The Hague – September 25th, 2023

European Symposium On Research In Computer Security

Attacking Logo-based Phishing Website Detectors with Adversarial Perturbations

Jehyun Lee, Zhe Xin, Melanie Ng Pei See, Kanav Sabharwal, Giovanni Apruzzese, Dinil Mon Divakaran

UNIVERSITÄT LIECHTENSTEIN

Trustwave

Acronis

NUS National University of Singapore

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

WHAT?

1. We propose a **novel attack**

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

WHAT?

1. We propose a **novel attack**

2. We show that **it works**

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

WHAT?

1. We propose a **novel attack**

2. We show that **it works**

3. …against both state-of-the-art **systems** *and* **humans**

UNIVERSITÄT
LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

WHAT?

1. We propose a **novel attack**
2. We show that **it works**
3. …against both state-of-the-art **systems** *and* **humans**

WHY?

- **Phishing** websites are everywhere

UNIVERSITÄT
LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

WHAT?

1. We propose a **novel attack**
2. We show that **it works**
3. …against both state-of-the-art **systems** *and* **humans**

WHY?

- **Phishing** websites are everywhere
- **Countermeasure**: visual similarity techniques reliant on deep learning
  - Trendy in research [7] but also deployed in practice [50]

UNIVERSITÄT LIECHTENSTEIN

[7] Abdelnabi, S., Krombholz, K., Fritz, M.: *Visualphishnet: Zero-day phishing website detection by visual similarity*. ACM CCS (2020)
[50] Apruzzese, G., et al.: *"Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice*. IEEE SaTML (2023)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

WHAT?

1. We propose a **novel attack**

2. We show that **it works**

3. …against both state-of-the-art **systems** *and* **humans**

WHY?

- **Phishing** websites are everywhere
- **Countermeasure**: visual similarity techniques reliant on deep learning
  - Trendy in research [7] but also deployed in practice [50]
- **Problem**: the security of these defenses has not been scrutinized yet
  - Especially from a "human" perspective!

UNIVERSITÄT LIECHTENSTEIN

[7] Abdelnabi, S., Krombholz, K., Fritz, M.: *Visualphishnet: Zero-day phishing website detection by visual similarity*. ACM CCS (2020)
[50] Apruzzese, G., et al.: *"Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice*. IEEE SaTML (2023)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

WHAT?

1. We propose a **novel attack**

2. We show that **it works**

3. …against both state-of-the-art **systems** *and* **humans**

WHY?

- **Phishing** websites are everywhere
- **Countermeasure**: visual similarity techniques reliant on deep learning
  - Trendy in research [7] but also deployed in practice [50]
- **Problem**: the security of these defenses has not been scrutinized yet
  - Especially from a "human" perspective!

Disclaimer:
non-technical talk!

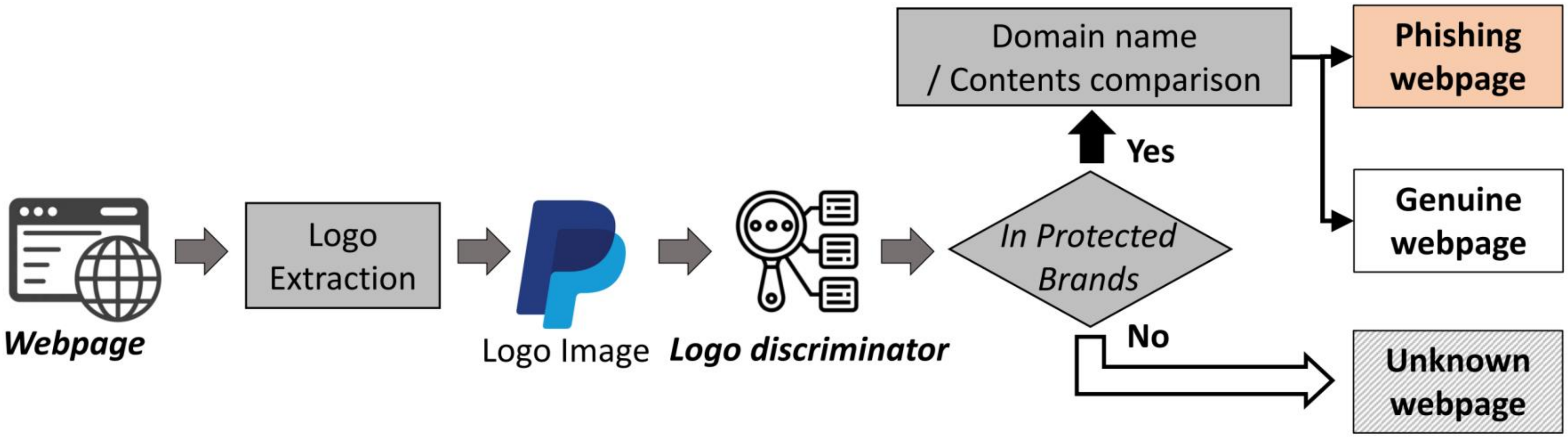
UNIVERSITÄT LIECHTENSTEIN

[7] Abdelnabi, S., Krombholz, K., Fritz, M.: *Visualphishnet: Zero-day phishing website detection by visual similarity*. ACM CCS (2020)
[50] Apruzzese, G., et al.: *"Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice*. IEEE SaTML (2023)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Logo-based Phishing Website Detection

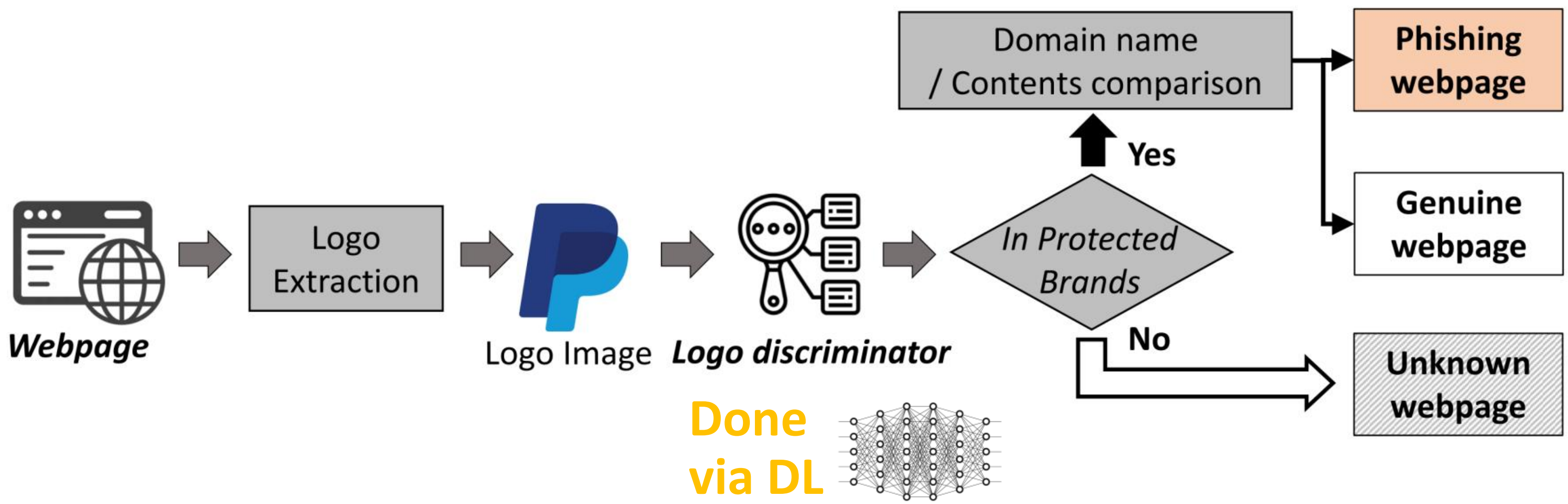
in a nutshell



Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Logo-based Phishing Website Detection

