Securing Tiny Transformer-based Computer Vision Models: Evaluating Real-World Patch Attacks

Securing Tiny Transformer-based Computer Vision Models: Evaluating Real-World Patch Attacks

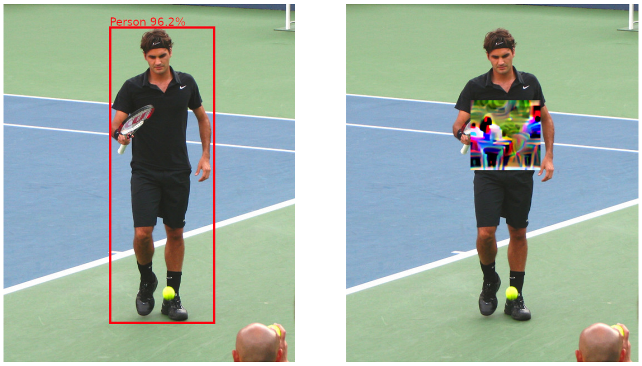
Andrea Mattei[†], Moritz Scherer[†], Cristian Cioflan[†], Michele Magno[†], Luca Benini[†‡]
[†]Dept. of Information Technology and Electrical Engineering, ETH Zürich, Switzerland
[‡]Dept. of Electrical, Electronic and Information Engineering, University of Bologna, Italy

***Abstract*— Transformers have significantly impacted the field of Computer Vision (CV) and the Internet of Things (IoT), surpassing Convolutional Neural Networks (CNN) in various tasks. However, ensuring the security of CV models for critical real-world IoT applications such as autonomous driving, surveillance, and biomedical technologies is crucial. The adversarial robustness of these models has become a key research area, especially for edge processing. This work evaluates the robustness of Swin tiny and ConvNeXt tiny, specifically focusing on real-world patch attacks in Object Detection scenarios. To ensure a fair comparison, we establish a level playing field between Transformer-based and CNN architectures, examining their vulnerabilities and potential defenses. Experimental results demonstrate the susceptibility of the Swin tiny and ConvNeXt tiny models to patch attacks, resulting in a significant decrease in average precision (AP) for the ”Person” class. When trained adversarial patches were applied, the AP drops to 12.8% and 15.2% for Swin tiny and ConvNeXt tiny models, respectively, highlighting their vulnerability to these attacks. This paper contributes to securing CV models on IoT vision devices, providing insights into the robustness of transformer-based architectures against real-world attacks, and advancing the field of adversarial robustness in embedded computer vision.**

I. Introduction

In the past decade, Convolutional Neural Networks (CNNs) have evolved remarkably, transforming from a niche research topic to the most widely-used model architecture for visual recognition tasks in academia and industry [1]. The advancements in CNNs have revolutionized the field of Computer Vision (CV), enabling breakthroughs in image classification and object detection [2]. However, recent developments in transformer-based architectures have started to gain significant traction within the CV community, leading to the emergence of novel state-of-the-art models that rely on the attention mechanism [3]–[7].

Among these transformer variants, vision and Swin transformers [4], [7] have made a particularly profound impact, giving rise to a plethora of derived models that achieve remarkable performance on image classification and object detection tasks. The Attention mechanism employed by these models allows them to capture complex spatial and long-range dependencies in visual data, leading to enhanced representation learning and improved performance [8].



(a) Unpatched Image

(b) Patched Image

Fig. 1: Example of patch attack successfully used for hiding people

While the measured performance of these transformer-based models is undeniably impressive, it is equally important to consider their robustness in real-world scenarios, especially for critical applications such as autonomous driving and biomedical technologies [9], [10]. Ensuring the security and adversarial robustness of CV models has become an increasingly important research area, as the deployment of Artificial Intelligence (AI) systems in safety-critical contexts necessitates the ability to withstand attacks and maintain reliable performance [11].

This work investigates the adversarial robustness of transformer-based models, specifically in real-world patch attacks, also known as model evasion attacks, for object detection [12]. Patch attacks, as shown in Figure 1, are particularly relevant in High-Performance Computing (HPC) edge Internet of Things (IoT) applications, where limited computational resources and strict latency constraints make them appealing to attackers [11]. An attacker can manipulate the perception system by exploiting vulnerabilities in the model's response to adversarial patches, potentially leading to severe consequences in safety-critical scenarios [13].

