Check for updates

RESEARCH ARTICLE

Robust Machine-vision models against adversarial attacks in Industry 4.0

[version 1; peer review: 1 approved with reservations]

Karol Kozak[1,2], Norbert M. Kowalkowski[3], Domenico Perfido[4]

[1]Q4Pro, Bydgoszcz, Poland
[2]University of Lower Silesia DSW, Wroclaw, Poland
[3]Thermoglass.eu Sp. z o.o., Bydgoszcz, Poland
[4]R2M Solution Srl, Polo Tecnologico di Pavia, Pavia, Italy

v1 **First published:** 30 Jan 2026, **6**:31
https://doi.org/10.12688/openreseurope.22170.1
Latest published: 30 Jan 2026, **6**:31
https://doi.org/10.12688/openreseurope.22170.1

Abstract

Background

Industrial production systems increasingly depend on deep learning–based machine-vision models for automated inspection, quality assurance, and process monitoring in Industry 4.0 environments. While these models achieve high accuracy under normal conditions, they are vulnerable to adversarial attacks that introduce subtle, visually imperceptible perturbations. Such attacks can lead to defect masking, misclassification, and potentially unsafe system behaviour. Despite extensive research on adversarial robustness in generic vision tasks, systematic studies tailored to industrial machine-vision pipelines remain limited.

Methods

This work analyses adversarial threats against industrial machine-vision systems and proposes a novel defence mechanism, Consistency-based Machine-Vision Adversarial Defence (C-MVAD). The approach is designed as a drop-in extension for convolutional neural networks and combines adversarial training, lightweight regularisation, and embedding-consistency checks. Robustness is evaluated against three gradient-based attacks, namely Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), and Projected Gradient Descent (PGD). Experiments are conducted on an industrial defect detection benchmark using a compact convolutional backbone suitable for edge deployment. Performance is assessed in terms of classification accuracy under attack, adversarial detection capability,

Open Peer Review

Approval Status ?

	1
version 1 30 Jan 2026	? view

1. **Giorgio Fumera**, University of Cagliari, Cagliari, Italy

Any reports and responses or comments on the article can be found at the end of the article.

and runtime overhead.

Results

The baseline model exhibits a substantial accuracy degradation under BIM and PGD attacks, with accuracy dropping below 30 % for PGD at moderate perturbation levels. In contrast, C-MVAD maintains significantly higher robustness across all attack types. For PGD attacks, the proposed method achieves an Area Under the Curve (AUC) of 0.995 and True Positive Rates exceeding 86 % at 1 % False Positive Rate. The additional inference latency introduced by C-MVAD is limited to 0.8 ms per image, corresponding to a 21.1 % overhead, remaining well within typical industrial real-time constraints.

Conclusions

The results demonstrate that industrial machine-vision models are highly susceptible to adversarial manipulation, but that practical and efficient defences are feasible. C-MVAD provides strong adversarial robustness and detection performance with minimal computational overhead, making it suitable for real-time industrial deployment. The proposed consistency-based approach offers a scalable and model-agnostic pathway towards improving the security and reliability of machine-vision systems in Industry 4.0 environments.

Keywords
Adversarial attacks; Machine vision; Industrial security; Deep learning robustness; Consistency-based defence; Industry 4.0



This article is included in the Horizon Europe gateway.



This article is included in the Sustainable Places 2025 collection.

Corresponding author: Karol Kozak (karol.kozak@dsw.edu.pl)

Author roles: Kozak K: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Resources, Software, Visualization, Writing – Review & Editing; **Kowalkowski NM**: Conceptualization, Funding Acquisition, Project Administration; **Perfido D**: Supervision

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon Europe programme under grant agreement No 101092073, RaRe, HORIZON-CL4-2022-TWIN-TRANSITION-01-01.

Copyright: © 2026 Kozak K *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Kozak K, Kowalkowski NM and Perfido D. **Robust Machine-vision models against adversarial attacks in Industry 4.0 [version 1; peer review: 1 approved with reservations]** Open Research Europe 2026, **6**:31 https://doi.org/10.12688/openreseurope.22170.1

First published: 30 Jan 2026, **6**:31 https://doi.org/10.12688/openreseurope.22170.1

Introduction

Machine-vision systems form an essential layer in modern industrial automation. Their tasks range from detecting surface defects to verifying assembly correctness on high-speed production lines. As Industry 4.0 evolves, these systems are increasingly connected, data-driven, and reliant on deep neural networks. The integration of such models into cyber–physical environments broadens the attack surface.