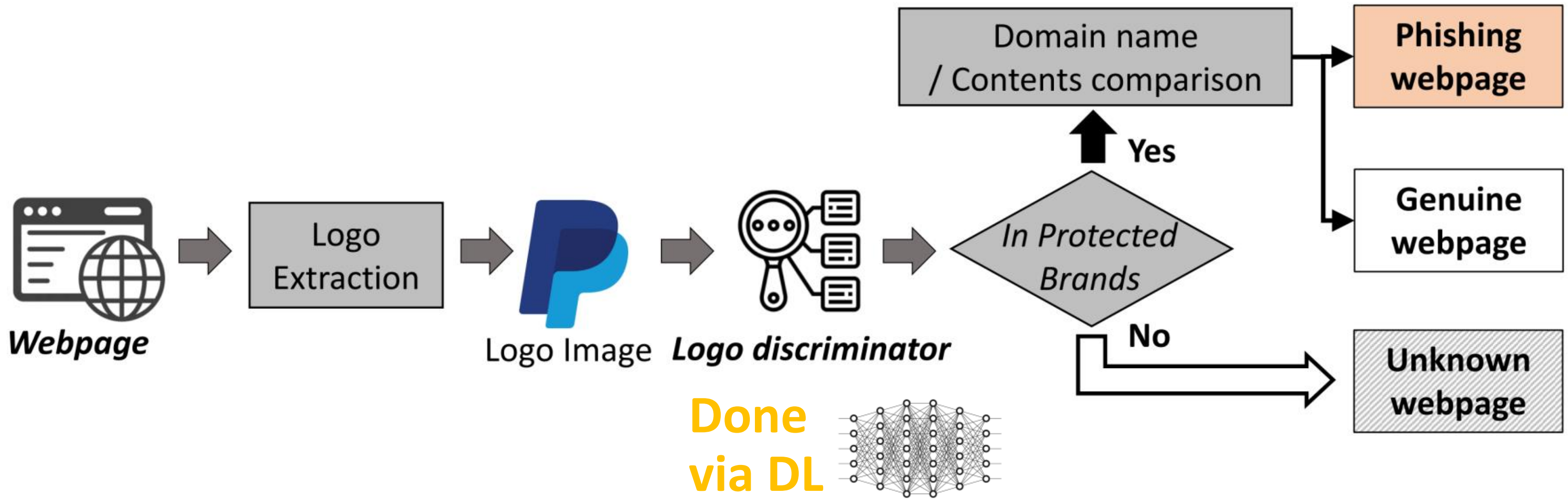
in a nutshell



Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Logo-based Phishing Website Detection

in a nutshell

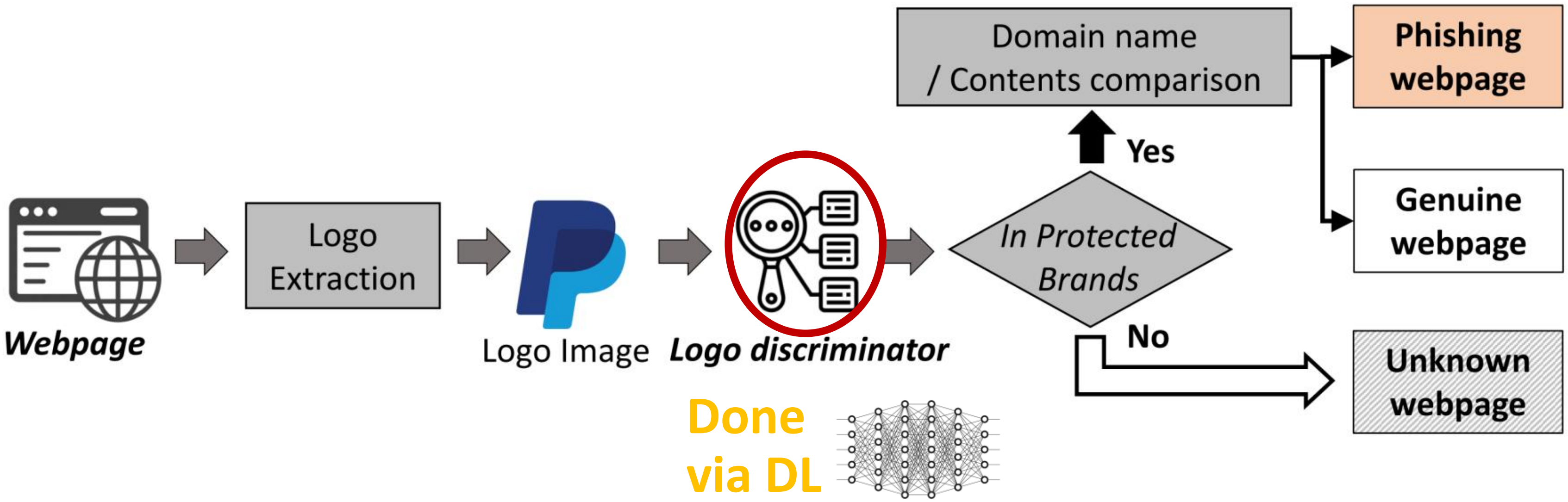


Problem: these systems are tweaked to minimize false positives.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Logo-based Phishing Website Detection

in a nutshell



Problem: these systems are tweaked to minimize false positives.

We focus on the Logo-discriminator.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Our attack: adversarial logos

Intuition: create an adversarial logo that is (i) minimally altered w.r.t. its original variant; and that (ii) misleads the logo discriminator.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Our attack: adversarial logos

Intuition: create an adversarial logo that is (i) minimally altered w.r.t. its original variant; and that (ii) misleads the logo discriminator.

1. **Knowledge:**

2. **Capabilities:**

3. **Strategy:**

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Our attack: adversarial logos

Intuition: create an adversarial logo that is (i) minimally altered w.r.t. its original variant; and that (ii) misleads the logo discriminator.

No knowledge of the DL model is required!

1. **Knowledge:**
   - the attacker expects the detector to have the "phished" brand(s) in the protected set (and that its logos are inspected)
2. **Capabilities:**
3. **Strategy:**

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Our attack: adversarial logos

Intuition: create an adversarial logo that is (i) minimally altered w.r.t. its original variant; and that (ii) misleads the logo discriminator.

1. **Knowledge:**
   - the attacker expects the detector to have the "phished" brand(s) in the protected set (and that its logos are inspected)

No knowledge of the DL model is required!

2. **Capabilities:**
   - the attacker can observe the decision of the detector
   - the attacker can manipulate their phishing webpages

The attacker can do nothing to the training data.

