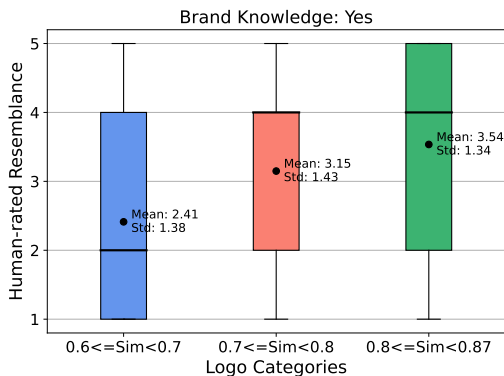


Figure 8: Main Results of our Logo-level User Study— Distribution of the human-rated resemblance (higher is better) across adversarial logos (targeting the 18 brands in Table ??) crafted with LogoMorph. Each bin corresponds to logos that achieve a given similarity (according to PhishIntention) with the “original” logo of the targeted brand. We only report the answers of participants who are familiar with the targeted brand.

(i.e., they answered “Yes” to the brand-knowledge question). These responses count for about 90% of all valid responses. Figure 8 shows the aggregated “human-rated resemblance” (y-axis) across all logos within a given category (boxplots). We can see that the human-rated resemblance is highest for logos within the *very similar* category: the average rating is 3.54, and the standard deviation is 1.34, and a t-test confirms that such a rating is statistically significantly ($p < 0.001$) superior than the middle-point of 3. Hence, we can claim that *logos crafted with LogoMorph having $0.8 \leq \text{Sim} < 0.87$ tend to preserve the characteristics of the targeted brand* (at least according to our participants, and w.r.t. our considered 18 brands). This suggests that, from an attacker’s viewpoint, it is wise to use LogoMorph by picking adversarial logos having $0.8 \leq \text{Sim} < 0.87$ to embed in the web pages. (Indeed, this is what we have done in the user study discussed in §??.) For the *somewhat similar* logos ($0.7 \leq \text{Sim} < 0.8$), the average rating is 3.15 (i.e., also above the middle-point 3), indicating a positive resemblance of the target brand. Interestingly, however, for some users, even logos having a lower similarity can resemble (to some extent) the targeted brand, as shown by the wide whiskers in the blue ($0.6 \leq \text{Sim} < 0.7$) boxplot.

Per-Brand Analysis. We find it instructive to analyze the human-rated resemblance on a brand-by-brand basis. To this purpose, we show in Figure 9 the human-rated resemblance of the *best* category of logos for each brand. We can see that, in general, the human-rated resemblance is highest for the logos having $0.8 \leq \text{Sim} < 0.87$. However, for three brands (Google, Yahoo, Instagram), our participants rated logos having $0.7 \leq \text{Sim} < 0.8$ with a higher resemblance than those having $0.8 \leq \text{Sim} < 0.87$. We find this result intriguing, suggesting that *the similarity metric used by PhishIntention (to compare any logo with a “benign” logo of a given brand) may not*

fully align with the way human users perceive whether a logo resembles its intended brand. This result may also be inspiring for future research, since it may drive the development of novel metrics to compute the similarity. Nevertheless, another interesting result shown in Figure 9 is that, even though the overall human-rated resemblance (for $0.8 \leq \text{Sim} < 0.87$) has an average of 3.54 (see Fig. 8), the adversarial logos of some brands exhibit a remarkably high resemblance. For instance, the average human-rated resemblance for AT&T is 4.86, whereas the one for eBay and PayPal is above 4.

Knowledge of the Brand. We also report the results of our user study by considering our full responses (i.e., without removing the responses of users who do not know the brand), shown in Figure 10; and those of users who do not know the targeted brand, shown in Figure 11. For Figure 11, the green and orange boxplots each include 171 responses and the blue one has 170. For Figure 10 the green and orange boxplots each have 1,800 responses and the blue one has 1,600. Given that only few users did not know our brands (which we chose among those being most popular in the U.S.—*thereby proving that our choice of brands to consider in our main paper is well-founded*), there is a negligible difference between Figure 8 and Figure 10. Finally, by turning the attention at Figure 11, we see that these users were mostly unsure (which makes sense, since they did not know the brand).

