BadLogo: A Physically Realizable Adversarial Sticker for Evaluating the Robustness of Face Recognition Models

Fuqi Qi
Xidian University
Xi'an, China
fqqi_1@stu.xidian.edu.cn

Haichang Gao✉
Xidian University
Xi'an, China
hchgao@xidian.edu.cn

Boling Li
Xidian University
Xi'an, China
23031212152@stu.xidian.edu.cn

Shiping Guo
Xidian University
Xi'an, China
spguo@stu.xidian.edu.cn

Yuming Zheng
Xidian University
Xi'an, China
zymcore@stu.xidian.edu.cn

Bingqian Zhou
Xidian University
Xi'an, China
zbqxidian@stu.xidian.edu.cn

***Abstract*—Deep learning-based face recognition systems are increasingly deployed in security-critical applications, yet remain vulnerable to adversarial attacks. Existing physical attacks often lack stealth or realism, limiting their utility for evaluating real-world robustness. To address this, we propose BadLogo, a physically realizable, sticker-based perturbation that balances attack effectiveness and visual plausibility. We also introduce PLOF-GAN, an attack model employing a two-stage pretraining strategy to progressively transform benign logos into high-quality adversarial patches. An improved 3D mapping algorithm ensures precise alignment with facial geometry, enhancing the realism of physical deployment. To quantitatively assess stealthiness and visual plausibility, we design a new metric, BROI, which measures both region occlusion and semantic coherence. Whether in evasion or impersonation attack scenarios, extensive experiments on identity verification and identification tasks demonstrate that BadLogo achieves high attack success rates with superior stealthiness compared to existing methods. Notably, our approach exhibits strong transferability across multiple face recognition models and generalizes to object detection scenarios. By leveraging adversarial stickers, this work reveals critical vulnerabilities in face recognition systems under physical-world conditions and offers new perspectives for developing more robust models.**

***Index Terms*—face recognition, adversarial patch, robustness evaluation, physical adversarial attack**

I. Introduction

With the rapid advancement of artificial intelligence (AI), face recognition (FR) has been widely adopted in domains such as smartphone unlocking, security surveillance, identity verification, and financial transactions [1] [2], significantly improving the automation and convenience of identity authentication. Deep neural networks (DNNs), due to their high accuracy and efficiency in processing large-scale datasets, have become the backbone of FR systems. Numerous high-performance models [3]–[8] have thus been rapidly deployed.

✉ Corresponding author.

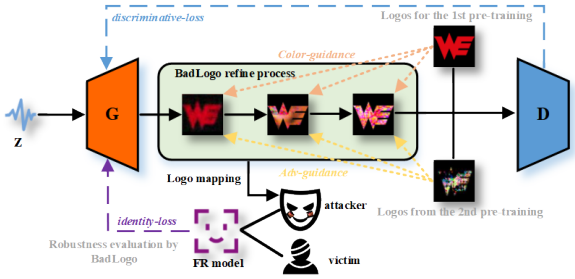

Fig. 1: Schematic diagram of our method. The first stage of pre-training provides color guidance for the generator and the second stage of pre-training provides adversarial guidace for the generator.

However, DNNs have been shown to be inherently vulnerable to adversarial attacks [9], where small, malicious perturbations lead to incorrect identity predictions. These adversarial examples raise serious concerns about the reliability of FR systems in real-world applications.

Conversely, adversarial attacks—especially those implemented in the physical world—serve as practical tools for evaluating DNN robustness [10] [11]. Given the security-critical nature of FR deployments, simulating physical-world threats is essential for assessing their resilience. Several works have proposed printed adversarial patches applied to key regions [12] [13], while others exploit projected light patterns [14] [15] or adversarial makeup [16] [17] to fool models. These efforts have exposed important security flaws and informed subsequent improvements.

Nonetheless, existing robustness evaluation methods based on physical adversarial attacks have significant limitations. Many works overlook real-world FR constraints, resulting in attacks that lack stealth and plausibility. In practical settings,

FR systems often integrate liveness detection [18], human-computer interaction [19], and backend human review, making visibly unnatural or oversized perturbations easy to detect. Moreover, current optimization strategies are inefficient, requiring excessive training while producing visually uncontrolled and computationally expensive outputs. Finally, most methods generalize poorly, performing well only under constrained conditions and struggling to transfer across different models, environments, or tasks.

To enhance the effectiveness of adversarial attacks and robustness evaluation for FR models, this paper proposes BadLogo, a novel, covert, and physically realizable perturbation, along with a corresponding generative model, PLOF (Put Logo on Face)-GAN, as shown in Figure 1. We leverage aesthetically pleasing and semantically meaningful logos as perturbation templates and employ a two-stage pre-trained generative adversarial network to efficiently generate high-quality perturbations. This approach enables a stealthy, plausible, and physically realizable evaluation of the adversarial robustness of FR models in real-world applications. Our method exhibits strong generalization, making it applicable to various FR models and even transferable to object detection tasks. This work highlights the security threats posed by sticker-based perturbations to FR systems and provides valuable insights for future robustness improvements and application security. The main contributions of this paper are outlined as follows:

- Our work fully considers the real-world application scenarios of FR models. We introduce a novel sticker-based perturbation using common tattoo stickers as the medium, providing a physically realizable, reasonable, and inconspicuous method to evaluate the robustness of FR models in practical settings.
- This paper proposes a GAN-based attack model with a two-stage pretraining strategy for generating BadLogo. By incorporating semantic pretraining and adversarial fine-tuning, we address the challenges of attack model training and achieve a balance between the visual semantics and adversarial effectiveness of the perturbation.
- We develop an improved 3D mapping algorithm to enable automatic alignment of sticker perturbations with facial features while simulating realistic 3D effects when applied to the face. Additionally, we introduce a novel evaluation metric specifically designed to assess the reasonability and imperceptibility of the perturbations.
- Our method demonstrates strong transferability across different models, environments, and tasks. The robustness evaluation highlights the pervasive threat posed by sticker perturbations, offering insights for enhancing the security of future applications.

II. Related works

A. Adversarial attacks for model robustness evaluation

Adversarial attacks are a widely used empirical method for evaluating the robustness of DNNs. By introducing carefully designed perturbations, researchers assess how DNNs respond to malicious input variations, revealing their vulnerabilities and guiding defense improvements [9] [20]. These attacks can be grouped into three categories: gradient-based methods (e.g., FGSM [9], BIM [21], PGD [22], MIM [23], C&W [24], ZOO [25]), search-based methods using black-box optimization [26], [27], and generation-based methods employing generative models like VAE [28], GAN [29] [30], or diffusion models [31]. While effective in digital domains, these methods typically apply global, norm-constrained perturbations to preserve imperceptibility.

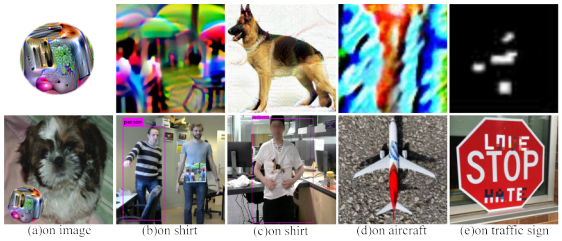

Fig. 2: Some samples of different adversarial patches in object detection tasks.

Subsequent research has explored localized and sparse attacks. Techniques like One Pixel Attack [32], JSMA [33], and SparseFool [34] relax imperceptibility constraints by modifying a small number of salient pixels while maintaining stealth and effectiveness.

To improve physical-world applicability and implementability, adversarial patches were introduced. Initial designs are universal patches [35] [36], followed by task-specific adaptations for object detection [12] [37]–[42], shown in Figure 2. Although effective, their large and visible nature often compromises stealth. Recent efforts focus on naturalistic patch generation using perceptual constraints [43] [44] to improve plausibility.

Beyond attack-based evaluations, certification-based techniques—such as ReLUplex [45], linear relaxation [46], and randomized smoothing [47]—provide formal robustness guarantees. However, these are often computationally expensive and architecture-bound, whereas attack-based evaluations are more adaptable to real-world tasks like face recognition, which are highly sensitive to environmental factors.