3. **Strategy:**

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Our attack: adversarial logos

Intuition: create an adversarial logo that is (i) minimally altered w.r.t. its original variant; and that (ii) misleads the logo discriminator.

No knowledge of the DL model is required!

1. **Knowledge:**
   - the attacker expects the detector to have the "phished" brand(s) in the protected set (and that its logos are inspected)
2. **Capabilities:**
   - the attacker can observe the decision of the detector
   - the attacker can manipulate their phishing webpages

The attacker can do nothing to the training data.

3. **Strategy:** Manipulate the logo so that the discriminator has a lower confidence → the detector will default to a "unknown webpage"

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evaluation: Discriminators

- We propose two novel methods for logo-identification: ViT and Swin
  - Both ViT and Swin leverage transformers [23, 36].

We are the first to use transformers for logo-identification (ttbook)

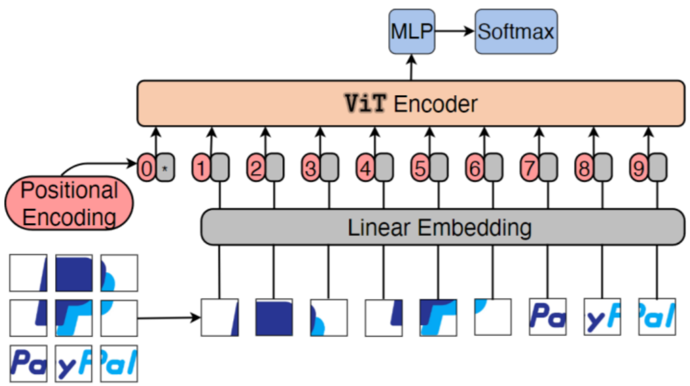


Fig. 2: ViT-based Model Architecture

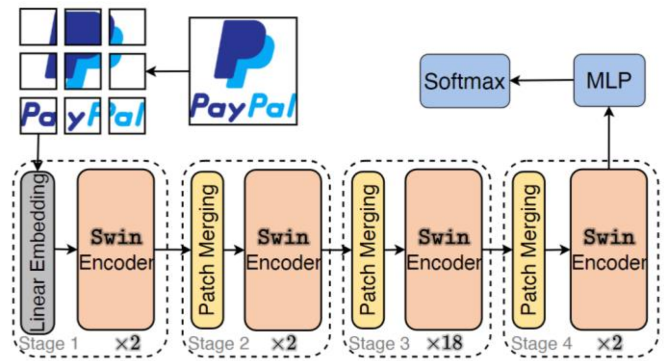


Fig. 3: Swin-based Model Architecture

[23] Dosovitskiy, A., et al.: *An image isworth 16x16 words: Transformers for image recognition at scale.* arXiv:2010.11929 (2020)
[36] Liu, Z., et al. : Swin transformer: Hierarchical vision transformer using shifted windows. IEEE/CVF ICCV (2021)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evaluation: Discriminators

- We propose two novel methods for logo-identification: ViT and Swin
  - Both ViT and Swin leverage transformers [23, 36].

We are the first to use transformers for logo-identification (ttbook)

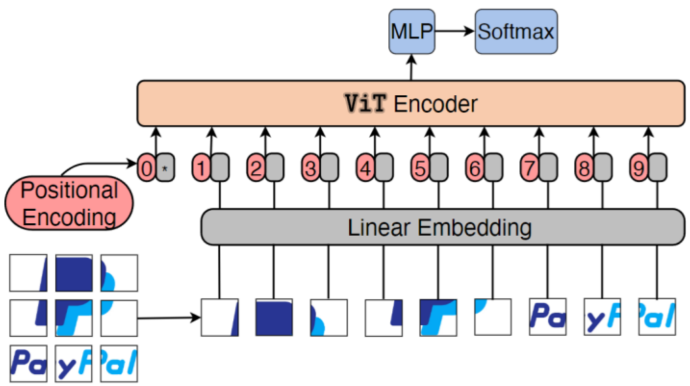


Fig. 2: ViT-based Model Architecture

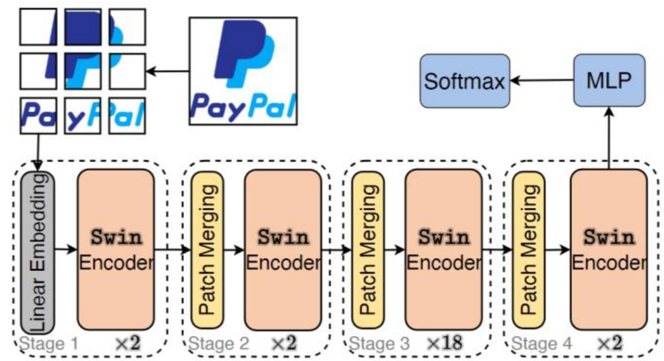


Fig. 3: Swin-based Model Architecture

- We will show that these methods reach state-of-the-art performance (currently obtained by Siamese networks [34])

UNIVERSITÄT LIECHTENSTEIN

[23] Dosovitskiy, A., et al.: *An image isworth 16x16 words: Transformers for image recognition at scale.* arXiv:2010.11929 (2020)
[36] Liu, Z., et al. : Swin transformer: Hierarchical vision transformer using shifted windows. IEEE/CVF ICCV (2021)
[34]: Lin, Y., et al.: *Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages.* USENIX Security (2021)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evaluation: Discriminators

- We propose two novel methods for logo-identification: ViT and Swin
  - Both ViT and Swin leverage transformers [23, 36].

We are the first to use transformers for logo-identification (ttbook)

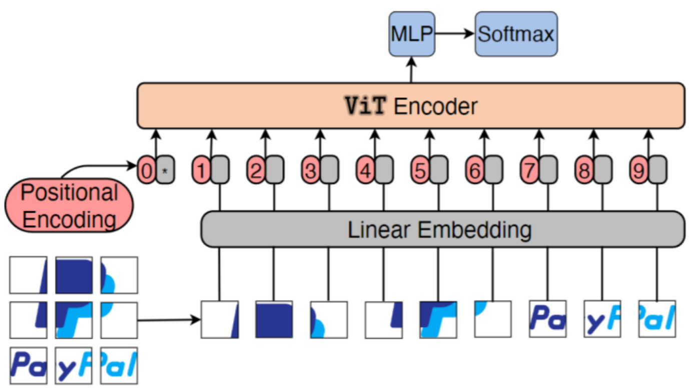


Fig. 2: ViT-based Model Architecture

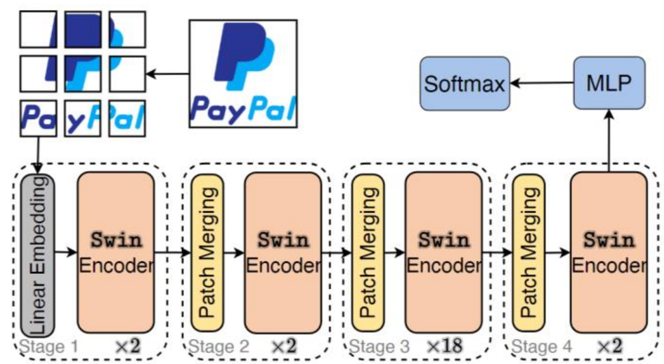


Fig. 3: Swin-based Model Architecture

- We will show that these methods reach state-of-the-art performance (currently obtained by Siamese networks [34])
  - Siamese networks have been assessed in white-box settings

...but our attacker is not a white-box!