This paper focuses on adversarial robustness in transformer-based object detection algorithms for IoT vision embedded systems. Firstly, we implement and evaluate patch attacks against state-of-the-art transformer-based models, specifically focusing on their vulnerability to adversarial patches. This analysis provides insights into the susceptibility of these models to targeted attacks and the potential impact on their performance. Secondly, we evaluate the effectiveness of adversarial training as a defense mechanism against patch attacks in tiny transformer-based object detection algorithms. By subjecting the models to adversarial training, we investigate whether their robustness can be enhanced to better withstand adversarial patch attacks. Finally, we conduct measurements and comparisons to assess the robustness of both CNN- and transformer-based algorithms under patch attacks and adversarial training. Through these contributions, we aim to improve the understanding of the vulnerabilities and defenses of transformer-based models in the face of patch-based adversarial attacks, providing valuable insights into the robustness of these algorithms in real-world scenarios. The scientific contribution can be summarized as follow:

- Implementation & evaluation of patch attacks against state-of-the-art transformer-based object detection algorithms;
- Evaluation of Adversarial Training for patch attacks against tiny transformer-based object detection algorithm such as Swin tiny with as few as $28\,\mathrm{M}$ parameters;
- Measurements and comparisons for robustness of CNN- and transformer-based algorithms under patch attacks and adversarial training.

II. Related Work

In recent years, transformer-based architectures have revolutionized the field of CV by surpassing the performance of CNNs in various tasks. However, the security and robustness of CV models have become crucial considerations, particularly for critical real-world applications such as autonomous driving and biomedical technologies [14]. Consequently, research into the adversarial robustness of these models has gained significant importance. In this section, we explore relevant work in the field, focusing on four key subsections: transformers for CV, adversarial attacks, robustness of vision transformers, and defence against adversarial patches. By examining the literature in these areas, we aim to comprehensively understand the advancements, challenges, and strategies related to adversarial robustness in the context of transformer-based models for CV.

A. Transformers for Computer Vision

Following their success in the field of Natural Language Processing (NLP) [3], transformers, have recently been adapted for CV tasks. Nevertheless, designing a transformer network for image processing requires several adaptations. Unlike human-written language in NLP, images represent inputs with larger dimensionality, which makes it difficult to process high-resolution images with transformer networks due to the proportional increase in parameters and memory. Therefore, the majority of recently proposed versions of transformers for CV use either low-resolution images [15] or reduce the size of the feature maps [4], [5] leading into the transformer network. Despite these constraints, the introduction of transformers in the CV field thus allowed for establishing new state-of-the-art results [16][8] in object detection tasks, as done on the well-established COCO dataset [17].

B. Adversarial Attacks

Driven by the widespread adoption of machine learning in real-world applications, research on adversarial attacks against deep learning models focuses on evaluating the robustness of these models under artificially perturbed inputs [18]. An adversarial attack can be defined as a perturbation that, added to the input of a model, maliciously modifies the network's output. Generally, the perturbations are constrained in intensity and/or localization to make them less identifiable and more portable. Although many forms of perturbation are not practically reproducible in an authentic environment, so-called patch attacks have proven effective in realistic settings.

Originally introduced by Brown et al. [19] for the image classification task, patch attacks attempt to achieve misclassification through malicious perturbations with a constrained input image size. This category of attacks has proven a viable way to craft adversarial examples applicable to real-world cases. Generally, a patch is trained over multiple images through backpropagation, and, to enhance its effectiveness, random transformations sampled from a distribution are applied to the patch before its application to the input image. Such a technique is Expectation-over-Transformation (EoT) [20].

Since patches are local perturbations, this category of attacks has also been extended to object detection models that intrinsically perform a local analysis of the input images. The effectiveness of this strategy was proven by [21], an attack that crafts malicious textures for stop signs, capable of causing their misdetection or misclassification by a Faster R-CNN model [22]. This technique was further improved by Chen et al. [12] by applying the EoT algorithm to the training procedure to obtain more robust textures.

Thys et al. [23] introduced another application of patch attacks in object detection: evasion of people from surveillance systems. The produced patch can hide a person from a YOLOv2 [24] detector by degrading the objectness score of the target bounding box. Using appropriate transformations that mimic fabric creases, it is possible to craft patches printable on pieces of clothing like T-shirts [13] or jumpers [25]. Patches can be placed not only on the subjects' clothes, but they can also be applied directly on the camera lenses. This method relies on the production of translucent patches [26] applied on the lenses, which can be mistaken for benign scratches or dirt, yet lead to the misdetection of target objects.

C. Robustness of Vision Transformers

To increase the robustness of CV systems against adversarial attacks, Vision Transformers (ViTs) [4] have been intensively studied. The original ViT has shown very limited or no advantages in terms of robustness and transferability to other models over CNNs for neither noise-based attacks [27] [28], nor for patch-based ones [29], [30]. However, it is worth mentioning that these works compare two very different architectures: ViT against ResNet-based [31] networks. Therefore, they can offer only a limited overview of the real merits of the self-attention networks concerning additional factors such as the use of pre-trained networks or various architectural components, thus an extended analysis could only prove beneficial.

