# Interpretation Conflict in Helmet Recognition under Adversarial Attack

**He Wen, Simaan AbouRizk**

Department of Civil and Environmental Engineering, University of Alberta, Canada
hwen7@ualberta.ca, abourizk@ualberta.ca

**Abstract –**

**Humans and Artificial Intelligence (AI) may have observation and interpretation conflicts in collaborative interaction. The adversarial samples make such conflicts more likely to occur in the field of image recognition. However, few studies have been seen combining the human-AI conflict and adversarial attack. This study presents the interpretation conflict due to adversarial samples in the helmet recognition task. A simulation also has been conducted to illustrate this problem. The results show that it should be prudent for the construction industry to land AI applications due to adversarial attacks on image recognition; the adversarial samples easily trigger interpretation conflicts, for example, the logo, graffiti, sticker, and text on helmets; lean construction should be propagated for the preconditions for AI applications.**

**Keywords** –

**Human-AI conflict; Adversarial attack; Risk; Cross-entropy; Distance**

## 1. Introduction

Artificial intelligence (AI) and machine learning have enabled numerous creative applications in construction operations [1]. Typical examples include face recognition, personnel positioning, and violation detection [2], [3], contributing to safety management and operations. More specifically, helmet detection is one of the mature experiments [4], [5], since the safety helmet is the most essential and mandated personal protective equipment (PPE), and violation is still occasional or even expected. For example, a recent survey in California shows that 36% of construction workers struggle to ensure they consistently wear PPE [6], even if they are all well-equipped at daily check-in. Similar studies also indicate that on-site supervision and enforcement are required but time/effort consuming [7], [8].

Fortunately, computer vision and image recognition with deep learning facilitate this task instead of human inspection by vision (Figure 1), integrating body detection and personnel location [9], [10], [11]. As the pioneer field of AI, image recognition of helmets has constantly improved its accuracy in academic experiments [12], especially with the version update of the algorithm of You Only Look Once (YOLO) [13]. While the majority of research findings boast an accuracy rate exceeding 90%, the authors endeavoured to replicate these experiments utilizing algorithms outlined in published papers and publicly available construction site images, however, the accuracy was still unsatisfactory.


Figure 1. Worker location and helmet detection [10]

One significant cause of such accuracy problems is the samples in field applications often have some noise or are heavily polluted. For example, in the helmet recognition task, the logo, graffiti, sticker, and text might be considered the adversarial samples (Figure 2), or even the light and shadow may manipulate the results. This is the phenomenon of adversarial attacks [14]. The deep learning neural network misclassifies the adversarial sample by adding an imperceptible perturbation to the original image [15]. Indeed, the problem of adversarial attacks in image recognition has received long-term attention and research, and feasible countermeasures have been proposed [16], [17]. However, in the field of helmet detection, many studies do not mention this issue.


Figure 2. Adversarial samples of helmets

Once the adversarial sample misleads the AI, it occurs a typical human-AI conflict [18], both observation conflict and interpretation conflict. This can trigger false alerts to workers or false violation records. On the other hand, it may also miss the detection of the helmet. In a critical environment, such a situation might trigger prudent risks. Therefore, this study aims to present this problem, alert the practitioners about this risk, and then demonstrate it through a simulation. The novelty of this paper is:

i. The mathematical expression of interpretation conflict in image recognition.
ii. A measurement of the conflict based on the vector distance and cross-entropy.
iii. The combination of helmet recognition and adversarial attack.

A reminder to the readers of this article: Section 2 describes the problem in mathematical expressions; Section 3 presents the simulation of a case; Section 4 summarizes the discussion of the simulation results and solutions to the proposed problem; Section 5 includes the remarkable conclusions, contributions, and limitations.

## 2. Problem statement

For computer vision, AI regards a picture as a $height \times width \times channel$ matrix, usually with a basic kernel of 3 $pixel \times 3\ pixel \times 3\ RGB$ [19], where RGB means red, green, and blue. Then the matrix is converted to a high-dimensional column vector by the three channels. On the other side, humans do not yet know how the brain works, from seeing a picture to recognizing the classification of the picture, at least not mathematically. Therefore, assume that the same is true for humans, and the human observation is also converted into a vector, then the variable of observation difference (VOD) for observation conflict can be expressed as [20]:

$$VOD = X_A - X_H \quad (1)$$

Where is $X_A$ the AI vector and $X_H$ is the human vector.