UNIVERSITÄT LIECHTENSTEIN

[23] Dosovitskiy, A., et al.: *An image isworth 16x16 words: Transformers for image recognition at scale.* arXiv:2010.11929 (2020)
[36] Liu, Z., et al. : Swin transformer: Hierarchical vision transformer using shifted windows. IEEE/CVF ICCV (2021)
[34]: Lin, Y., et al.: *Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages.* USENIX Security (2021)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evaluation: Attack

We are inspired by "GAN"

- Our attack applies a "Generative Adversarial Perturbations" (GAP)

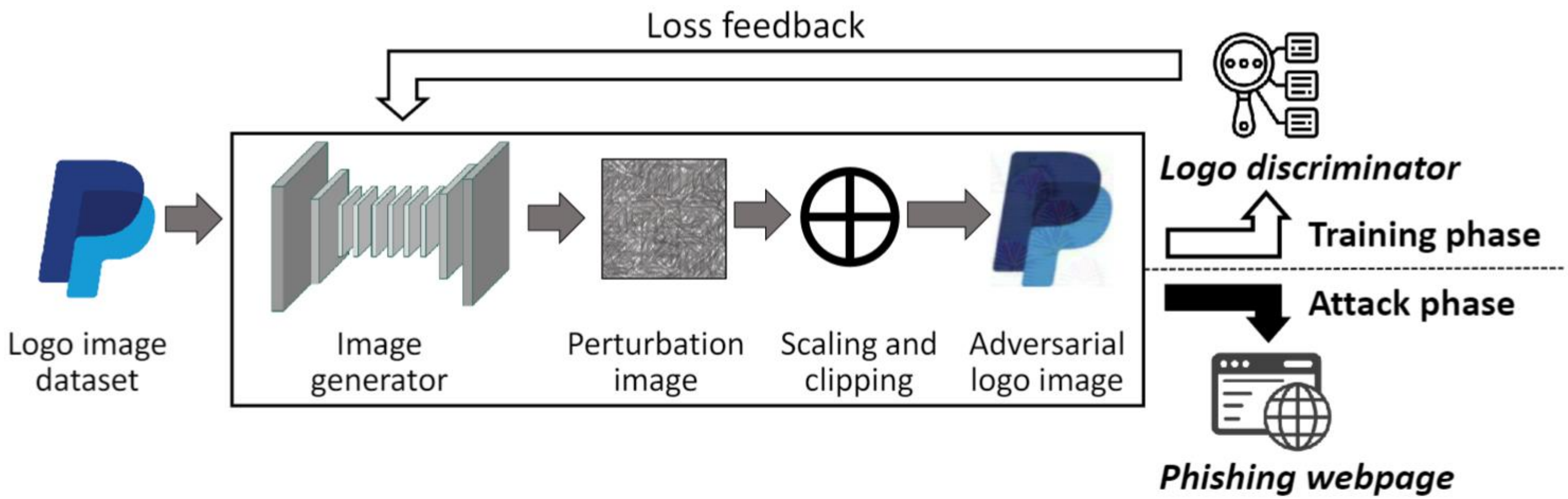


Fig. 4: Generative adversarial perturbation workflow

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evaluation: Attack

We are inspired by "GAN"

- Our attack applies a "Generative Adversarial Perturbations" (GAP)

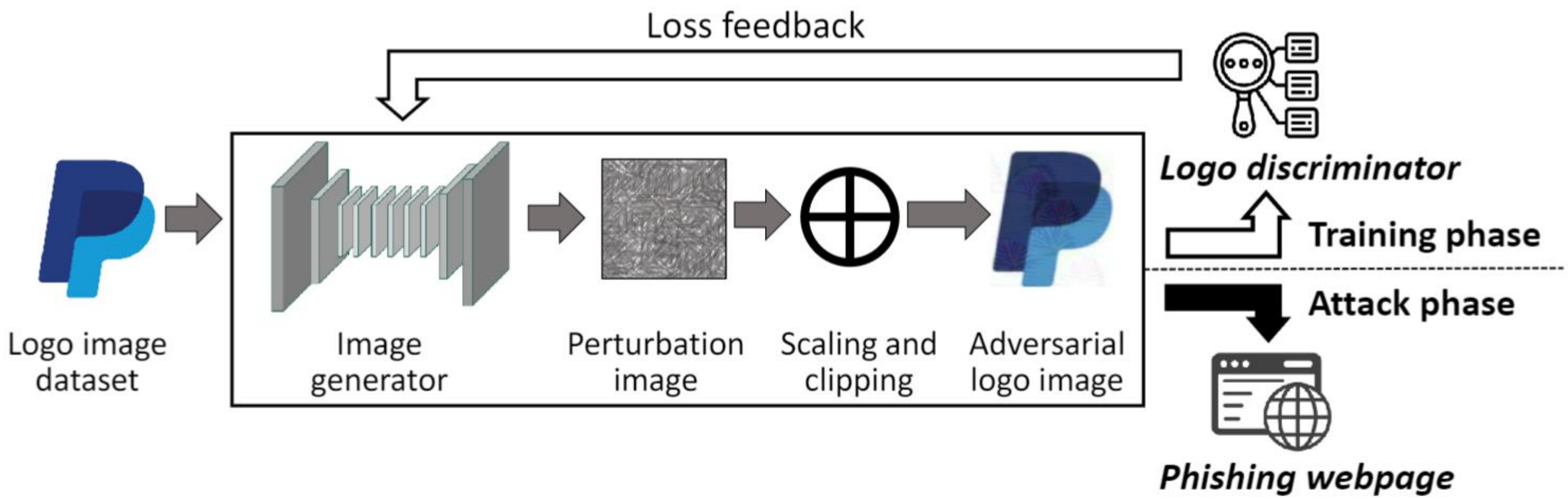


Fig. 4: Generative adversarial perturbation workflow

- The GAP automatically "learns" to craft adversarial logos that mislead the logo discriminator – while being minimally altered.