For face recognition, early robustness evaluations applied digital attacks [10] [11], their lack of physical feasibility limited relevance to real-world scenarios [48]. Thus, physical

Fig. 3: Some samples of physical attacks on FR models based on patches.

attacks based on adversarial patches tailored to FR task have emerged, which is shown in Figure 3. These include patch-based occlusion methods [13] [49] [50] [51], adversarial accessories [52] [53], and full-face 3D mask attacks [54] [55]. Despite effectiveness, they often suffer from poor realism, implausible visual design, or excessive size—making them prone to detection by liveness detection [18] or man-machine interactive systems [19].

Other approaches manipulate facial makeup [16] [17] [56], [57] or adversarial lighting [14] [15], but these often require exaggerated or constrained environmental settings, limiting their practicality.

B. GAN and its applications

GAN [58] is a classical generative model, which has been widely applied in tasks such as image generation [59], image inpainting [60], and style transfer [61]. Leveraging their adversarial nature, many studies [29] [30] have adopted GANs to generate adversarial examples.

However, the intrinsic design of GANs leads to significant training challenges, such as gradient vanishing, mode collapse, and instability due to the non-convex nature of their minimax optimization [62] [63]. Stabilization strategies like WGAN [64] and WGAN-GP [65] improve convergence, yet generating high-quality adversarial examples using GANs remains non-trivial, especially for adversarial attack tasks [66].

III. The proposal of BadLogo

Adversarial patch perturbations often suffer from semantic inconsistency or deficiency. For instance, in Figure 2(a), a toaster-shaped patch applied to a banana introduces semantic conflict, while the patch in Figure 2(e) overlays black-and-white text irrelevant to the original traffic sign, disrupting both human and model perception. Some patches, such as those in Figures 2(b)(d), either lack semantic meaning or fail to resemble plausible objects like accessories or clothing. Although the dog-shaped patch in Figure 2(c) improves visual plausibility, it has negligible attack effectiveness, undermining its value for robustness evaluation.

This semantic issue becomes more pronounced in FR tasks, where the human face dominates the image space, making added patches highly conspicuous. In Figure 3, many patches either exhibit facial artifacts, vague semantics, or constrained shapes (e.g., those attached to eyeglass frames). Some methods incorporate facial features into masks to enhance plausibility, but they often lack contextual appropriateness and fail against liveness detection [18] [19], especially when occluding the eyes.

To address these issues, we propose BadLogo, a tattoo-style adversarial patch designed to resemble event-related logo stickers commonly seen on human faces during public gatherings, which is shown in Figure 4. This design offers strong semantic plausibility and physical feasibility, making it suitable for evaluating FR robustness in real-world settings. Unlike previous methods such as Meaningful Sticker [53],

Fig. 4: Tattoo sticker logos that match the physical domain environment.

BadLogo adopts publicly recognizable logos with clean contours and simple, uniform colors, which are then adversarially fine-tuned. These characteristics enhance both stealth and attack effectiveness, providing a more realistic benchmark for FR system robustness.

IV. The framework of PLOF-GAN

GANs are naturally suited for adversarial tasks due to their two-player structure. However, when applied to adversarial patch generation, training instability is exacerbated by multi-objective optimization. Specifically, the generator must satisfy two conflicting goals: fooling the discriminator to ensure semantic plausibility, and misleading the target FR model to induce misclassification. The resulting gradient feedback from both networks often diverges, causing unstable parameter updates and hindering convergence. Consequently, traditional GAN-based methods struggle to balance visual realism with attack effectiveness, making it difficult to generate patches that are both stealthy and adversarially effective.

To address these issues, we propose PLOF-GAN, a BadLogo generation framework built upon WGAN [64]. As shown in Figure 5, our method introduces a novel two-stage training strategy and a redesigned attack pipeline. Our two-stage pre-training strategy ensures the color semantic learning capability and adversarial texture generation capability of PLOF-GAN respectively. Our proposed fine-tuning-based attack strategy balances the training of attack models and accelerates convergence. Overall, our approach enables PLOF-GAN to quickly generate BadLogo with high visual fidelity and strong attack performance, which improves the stealthiness, rationality and efficiency of robustness assessment based on adversarial attacks.

A. The first pre-training stage

The first phase aims to pre-train PLOF-GAN to learn the color semantic information of real-world logos. As illustrated in Figure 5(a), given a real club logo denoted by P_0, characterized by a single color c_0, we standardize the logo to a fixed size of 40×40 and extract its contour as a binary mask M_0. To construct a diverse yet consistent training set, we introduce slight perturbations to the RGB value c_0, creating a set of color variants. These perturbed colors are then blended with 100 texture images collected from Google Images to generate a set of background patterns, denoted as C_0. By filling the mask M_0 with these patterns, we synthesize a dataset of realistic-

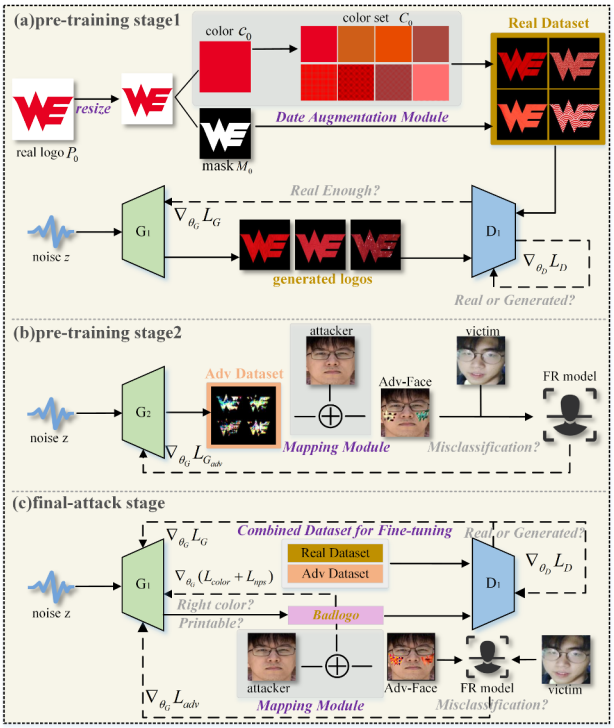

Fig. 5: The pipelines of BadLogo generation and PLOF-GAN training. The combined dataset provides both the semantic features of the real logo and the adversarial features during the fine-tuning process. G_2 has the same network structure as G_1 but is trained with reset parameters. G_1 and D_1 will be refined in the final attack. $\oplus$ means mapping algorithm.

looking logos, referred to as $RealDataset$, which serves as the training data for this pre-training phase.

Indeed, our approach uses a data augmentation strategy to enhance the diversity of the results generated by the model. During this phase, the discriminator treats samples from $RealDataset$ as real, and those generated by the generator as fake. The generator and discriminator are trained in an adversarial manner, following the standard minimax loss defined in Equations (1)–(3):

$$L_D = \mathbb{E}_{x\sim p_1}[\log D(x)] + \mathbb{E}_{x\sim p_g}[\log(1 - D(G(x)))] \quad (1)$$

$$L_G = E_{z\sim p_z}[\log(1 - D(G(z)))] \quad (2)$$

$$\begin{aligned}\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] \\ + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]\end{aligned} \quad (3)$$

Where p_1 represents the distribution of $RealDataset$, p_z is the prior distribution over the latent space (uniformly sampled from the interval [-1, 1]) and p_g represents the distribution of generated samples in each iteration. Note that, the discriminator is to optimized to maximize L_D, while the generator is to minimize L_G, thus enabling the generator to learn the underlying structure and texture of realistic logos.

B. The second pre-training stage

As shown in Figure 5(b), our second phase of pre-training introduces an additional generator G_2 with the same architecture as G_1, trained independently to generate adversarial patches $\hat{p}$ that directly attack the FR model. These logos are applied to the cheek region of input face images x through a mapping function, which will be detailed in Section 4.4. During training, the target FR model is treated as a fixed feature extractor $F()$, and the adversarial objective is constructed in the embedding space. Specifically, the generator is optimized using the following feature-level loss:

$$L_{G_{adv}} = \begin{cases} |F(x+\hat{p}) - F(x)| & \text{untargeted} \\ -|F(x+\hat{p}) - F(x_t)| & \text{targeted} \end{cases} \quad (4)$$

Here, x_t denotes the image of the target identity in impersonation (targeted) attacks. For untargeted attacks, the generator is encouraged to push $x+\hat{p}$ away from its original identity, while for targeted attacks, it is guided to move the embedding of $x+\hat{p}$ closer to x_t.