D. Defense Against Adversarial Patches

With the increased performance of patch attacks came the requirement to maintain accuracy against adversarial patches. Such a task is especially challenging due to the apparent randomness of the patches in terms of pattern, dimension relative to the target image, and localization within the image. Therefore, an effective defense mechanism should preserve the model's accuracy without requiring prior knowledge about the attack while introducing minimal computational overheads.

Conceptually, the approaches to robust classification or object detection models against patch attacks can be categorized into two classes. The first set of methods, commonly known as adversarial training, relies on heuristically determining the location of patches and eliminating them before performing object detection. Ji et al. [32] adapt the YOLOv2 [24] network, introducing an additional *patch* class that the model can detect independently from the other objects in the image, thus effectively separating the valuable information from the malicious attack. In Segment and Complete [33], the authors introduce an additional stage in the object detection pipeline, firstly determining the adversarial patch location and masking it at the pixel level, followed by performing object detection on the so-cleaned image. Albeit empirically efficient, these methods are prone to be rendered ineffective by adaptive attackers, as shown by Chiang et al. [34] in the context of image classification.

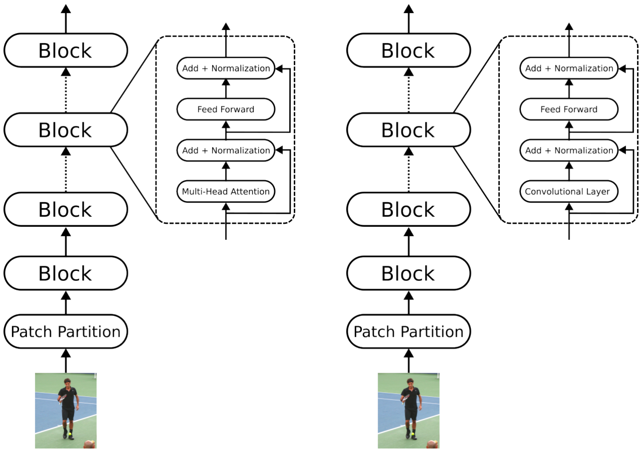
Conversely, certified defense mechanisms aim to maximize the provable robustness of a system in generic settings. Han el at. [35] leverage that patched images cluster Superfically Important Neurons (SINs), thus proposing a certified defense relying on SIN-based sparsification. Xiang et al. [36] draw intuition from the aforementioned first class of techniques, proposing a patch-agnostic masking method that eliminates adversarial pixels, with the robustness intrinsically certifiable by additionally removing duplicate bounding boxes for detected objects. Although efficient in eliminating patch attacks, these techniques increase the computational cost during inference, thus limiting their applicability in real-time settings.

This paper significantly contributes to the field by assessing the adversarial robustness of transformer-based models against real-world patch attacks and comparing their performance with traditional CNN architectures. This paper explores the adversarial robustness of transformer-based models through implementing and evaluating patch attacks, assessing adversarial training, and comprehensive measurements and comparisons. These contributions deepen our understanding of the strengths and weaknesses of self-attention mechanisms, advancing the field of adversarial robustness in CV. The findings of this study hold great potential for enhancing the security and reliability of CV models deployed on edge devices, thereby ensuring their effectiveness and trustworthiness in real-world applications.

III. Methodology

A. Swin Transformer and ConvNeXt

One of the recent best-performing ViT models is the Swin transformer [7], shown in Figure 2a. Instead of using regular self-attention, Swin restricts the span of its self-attention mechanism (i.e. for each patch only neighboring ones are considered) in a similar way to convolutional layers. The model starts initially with smaller patches than a standard ViT (4×4 px patches versus 16×16 px ones). After each block, patches are merged together into progressively bigger ones. The result is a hierarchical structure akin to dilated CNNs, allowing higher resolution analysis while minimizing the computational cost.



(a) Swin Transformer (b) ConvNeXt

Fig. 2: Architecture of analysed backbones.

Given the similarity of the Swin transformer to CNNs, ConvNeXt [37], depicted in Figure 2b, proposed the same architecture with standard convolutional layers instead of the Self-Attention ones. Insights from the vision transformers were also used to model the structure of the convolutional layers by, for example, using a bigger kernel (7×7) than the typical 3×3 kernel used in standard ResNets. The training procedure along the image augmentation transforms was harmonized with the ones of the Swin transformer. The resulting architecture has proven to be very competitive in terms of performance compared to vision transformers, with mean average precision close to the latter on the COCO dataset [17].