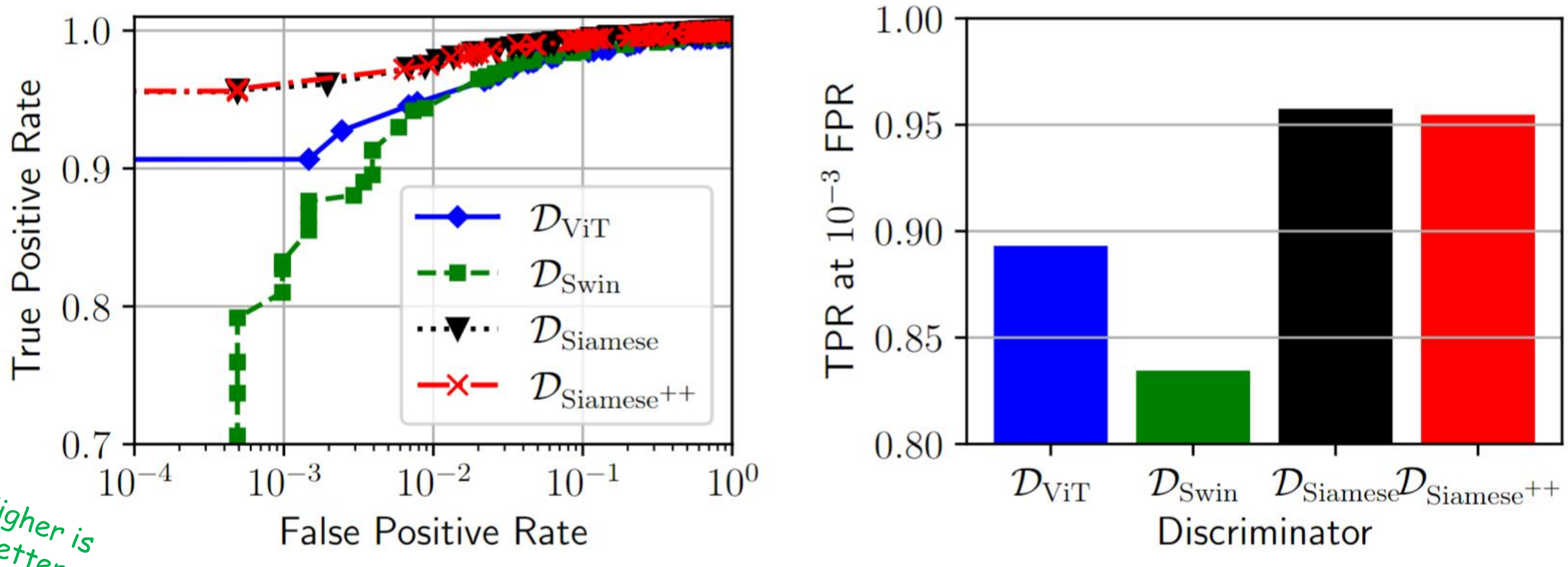
We will assess the cross-model transferability of our adversarial logos!

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Baseline

$\mathcal{D}_{Siamese++}$ is a "robust" version of Siamese networks



Higher is better

(a) ROC curves **(b)** TPR at 10^{-3} FPR

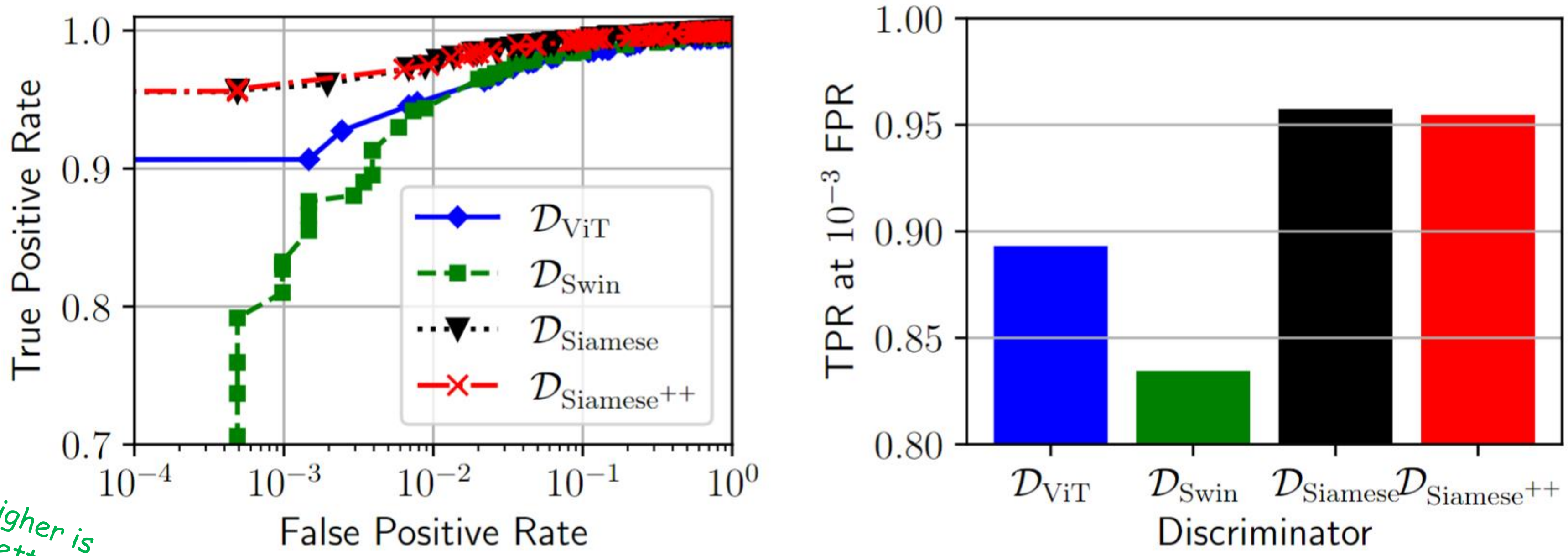
Our baselines are trained to identify 181 brands (~28k logos)

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Baseline

$\mathcal{D}_{Siamese++}$ is a "robust" version of Siamese networks



Higher is better

(a) ROC curves

(b) TPR at 10^{-3} FPR

Takeaways:

1. Our baselines "work well" (in the absence of attacks!)
2. ViT and Swin are slightly worse than Siamese…

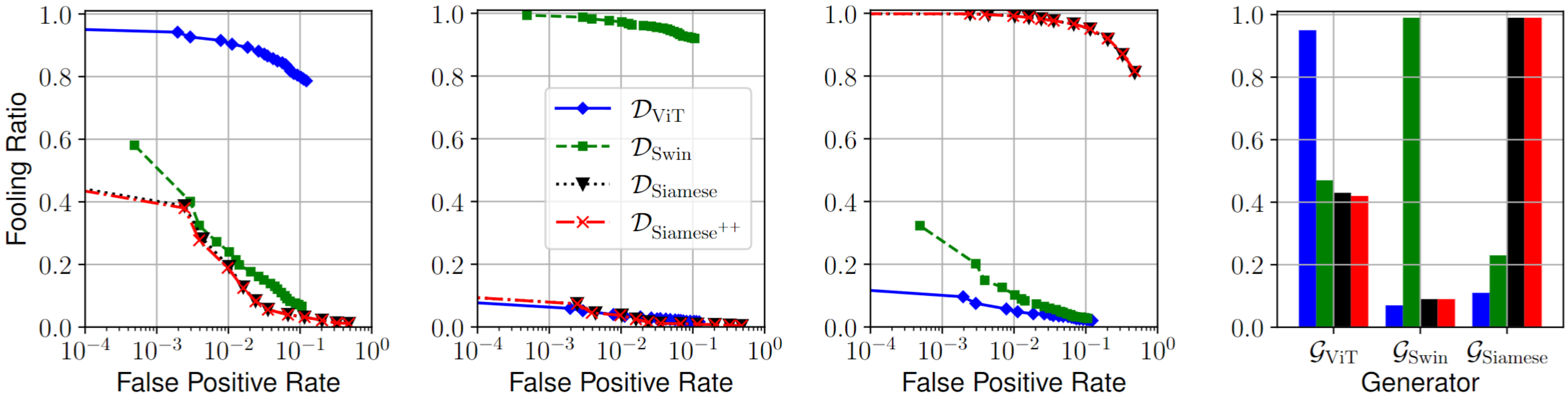
Our baselines are trained to identify 181 brands (~28k logos)