This phase plays a critical role in the whole attack pipeline of PLOF-GAN. It provides a model-aware initialization for the adversarial objective, allowing the generator to learn meaningful attack directions in embedding space before being regularized by visual constraints. More importantly, it generates a dataset $AdvDataset$, composed of successful adversarial logos, that is later integrated into the final training stage to promote convergence and stability.

C. The final attack stage

In the final attack phase, both $RealDataset$ and $AdvDataset$ are utilized to jointly fine-tune PLOF-GAN, which is shown in Figure 5(c). $RealDataset$ continues to serve as a reference for visual authenticity, helping D_1 maintain its capability to distinguish between visually plausible and implausible logos. The inclusion of $AdvDataset$ serves as a gradient regularization mechanism. By mixing $AdvDataset$ into the discriminator's 'real' input space, the overall training gradient passed to G_1 becomes more tolerant to adversarial semantics. Furthermore, the generator receives more constructive feedback of adversarial information, achieving better attack performance without compromising visual fidelity.

The fine-tuning strategy employed in the attack phase is guided by the multivariate loss, as shown in Equation (5) and (6).

$$L_D = E_{x\sim p_2}[\log D(x)] + E_{x\sim p_g}[\log(1 - D(G(x)))] \quad (5)$$

$$\begin{aligned}L_G = -\mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \\ + \lambda_1 L_{G_{\text{adv}}} + \lambda_2 L_{\text{color}} + \lambda_3 L_{\text{nps}}\end{aligned} \quad (6)$$

Here, p_2 represents the mixture distribution of the combined dataset, p_g denotes the distribution of generated logos

Algorithm 1 Pseudocode for BadLogo Generation and PLOF-GAN Training

for each real logo P_0 **do**
  Generate color variants from c_0
  Combine with textures to form C_0
  Fill M_0 to obtain *Real Dataset*
end for
for each training iteration in Phase 1 **do**
  Sample $z \sim p_z$
  Generate logo using $G_1(z)$
  Update D_1 using *Real Dataset* (Eq. 1)
  Update G_1 using pre-training loss (Eq. 2)
end for
for each training iteration in Phase 2 **do**
  Sample $z \sim p_z$ and image x
  Generate adversarial logo $\hat{p} = G_2(z)$
  Perform 3D mapping of $\hat{p}$ onto x to get x_p
  Compute adversarial loss $L_{G_{\text{adv}}}$ (Eq. 4)
  Update G_2 to minimize $L_{G_{\text{adv}}}$
  Store $\hat{p}$, which has already lead to misclassification, into *Adv Dataset*
end for
for each training iteration in Attack Phase **do**
  Sample $z \sim p_z$ and image x
  Generate patch $p = G_1(z)$
  Apply 3D mapping to embed p onto x to obtain x_p
  Train D_1 with *Real Dataset* $\cup$ *Adv Dataset* (Eq. 5)
  Compute total loss L_G (Eq. 6) including:
    $L_{G_{\text{adv}}}$ (feature), L_{color} (Eq. 7), L_{nps} (Eq. 8)
  Update G_1 by minimizing L_G
end for

produced by G_1, and p_z is a uniform latent distribution over the interval $[-1, 1]$. As shown in Equation (7) and (8), our approach introduces two regularization terms. The generator is optimized not only to fool the discriminator but also to maintain adversarial effectiveness and physical realizability.

$$L_{\text{color}} = \sum_{i,j} \sqrt{(p_{i,j} - p_{i+1,j})^2 + (p_{i,j} - p_{i,j+1})^2} + |p_{i,j} - c_0| \tag{7}$$

The color loss L_{color} encourages local smoothness and suppresses deviation from the base color c_0. For Equation (7), $p_{i,j}$ is a pixel in the logo, $p_{i+1,j}$ and $p_{i,j+1}$ represent two neighboring pixels. The first term in this loss is used for pixel smoothing and the second term is used to reduce modifications to the logo color to ensure color semantic features of the initial logo template.

$$L_{nps}(P) = \sum_{p_{\text{logo}} \in P} \min_{c_{\text{print}} \in C} |p_{\text{logo}} - c_{\text{print}}| \tag{8}$$

The non-printable score loss L_{nps} penalizes RGB values that are hard to reproduce using common printing devices, thereby ensuring the generated logos remain physically realizable in real-world deployment. To accurately capture the color space of the printing device, we photograph a color palette derived from a uniform sampling across the RGB color space under standardized lighting conditions. This study employs a clustering algorithm to distill 256 representative triplets as cluster centers from the space of RGB triplets thus acquired. Each selected center replaces the corresponding RGB triplets within its cluster. Throughout the model optimization phase, only these 256 RGB triplets are engaged in the computation of the non-printable score. For Equation (8), C represents the 256 cluster centers of the printer color space, $c_{\text{print}} \subset [0, 1]^3$ represents the RGB triplet of any printable color in the printer color space, p_{logo} represents the pixels in the logo P.

Together, this final phase enables G_1 to synthesize BadLogo that are visually consistent with real-world designs and semantically potent against FR models while being suitable for physical implementation. The pseudocode of BadLogo generation and PLOF-GAN training is shown in Algorithm 1.

D. The design of 3D mapping algorithm

To enhance the realism and physical plausibility of the adversarial logo placement, we design a 3D geometry-aware mapping algorithm, illustrated in Figure 6, which simulates the appearance of a real sticker-like logo adhered to the human face. This process automatically determines the optimal logo position and orientation based on facial geometry and emulates 3D projection and deformation effects, ensuring the effectiveness under real-world imaging conditions, including changes in pose and illumination.

Given a generated logo and the attacker face, we first employ a facial landmark detector to identify key facial regions, particularly focusing on the cheek area as the target placement zone. The image is then passed through a 3D face reconstruction module based on the 3D Morphable Model (3DMM) [67], yielding a 3D mesh of the attacker's face surface. The logo is then projected onto the 3D cheek surface to obtain a corresponding 3D logo mask. This step aligns the logo with local curvature and orientation, ensuring that the logo adheres naturally to the topology of the face. After 3D placement, the logo-face mesh is re-projected into the 2D image plane, generating a spatially consistent 2D logo mask. This mask is then overlaid onto the original 2D face image, resulting in a realistic adversarial image of the attacker with the logo.

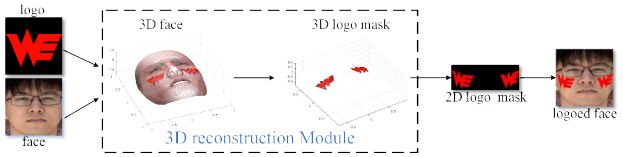

Fig. 6: The pipeline of our 3D mapping algorithm

E. Design of indicators for stealthiness and reasonableness

Previous works often assess the stealthiness and plausibility of adversarial patches using subjective criteria, making it difficult to compare patches of varying forms and sizes fairly. To address this, we introduce BROI (Blocked Regions of Interest), a semantic-aware metric that quantifies the proportion of a patch overlapping with key anatomical facial regions, offering a principled evaluation of stealthiness in practical deployments.

BROI is computed using CelebAMask-HQ [68], which provides pixel-level facial annotations. We employ BiSeNet [69], a lightweight segmentation network, pre-trained on CelebAMask-HQ, to divide input face image into 19 regions. From these, we extract five regions most critical for facial perception and recognition: eyes, nose, mouth, forehead, and cheeks. The eye region includes both eyes and glasses; the forehead is defined as the skin area above the eyebrows; cheeks refer to skin below the eyes. For each region R_i and the binary patch mask M, we compute the Intersection over Union (IoU) to define the BROI score:

$$BROI_i = \frac{|R_i \cap M|}{|R_i|} \tag{9}$$

$BROI_i \in [0, 1]$ measures the relative coverage of the patch over region. A higher value indicates more significant overlap, potentially implying lower stealthiness or reduced plausibility if the region is semantically sensitive (e.g., eyes or mouth). By aggregating the BROI values across all five regions or reporting them individually, this metric enables comprehensive and interpretable comparisons across different patch designs.