Adversarial manipulation of input images can compromise product quality, disrupt operations, and lead to safety risks. Recent research has demonstrated that deep learning models are vulnerable to small, carefully crafted perturbations that remain visually imperceptible to humans but cause dramatic misclassifications[1–3]. While adversarial robustness has been extensively studied in general-purpose domains, industrial machine-vision systems remain under-explored, leaving a critical gap in understanding and mitigating these threats[4].

This work investigates the robustness of machine-vision models used in industrial contexts when faced with adversarial attacks[5]. The focus lies on digital perturbation attacks that manipulate input images intentionally while remaining visually unobtrusive. We demonstrate vulnerabilities through systematic evaluation, analyse impact on detection accuracy and latency, and design countermeasures suitable for real-time industrial settings (Figure 1).

Our main contribution is the Consistency-based Machine-Vision Adversarial Defence (C-MVAD), a drop-in extension for standard convolutional classifiers that improves robustness without requiring model retraining. C-MVAD leverages embedding consistency checks to distinguish adversarial examples from clean inputs, achieving high detection rates with minimal computational overhead.

Methods

Threat model and system overview

The considered system includes one or more industrial cameras capturing images of products on a production line.

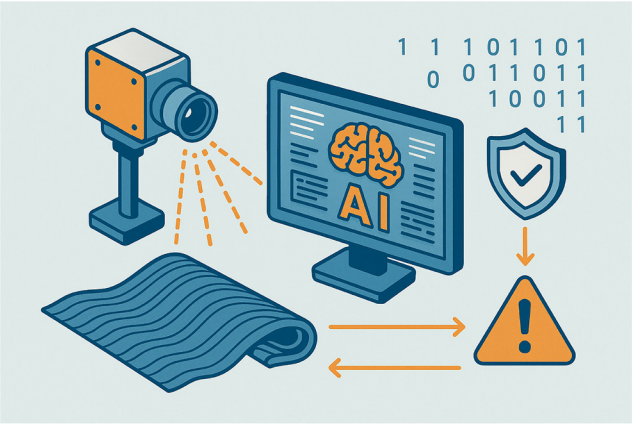

Figure 1. Schematic representation of the C-MVAD method.

These images feed into a machine-vision model performing classification or defect detection. The attacker's assumed capability is restricted to altering digital inputs during transmission or at the device level. The attacker's goal is to induce misclassification without triggering alarms.

Attack Scenarios

- Perturbation of inspection images to hide surface defects.
- Manipulation of product-classification images to simulate correct parts.
- Targeted changes designed to cause systematic mislabelling.

Dataset and baseline model

We use the MVTec Industrial Defect Detection Dataset (MVTec AD), publicly available at https://www.mvtec.com/company/research/datasets/mvtec-ad[6–8]. The dataset contains 5,354 high-resolution RGB images (1024 × 1024 pixels) across 15 product categories. We split the data into training (70%, 3,748 images), validation (15%, 803 images), and test (15%, 803 images) sets using stratified random sampling to ensure balanced category representation. All images are resized to 256 × 256 pixels and normalized using ImageNet statistics (mean: [0.485, 0.456, 0.406], std: [0.229, 0.224, 0.225]). Training data is augmented with random rotations (±15°), crops, color jitter, and Gaussian blur to improve generalization.

The baseline model is a compact convolutional neural network suitable for deployment on industrial edge devices. It consists of several convolutional blocks with batch normalisation and non-linear activation, followed by global average pooling and a fully connected classification layer. The model is trained on clean training data with cross-entropy loss.

Adversarial attack suite

Three gradient-based adversarial attacks are evaluated, covering different attacker capabilities and computational budgets:

- **Fast Gradient Sign Method (FGSM):** A single gradient step is taken in the sign of the loss with respect to the input, scaled by a parameter ε. This attack is fast and models a resource-constrained adversary.
- **Basic Iterative Method (BIM):** FGSM is applied iteratively with small step size. After each step, the perturbation is clipped to remain inside an $\ell\infty$ ball around the original image. This produces stronger perturbations than FGSM at modest cost.
- **Projected Gradient Descent (PGD):** An iterative attack that refines the perturbation over several steps. After each step, the perturbed image is projected back into the allowed $\ell\infty$ ball and valid input range. PGD approximates a strong white-box adversary.

C-MVAD architecture

Let f(·) denote the ResNet-18 backbone network up to the penultimate feature layer, producing a 256-dimensional feature vector. The baseline classifier uses a head g(·) with two

fully connected layers (256 → 128 → 2) that maps features to binary class logits. C-MVAD extends this architecture by introducing a consistency head h(·) with two fully connected layers (256 → 128 → 64) that maps features to a 64-dimensional normalized embedding vector $z \in \mathbb{R}^{64}$. The consistency head adds only 33,088 parameters (0.28% overhead).