In addition, AI further performs deep learning with the convolutional neural network (CNN) to get the score, then applies the Softmax function to transfer the score to the classification probability $\hat{y}_A$. The last step is to conclude the classification result $y_A$ through the cross-entropy function, where $y_A$ is a one-hot vector. Naturally, when humans see an image, they estimate the probability $\hat{y}_H$ for a limited number of classifications, and then get the result $y_H$, which can also be expressed as a one-hot vector. Usually, humans could recognize their classification result $y_H$ immediately. An example of recognizing a helmet is shown in Figure 3. Thus, the variable of interpretation difference (VID) [18], which is the interpretation conflict, can be simplified as the difference between two $n \times 1$ one-hot vectors:

$$VID = Y_A - Y_H \quad (2)$$

Where $n$ is the number of classifications. When $VID = 0_{n\times1}$, there is no interpretation conflict; when $VID \neq 0_{n\times1}$, there is an interpretation conflict.

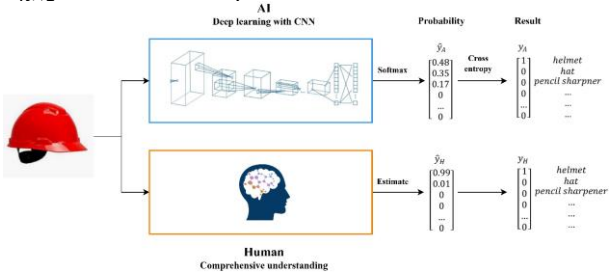


Figure 3. An example of how humans and AI recognize a picture

Then

$$VID \propto d = cross\ entropy(\hat{y}_A, y_H) \quad (3)$$

Where $d$ is the distance between $\hat{y}_A$ and $y_H$. This distance could be applied to measure the interpretation conflict under various noises, including adversarial samples.

Due to the improvement of AI learning ability, the accuracy rate has been increasing for the recognition task of standard samples, which is close to human cognition, reaching above 80% accuracy in 50-150ms [21], [22]. However, adversarial samples have a greater chance of interpretation conflict. As described in the Introduction, a small perturbation is added to the picture. Typically, a perturbation involves increasing or decreasing small values to/from each pixel of the image. Then humans cannot tell the difference between before and after, and get the same classification result. However, AI may give an unexpected result; for example, Figure 4 shows that AI recognizes a helmet as a pencil sharpener under adversarial attack. Here the $VID = [-1,0,1,0, \dots ,0]^T$.

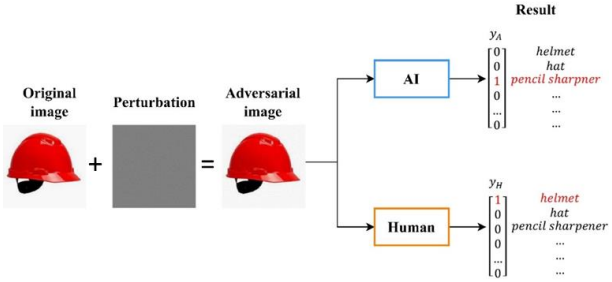


Figure 4. An example of interpretation conflict under adversarial attack

Therefore, the problems to be presented and solved in this study are: Does AI accurately detect whether workers are wearing helmets on construction sites, and do adversarial samples potentially trigger interpretation conflicts between human supervisors and AI?

## 3. Simulation and results

Though helmet identification is a typical application of image recognition with high accuracy, under adversarial attack, it may show unpredictable errors. Therefore, the simulation is designed as the following major steps to present this problem (Figure 5).