V. Experiments and Analysis

A. Experimental settings

Face recognition involves two primary tasks: 1:1 identity verification, which determines whether a pair of images belong to the same identity, and 1:N identification, where the model assigns each test image to one of N known identities. Real-world applications of 1:1 include smartphone unlocking and border control, while 1:N is commonly used in access control and attendance systems. We evaluate model robustness against physical adversarial patches under both tasks, treating 1:N as a closed-set scenario where test identities are all present in the gallery.

We adopt two representative FR models: ArcFace [8][1] and FaceNet [4][2], both of which are pre-trained on CASIA-WebFace [70], including 10,575 identities and 494,414 images.

All of our experiments were conducted on LFW [71], CelebA [72], and a self-collected Lab dataset comprising 20 individuals from our scientific research team with 15 images each. To carry digital and physical attacks, we constructed CelebA-Mixed and LFW-Mixed by combining 480 identities (with more than 10 images) from CelebA/LFW with the 20 Lab identities. For verification task, we randomly sampled 5,000 positive and 5,000 negative pairs per dataset. Both models achieved 99% accuracy. In attack evaluations, positive pairs were used for evasion (untargeted) attacks, while negative pairs supported impersonation (targeted) attacks. For identification task, 10 images per identity were used for training, and the rest for testing and attack. An MLP classifier was appended to each model's feature extractor and fine-tuned to maintain 99% accuracy.

We compared our BadLogo with four representative patch-based baselines: AdvHat [13], AdvGlass [52], FaceAdv [51] (a shape-constrained patch adapted from [37] with triangle, square, and circular shapes), and Meaningful Sticker [53]. For fairness, BadLogo and Meaningful Sticker used the same logo template in all main experiments; alternative logos are evaluated in ablation studies. We excluded 3D-mask-based methods [54] [55] due to their large perturbation areas and reliance on specialized hardware, which fall outside the scope of sticker-style physical patch attacks.

We used three metrics to evaluate performance: Attack Success Rate (ASR), BROI, and Time. With comparable ASRs, the lower the BROI and Time, the more valid and persuasive the robustness assessment.

- ASR measures the percentage of misclassified samples among all test images, evaluating the attack capability.
- BROI quantifies the proportion of the face obscured by the patch, reflecting its concealment. Larger BROI values may improve attack success but reduce stealth and rationality, approaching the effect of masks in presentation attacks and violating the adversarial patch constraint.
- Time represents the average time (in milliseconds) required to attack a single face image, calculated by dividing the total time (including pretraining and attack iterations) by the number of attacked samples.

All experiments were implemented in PyTorch and conducted on a system with four NVIDIA GeForce RTX 3090 (24GB) GPUs and an Intel Xeon Silver 4210 CPU. We used a batch size of 64, with a maximum of 20 attack epochs. If

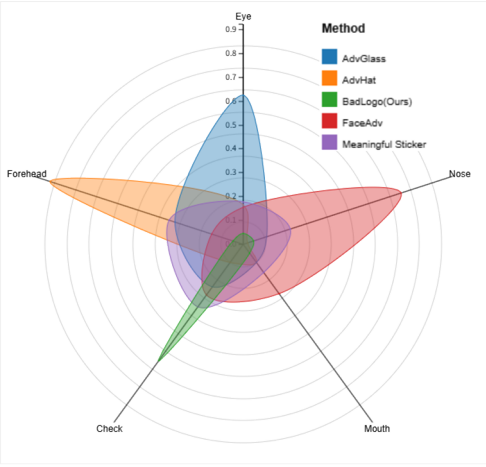

Fig. 7: BROI of generated patches of different attack methods

[1] https://github.com/deepinsight/insightface
[2] https://github.com/timesler/facenet-pytorch

TABLE I: Results of digital evasion attacks to verify the robustness of FR models.

Dataset	Method	1:1 ArcFace		1:1 FaceNet		1:N ArcFace		1:N FaceNet	
		ASR	Time	ASR	Time	ASR	Time	ASR	Time
CelebA-mixed	AdvHat	0.944	505.8	0.929	577.8	0.871	967.3	0.878	962.0
	AdvGlass	0.841	752.6	0.820	865.8	0.778	1009.2	0.783	1013.8
	FaceAdv	0.889	820.8	0.871	904.7	0.823	1271.3	0.824	1282.3
	Meaningful Sticker	0.594	**259.8**	0.575	**274.9**	0.408	**374.7**	0.415	**379.3**
	BadLogo(Ours)	**0.995**	436.9	**0.969**	482.6	**0.930**	640.9	**0.936**	644.6
LFW-mixed	AdvHat	0.965	498.2	0.940	475.1	0.899	962.5	0.904	966.3
	AdvGlass	0.907	748.1	0.894	861.3	0.815	1007.3	0.818	1006.9
	FaceAdv	0.891	819.6	0.876	897.3	0.825	1268.4	0.829	1269.0
	Meaningful Sticker	0.602	**259.9**	0.652	**294.9**	0.539	**374.5**	0.545	**382.0**
	BadLogo(Ours)	**0.997**	430.4	**0.977**	479.3	**0.952**	637.3	**0.959**	636.8

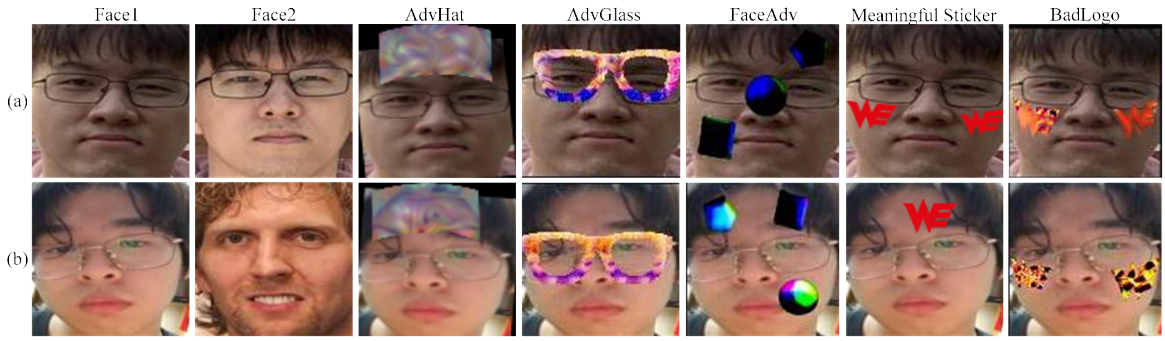

Fig. 8: The samples of adversarial faces with different generated patches in the digital attack. (a) shows some examples in the evasion attacks, and (b) shows some examples in the impersonation attacks.

the attack failed to succeed within 20 epochs, it was marked as a failure. In the final attack stage of PLOF-GAN, we set $\lambda_1 = 1$, $\lambda_2 = 0.2$, $\lambda_3 = 0.01$.

B. Experiments Results

We uniformly computed the BROI values for all methods in the digital domain. Note that the BROI values calculated in this paper are averaged over the entire dataset. Figure 7 show the BROI of generated patches of all the attack methods.

BadLogo achieves the lowest BROI score by confining its logo to the cheek area, thereby avoiding key facial features. In contrast, other methods cause more significant occlusion. AdvHat covers most of the forehead and partially blocks the eyes. AdvGlass occludes over 60% of the eye region, around 20% of the forehead, and part of the cheeks—largely due to its lack of geometric alignment, which leads to poorly positioned eyeglass frames that deviate from realistic wearing patterns. FaceAdv places three patches around the nose, heavily occluding the central facial area. Meaningful Sticker, optimized through evolutionary search, distributes perturbations more evenly across facial regions, yielding a moderately balanced high BROI score.