Takeaway. From this user study, we have learned three lessons. First, as long as the adversarial logos can bypass the detector, choosing logos with a higher similarity is the most sensible way to use LogoMorph. Second, the similarity metric used by PhishIntention may not always align with the way human users perceive whether a logo resembles its brand. Third, more than 90% of our responses correspond to brands known by our participants—thereby justifying our selection of the brands covered in our main paper.

Comparison with PhishGAP [8]. We stress that all these results only capture how well a given user rates the *logo in isolation*—and cannot be used to determine if users are “fooled” by such logos (since the logos must be put in webpages). This is why, in our main paper (§??), we have carried out another user study showing entire webpages to our participants. We note that PhishGAP [8] also carried out user studies at the logo level. Specifically, the authors of [8] assessed how well their proposed adversarial logos could fool humans via a questionnaire containing various pairs of logos: an “adversarial logo” (crafted via PhishGAP [8]) and an “original” logo. Participants (i.e., 30 university students, and 287 MTurk workers) were asked to rate how similar the two logos were: ideally, the more similar they were, the more likely that the “adversarial logo” could have fooled a human user. The results showed that, on average, users rated the pairs above the middle-point, demonstrating that PhishGAP’s logos can deceive humans. However, the major limitation of this study is that it only focuses on the logo level: as we showed

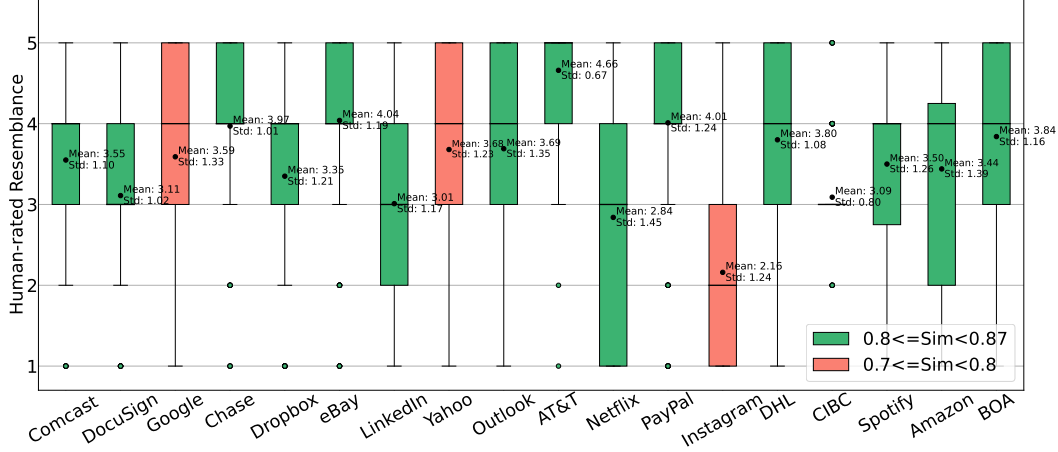


Figure 9: **Per-brand Results of our Logo-level User Study**—We show, for each targeted brand, the “best category” of adversarial logos, i.e., those with the highest human-rated resemblance. For three brands (Google, Yahoo, Instagram), our participants deem that logos with a lower similarity (according to PhishIntention) have a better resemblance with the targeted brands.

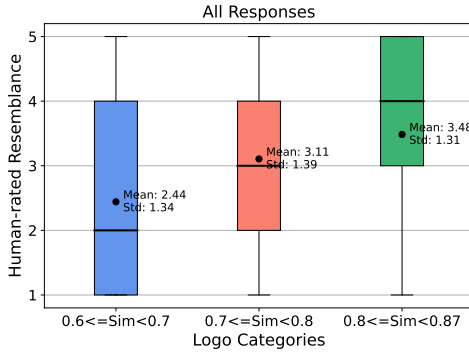


Figure 10: **All Responses of Our Logo-Level User Study**—We consider the responses of those users who know the targeted brands, and those who do not (combining Figures 8 and 11).