These two models are used as backbones for extracting features from images. For performing object detection, it is necessary to have an additional model that uses these features to generate predictions. For both models the original authors proposed to use a mask R-CNN architecture [38], shown in Figure 3, or its derivative cascade R-CNN [39], obtaining state-of-the-art results on the COCO dataset [17]. This architecture features two separate stages. After the backbone extracts the features from the input image, the first stage of the network, called Region Proposal Network (RPN), generates a set proposal, i.e. boxes that have a probability ($objectness$) of having an object inside. The RPN is a CNN-base feature pyramid network that extracts and refines a hierarchical set of feature maps into boxes. The proposal is first filtered with the Non-Maximum Suppression (NMS) mechanism and passed to the second stage called the Region of Interest (RoI) pooler,

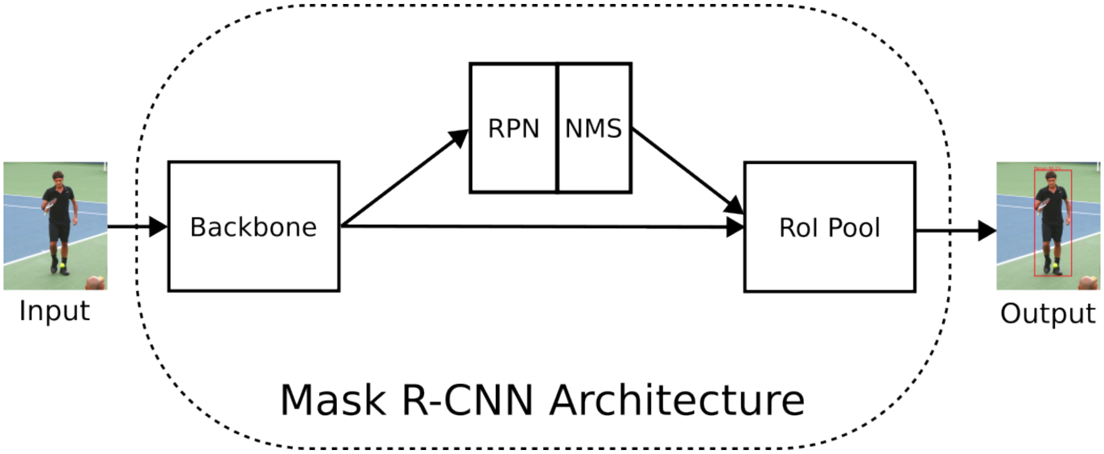


Fig. 3: Mask R-CNN architecture

which selects and classifies the output bounding boxes.

B. Attack Details

As part of our contributions, we evaluated patch attacks, building upon the attack proposed by Wu et al. [25]. In their work, the attack targets explicitly mask R-CNN architectures, generating adversarial patches that are easily printable. Taking inspiration from this previous research, we adapted the attack methodology to assess the vulnerability of state-of-the-art transformer-based object detection algorithms to similar patch-based attacks. The attack loop and its effectiveness are depicted in Figure 4. By leveraging the insights gained from the work of Wu et al. [25] and applying them to transformer-based models, we aim to advance our understanding of the robustness of these models against patch attacks and contribute to the development of more secure CV systems.

Patches are applied to the input image in the preprocessing stage. After that, they are fed to the backbone and the RPN stage like in the normal inference loop. After that, the proposals, instead of being passed to the RoI pooler stage, are used by the attacker to refine the patch. Specifically, the goal is to saturate the NMS mechanism with false positives in the area of the target object. These false proposals degrade the $objectness$ score of true ones and induce the NMS filter to drop the latter. As shown by Thys et al. [23], this approach results in more effective patches than attacking the target class's classification score or mixed objectness-class classification scores. Given a victim detector D, the core loss function of the attack is defined as:

$$L_{core} = \mathbb{E}_j \left(\sum_s \max\{obj(D_{s,\theta}(x_j, P)) + 1, 0\}^2 \right) \tag{1}$$

Generalization and printability of the patches is achieved in two ways:

- Using an auxiliary total variation regularization loss L_{tv} that ensures a smooth transition of colors across pixels. The importance of the regularization is regulated with a coefficient α such that:

$$L = L_{core} + \alpha L_{tv} \tag{2}$$

- Applying a set of random transformations on the patch before inserting it on the image during training. The random transformations used for this attack are rotation, translation, scaling, variation of brightness, and contrast.