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Attack

Higher = stronger attack



(a) $\mathcal{G}_{\mathrm{ViT}}$ **(b)** $\mathcal{G}_{\mathrm{Swin}}$ **(c)** $\mathcal{G}_{\mathrm{Siamese}}$ **(d)** at 10^{-3} FPR

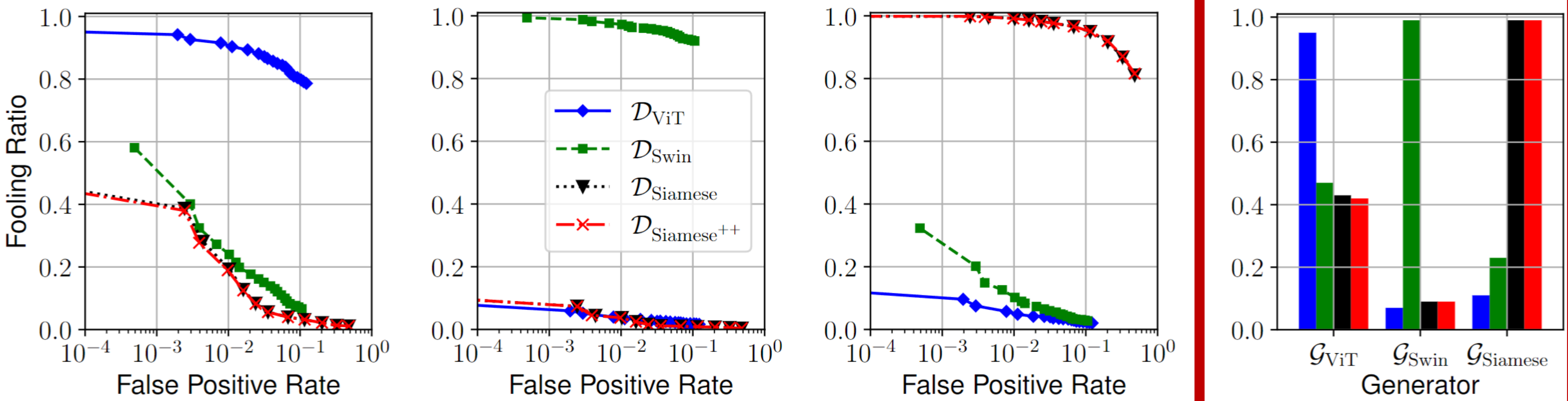
E.g.: G_{ViT} denotes the GAN trained to evade D_{ViT}

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Attack

Higher = stronger attack



(a) $\mathcal{G}_{\mathrm{ViT}}$ **(b)** $\mathcal{G}_{\mathrm{Swin}}$ **(c)** $\mathcal{G}_{\mathrm{Siamese}}$ **(d)** at 10^{-3} FPR

E.g.: G_{ViT} denotes the GAN trained to evade D_{ViT}

Takeaways:

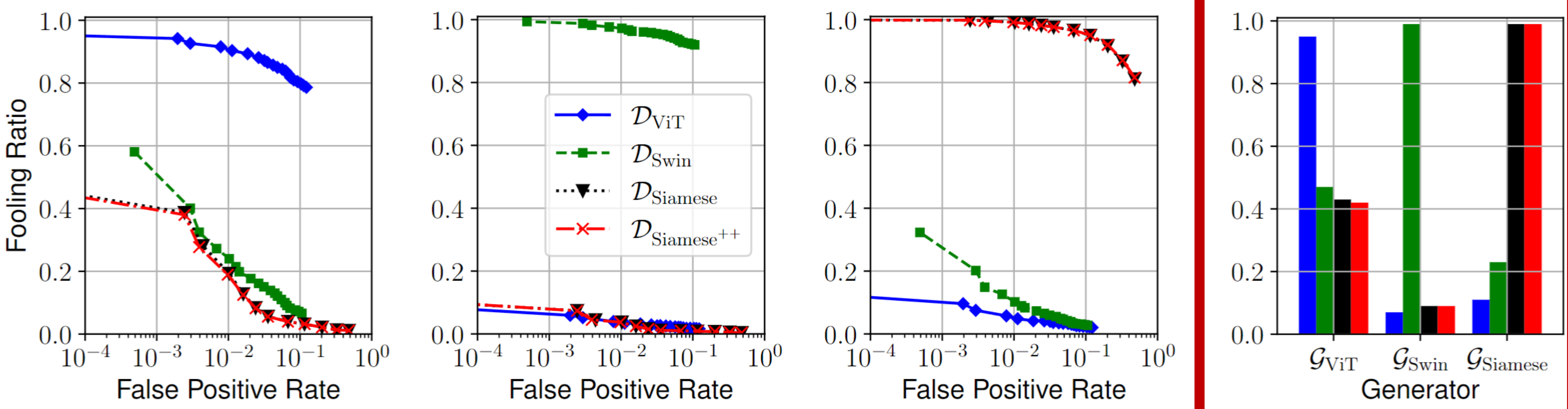
1. When the attacker and defender use the same model, the attack is ~100% effective
2. ViT is the “more robust” detector! (if the attacker is blind)

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Attack

Higher = stronger attack



(a) $\mathcal{G}_{\text{ViT}}$ **(b)** $\mathcal{G}_{\text{Swin}}$ **(c)** $\mathcal{G}_{\text{Siamese}}$ **(d)** at 10^{-3} FPR

E.g.: G_{ViT} denotes the GAN trained to evade D_{ViT}

Takeaways:

1. When the attacker and defender use the same model, the attack is ~100% effective
2. ViT is the "more robust" detector! (if the attacker is blind)

Table 1: Training time for the perturbation generators

	$\mathcal{G}_{\text{ViT}}$	$\mathcal{G}_{\text{Swin}}$	$\mathcal{G}_{\text{Siamese}}$
Avg. training time per epoch (min.)	12	23	8
No. of epochs for 0.9 fooling ratio	62	12	1
Training time for 0.9 fooling ratio (min.)	744	277	8

Training G_{ViT} is very expensive!

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Humans?

- We ask ourselves the following research question (RQ):

> Given a pair of logos (i.e., an 'original' one, and an 'adversarial' one), can the human spot any difference?

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Humans?

- We ask ourselves the following research question (RQ):

> Given a pair of logos (i.e., an 'original' one, and an 'adversarial' one), can the human spot any difference?

- We carry out two user-studies to answer our RQ:
  - **Vertical Study**: small population (N=30) of similar users; 10 questions, but different for every participant.
  - **Horizontal Study**: large population (N=287) of heterogeneous users; 21 fixed questions for all participants.

Yes, we added control questions and attention checks!

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Humans?

Look at these two images for no more than 5 seconds, and then answer the similarity question.

Logo A

Logo B

On a scale from 1 to 5, how similar do you think these two logos are? *

	1	2	3	4	5	
Very Different	○	○	○	○	○	Very Similar

Look at these two images for no more than 5 seconds, and then answer the similarity question.