Overall, BadLogo's minimal occlusion—especially of five senses area—demonstrates the effectiveness of using small, well-positioned logos. Its design improves stealth and benefits from the proposed 3D mapping algorithm, ensuring accurate alignment across digital and physical domains. In contrast, competing methods exhibit larger misalignments, reducing their transferability in real-world settings. Consequently, the BROI metric not only quantifies occlusion but also supports the reliability and practicality of robustness evaluation.

1) Evasion attacks in the digital world: For the 1:1 identity verification task, we performed evasion attacks on 5,000 positive face pairs. Since the model compares only two images, small perturbations can easily reduce similarity below the threshold. As shown in Table I, all methods except Meaningful Sticker achieved ASR above 75% in most settings. Notably, BadLogo consistently exceeded 90% ASR, demonstrating superior attack effectiveness.

In the 1:N identification task, our method again achieved the highest ASR across all configurations. While Meaningful Sticker had the shortest runtime—due to the absence of optimization—it also performed the worst, with ASR often below 50%. In contrast, BadLogo required only twice the time but outperformed all optimization-based methods in both ASR and convergence speed, offering a favorable trade-off between effectiveness and efficiency.

Beyond ASR, Table I reveals further insights. All methods (except FaceAdv) achieved higher ASR on LFW-Mixed than on CelebA-Mixed, likely due to LFW's lower image quality and lighting, which amplify the impact of colorful perturbations. FaceAdv's poor performance stems from large black patch regions (Figure 8(a)), a consequence of lacking color constraints during training.

TABLE II: Results of digital impersonation attacks on CelebA-mixed dataset to verify the robustness of FR models.

Method	1:1 ArcFace		1:1 FaceNet		1:N ArcFace		1:N FaceNet	
	ASR	Time	ASR	Time	ASR	Time	ASR	Time
AdvHat	**0.591**	1285.8	**0.622**	1263.4	**0.642**	1047.0	**0.655**	1026.5
AdvGlass	0.414	1438.7	0.441	1419.2	0.468	1124.3	0.477	1103.9
FaceAdv	0.363	1364.4	0.385	1340.7	0.406	1195.2	0.415	1156.9
Meaningful Sticker	0.052	**382.8**	0.049	**379.4**	0.148	**383.7**	0.152	**383.6**
BadLogo	0.556	812.7	0.582	807.0	0.627	743.5	0.637	727.1

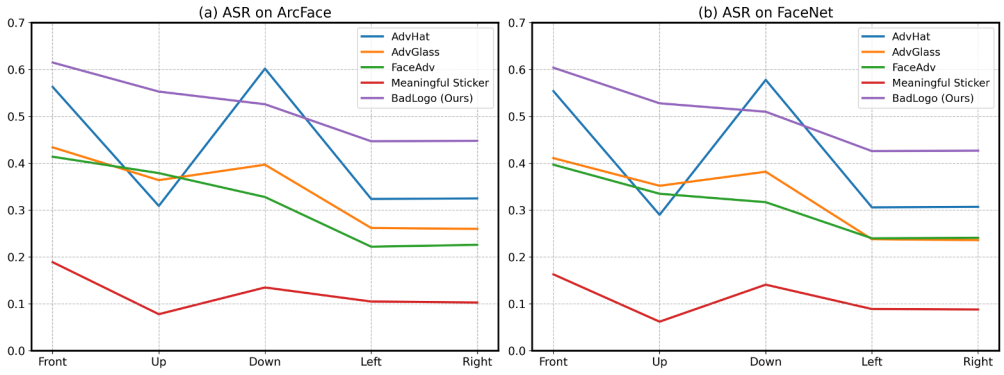

Fig. 9: Results of physical domain evasion attacks on the 1:1 verification task.

In terms of model robustness, FaceNet outperformed ArcFace in the verification task, benefiting from triplet loss and superior feature separation. However, in the identification task, the appended MLP classifier tended to overfit, reducing the performance gap. Overall, evasion attacks in 1:1 tasks were more effective and efficient than impersonation attacks in 1:N tasks, since the former only requires reducing similarity, whereas the latter involves crossing more complex classification boundaries. This highlights the better robustness on identification tasks under adversarial patch attacks.

2) Impersonation attacks in the digital world: Impersonation attacks were performed on CelebA-Mixed. For the 1:1 verification task, 5,000 negative pairs were used; for 1:N identification, all test samples served as attackers, with randomly selected identities in training set as targets. Results are shown in Figure 8(b) and Table II.

Compared to evasion, impersonation attack is inherently more difficult, as it requires increasing similarity between non-matching pairs—often harder than merely crossing the classifier's decision boundary. This resulted in a notable ASR drop and nearly doubled runtime. While AdvHat achieved the highest ASR, slightly outperforming BadLogo, our method showed clear advantages in efficiency, recording the lowest time cost among all optimization-based methods.

Impersonation relies on maximizing feature similarity between adversarial and target faces. Large-area patches are more likely to achieve this by altering key facial regions. AdvHat, for instance, covers the forehead, eyebrows, and eye sockets—regions highly influential for recognition, thus attaining a high ASR. However, its patches introduce visibly artificial features, making them easily detectable to human observers.

BadLogo incorporates a color-consistency loss to preserve the natural appearance and ensure the patch remains faithful to the original logo. Its compact design, restricted to the cheeks, avoids critical facial features while maintaining semantic plausibility. AdvGlass, lacking a face-aware mapping algorithm, often produces misaligned or floating glasses frames, leading to semantic inconsistencies and reduced transferability to the physical world.

As shown in Figure 7 and Figure 8, BadLogo achieves the smallest BROI, minimizing occlusion while maintaining competitive ASR. Other methods invariably cover more sensitive regions. When ASRs are similar, BadLogo consistently outperforms others in both time and BROI, highlighting its strengths in stealth and practical deployment.

In summary, BadLogo offers a balanced trade-off between adversarial strength, computational efficiency, visual realism, and stealth. These qualities make it particularly effective for evaluating the real-world robustness of FR models under subtle and semantically coherent attacks.

3) Attacks in the physical world: To evaluate real-world FR model robustness, we conducted physical attacks using participants from the Lab dataset. Adversarial patches generated digitally were printed and affixed to attackers' faces. During testing, each participant continuously adjusted their head pose while facing the camera, and 1,000 frames per pose were recorded and processed for identity matching. In the 1:1 verification task, we evaluated both evasion (20 positive pairs) and impersonation (20 negative pairs, each pairing an attacker with a different individual).

As shown in Figure 9, BadLogo consistently achieved the

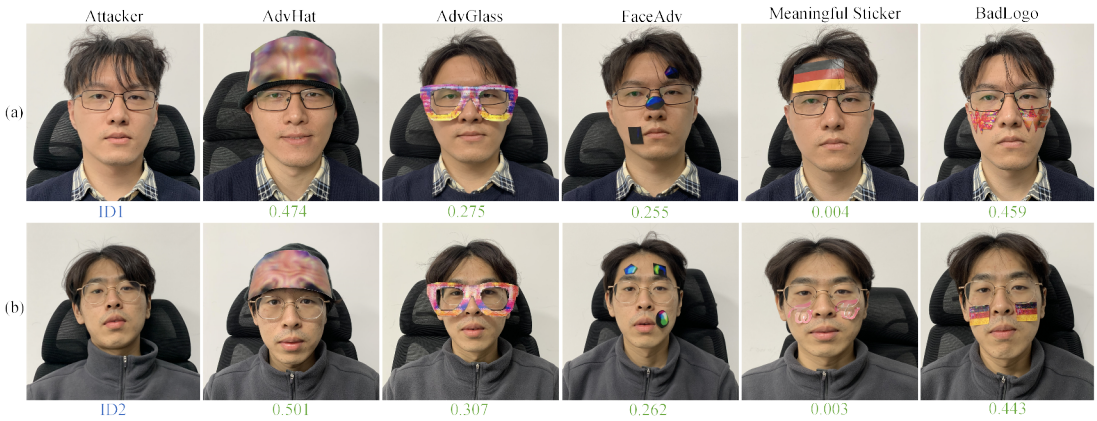

Fig. 10: Results of a physical domain impersonation attack on the 1:1 verification task. Row (a) represents the result of the attacker ID0 imitating the victim ID1, and row (b) the result of the attacker ID1 imitating the victim ID0. The data in green color indicates the ASR of the impersonation attack.

highest ASR across most settings. Facial pose had a notable impact: for most methods, ASR peaked when the face faced the camera directly and dropped with increased deviation. An exception was AdvHat, whose forehead-mounted patch became more exposed in downward poses, boosting ASR. BadLogo's symmetrical cheek patches, optimized bilaterally, maintained stable performance across varied head orientations, demonstrating greater robustness to pose variation. These results underscore BadLogo's superior physical-world adaptability, making it a reliable tool for evaluating FR model robustness under real-world conditions.