C. Adversarial Training

One of the main goals of this paper is to evaluate the robustness of state-of-the-art computer vision by training adversarial patches using an Adam optimizer [40] with a starting learning rate of $\lambda = 3 \times 10^{-2}$.. All the patches were trained for 120000 iterations on the COCO dataset. After being normalized w.r.t. the dataset, images were fed one-by-one (i.e. batch size of 1) in order to have an equal comparison between all models, including bigger ones. The target class is the same as the original paper (i.e. ”Person”).

IV. Implementation

A. Implementation of Attack Loop

The similarity between the two models provides a unique opportunity to analyze fairly the effect of self-attention on the robustness of a detector. Therefore, one of our focuses in implementing the attack is ensuring that tested detector architectures share all the same components but the backbones and the attack loop are identical.

Both original implementations of the detectors were based on the library MMDet [41]. The rest of the detection loop is built atop the maskR-CNN benchmarking framework [42] and its implementation of mask R-CNN.

We used also the weights provided by the original authors. By doing so we provide a fair comparison w.r.t. the training of the detectors. All the core models (i.e. the backbones of the detectors) were pre-trained on ImageNet [43] while the final architecture was fine-tuned on the COCO dataset [17]. The

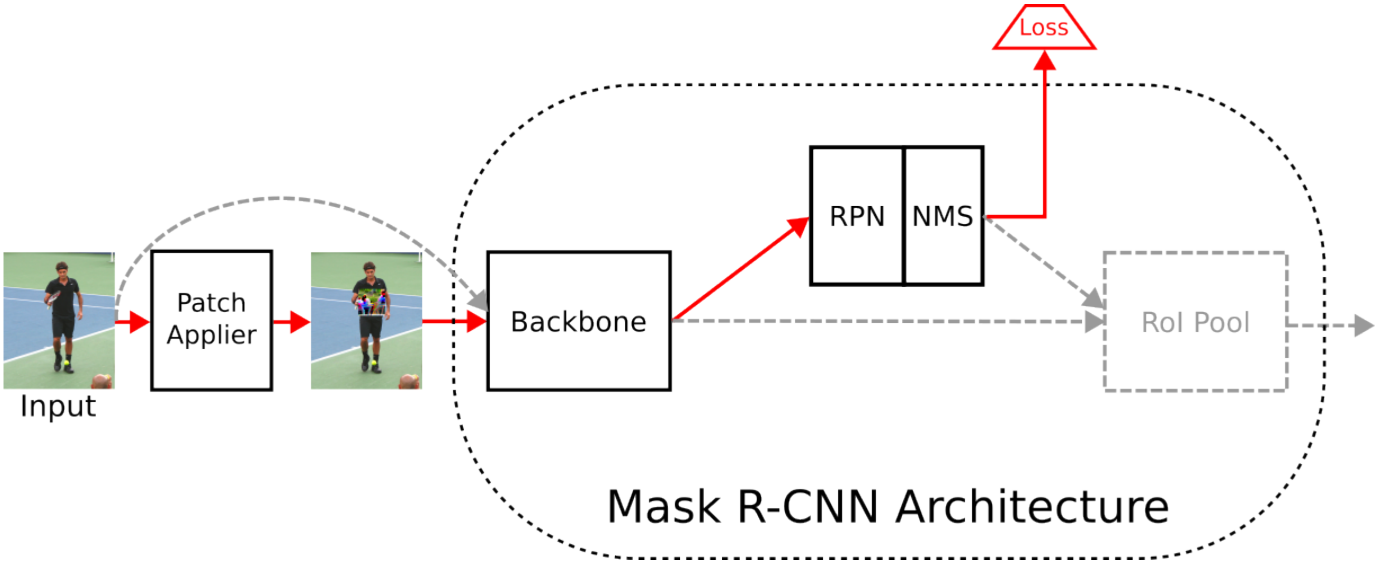


Fig. 4: Training loop (red path) of the attack compared with the detector training loop (grey one).

COCO dataset is also the reference dataset for training and testing the patches.