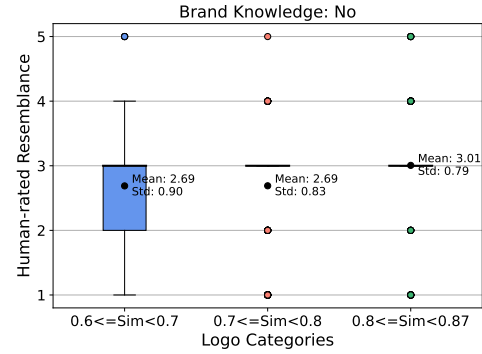


Figure 11: **Responses from of Users Who Do Not Know the Brands**—Less than 10% of our responses are made by users who do not know the brand represented by the (adversarial) logo.

in §??, only 5.5% of the logos generated by PhishGAP can bypass PhishIntention end-to-end. Hence, despite showing that the logos of PhishGAP can deceive humans, human users would not see these logos in the first place since the majority would be blocked by Phishintention. In contrast, our main user study (in §??) is done by selecting adversarial webpages that bypassed Phishintention end-to-end, thereby allowing to gauge the real threat of LogoMorph to human users.

G Extended Logo Experiments

Insofar, our evaluation revolved around a subset of 18 brands included in PhishIntention’s dataset. Here, we expand the our assessment to cover more brands, demonstrating that LogoMorph is broadly applicable to a variety of brands and their logos. We regard the 18 brands used in §?? as the “*Main Set*”, and then introduce an “*Expanded Set*” of 92 brands (110 brands in total). For the Expanded Set, we first include the top 50 most impersonated phishing attack brands [12] (ex-

cluding those that are not in PhishIntention’s protected brand list). Then we fill out the rest of the slots based on PhishIntention’s protected brand list. Among the 92 brands, 17 have image-only logos⁵, and 75 have text-only logos.

Attacking the Main and Expanded Sets. We replicate the white-box experiment in §?? and §?? to cover our 110 brands. For image logos, we use LogoMorph to generate 200 adversarial logos for each brand. For text logos, we follow the selection procedure (§??) by first generating 2,556 candidate fonts for each brand, and then selecting the top 200 fonts that bypass the similarity threshold to generate logos. The results are presented in Figure 12. The 110 brands are sorted based on the number of successful adversarial logos identified by LogoMorph against PhishIntention’s logo discriminator (logo-level). The orange color represents the Man Set and the blue color represents the Expanded Set. The result shows that the 18 brands used in the main paper have sufficient diversity, cov-

⁵For brands like Capital One Bank, differentiating the text and image parts of the logo is challenging, and thus we regard the logo as image-only.

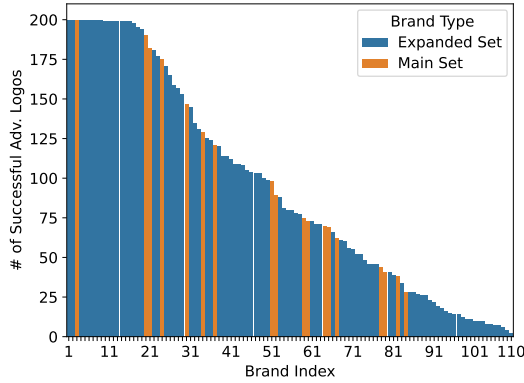


Figure 12: **Successful Adv. Logos Per Brand (110 Brands)** —We sorted the 110 brands on the x-axis based on the number of successful adversarial logos identified by LogoMorph (out of 200 candidate logos tested against PhishIntention).

ering both easy-to-attack and hard-to-attack brands. For 50 of these brands, LogoMorph identified more than 100 successful logos. For all 110 brands (100%), LogoMorph identified at least one viable attack logo for the attacker (from 200 candidate logos). In a white-box setting, the attacker would hence always find a suitable logo for each of these 110 brands.

Brands with “Few-shot” Logos. Intuitively, the attack difficulty may be related to the number of different logos each brand has. For example, brands with more logos may allow PhishIntention to learn a more robust embedding for the brand (i.e., harder to attack). We plot the result for the 110 brands in Figure 13: each point represents a given brand, the y-axis reports the number of successful logos, and the x-axis the number of logos included in PhishIntention’s training set for that brand. First, we see that brands with fewer logos (e.g., less than 10) are not necessarily easier or harder to attack since the number of successful attack logos varies between 0–200. Second, brands with more logos (e.g., more than 25) seem to be harder to attack. We run a Pearson correlation analysis on these two variables and resulting $r = -0.12$, $p = 0.21$. While it exhibits a negative correlation, the high p value (> 0.05) means the observed correlation is not statistically significant.

Takeaway. There is no significant evidence supporting the fact that the number of logos included in the training set for a brand is a significant factor in predicting the effectiveness of using LogoMorph against such a brand.

