

# Random Position Adversarial Patch for Vision Transformers

Mingzhen Shao  
 Kahlert School of Computing  
 University of Utah  
 shao@cs.utah.edu

## Abstract

*Previous studies have shown the vulnerability of vision transformers to adversarial patches, but these studies all rely on a critical assumption: the attack patches must be perfectly aligned with the patches used for linear projection in vision transformers. Due to this stringent requirement, deploying adversarial patches for vision transformers in the physical world becomes impractical, unlike their effectiveness on CNNs. This paper proposes a novel method for generating an adversarial patch (G-Patch) that overcomes the alignment constraint, allowing the patch to launch a targeted attack at any position within the field of view. Specifically, instead of directly optimizing the patch using gradients, we employ a GAN-like structure to generate the adversarial patch. Our experiments show the effectiveness of the adversarial patch in achieving universal attacks on vision transformers, both in digital and physical-world scenarios. Additionally, further analysis reveals that the generated adversarial patch exhibits robustness to brightness restriction, color transfer, and random noise. Real-world attack experiments validate the effectiveness of the G-Patch to launch robust attacks even under some very challenging conditions.*

## 1. Introduction

Recently, vision transformers (ViTs) have garnered significant attention due to their impressive performance [3–5, 9, 11, 14, 25, 30] and their ability to surpass convolutional neural networks (CNNs) in various domains. This remarkable performance has spurred interest in examining the robustness of ViTs, particularly considering the well-known vulnerability of CNNs to adversarial attacks [1, 17, 20, 23].

Drawing from the lessons learned with CNNs, adversarial attacks can be classified as image-dependent or image-independent. An image-dependent attack typically makes minute modifications to the source image [8, 15, 28]. These approaches have the drawback of being tailored to specific source images and limited in their deployment in the physical domain. Moreover, vision transformers have demon-

strated remarkable robustness against these types of attacks [1], showing their resilience when facing such adversarial attacks.

In contrast, the image-independent attack aims to create a patch that can be put alongside the target, without prior knowledge of lighting conditions, camera angles, or even elements present in the scene [2]. Adversarial patches have proven highly effective against CNNs, as they can be positioned anywhere within the classifier’s field of view to launch an attack. Astonishingly, they only require 10% of the pixels in the input image to deceive powerful CNN models like ResNet50 [2].

Unlike CNNs, vision transformers treat the input image as a sequence of image patches. To carry out an adversarial patch attack, a commonly employed approach is to substitute certain input image patches with adversarial samples [7, 10, 12]. These studies have shown the heightened vulnerability of vision transformers to adversarial patches. However, all the experiments conducted so far have been limited to the digital domain to accurately locate the adversarial patches. Gu *et al.* demonstrated that even a slight shift of a single pixel could dramatically decrease the attack success rate.

To overcome these strict limitations and enable physical-world deployment, we propose a novel approach that uses a GAN-like structure to generate universal and targeted adversarial patches (G-Patch). Our model consists of three main components: the generator, deployer, and discriminator. The generator is responsible for creating an adversarial patch. The deployer then attaches the patch to a random position on the source image. Finally, the victim network acts as the discriminator, providing predictions based on the modified image. Notably, unlike traditional GAN setups, the discriminator (victim network) remains unaltered throughout the training process.

Our experiments demonstrate that the generated adversarial patches can successfully launch attacks on various victim models at any position within the field of view. These patches achieve a high attack success rate of over 90% while maintaining a small size of approximately 10% of the

source image (on ViT-B/16). Further analysis reveals that the generated adversarial patches exhibit strong robustness to brightness restriction, color transfer, and random noise. This robustness to different distributions enhances their effectiveness during physical-world deployment. To assess their practical performance, we printed and positioned the adversarial patches in real-world settings. The results show that the patches consistently perform robustly in the physical world. To the best of our knowledge, our research is the first endeavor to achieve random position adversarial patch attacks on vision transformers.

Our contributions can be summarized as follows:

- We propose a new model to generate random position adversarial patches for vision transformers, which can launch targeted attacks at any position within the field of view.
- We show that the adversarial patches generated for vision transformers exhibit strong robustness to brightness restriction, color transfer, and random noise.
- We demonstrate that the generated adversarial patch can be robustly deployed in the physical world.