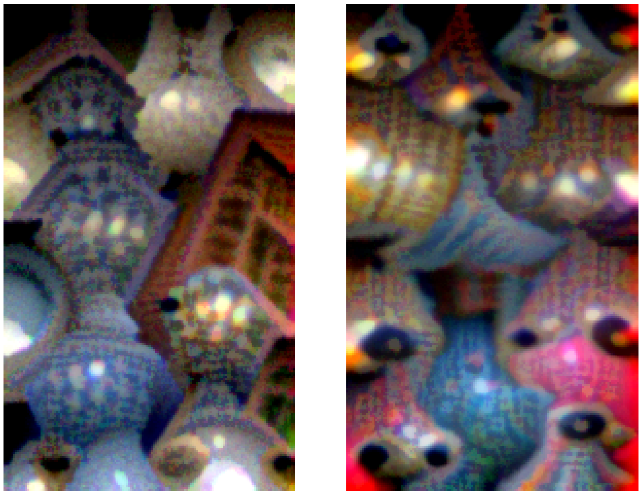
The attack starts from a randomly initialized 250×150 px patch. We used the annotation of the dataset to localize the people inside the images. Then the patches were applied with a random offset between 0 and 0.05 of the image height on y-axis from the center of the target bounding box. Additionally, the patches were scaled from their base sizes accordingly to the ones of their targets. To make it more robust, the patch is also randomly rotated with an angle in the interval $[-0.028\pi, 0.028\pi]$, while the contrast and the brightness are scaled by a random factor draw respectively in the interval $[0.9, 1.1]$ and $[0.8, 1.2]$.

V. Results

In the realm of object detection, the commonly used performance metric is Average Precision (AP) per class, which provides a balanced assessment of precision and recall. However, when evaluating physical attacks, success rates using a fixed detection threshold are often preferred due to their interpretability. Nevertheless, manually selecting appropriate thresholds for computing success rates can be challenging. To address this challenge and ensure a comprehensive evaluation, we report both the AP, which averages over all confidence thresholds, and success rates in our experimental results section. This section presents the AP results for Swin tiny and ConvNeXt tiny providing insights into their performance under the developed patch attacks and with a random patch of the same size. By considering both the AP and success rates, we thoroughly analyze the models' robustness against adversarial patches.

A. Adversarial Attack

All experiments have been performed on RTX-1080Ti GPU with 12 GB of memory and an Intel Xeon E5-2640 CPU.



(a) Swin Transformer (b) ConvNeXt

Fig. 5: Example of trained patches

Figure 5 presents the trained patches of base size 250×150 px for Mask R-CNN detectors with backbones Swin tiny and ConvNeXt tiny, each comprising of 28 M parameters. Table I presents the AP results at various Intersection over Union (IoU) thresholds for the "Person" class when different patches are applied.

When evaluating the Swin tiny model, the AP for clean images is 50.3%. However, when a trained adversarial patch is applied, the AP drops significantly to 12.8 %, indicating the model's vulnerability to the patch attack. Similarly, the ConvNeXt tiny model achieves an AP of 51.4 % on clean images, which decreases to 15.2 % when subjected to the trained patch.

Interestingly, when a random patch is applied instead of the trained patch, both models experience a less severe decline

Patch Victim	Clean	Swin Tiny	ConvNeXt Tiny	Random
Swin Tiny	50.3 %	12.8 %	18.4 %	36.4 %
ConvNeXt Tiny	51.4 %	15.2 %	6.7 %	34.8 %

TABLE I: The AP@[IoU=0.5:0.95] on the class ”Person” when the patches are applied

in performance. The AP for the Swin tiny model decreases to 18.4 %, while the ConvNeXt tiny model reaches an AP of 6.7 %. These results suggest that the trained patch is more effective at fooling the models than randomly generated patches.

Overall, the table highlights the impact of patch attacks on the performance of the models, demonstrating their vulnerability to carefully crafted adversarial patches and the varying degrees of success achieved by different patches.

The attack proved to actively temper the performances of both detectors as the trained patches were significantly more effective than a random one. Unlike previous analysis on the image classification task [29], the results show a significant difference in terms of robustness between a transformer-based architecture and a CNN. Moreover, the patches trained on the Swin transformer generalize more effectively.

VI. Conclusion

This paper investigated the adversarial robustness of transformer-based models against patch attacks in the context of Object Detection. Through our evaluation, we demonstrate the susceptibility of state-of-the-art Swin tiny and ConvNeXt tiny models to patch attacks, leading to a significant decrease in AP for the ”Person” class. We find that trained adversarial patches result in a substantial decline in AP, highlighting the vulnerability of these models to carefully crafted attacks. Our main contributions include the implementation and evaluation of patch attacks, an assessment of adversarial training for patch attacks, and comprehensive measurements and comparisons of robustness between CNN- and transformer-based models. These contributions enhance our understanding of the strengths and weaknesses of self-attention mechanisms and advance the field of adversarial robustness in computer vision.

Acknowledgement

The authors would like to thank *armasuisse Science & Technology* for funding this research.

References

[1] G. Cerutti, R. Andri, L. Cavigelli, E. Farella, M. Magno, and L. Benini, “Sound event detection with binary neural networks on tightly power-constrained iot devices,” in *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, 2020, pp. 19–24.