Logo A

Logo B

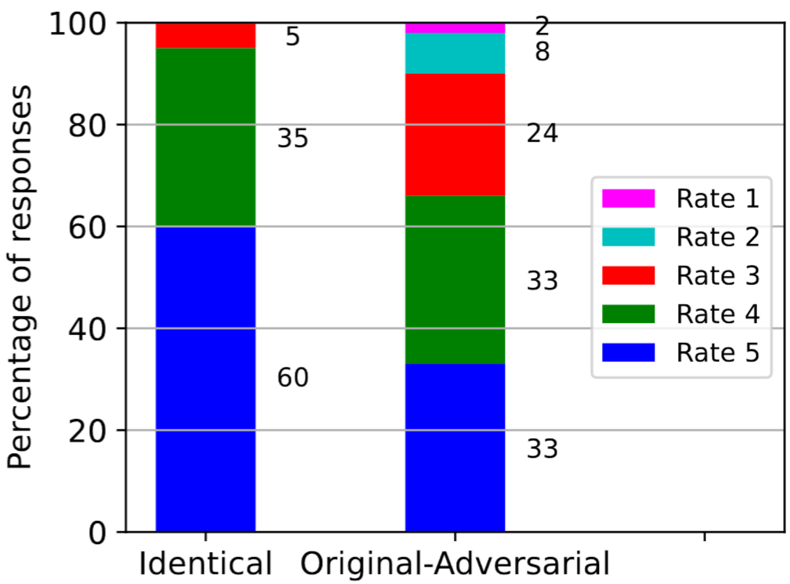
On a scale from 1 to 5, how similar do you think these two logos are? *

	1	2	3	4	5	
Very Different	○	○	○	○	○	Very Similar

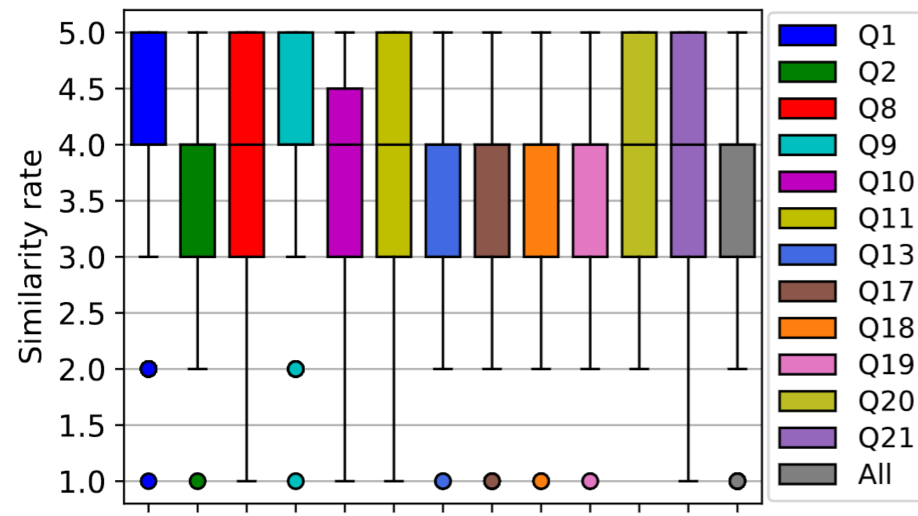
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Humans? Deceived!

- For every question, users had to say how “similar” the two logos were (5= very similar, 1= not similar at all)



(a) Vertical Study



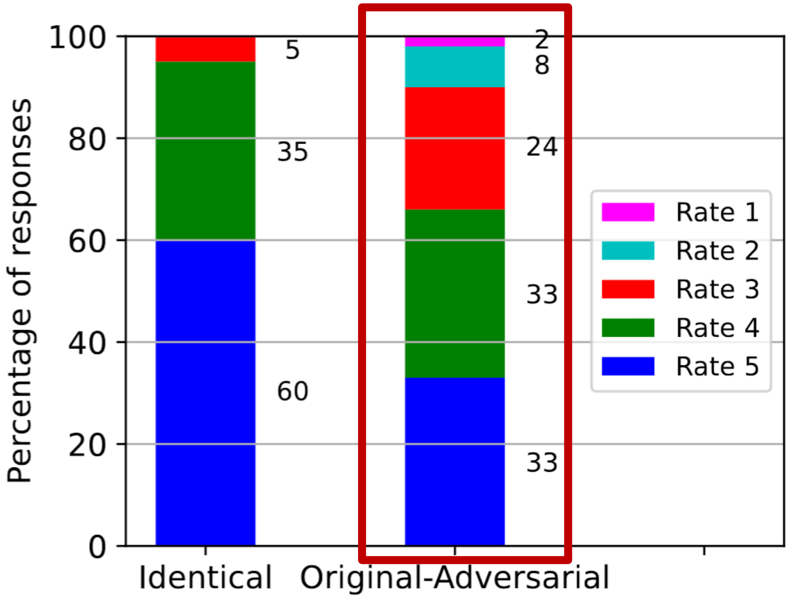
(b) Horizontal Study

UNIVERSITÄT LIECHTENSTEIN

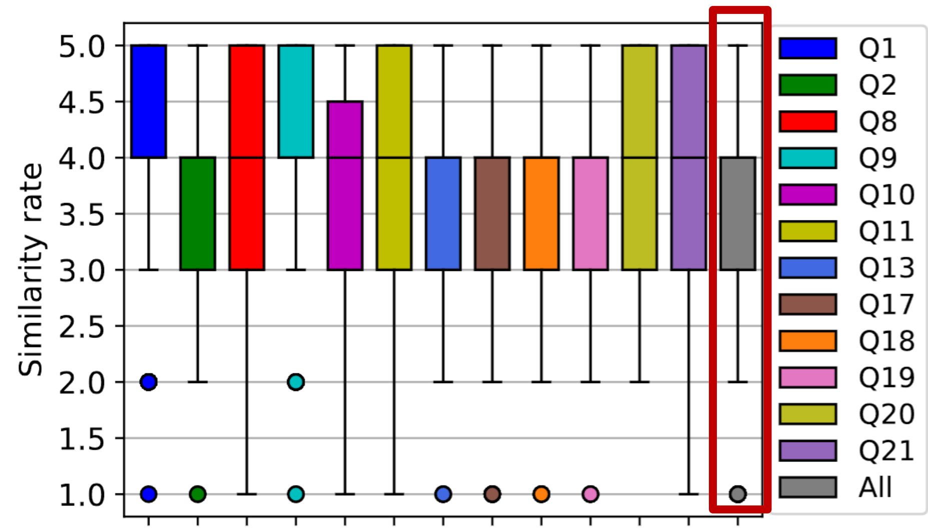
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Humans? Deceived!