## 2. Related work

### 2.1. Vision transformer

The transformer was first introduced by Vaswani *et al.* [26] for NLP tasks. Following the success in NLP, Dosoviskiy *et al.* [5] proposed the vision transformer (ViT) that leveraged non-overlapping patches as tokens input to a similar attention based architecture. Since then, numerous models have been proposed to improve the performance of vision transformer models. Touvron *et al.* [25] introduced a teacher-student strategy in their DeiT models that dramatically reduced the pre-training request. Liu *et al.* [14] proposed the SWIN transformer using the shifted windowing scheme that achieves greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection.

Vision transformers have also been used in different vision tasks, including zero-shot classification [18], captioning [13], and image generation [19]. Due to the great success of vision transformers on vision tasks, many researchers have committed to analyzing the models' robustness to adversarial attacks.

### 2.2. Adversarial attacks

Adversarial attacks are widely employed to deceive deep learning models, resulting in remarkable successes. The first adversarial attack for deep learning was introduced by szegedy *et al.* [24].

Since their seminal work, numerous researchers have devised increasingly efficient techniques for generating adversarial attacks. In computer vision tasks, adversarial attacks can be classified into two types, depending on their reliance on the input image.

The first type is image-dependent adversarial attacks, which typically make minute modifications to the source image. These attacks employ various optimization strategies such as the Fast Gradient Sign Method (FGSM) [8], Projected Gradient Descent (PGD) [15], and Skip Gradient (SGD) [28]. However, these approaches often exhibit weaknesses due to their design being tailored to specific source images or limited to the digital domain.

In contrast, the second type of attack, image-independent attacks, uses an additional object (patch) to eliminate the requirement of relying on the input image. The patch is trained to create an attack without prior knowledge of the other items within the scene. The first image-independent attack approach was proposed by Brown *et al.* [2]. They used gradient-based optimization to iteratively update the pixel values of the patch in the source image to find the optimal values that can cause the target model to misclassify the object (AdvPatch). This patch can be placed anywhere within the field of view of a classifier and launch an attack. Since then, many studies have followed the same strategy to develop patches for physical attacks aimed at deceiving classifiers or object detectors, such as traffic signs [6], cloaks [29], or vehicles [31]. These successes have sparked researchers' interest in applying the same methods to vision transformers.

### 2.3. Robustness of vision transformer

Shortly after the introduction of vision transformers, several researchers [1, 16, 22] conducted studies demonstrating the superior robustness of vision transformers compared to CNNs when the entire image is perturbed with adversarial perturbations. However, subsequent research by Fu *et al.* [7] explored the vulnerability of vision transformers to patch attacks and found that vision transformers are more susceptible to such attacks compared to CNNs. Additionally, Gu *et al.* [10] further showed that whereas vision transformers are generally resilient to patch-based natural attacks, they are more vulnerable to adversarial patch attacks when compared to comparable CNNs.

All the preceding studies used a generation method similar to the one employed for CNNs, which involves replacing certain input patches with random noise and uses gradient-based optimization to iteratively update the pixel values, aiming to find the optimal values that can deceive the target model. However, unlike adversarial patches for CNNs, the patches they obtained for vision transformers must be precisely aligned with the input image patches. Gu *et al.* demonstrated that even a slight shift of a single pixel could