[2] F. Conti, D. Palossi, R. Andri, M. Magno, and L. Benini, “Accelerated visual context classification on a low-power smartwatch,” *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 19–30, 2016.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is All you Need,” in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Jun. 2021, arXiv:2010.11929 [cs]. [Online]. Available: http://arxiv.org/abs/2010.11929

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-End Object Detection with Transformers,” arXiv, Tech. Rep. arXiv:2005.12872, May 2020, arXiv:2005.12872 [cs] type: article. [Online]. Available: http://arxiv.org/abs/2005.12872

[6] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” arXiv, Tech. Rep. arXiv:2012.12877, Jan. 2021, arXiv:2012.12877 [cs] type: article. [Online]. Available: http://arxiv.org/abs/2012.12877

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” arXiv, Tech. Rep. arXiv:2103.14030, Aug. 2021, arXiv:2103.14030 [cs]. [Online]. Available: http://arxiv.org/abs/2103.14030

[8] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, “DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection,” Jul. 2022, arXiv:2203.03605 [cs]. [Online]. Available: http://arxiv.org/abs/2203.03605

[9] X. Luo, J. Zhang, K. Yang, A. Roitberg, K. Peng, and R. Stiefelhagen, “Towards robust semantic segmentation of accident scenes via multi-source mixed sampling and meta-learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4429–4439.

[10] L. Thangiah, C. Ramanathan, and L. S. Chodisetty, “Distribution transformer condition monitoring based on edge intelligence for industrial iot,” in *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*. IEEE, 2019, pp. 733–736.

[11] A. Singh and B. Sikdar, “Black-box adversarial attack for deep learning classifiers in iot applications,” in *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)*. IEEE, 2022, pp. 1–6.

[12] S.-T. Chen, C. Cornelius, J. Martin, and D. H. Chau, “ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector,” *arXiv:1804.05810 [cs, stat]*, vol. 11051, pp. 52–68, 2019, arXiv: 1804.05810. [Online]. Available: http://arxiv.org/abs/1804.05810

[13] 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,” *arXiv:1910.11099 [cs]*, Jul. 2020, arXiv: 1910.11099. [Online]. Available: http://arxiv.org/abs/1910.11099

[14] M. Magno, G. A. Salvatore, P. Jokic, and L. Benini, “Self-sustainable smart ring for long-term monitoring of blood oxygenation,” *IEEE access*, vol. 7, pp. 115 400–115 408, 2019.

[15] N. Parmar, A. Vaswani, J. Uszkoreit, Kaiser, N. Shazeer, A. Ku, and D. Tran, “Image Transformer,” arXiv, Tech. Rep. arXiv:1802.05751, Jun. 2018, arXiv:1802.05751 [cs] type: article. [Online]. Available: http://arxiv.org/abs/1802.05751

[16] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, “Swin Transformer V2: Scaling Up Capacity and Resolution,” Apr. 2022, arXiv:2111.09883 [cs]. [Online]. Available: http://arxiv.org/abs/2111.09883

[17] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, “Microsoft COCO: Common Objects in Context,” May 2014. [Online]. Available: https://arxiv.org/abs/1405.0312v3

[18] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv:1312.6199 [cs]*, Feb. 2014, arXiv: 1312.6199. [Online]. Available: http://arxiv.org/abs/1312.6199

[19] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, “Adversarial Patch,” *arXiv:1712.09665 [cs]*, May 2018, arXiv: 1712.09665. [Online]. Available: http://arxiv.org/abs/1712.09665

[20] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, “Synthesizing Robust Adversarial Examples,” *arXiv:1707.07397 [cs]*, Jun. 2018, arXiv: 1707.07397. [Online]. Available: http://arxiv.org/abs/1707.07397

[21] J. Lu, H. Sibai, and E. Fabry, “Adversarial Examples that Fool Detectors,” *arXiv:1712.02494 [cs]*, Dec. 2017, arXiv: 1712.02494. [Online]. Available: http://arxiv.org/abs/1712.02494

[22] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” arXiv,

Tech. Rep. arXiv:1506.01497, Jan. 2016, arXiv:1506.01497 [cs] type: article. [Online]. Available: http://arxiv.org/abs/1506.01497

[23] S. Thys, W. Van Ranst, and T. Goedemé, "Fooling automated surveillance cameras: adversarial patches to attack person detection," *arXiv:1904.08653 [cs]*, Apr. 2019, arXiv: 1904.08653. [Online]. Available: http://arxiv.org/abs/1904.08653

[24] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," arXiv, Tech. Rep. arXiv:1612.08242, Dec. 2016, arXiv:1612.08242 [cs] type: article. [Online]. Available: http://arxiv.org/abs/1612.08242

[25] Z. Wu, S.-N. Lim, L. Davis, and T. Goldstein, "Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors," *arXiv:1910.14667 [cs, math]*, Jul. 2020, arXiv: 1910.14667. [Online]. Available: http://arxiv.org/abs/1910.14667

[26] A. Zolfi, M. Kravchik, Y. Elovici, and A. Shabtai, "The Translucent Patch: A Physical and Universal Attack on Object Detectors," 2021, pp. 15 232–15 241.