Figure 10 further illustrates the trade-off between effectiveness and reasonableness of robustness assessment methods. Although AdvHat achieved ASR close to BadLogo, its large patch size and heavy occlusion of key facial features undermine practical usability, potentially triggering liveness detection. AdvGlass, lacking alignment mechanisms and consideration of whether the attacker wears glasses, often produced misaligned frames and semantic inconsistencies. Meaningful Sticker, without adversarial optimization, failed in impersonation attacks due to its inability to enhance feature similarity. FaceAdv also underperformed, largely due to poor pretraining, resulting in unnatural black patches with low semantic relevance.

BadLogo employs a two-stage design: pretraining on real logos instills authentic semantic and color priors, while fine-tuning preserves natural appearance. This ensures that patches remain both visually plausible and adversarially effective.

Overall, BadLogo achieves a balanced combination of attack ability, visual stealth, rationality, and practical deployability, making it well-suited for robustness evaluation under constrained, real-world conditions. Its strong performance in physical-domain attacks—especially in contexts such as stadiums, concerts, and other large-scale public security scenarios—demonstrates its value for assessing and improving the resilience of deployed face recognition systems.

C. Ablation experiments

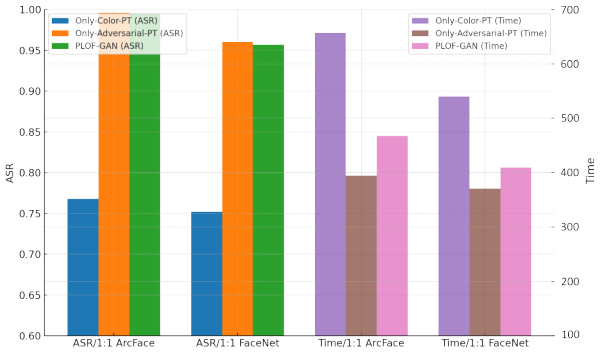

Fig. 11: Results of ablation experiments on the training strategy.

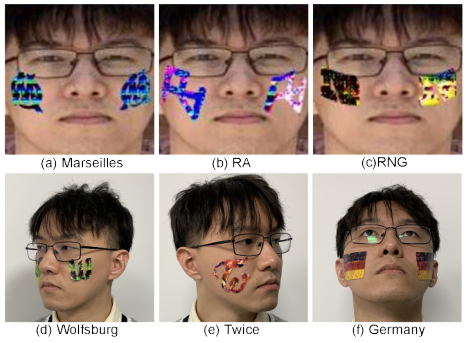

Fig. 12: Some examples of BadLogo generated in the evasion attacks for the identity verification task in both digital and physical world.

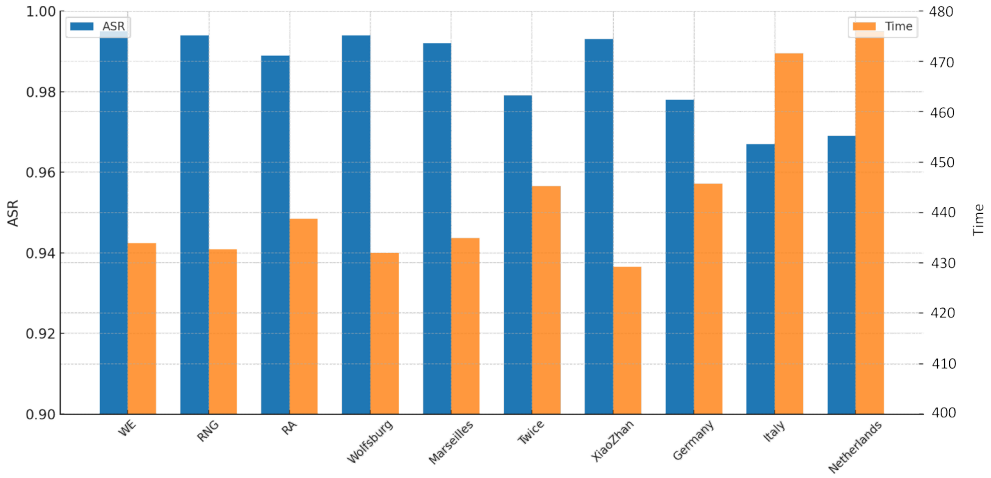

Fig. 13: The results of BadLogo using different logo pattern in the evasion attacks for the 1:1 identity verification task on ArcFace.

1) Training strategy: To evaluate the impact of the two-stage pretraining strategy in the PLOF-GAN pipeline, we conducted an ablation study comprising three training configurations: (1) the full three-stage pipeline, which includes color pretraining on real logos, adversarial pretraining using target model feedback, and final attack optimization; (2) Only-Color-PT, which omits adversarial pretraining; and (3) Only-Adversarial-PT, which skips color pretraining. Each configuration was used to generate BadLogo and carry out evasion attacks under a 1:1 identity verification setting using the CelebA-Mixed dataset.

As shown in Figure 11, Only-Adversarial-PT achieved the shortest runtime, slightly outperforming the full pipeline, whereas Only-Color-PT was significantly slower and frequently failed to converge. In terms of ASR, Only-Adversarial-PT performed comparably to the full version and notably outperformed Only-Color-PT. These results suggest that adversarial pretraining is essential for efficient convergence and effective attack generation, while color pretraining alone is insufficient to support adversarial feature learning. Without adversarial initialization, the semantic gap between real logos and adversarial patches introduces training instability—manifesting as gradient vanishing or mode collapse—ultimately degrading convergence and attack performance.

Moreover, as illustrated in Figure 4, color pretraining enhances the realism of generated logos, which is crucial for stealth and physical-world deployment. Adversarial pretraining, in turn, embeds discriminative features and accelerates convergence. Together, the two-stage strategy balances visual plausibility, attack effectiveness, and training stability, reinforcing the practicality and robustness evaluation value of PLOF-GAN under real-world constraints.

2) Logo Pattern: To assess the impact of logo shape, color, and size on attack performance, we constructed a logo dataset comprising five team logos, two celebrity logos, and three national flags, as shown in Figure 3). Each logo was fine-tuned by removing borders and text, retaining only core visual elements, and then used to train PLOF-GAN. Figure 12 show some examples of different logos and Figure 13 shows the result of evasion attack for the verification task on CelebA-Mixed.

Results show that logo appearance had minimal influence on ASR and training efficiency. Despite differences in geometry and color, all ten logos achieved similar performance. A slight drop was observed for the German and Italian flags—likely due to large white regions, which imposed stricter constraints under the color-consistency loss and slightly slowed convergence. However, overall variation in ASR and training time remained small.

These findings indicate that BadLogo's effectiveness is mainly attributed to the generative capacity of PLOF-GAN.

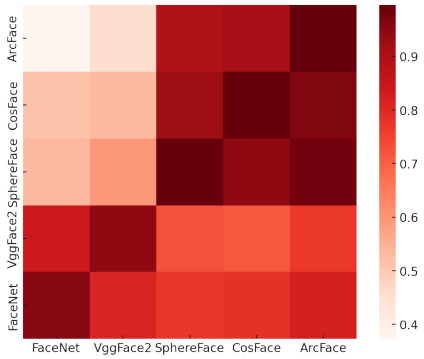

Fig. 14: The ASR results of transferability attacks. The horizontal axis indicates that the model is known, and the vertical axis indicates that the model is unknown. The diagonal is self-attack, which is the same as experiments in Section 5.2.