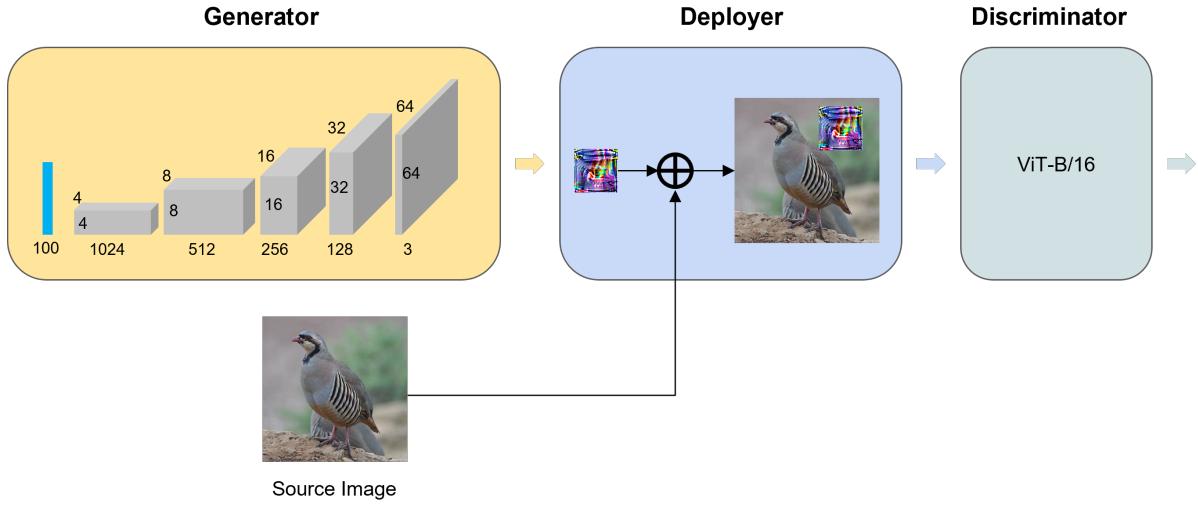


Figure 1. Network

dramatically decrease the attack success rate. The strict requirement posed a significant challenge to the practicality of using adversarial patches in real-world scenarios, as misalignment between attack patches and image patches is commonly encountered due to various factors. Consequently, it is crucial to develop methods for creating adversarial patches that account for these realistic conditions where misalignment can occur.

### 3. Methods

#### 3.1. Network

We introduce a GAN-like model for generating the adversarial patch. Instead of relying on direct gradient optimization of a random initial patch, our approach employs a five-layer convolutional network to transform a random noise into the desired adversarial patch. The structure of our model is shown in Figure 1.

The model can be divided into three functional parts: generator, deployer, and discriminator.

**Generator:** The generator consists of five convolutional layers, each accompanied by batch normalization and ReLU activation layers. The first convolutional layer is responsible for projecting and reshaping the random input vector into a four-dimensional tensor. The next four convolutional layers progressively upsample and refine the feature maps, capturing more complex patterns and details. The last convolutional layer is followed by a threshold layer instead of the batch normalization and ReLU layers. This threshold layer ensures that the output values of the generator are limited to a specific range. The threshold layer is defined as follows:

$$Th(x) = k * \tanh(x) + k \quad (1)$$

where  $k$  is a hyperparameter to adjust the range of the output and  $\tanh(x)$  applies the hyperbolic tangent function element-wise. We add  $k$  here to ensure that all values in the output remain above 0.

In default, we use  $k = 0.5$  to scale the range of the patch to 1, and for the following experiments, we use different  $k$  to achieve brightness restriction. By changing the kernel size and stride of different convolutional layers, we can change the size of the output adversarial patch.

**Deployer:** Given an image  $x \in [0, 1]^{w \times h \times c}$  with class  $y$  and the generated adversarial patch  $p$ . We use Algorithm 1 to generate a random binary mask  $M$  with the same shape of  $x$ .

---

#### Algorithm 1 Mask generation

---

**Input** source image:  $x$ , adversarial patch:  $p$

```

1:  $M \leftarrow \text{zeros}(x)$             $\triangleright$  all-zero mask  $M$  with shape  $x$ 
2:  $k \leftarrow \text{random.randint}(0, M[0] - p[0])$ 
3:  $l \leftarrow \text{random.randint}(0, M[1] - p[1])$             $\triangleright$  random
   position  $k, l$  within the range of  $M$ 
4: for  $i \leftarrow k$  to  $k + p[0] - 1$  do
5:   for  $j \leftarrow l$  to  $l + p[1] - 1$  do
6:      $M[i, j] \leftarrow 1$        $\triangleright$  mask only includes elements
       within shape  $p$  at position  $(k, l)$ 
7:   end for
8: end for
9: return  $M$ 

```

---

Then a modified image is generated by the deploy function  $T(p, x)$ :

$$T(p, x) = M * p + (1 - M) * x \quad (2)$$

The modified image is used as input for the discriminator.

**Discriminator:** The discriminator in our network is composed of the victim network (ViT-B/16 in the figure). It can be replaced with different models to generate adversarial patches for different victim networks. Unlike traditional GANs, the discriminator in our network remains unaltered throughout the training process.

For a targeted attack, the final loss of our network can be formed as follows:

$$L = \log(\text{softmax}(Pr(\hat{y}|T(p, x)))) \quad (3)$$

where the  $\hat{y}$  is the target class and  $\hat{y} \neq y$ ,  $Pr$  is the prediction of the discriminator with respect to class  $\hat{y}$ .