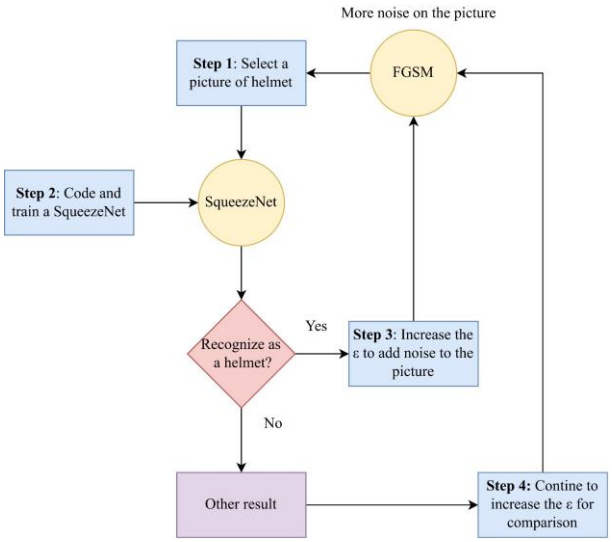


Figure 5. Simulation procedure

**Step 1**: Randomly search and choose a helmet picture from the public Internet as the test sample.

**Step 2**: Code the program in MATLAB r2022a to train and test the image with SqueezeNet and the Fast Gradient Sign Method (FGSM) [23].

SqueezeNet is a pre-trained neural network with plenty of classification labels and a relatively small computational occupation suitable for academic simulation [24]. This study added more training samples to optimize the SqueezeNet to identify the test examples.

200 training samples of safety helmets from the public Internet are added (Figure 6), and the label is marked "safety helmet" (Hereinafter referred to as "helmet") to distinguish the "crash helmet" in the original SqueezeNet.

Figure 6. Example of training samples

The FGSM is a mature technique for generating adversarial samples [15], and it has

$$X_{adv} = X + \varepsilon * sign(\nabla_X L(X,T)) \quad (4)$$

Where $X$ is the original image vector, $X_{adv}$ is the adversarial image vector, $\nabla_X L(X,T)$ is the gradient of the loss function $L$ to the targeted label $T$; $\varepsilon$ controls the size of the push and the adversarial strength, which means that the larger the $\varepsilon$ value, the greater the perturbation.

**Step 3**: Increase $\varepsilon$ gradually until the classification result changes from "helmet" to another label. The procedure is designed to trigger an interpretation conflict.

**Step 4**: Continue to increase $\varepsilon$ to generate enough conflict results for comparison and discussion. After the simulation, the results are shown in Figure 7.

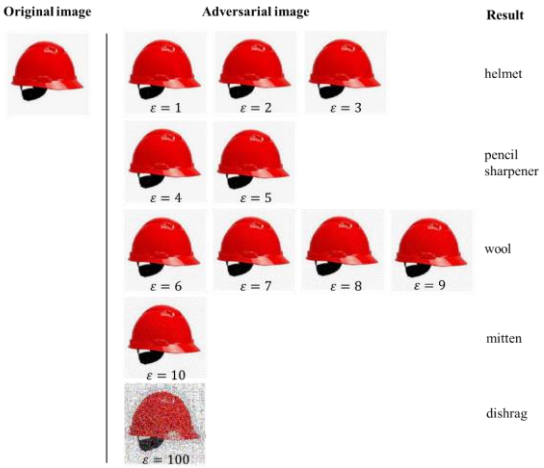


Figure 7. Original image and adversarial images

As the attack strength increases, in other words, the noise increases, it misleads the AI to recognize the helmet as "pencil sharpener", "wool", "mitten", and "dishrag". The relation between conflict measurement (distance $d$) and attack strength (control parameter $\varepsilon$) is shown in Figure 8.

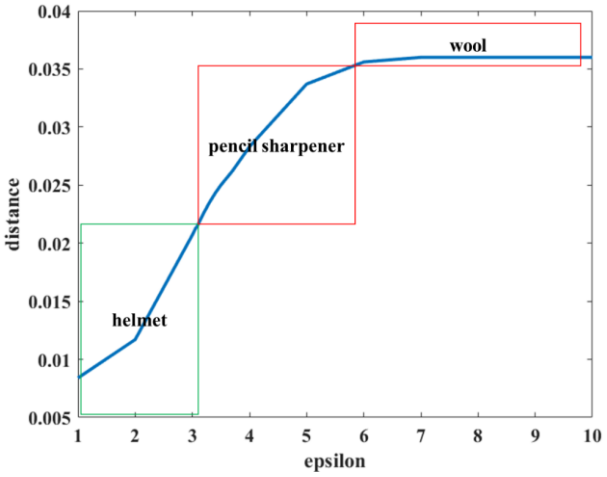


Figure 8. The relation between conflict measurement and attack strength

## 4. Discussion

The simulation results presented in this study shed light on the vulnerability of AI-based image recognition systems to adversarial attacks, particularly in helmet detection on construction sites. While this issue is present across all PPE detection scenarios, safety helmets are particularly vulnerable to adversarial attacks compared to safety vests and shoes. This susceptibility has been confirmed through extensive experimentation with various items, leading the authors to highlight it as a typical scenario in helmet detection.