(a)logo on the front (b)logo on the back (c)logo on car window (d)logo on car door

Fig. 15: Some examples of physical dodging attacks on the object models

The two-stage pretraining framework enables PLOF-GAN to robustly adapt to diverse logo styles, offering flexibility in selecting patches tailored to different identities or deployment scenarios. This compatibility enhances the practical utility of BadLogo in evaluating model robustness under varied and realistic physical constraints, making it suitable for stealthy, context-aware adversarial testing in real-world systems.

D. Testing attack transferability

1) Transferability for different FR models: While earlier experiments assumed a white-box setting with full model access, such assumptions rarely hold in real-world deployments. To assess black-box robustness and transferability, we conducted evasion attacks on the 1:1 verification task under a black-box scenario.

In each trial, one FR model was randomly selected as the source model to generate BadLogo in an open-set setting. The resulting logos were then affixed to the attacker's face and evaluated on unseen target models. As shown in Figure 14, BadLogo exhibited strong transferability across diverse architectures.

Specifically, due to shared architecture and similar loss design, SphereFace, CosFace, and ArcFace formed a highly compatible group, with transfer ASR reaching up to 90%. On FaceNet and VGGFace2, which differ structurally but use similar loss functions, ASR remained above 70%. Even under substantial architectural divergence, BadLogo consistently achieved ASR over 40%, indicating strong generalization.

These results confirm that BadLogo—though optimized under white-box conditions—maintains adversarial effectiveness in black-box settings. This high level of generalization stems from the robust priors learned through PLOF-GAN, reinforcing BadLogo's applicability for evaluating FR model robustness under realistic, restricted-access scenarios.

2) Transferability for different vision Tasks: To assess the generalizability and task adaptability of our approach, we applied BadLogo to object detection tasks, including human and vehicle detection. We collected full-body images (front/back views) of lab members and multi-angle images of 30 vehicles as training data for PLOF-GAN. The target models were YOLO-V2[3] and YOLO-V5[4], both trained on the VOC2007 dataset.

Unlike the rigid alignment in face recognition, we introduced TPS transformations [39] to simulate non-rigid deformations of logos on clothing and affine transformations [13] for rigid placements on vehicle surfaces (e.g., doors, windows). To adapt to detection models—which require both localization and classification—we modified the adversarial loss in PLOF-GAN to jointly reduce detection confidence and bounding box alignment (Equation 10), using IoU as the alignment error metric.

$$L_{adv} = IOU(B(x + logo), B(x)) + |F(x + logo) - F(x)| \tag{10}$$

TABLE III: ASR of BadLogo on different object detection tasks

Model	Human Detection Task	Vehicle Detection Task
YOLO-V2	0.789	0.856
YOLO-V5	0.724	0.781

As shown in Figure 15, a vest logo or license plate patch enabled evasion, while unpatched individuals or vehicles remained detectable. These results highlight BadLogo's effectiveness even without detection-specific tuning.

[3]YOLO-V2 code: https://pjreddie.com/darknet/yolo/
[4]YOLO-V5 code: https://github.com/ultralytics/yolov5

Despite being designed for face recognition, BadLogo required only minimal adaptation to perform well in object detection tasks, as shown in Table III. The core reason for the attack's effectiveness lies in the dual-objective adversarial loss function, L_{adv}, that we designed for PLOF-GAN, as defined in Equation (10). This loss function simultaneously attacks two critical components of a detection model. The framework optimizes the patch's texture to include adversarial patterns that interfere with the model's feature extractor. When these adversarial logos are placed on a person or vehicle, they disrupt the high-level semantic features the model has learned to recognize, thereby significantly lowering the detector's confidence score for the target class (e.g., "person" or "car"). The loss function also incorporates an attack on the bounding box prediction by minimizing the IOU between the predicted and ground-truth boxes. This misleads the model's localization capabilities, causing it to fail in accurately framing the object even if it senses its presence, leading to a detection failure.

Despite the limited number of object models in the experiments, we still believe that BadLogo and PLOF-GAN are also effective for other object detection models. This is because our attack framework does not exploit the architectural flaws of a specific type of YOLO model, but rather exploits a general vulnerability of deep vision models in the feature extraction and decision stage.

This demonstrates the cross-task robustness and strong transferability of BadLogo, making it a practical tool for universal adversarial robustness evaluation in diverse real-world scenarios.

VI. Discussion

A key challenge in adversarial attacks is generating perturbations that are both semantically meaningful and adversarially effective. To address this, we propose a two-stage pretraining strategy: the generator is first trained to capture the shape and color characteristics of real logos, and then gradually adapted to the adversarial distribution by introducing adversarial samples into the discriminator. This staged design reduces optimization complexity, accelerates convergence, and enables BadLogo to balance attack strength with visual plausibility.

Unlike prior works relying on large-area patches that resemble presentation attack masks [54] [55], BadLogo remains inconspicuous in real-world scenarios. Such large patches are easily flagged by liveness detection systems [19] [18]. To quantitatively assess stealthiness, we introduce the BROI metric, which measures the facial area occluded by the patch. As shown in Figure 8, BadLogo consistently yields lower BROI values, indicating better concealment.

In contrast to abstract or synthetic patch designs, BadLogo leverages real, semantically meaningful logos—from sports teams, fan clubs, celebrities, national flags, to brands like Nike or Adidas. These elements are commonly worn in public settings, making their use in accessories or face decorations both natural and socially acceptable. This context-aware design allows BadLogo to blend seamlessly into real-world environments, such as concerts or stadiums, where face recognition is increasingly used for identity verification. Our method can stealthily bypass face recognition systems without drawing attention, thereby uncovering critical vulnerabilities that are difficult to detect through conventional testing.

By generating realistic, stealthy adversarial logos, BadLogo reveals security vulnerabilities that conventional methods may overlook. It provides a practical and actionable framework for evaluating face recognition robustness in deployed, high-security scenarios. The ability to simulate plausible attacks offers valuable insights for building more resilient and secure FR systems. We believe this application-oriented perspective significantly enhances the broader impact of this research, and offers practical guidance for both system designers and security practitioners in mitigating real-world adversarial risks.

VII. Conclusion

In this paper, we propose BadLogo, a physically realizable adversarial patch that leverages real-world semantic logos to evaluate the robustness of FR models. Unlike conventional approaches, BadLogo emphasizes both visual plausibility and attack efficiency, aligning with realistic deployment scenarios where stealth is critical. We introduce PLOF-GAN, a two-stage pretraining framework that improves training stability and convergence, along with a 3D mapping algorithm for automatic, accurate face alignment. Extensive experiments—covering digital and physical attacks, impersonation and evasion tasks, ablation studies, and black-box transfers—demonstrate the effectiveness and robustness of BadLogo across diverse settings. We further validate its task generalization by extending it to object detection models.

Beyond technical performance, BadLogo has strong real-world relevance. By embedding adversarial perturbations into commonly seen logos, our method reveals vulnerabilities in FR systems that are difficult to detect through traditional evaluation. The proposed BROI metric further enables principled assessment of patch stealthiness.

In summary, BadLogo functions as both a powerful attack mechanism and a practical evaluation framework for FR and other visual systems in the physical world. Our work provides lessons for future security enhancements in FR models.

Acknowledgment

The authors wish to thank the editors and anonymous reviewers for their valuable comments and helpful suggestions which greatly improved the paper's quality. This work was supported in part by the National Key R&D Program of China (2023YFB3107505), in part by Shaanxi Natural Science Funds for Distinguished Young Scholars (2023-JC-JQ-52), in part by the Natural Science Foundation of China (62302371), and in part by the Postdoctoral Fellowship Program of CPSF (GZC20232035).

References

[1] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Computer Vision and Image Understanding*, vol. 189, p. 102805, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314219301183

[2] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220316945

[3] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, Cham, 2016, pp. 499–515.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*. IEEE, June 2015, pp. 815–823.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, IEEE. IEEE, 2018, pp. 67–74.

[6] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2017, pp. 212–220.

[7] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2018, pp. 5265–5274.

[8] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 5962–5979, 2022.

[9] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[10] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa, "Unravelling robustness of deep learning based face recognition against adversarial attacks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1. Conference Organizer, 2018.