## 4. Experimental results and analysis

In this section, we first provide detailed information about the experimental setup used in our study. Next, we show the adversarial patches generated by our proposed model and evaluate their performance on various victim networks. Our results highlight the effectiveness of these patches in launching attacks from any position within the field of view. Furthermore, we conduct an in-depth analysis of the robustness of the generated adversarial patches. We investigate their resilience to brightness restriction, color transfer, and random noise, providing insights into their stability and effectiveness. Lastly, we validate the practical applicability of the generated adversarial patches by physically printing and placing them in real-world scenarios. This empirical evaluation demonstrates that the patches can effectively deceive vision transformer based systems in complex physical environments.

### 4.1. Experimental setup

In our experiments, we use the weights and the shared models from the *Pytorch Image models* [27] repository. These models are trained on the ImageNet1K dataset. To ensure the optimal performance, we conduct training for each configuration over 40 epochs, selecting the patch that achieves the highest performance as the final output patch. The input images used in our experiments have dimensions of 224x224, with pixel values from 0 to 1.

To assess the attack success rate (ASR), we begin by assembling a collection of images that are accurately classified by the models. The total number of these collected images is denoted as  $P$ . we apply the adversarial attack patch to this set of images and determine the number of images, denoted as  $Q$ , that are classified as the target class. The ASR is then defined as  $\frac{Q}{P}$ , serving as a metric to measure the effectiveness of the attack.

In order to evaluate the patch’s performance in the physical world, we use an HP laser printer to print the adversarial patches on A4 paper. Then we position the printed adversarial patch alongside the target object and capture

photographs using a Google Pixel 6a smartphone. This physical-world test incorporates various real-world factors such as camera angle changes, lighting variations, and different types of noise. By subjecting the generated patch to these real-world conditions, we are able to comprehensively evaluate its practical effectiveness in real-world scenarios.

### 4.2. Performance of generated patches

We select the ViT [5] and SWIN Transformer [14] as the fundamental victim network structures in our study. These two networks represent the key variations in patch handling within vision transformers: the ViT used fixed, non-overlapping patches, whereas the SWIN transformer incorporates shifted patch sizes.

The attack success rates of different vision transformers with different patch sizes are summarized in Table 1.

Models	Patch size		
	48x48	64x64	80x80
ViT-B/16	6.7%	69.6%	97.1%
ViT-L/16	2.7%	64.3%	88.7%
SWIN-B/16	59.5%	96.8%	99.6%

Table 1. ASR of generated patches on different vision transformers

We first observe that regardless of the architecture or depth of the vision transformers, the generated adversarial patches can achieve a high attack success rate even with a relatively small size ( 10% of the source image). This finding demonstrates the effectiveness of the generated adversarial patches that can launch attacks from any position of the source image.

Despite using a distinct method for generating adversarial patches, we are not surprised to discover a strong correlation between the patch size and its performance. Specifically, larger patches exhibit a higher attack success rate.

When comparing the results on the smaller models (ViT-B/16) to the larger models (ViT-L/16), we observe a significant disparity in their robustness across all patch sizes. This observation serves as evidence that the larger models possess inherent advantages in terms of defending against such attacks.

Furthermore, we confirm that the SWIN-B/16 shows a notably high ASR compared to the same-sized ViT-B/16. This outcome corroborates Shao *et al.*’s observation [22] that emphasizing low-level features in vision transformers can improve their overall performance but may have a detrimental impact on adversarial robustness.

In Figure 2, we present the adversarial patches generated for different victim models. The patch created for the SWIN model displays a remarkable dissimilarity compared to those for ViT models. This striking disparity serves

as an explanation for the ineffective transfer of adversarial patches between different models.

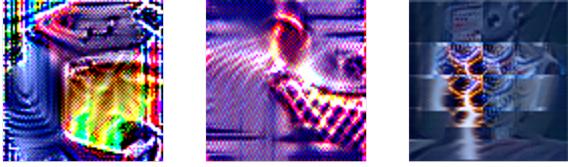


Figure 2. Patches for different networks with size 80x80 (left: ViT-B/16, middle: ViT-L/16, right: SWIN-B/16)

Some modified images created for the ViT-B/16 are shown in Figure 3. These images demonstrate the flexibility of our patch placement methodology, as the adversarial patch is positioned randomly on the source image, regardless of its specific position or alignment.