[27] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, "Understanding Robustness of Transformers for Image Classification," *arXiv:2103.14586 [cs]*, Oct. 2021, arXiv: 2103.14586. [Online]. Available: http://arxiv.org/abs/2103.14586

[28] K. Mahmood, R. Mahmood, and M. van Dijk, "On the Robustness of Vision Transformers to Adversarial Examples," arXiv, Tech. Rep. arXiv:2104.02610, Jun. 2021, arXiv:2104.02610 [cs] type: article. [Online]. Available: http://arxiv.org/abs/2104.02610

[29] Y. Bai, J. Mei, A. Yuille, and C. Xie, "Are Transformers More Robust Than CNNs?" arXiv, Tech. Rep. arXiv:2111.05464, Nov. 2021, arXiv:2111.05464 [cs] type: article. [Online]. Available: http://arxiv.org/abs/2111.05464

[30] Y. Fu, S. Zhang, S. Wu, C. Wan, and Y. Lin, "Patch-Fool: Are Vision Transformers Always Robust Against Adversarial Perturbations?" *arXiv:2203.08392 [cs]*, Apr. 2022, arXiv: 2203.08392. [Online]. Available: http://arxiv.org/abs/2203.08392

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," arXiv, Tech. Rep. arXiv:1512.03385, Dec. 2015, arXiv:1512.03385 [cs] type: article. [Online]. Available: http://arxiv.org/abs/1512.03385

[32] N. Ji, Y. Feng, H. Xie, X. Xiang, and N. Liu, "Adversarial YOLO: Defense Human Detection Patch Attacks via Detecting Adversarial Patches," *arXiv:2103.08860 [cs]*, Mar. 2021, arXiv: 2103.08860. [Online]. Available: http://arxiv.org/abs/2103.08860

[33] J. Liu, A. Levine, C. P. Lau, R. Chellappa, and S. Feizi, "Segment and Complete: Defending Object Detectors against Adversarial Patch Attacks with Robust Patch Detection," *arXiv:2112.04532 [cs, eess]*, May 2022, arXiv: 2112.04532. [Online]. Available: http://arxiv.org/abs/2112.04532

[34] P.-Y. Chiang, R. Ni, A. Abdelkader, C. Zhu, C. Studer, and T. Goldstein, "Certified Defenses for Adversarial Patches," *arXiv:2003.06693 [cs, stat]*, Sep. 2020, arXiv: 2003.06693. [Online]. Available: http://arxiv.org/abs/2003.06693

[35] H. Han, K. Xu, X. Hu, X. Chen, L. Liang, Z. Du, Q. Guo, Y. Wang, and Y. Chen, "ScaleCert: Scalable Certified Defense against Adversarial Patches with Sparse Superficial Layers," Nov. 2021, arXiv:2110.14120 [cs]. [Online]. Available: http://arxiv.org/abs/2110.14120

[36] C. Xiang, A. Valtchanov, S. Mahloujifar, and P. Mittal, "ObjectSeeker: Certifiably Robust Object Detection against Patch Hiding Attacks via Patch-agnostic Masking," Feb. 2022, arXiv:2202.01811 [cs]. [Online]. Available: http://arxiv.org/abs/2202.01811

[37] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," arXiv, Tech. Rep. arXiv:2201.03545, Mar. 2022, arXiv:2201.03545 [cs]. [Online]. Available: http://arxiv.org/abs/2201.03545

[38] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," Mar. 2017. [Online]. Available: https://arxiv.org/abs/1703.06870v3

[39] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into High Quality Object Detection," Dec. 2017. [Online]. Available: https://arxiv.org/abs/1712.00726v1

[40] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv, Tech. Rep. arXiv:1412.6980, Jan. 2017, arXiv:1412.6980 [cs] type: article. [Online]. Available: http://arxiv.org/abs/1412.6980

[41] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open MMLab Detection Toolbox and Benchmark," Jun. 2019, arXiv:1906.07155 [cs, eess]. [Online]. Available: http://arxiv.org/abs/1906.07155

[42] F. Massa and R. Girshick, "maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch," 2018. [Online]. Available: https://github.com/facebookresearch/maskrcnn-benchmark

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255, iSSN: 1063-6919.