By systematically increasing the attack strength, as controlled by the parameter $\varepsilon$ in the FGSM, it is observed a notable increase in interpretation conflicts, manifested as misclassifications by the AI system. One of the key findings of this study is the progressive deterioration in the AI system's performance as the attack strength increases. Initially, the AI system accurately identifies safety helmets from the test samples, achieving high classification accuracy. However, as the adversarial perturbations intensify, the system's confidence diminishes, ultimately resulting in misclassifications. Notably, the misclassifications observed in the simulation ranged from plausible but incorrect labels such as "pencil sharpener" to more evidently erroneous labels like "wool" or "mitten." This progression highlights the escalating confusion and uncertainty introduced by adversarial attacks.

The relationship between the conflict measurement (distance $d$) and attack strength ($\varepsilon$) provides valuable insights into the vulnerability of the AI system. As depicted in Figure 8, there is a clear positive correlation between attack strength and conflict measurement, indicating that stronger adversarial perturbations lead to greater deviation between the AI system's classification and the ground truth. This observation underscores the sensitivity of AI systems to subtle changes in input data, which can be exploited to induce interpretation conflicts and undermine the system's reliability.

From the simulation results, the mitigation strategies can be induced. One solution is adversarial training with adversarial samples, for example, enabling recurring training based on false positive data identified from construction sites.

In addition, model robustness evaluation would encourage the model to learn robust representations that are resilient to adversarial perturbations. It tunes model sensitivity to have a higher tolerance for various types of image qualities.

Furthermore, implementing defense mechanisms with adversarial sample detection can help mitigate adversarial attacks, since input data pre-processing can improve the sample quality.

## 5. Conclusions

This study points out the problem of helmet recognition under adversarial attack in the construction industry, which is a matter of deep concern with observation and interpretation conflict. The distance between AI prediction and human cognition could measure the human-AI conflict. This reminds practitioners not to mindlessly launch new AI applications and ignore the weaknesses and defects of AI technology itself.

This study underscores the discrepancy between AI-based image recognition systems and human perception. This interpretation conflict raises important questions regarding the limitations of current AI technologies and the need for further research to bridge the gap between AI and human cognition. Moreover, the findings of this study have significant implications for the deployment of AI-based safety monitoring systems in real-world contexts. The susceptibility of these systems to adversarial attacks underscores the importance of rigorous testing and validation procedures to assess their robustness and reliability. Lastly, promoting education and awareness initiatives of adversarial attacks can increase understanding of the capabilities, limitations, and risks associated with AI technologies. This is also the main intention of this study.

Thus, this study serves as a reminder to both industry and academia to consider the diverse array of environmental disturbances present at construction sites when employing AI technology. It underscores the importance of the construction site environment, since dirtiness, dim lighting, and outdated equipment/tools can potentially create adversarial samples. As a result, it suggests the preconditions for AI application, for example, maintaining cleanliness, ensuring adequate lighting, regularly maintaining equipment/tools/PPE, and adhering to standardization protocols for safety signs. They are equally vital for enhancing the precision of image recognition. Moreover, these are also the basic requirements for lean construction management.

Despite the contributions and insights, several limitations must be acknowledged to ensure a comprehensive understanding. Firstly, the simulation environment employed in this study inherently simplifies the complexity of real-world scenarios, for example, the training and test samples are from the public Internet, not real construction sites. Moreover, the generalizability of the findings and proposed solutions may be constrained by the specific characteristics of the AI models, datasets, and application domains. Also, the efficacy of the proposed solutions may vary depending on factors such as the architecture of the AI system, the nature of the adversarial attacks, and the diversity of the input data.

Therefore, future research should aim to address these limitations and explore new approaches to enhance the

robustness and reliability of AI systems in safety-critical applications, such as helmet recognition in this study. The research and practice of AI reliability are full of challenges and encourage further exploration.