Figure 3. Modified images with random patch position

### 4.3. Patch robustness analysis

Recent research [21] has demonstrated the remarkable robustness of adversarial patches designed for CNNs against brightness restriction, color transfer, and random noise. Inspired by this finding, we adopt a similar investigative approach to analyze the behavior of patches designed for vision transformers.

For our experiments, we use the patch size 80x80 to assess the impact of different features on the victim networks (ViT-B/16 and SWIN-B/16).

#### 4.3.1 Brightness restriction

Brightness restriction for the generated patches can be easily implemented by adjusting the  $k$  value within the threshold layer. The performance variations for different value ranges are shown in Figure 4. We observe a significantly higher robustness in the SWIN model compared to the ViT model when the brightness range is decreased. This disparity in robustness can be attributed to the vulnerability of

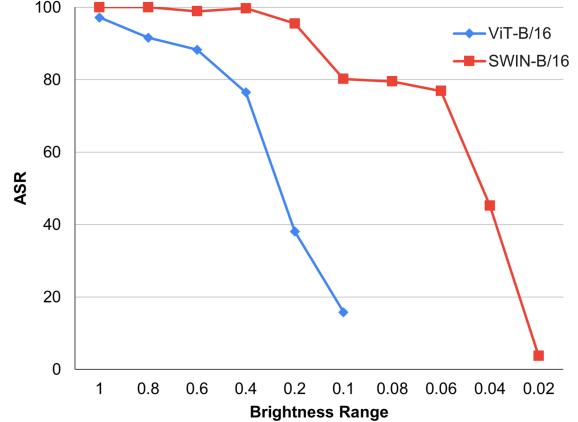


Figure 4. ASR with different brightness range

the SWIN model, as it requires comparatively less information to deceive the network. We find that even when the brightness range diminishes to just half of its original magnitude, the generated patch can still preserve over 80% of its ASR on ViT-B/16 model. The results show that patches designed for vision transformers exhibit a notable robustness to brightness restriction, similar to those observed in CNNs. Furthermore, the robustness of these patches is closely linked to the structure of the victim network. Some brightness-restricted patches and their brightness distributions are shown in Figure 5.

#### 4.3.2 Color transfer

In order to achieve color transfer, we add a  $\delta$  to all values within certain channels. However, performing this operation directly on the original patch (ranging from 0 to 1) can easily result in an overflow.

To address this issue, we use patches with brightness restriction (ranging from 0 to 0.8). By employing these patches, we ensure that overflow issues were avoided. This color transfer does not alter the texture distribution of the patch.

We present some color-transferred patches in Figure 6, and the performance between different color patches is shown in Table 2.

Models	Color		
	original color	color 1	color 2
ViT-B/16	91.6%	90.9%	90.7%
SWIN-B/16	99.6%	98.9%	99.7%

Table 2. ASR with different color transfer

We find that irrespective of the victim network, the generated adversarial patch consistently achieves nearly iden-