[11] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu, "Efficient decision-based black-box adversarial attacks on face recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7706–7714.

[12] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[13] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 819–826.

[14] M. Shen, Z. Liao, L. Zhu, K. Xu, and X. Du, "Vla: A practical visible light-based attack on face recognition systems in physical world," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–19, 2019.

[15] D.-L. Nguyen, S. S. Arora, Y. Wu, and H. Yang, "Adversarial light projection attacks on face recognition systems: A feasibility study," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 3548–3556.

[16] C.-S. Lin, C.-Y. Hsu, P.-Y. Chen, and C.-M. Yu, "Real-world adversarial examples involving makeup application," 2021. [Online]. Available: https://arxiv.org/abs/2109.03329

[17] N. Guetta, A. Shabtai, I. Singh, S. Momiyama, and Y. Elovici, "Dodging attack using carefully crafted natural makeup," 2021. [Online]. Available: https://arxiv.org/abs/2109.06467

[18] L. Li, Z. Xia, J. Wu, L. Yang, and H. Han, "Face presentation attack detection based on optical flow and texture analysis," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1455–1467, 2022.

[19] M. Shen, Y. Wei, Z. Liao, and L. Zhu, "Iritrack: Face presentation attack detection using iris tracking," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 2, pp. 1–21, 2021.

[20] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," pp. 427–436, 2015.

[21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015. [Online]. Available: https://arxiv.org/abs/1412.6572

[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2019. [Online]. Available: https://arxiv.org/abs/1706.06083

[23] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2018, pp. 9185–9193.

[24] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[25] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*. Conference Organizer, 2017, pp. 15–26.

[26] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *International conference on machine learning*, PMLR. Conference Organizer, 2018, pp. 2137–2146. [Online]. Available: https://proceedings.mlr.press/v80/ilyas18a.html

[27] M. Alzantot, Y. Sharma, S. Chakraborty, H. Zhang, C.-J. Hsieh, and M. B. Srivastava, "Genattack: Practical black-box attacks with gradient-free optimization," in *Proceedings of the genetic and evolutionary computation conference*. Conference Organizer, 2019, pp. 1111–1119.

[28] P. Tabacof, J. Tavares, and E. Valle, "Adversarial images for variational autoencoders," 2016. [Online]. Available: https://arxiv.org/abs/1612.00155

[29] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. Conference Organizer, 2018, pp. 3905–3911.

[30] S. Jandial, P. Mangla, S. Varshney, and V. Balasubramanian, "Advgan++: Harnessing latent layers for adversary generation," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 2045–2048.

[31] J. Chen, Y. Dai, and F. Huang, "Diffattack: Imperceptible and transferable audio adversarial attack via diffusion model," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[32] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[33] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy*, 2016, pp. 372–387.

[34] A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard, "Sparsefool: A few pixels make a big difference," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9079–9088.

[35] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," 2018. [Online]. Available: https://arxiv.org/abs/1712.09665

[36] D. Karmon, D. Zoran, and Y. Goldberg, "LaVAN: Localized and visible adversarial noise," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 2507–2515. [Online]. Available: https://proceedings.mlr.press/v80/karmon18a.html

[37] K. Li, D. Wang, W. Zhu, S. Li, Q. Wang, and X. Gao, "Physical adversarial patch attack for optical fine-grained aircraft recognition," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 436–448, 2025.

[38] S. Thys, W. V. Ranst, and T. Goedemé, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops(CVPRW)*, 2019, pp. 49–55.

[39] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin, "Adversarial t-shirt! evading person detectors in a physical world," in *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part v 16*, Springer. Springer, Cham, 2020, pp. 665–681.

[40] B. G. Doan, M. Xue, S. Ma, E. Abbasnejad, and D. C. Ranasinghe, "Tnt attacks! universal naturalistic adversarial patches against deep neural

network systems," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3816–3830, 2022.

[41] A. Guesmi, R. Ding, M. A. Hanif, I. Alouani, and M. Shafique, "Dap: A dynamic adversarial patch for evading person detectors," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 24 595–24 604.

[42] A. Du, B. Chen, T.-J. Chin, Y. W. Law, M. Sasdelli, R. Rajasegaran, and D. Campbell, "Physical adversarial attacks on an aerial imagery object detector," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 3798–3808.

[43] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao, "Perceptual-sensitive gan for generating adversarial patches," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01. Conference Organizer, 2019, pp. 1028–1035.

[44] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang, "Adversarial camouflage: Hiding physical-world attacks with natural styles," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 997–1005.

[45] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An efficient smt solver for verifying deep neural networks," in *Computer Aided Verification: 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I 30*, Springer. Springer, Cham, 2017, pp. 97–117.

[46] G. Singh, T. Gehr, M. Püschel, and M. Vechev, "An abstract domain for certifying neural networks," *Proceedings of the ACM on Programming Languages*, vol. 3, no. POPL, pp. 1–30, 2019.

[47] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 1310–1320. [Online]. Available: https://proceedings.mlr.press/v97/cohen19c.html

[48] Y. Zhong and W. Deng, "Towards transferable adversarial attack against deep face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1452–1466, 2021.

[49] M. Pautov, G. Melnikov, E. Kaziakhmedov, K. Kireev, and A. Petiushko, "On adversarial patches: Real-world attack on arcface-100 face recognition system," in *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, 2019, pp. 0391–0396.

[50] E. Kaziakhmedov, K. Kireev, G. Melnikov, M. Pautov, and A. Petiushko, "Real-world attack on mtcnn face detection system," in *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, 2019, pp. 0422–0427.

[51] M. Shen, H. Yu, L. Zhu, K. Xu, Q. Li, and X. Du, "Robust attacks on deep learning face recognition in the physical world," 2020. [Online]. Available: https://arxiv.org/abs/2011.13526

[52] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "A general framework for adversarial examples with objectives," *ACM Transactions on Privacy and Security (TOPS)*, vol. 22, no. 3, pp. 1–30, 2019.

[53] X. Wei, Y. Guo, and J. Yu, "Adversarial sticker: A stealthy attack method in the physical world," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 2711–2725, 2023.

[54] X. Yang, C. Liu, L. Xu, Y. Wang, Y. Dong, N. Chen, H. Su, and J. Zhu, "Towards effective adversarial textured 3d meshes on physical face recognition," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 4119–4128.

[55] X. Liu, F. Shen, J. Zhao, and C. Nie, "Eap: An effective black-box impersonation adversarial patch attack method on face recognition in the physical world," *Neurocomputing*, vol. 580, p. 127517, 2024.

[56] B. Yin, W. Wang, T. Yao, J. Guo, Z. Kong, S. Ding, J. Li, and C. Liu, "Adv-makeup: A new imperceptible and transferable attack on face recognition," 2021. [Online]. Available: https://arxiv.org/abs/2105.03162

[57] C.-S. Lin, C.-Y. Hsu, P.-Y. Chen, and C.-M. Yu, "Real-world adversarial examples via makeup," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 2854–2858.

[58] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[59] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016. [Online]. Available: https://arxiv.org/abs/1511.06434

[60] W. Wang, L. Niu, J. Zhang, X. Yang, and L. Zhang, "Dual-path image inpainting with auxiliary gan inversion," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 411–11 420.

[61] S. Yang, L. Jiang, Z. Liu, and C. C. Loy, "Pastiche master: Exemplar-based high-resolution portrait style transfer," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 7683–7692.

[62] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016.

[63] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually converge?" in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 3481–3490. [Online]. Available: https://proceedings.mlr.press/v80/mescheder18a.html

[64] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 214–223. [Online]. Available: https://proceedings.mlr.press/v70/arjovsky17a.html

[65] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[66] T. Zheng, C. Chen, and K. Ren, "Distributionally adversarial attack," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01. Conference Organizer, 2019, pp. 2253–2260.

[67] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 157–164.

[68] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5548–5557.

[69] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*. Springer, Cham, 2018, pp. 325–341.

[70] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014. [Online]. Available: https://arxiv.org/abs/1411.7923

[71] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. [Online]. Available: https://inria.hal.science/inria-00321923,

[72] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3730–3738.