- For every question, users had to say how “similar” the two logos were (5= very similar, 1= not similar at all)



(a) Vertical Study



(b) Horizontal Study

Takeaways:

1. Vertical Study: over 85% of participants rated >=3 similarity
2. Horizontal Study: the average similarity per question was >=3

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

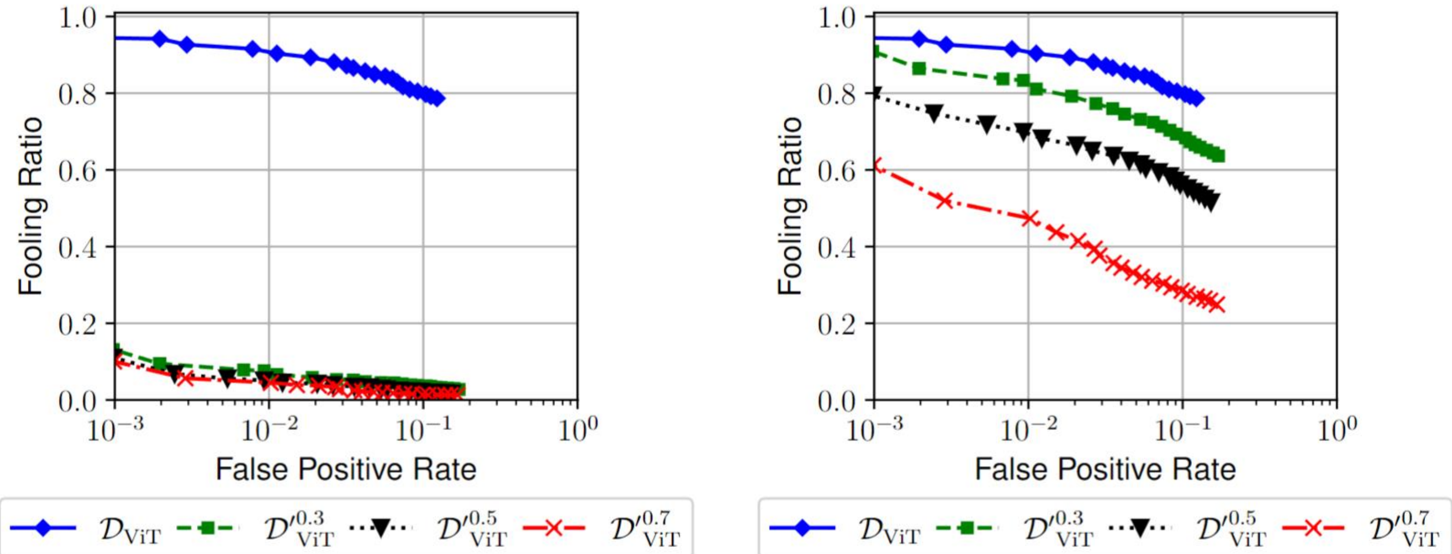
Countermeasures?

- Can adversarial logos be countered?
  - If so, can an adversary launch a counterattack?

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Countermeasures?

- Can adversarial logos be countered?
  - If so, can an adversary launch a counterattack?



(a) Against original generator $\mathcal{G}_{\mathrm{ViT}}$ **(b)** Against adaptive generators

Fig. 8: Performance of discriminator and generator due to adversarial training

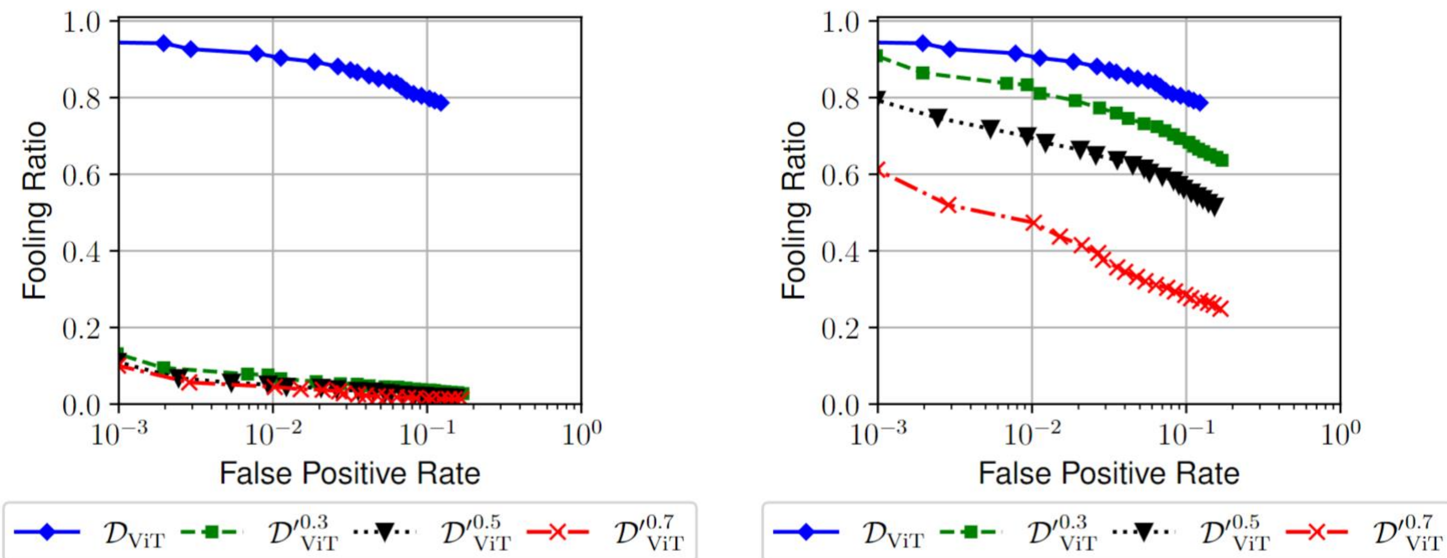
We use the logos generated by G_{ViT} for retraining

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Countermeasures?

- Can adversarial logos be countered? → **Yes** ☺
  - If so, can an adversary launch a counterattack? → **Yes** ☹



(a) Against original generator $\mathcal{G}_{\text{ViT}}$ **(b)** Against adaptive generators

Fig. 8: Performance of discriminator and generator due to adversarial training

We use the logos generated by $\mathcal{G}_{ViT}$ for retraining

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Conclusions

1. We proposed a **novel attack…**

2. We showed that **it works**

3. …against both state-of-the-art **systems** *and* **humans.**

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Conclusions

1. We propose a **novel attack…** …as well as two transformer-based methods for logo-identification.
2. We showed that **it works** …and that is realistically feasible…
3. …against both state-of-the-art **systems** *and* **humans.** …and that countermeasures exist, but can be countered;

…and also that our proposed transformer methods are more robust and more expensive to evade.

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Conclusions

1. We proposed a **novel attack…** *…as well as two transformer-based methods for logo-identification.*
2. We showed that **it works** *…and that is realistically feasible…*
3. …against both state-of-the-art **systems** *and* **humans.** *…and that countermeasures exist, but can be countered;*

…and also that our proposed transformer methods are more robust and more expensive to evade.

We focus on the Logo-discriminator.

Future research: consider other elements of a phishing detector, and assess the response of humans to the evasive samples!

UNIVERSITÄT LIECHTENSTEIN

All of our resources are publicly available [1]

[1]: https://github.com/JehLeeKR/Adversarial-phishing-logos

The Hague – September 25th, 2023

European Symposium On Research In Computer Security

Attacking Logo-based Phishing Website Detectors with Adversarial Perturbations

Jehyun Lee, Zhe Xin, Melanie Ng Pei See, Kanav Sabharwal,
Giovanni Apruzzese, Dinil Mon Divakaran

UNIVERSITÄT LIECHTENSTEIN

Trustwave®

Acronis

NUS National University of Singapore