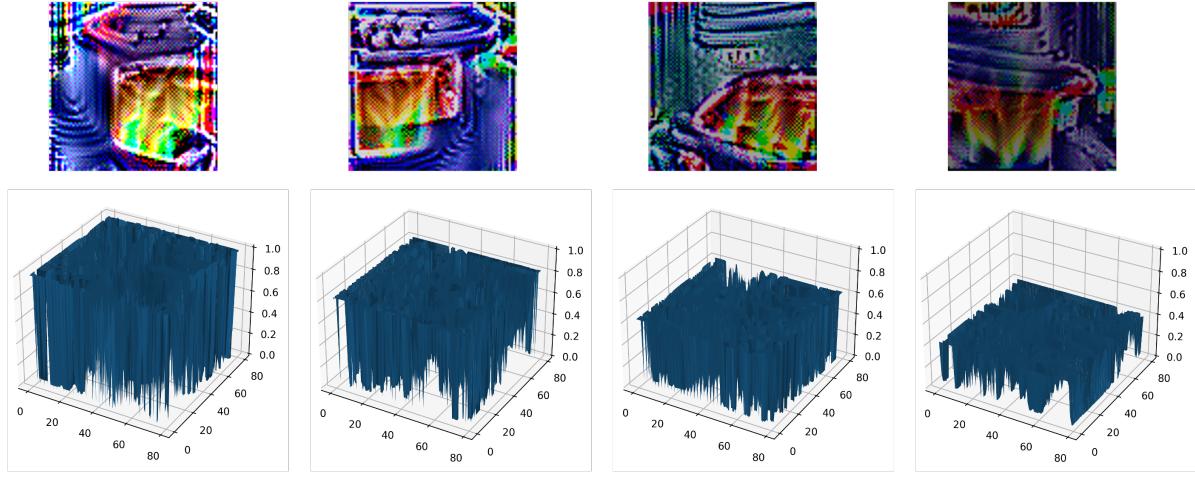


Figure 5. Patches with different brightness restriction and their brightness distribution



Figure 6. Different color transferred patch on ViT-B/16

tical ASR when subjected to different color transfers. This performance shows that the adversarial patch designed for version transformers does not rely on color information to deceive victim networks. This characteristic endows the patch with the capability to diminish its visual appeal through color transfer during deployment, akin to the approach employed by Shao [21] for CNNs.

#### 4.3.3 Random noise

In real-world deployments, the printed patch cannot be precisely the same as the digital version for various reasons (color accuracy of printer, carrier texture). To simulate the noise commonly encountered during real-world deployment, we generate random noise based on different signal-to-noise ratios (SNR). In order to avoid overflow when adding strong noise, we choose patches with a narrow brightness range (ranging from 0 to 0.6). The results of our experiments are shown in Table 3.

Models	SNR			
	10 dB	7 dB	5.2 dB	4 dB
ViT-B/16	83.4%	81.6%	69.1%	56.2%
SWIN-B/16	83.4%	81.6%	69.1%	56.2%

Table 3. ASR across different noise levels

We observe that the performance of the generated adversarial patch remains relatively stable even at an SNR of 7 dB (20% random color drift). This finding suggests that the generated adversarial patch exhibits robustness in dealing with the random noise typically encountered in real-world scenarios. The patch’s ability to maintain a high attack success rate in the presence of such noise further reinforces its effectiveness and practicality.

#### 4.4. Real-world attacks

The real-world deployability of adversarial patches designed for CNNs has been demonstrated in many works. However, due to the alignment problem, none of the adversarial patches designed for vision transformers has been deployed in the physical world before. Although our generated adversarial patches show the perfect position irrelative based on previous experiments, a valid concern remains regarding their robustness in real-world scenarios. In order to address this concern, we design several real-world deploy instances to show that the proposed attack patch can still work robustly in the physical world.

We have selected a range of scenarios, including indoor and outdoor environments, capturing images from different distances, angles, and lighting conditions. To ensure a more comprehensive evaluation, we use the ViT-B/16 model as the victim network instead of the easier SWIN-B/16 in our experiments. Figure 7 shows some figures and predictions.

We find that the generated patches show robust results in the physical world, even with brightness restriction. The top line in Figure 7 demonstrates the effectiveness of the generated adversarial patches in handling distortions caused by the inclination of the camera angle and even the bending of the printed patch. These results show that our generated patch can robustly launch attacks in the physical world without considering the position, lighting, and items in the