References

- [1] ABDELNABI, S., KROMBOLZ, K., AND FRITZ, M. Visualphishnet: Zero-day phishing website detection by visual similarity. In *Proc. of CCS* (2020).
- [2] APRUZZESE, G., CONTI, M., AND YUAN, Y. Spacephish: The evasion-space of adversarial attacks against phishing website detectors using machine learning. In *Proc. of ACSAC* (2022).

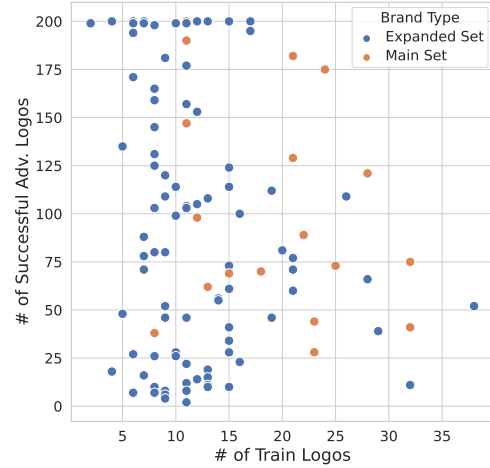


Figure 13: **# of Successful Adv. Logos vs. # Training Logos in PhishingIntention** —We show a scatter plot to examine the relationship between the number of training logos each brand has for PhishingIntention (X) and the number of successful adversarial logos identified by LogoMorph (Y).

Brand	# Success Logos	Rate
PayPal	165	0.83
Amazon	128	0.64
LinkedIn	128	0.64
BOA	93	0.47
DHL	81	0.41
ATT	79	0.40
Chase	78	0.39
Spotify	72	0.36
Dropbox	69	0.35
CICB	63	0.32
Outlook	30	0.15

Table 5: **Logo-Level Results for Image Logos**—Logos that bypass the *local surrogate* (out of 200 tested images per brand). We choose these logos to run a black-box attack against PhishIntention.

- [3] BAKI, S., AND VERMA, R. M. Sixteen years of phishing user studies: What have we learned? *IEEE Transactions on Dependable and Secure Computing* 20, 2 (2022), 1200–1212.
- [4] CARLINI, N., AND WAGNER, D. A. Towards evaluating the robustness of neural networks. In *Proc. of IEEE SP* (2017).
- [5] CHIEW, K. L., CHANG, E. H., TAN, C. L., ABDULLAH, J., AND YONG, K. S. C. Building standard offline anti-phishing dataset for benchmarking. *International Journal of Engineering & Technology* (2018).
- [6] DRAGANOVIC, A., DAMBRA, S., IUIT, J. A., ROUNDY, K., AND APRUZZESE, G. “do users fall for real adversarial phishing?” investigating the human response to evasive webpages. In *APWG 2023 eCrime Symposium* (2023).
- [7] HO, J., JAIN, A., AND ABBEEL, P. Denoising diffusion probabilistic models. In *Proc. of NeurIPS* (2020).
- [8] LEE, J., XIN, Z., SEE, M., SABHARWAL, K., APRUZZESE, G., AND DIVAKARAN, D. M. Attacking logo-based phishing website detectors with adversarial perturbations. In *Proc. of ESORICS* (2023).

- [9] LIN, Y., LIU, R., DIVAKARAN, D. M., NG, J. Y., CHAN, Q. Z., LU, Y., SI, Y., ZHANG, F., AND DONG, J. S. Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages. In *Proc. of USENIX Security* (2021).
- [10] LIU, R., LIN, Y., YANG, X., NG, S. H., DIVAKARAN, D. M., AND DONG, J. S. Inferring phishing intention via webpage appearance and dynamics: A deep vision based approach. In *Proc. of USENIX Security* (2022).
- [11] LUO, C. Understanding diffusion models: A unified perspective. In *arXiv preprint arXiv:2208.11970* (2022).
- [12] MORARU, A., AND DONAHUE, P. R. Top 50 most impersonated brands in phishing attacks and new tools you can use to protect your employees from them. <https://blog.cloudflare.com/50-most-impersonated-brands-protect-phishing>, 2023. Cloudflare.
- [13] NICHOL, A. Q., AND DHARIWAL, P. Improved denoising diffusion probabilistic models. In *Proc. of ICML* (2021).