## References


[1] M. Regona, T. Yigitcanlar, B. Xia, and R. Y. M. Li, "Opportunities and adoption challenges of AI in the construction industry: a PRISMA review," *Journal of open innovation: technology, market, and complexity*, vol. 8, no. 1, p. 45, 2022.

[2] J. Chai, H. Zeng, A. Li, and E. W. T. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," *Machine Learning with Applications*, vol. 6, no. August, p. 100134, 2021.

[3] W. Lin, J. Xue, C. Zhu, F. Jiang, and C. Ding, "An integrated application platform for safety management and control of smart power plants based on video image technology," in *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, IEEE, 2022, pp. 1436–1439.

[4] H. Peng and Z. Zhang, "Helmet Wearing Recognition of Construction Workers Using Convolutional Neural Network," *Wirel Commun Mob Comput*, vol. 2022, 2022.

[5] K. Li, X. Zhao, J. Bian, and M. Tan, "Automatic safety helmet wearing detection," *arXiv preprint arXiv:1802.00264*, 2018.

[6] A. N. Gattuso, "Common Issues of Compliance with Personal Protective Equipment for Construction Workers," 2021.

[7] T. K. M. Wong, S. S. Man, and A. H. S. Chan, "Critical factors for the use or non-use of personal protective equipment amongst construction workers," *Saf Sci*, vol. 126, p. 104663, 2020.

[8] A. D. Rafindadi, M. Napiah, I. Othman, H. Alarifi, U. Musa, and M. Muhammad, "Significant factors that influence the use and non-use of personal protective equipment (PPE) on construction sites—Supervisors' perspective," *Ain Shams Engineering Journal*, vol. 13, no. 3, p. 101619, 2022.

[9] L. Wang *et al.*, "Investigation Into Recognition Algorithm of Helmet Violation Based on YOLOv5-CBAM-DCN," *IEEE Access*, vol. 10, pp. 60622–60632, 2022.

[10] J. Shen, X. Xiong, Y. Li, W. He, P. Li, and X. Zheng, "Detecting safety helmet wearing on construction sites with bounding-box regression and deep transfer learning," *Computer-Aided Civil and Infrastructure Engineering*, vol. 36, no. 2, pp. 180–196, 2021.

[11] L. Wei, M. Cheng, M. Feng, and Z. Lijuan, "Research on recognition of safety helmet wearing of electric power construction personnel based on artificial intelligence technology," *J Phys Conf Ser*, vol. 1684, no. 1, 2020.

[12] N. D. Nath, A. H. Behzadan, and S. G. Paal, "Deep learning for site safety: Real-time detection of personal protective equipment," *Autom Constr*, vol. 112, p. 103085, 2020.

[13] M. I. B. Ahmed *et al.*, "Personal protective equipment detection: A deep-learning-based sustainable approach," *Sustainability*, vol. 15, no. 18, p. 13990, 2023.

[14] C. Szegedy *et al.*, "Intriguing properties of neural networks," *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, pp. 1–10, 2014.

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–11, 2015.

[16] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," *Applied Sciences*, vol. 9, no. 5, p. 909, 2019.

[17] H. Xu *et al.*, "Adversarial attacks and defenses in images, graphs and text: A review," *International Journal of Automation and Computing*, vol. 17, pp. 151–178, 2020.

[18] H. Wen, Md. T. Amin, F. Khan, S. Ahmed, S. Imtiaz, and E. Pistikopoulos, "Assessment of Situation Awareness Conflict Risk between Human and AI in Process System Operation," *Ind Eng Chem Res*, vol. 62, no. 9, pp. 4028–4038, Mar. 2023.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–14, 2015.

[20] H. Wen, Md. T. Amin, F. Khan, S. Ahmed, S. Imtiaz, and S. Pistikopoulos, "A methodology to assess human-automated system conflict from safety perspective," *Comput Chem Eng*, vol. 165, p. 107939, Jul. 2022.

[21] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun, "Deep image: Scaling up image recognition," *arXiv preprint arXiv:1501.02876*, vol. 7, no. 8, p. 4, 2015.

[22] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, no. 6582, pp. 520–522, 1996.

[23] M. H. Beale, M. T. Hagan, and H. B. Demuth, *Deep Learning Toolbox ™ User's Guide*. The MathWorks, Inc., 2023.

[24] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1–13, 2017.