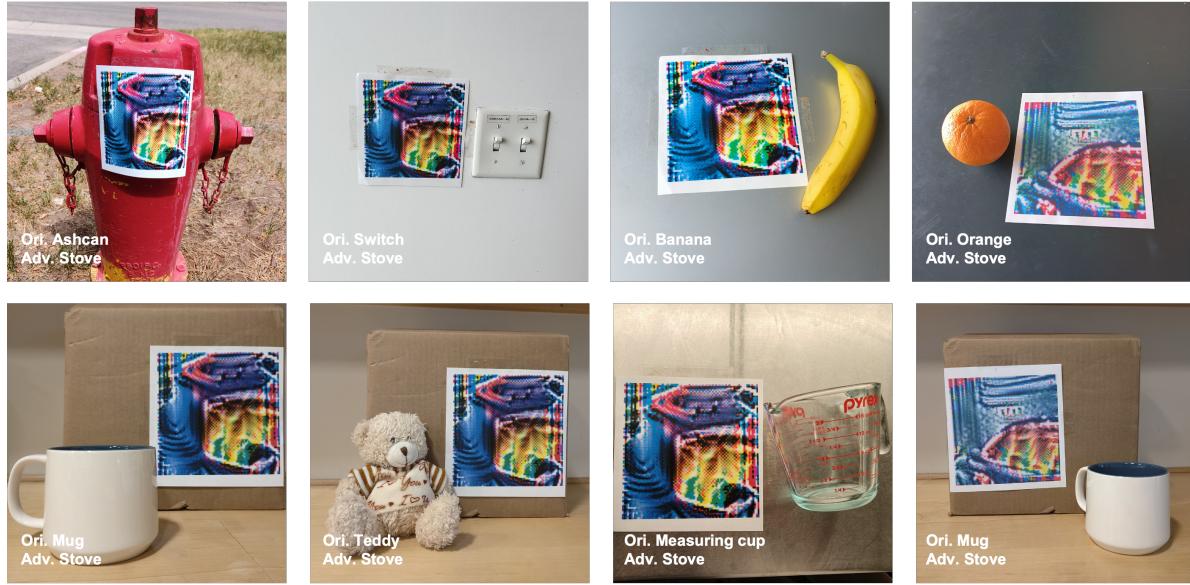


Figure 7. Prediction results in the physical world (top: outdoor lighting, bottom: indoor lighting)

field of view.

## 5. Conclusion

This paper introduces the G-Patch generating model, a novel approach for generating random position adversarial patches for vision transformers. The G-Patch successfully attacks on vision transformers from any position within the field of view, without requiring precise alignment or specific position. Furthermore, comprehensive analysis reveals that the G-Patch exhibits strong robustness against brightness restriction, color transfer, and random noise. These properties make the G-Patch highly resilient to various disturbances encountered in real-world scenarios. Real-world attack experiments validate the effectiveness of the G-Patch, showing its ability to launch robust attacks even under challenging conditions such as large camera angle inclinations and bending of printed patches.

## References

- [1] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10231–10241, 2021. [1](#), [2](#)
- [2] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. [1](#), [2](#)
- [3] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. [1](#)
- [4] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 589–598, 2021. [1](#)
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [1](#), [2](#), [4](#)
- [6] Ivan Evtimov, Kevin Eykholt, Earlene Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*, 2(3):4, 2017. [2](#)
- [7] Yonggan Fu, Shunyao Zhang, Shang Wu, Cheng Wan, and Yingyan Lin. Patch-fool: Are vision transformers always robust against adversarial perturbations? *arXiv preprint arXiv:2203.08392*, 2022. [1](#), [2](#)
- [8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. [1](#), [2](#)
- [9] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet’s clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021. [1](#)
- [10] Jindong Gu, Volker Tresp, and Yao Qin. Are vision transformers robust to patch perturbations? In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pages 404–421. Springer, 2022. [1](#), [2](#)

- [11] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021. 1
- [12] Ameya Joshi, Sai Charitha Akula, Gauri Jagatap, and Chinmay Hegde. A few adversarial tokens can break vision transformers. 1
- [13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2
- [14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1, 2, 4
- [15] Aleksander Madry, Aleksander Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial examples. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2
- [16] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2071–2081, 2022. 2
- [17] Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. Understanding and improving robustness of vision transformers through patch-based negative augmentation. *Advances in Neural Information Processing Systems*, 35:16276–16289, 2022. 1
- [18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2
- [19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2
- [20] Hadi Salman, Saachi Jain, Eric Wong, and Aleksander Madry. Certified patch robustness via smoothed vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15137–15147, 2022. 1
- [21] Mingzhen Shao. Brightness-restricted adversarial attack patch, 2023. 5, 6
- [22] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021. 2, 4
- [23] Yucheng Shi, Yahong Han, Yu-an Tan, and Xiaohui Kuang. Decision-based black-box attack against vision transformers via patch-wise adversarial removal. *Advances in Neural Information Processing Systems*, 35:12921–12933, 2022. 1
- [24] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2
- [25] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 1, 2
- [26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2
- [27] Ross Wightman. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>, 2019. 4
- [28] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. *arXiv preprint arXiv:2002.05990*, 2020. 1, 2
- [29] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 1–17. Springer, 2020. 2
- [30] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in Neural Information Processing Systems*, 34:30392–30400, 2021. 1
- [31] Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. Camou: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *International Conference on Learning Representations*, 2019. 2