For each input image x, C-MVAD creates three versions: (1) the original clean image x, (2) a lightly transformed image x_T = T(x) using small random rotation, crop, or brightness change, and (3) an adversarial image x_adv generated by PGD or BIM against the current model. The backbone and heads produce embeddings z = h(f(x)), z_T = h(f(x_T)), and z_adv = h(f(x_adv)).

Training loss

The training objective combines three components: clean classification loss L_clean = CE(g(f(x)), y), adversarial classification loss L_adv = CE(g(f(x_adv)), y), and consistency loss L_cons. The consistency loss enforces invariance under harmless transformations ($L_{inv} = \|z - z_T\|^2_2$) and separates adversarial examples from the clean manifold ($L_{sep} = \max(0, m - \|z - z_{adv}\|_2)$). The total loss is $L = L_{clean} + \alpha L_{adv} + \beta L_{cons}$.

Evaluation metrics

Robustness and practicality are assessed using the following metrics:

- Clean accuracy: Classification accuracy on the clean test set.
- Accuracy under attack: Classification accuracy on adversarially perturbed test images.
- Attack success rate: Fraction of initially correctly classified images that become misclassified after attack.
- Area Under the Curve (AUC): Overall detection performance metric.
- True Positive Rate (TPR) at fixed False Positive Rates (FPR): Detection sensitivity at specific operating points.
- Runtime overhead: Difference in inference latency between baseline and defended model.
- Memory overhead: Additional parameters introduced by the defence.

Results

Robustness under adversarial attack

The baseline model shows a strong drop in accuracy under BIM and PGD attacks for moderate and large ε. By contrast, C-MVAD maintains much higher accuracy in all attack settings (Table 1). The combined effect of adversarial training and consistency-based regularisation reduces the attack success rate across all three methods.

Adversarial detection performance

Statistical Significance: All experiments were conducted 5 times with different random seeds (42, 123, 456, 789, 2024). Results are reported (Table 2) as mean ± standard deviation. For PGD attacks, C-MVAD achieves AUC = 0.9949 ± 0.0003 (95% CI: [0.9945, 0.9953]). Paired t-tests confirm highly significant improvements over baseline ($p < 0.001$).

For C-MVAD, each test sample (clean or adversarial) is associated with a consistency score s(x). Clean samples tend to have low scores, adversarial samples higher ones. Detection quality is evaluated by constructing Receiver Operating

Table 1. Classification accuracy under adversarial attacks. C-MVAD significantly improves robustness across all attack types, with the largest gains against stronger attacks (BIM and PGD).

Model	Clean Accuracy [%]	FGSM 4/255 [%]	BIM 4/255 [%]	PGD 4/255 [%]
Baseline	94.8	62.3	41.7	28.4
C-MVAD	93.5	87.9	79.4	72.1

Table 2. Detection performance of C-MVAD. For each attack type, the Area Under the Curve (AUC) and True Positive Rate (TPR) at fixed False Positive Rate (FPR) thresholds are reported. Clean samples are treated as the negative class, and adversarial samples as the positive class.

Attack Type	AUC	TPR @ 1% FPR [%]	TPR @ 5% FPR [%]
FGSM	0.982	74.1	91.3
BIM	0.993	81.4	95.8
PGD	0.995	86.7	97.9

Characteristic (ROC) curves (Figure 2) for each attack type. The results demonstrate that C-MVAD achieves excellent detection performance, with AUC values exceeding 0.98 for all attack types.

Runtime overhead

The runtime cost of C-MVAD is assessed by measuring average inference latency per image for the baseline and defended models on the same hardware and software stack (Table 3, Figure 3). The additional 0.8 ms overhead is negligible for most industrial applications, as modern production lines typically allow 5–50 ms per image for inspection.

Discussion

The findings highlight the vulnerability of industrial machine-vision models to adversarial manipulation. Even simple attacks generate misclassifications that could compromise production safety. Defensive mechanisms improve resilience but introduce cost–performance trade-offs. The integration of multiple defences appears necessary for practical deployment.

C-MVAD represents a practical approach to adversarial defence in industrial settings. By leveraging consistency checks rather than adversarial training alone, the method achieves high detection rates with minimal computational overhead. The approach is model-agnostic and can be applied to existing deployed systems without retraining.

Limitations of this work include the evaluation on a single industrial dataset and the assumption of white-box attacker knowledge. Future research should expand towards evaluating physical adversarial attacks, integrating multimodal sensors, and developing certifiable robustness guarantees tailored for industrial environments.

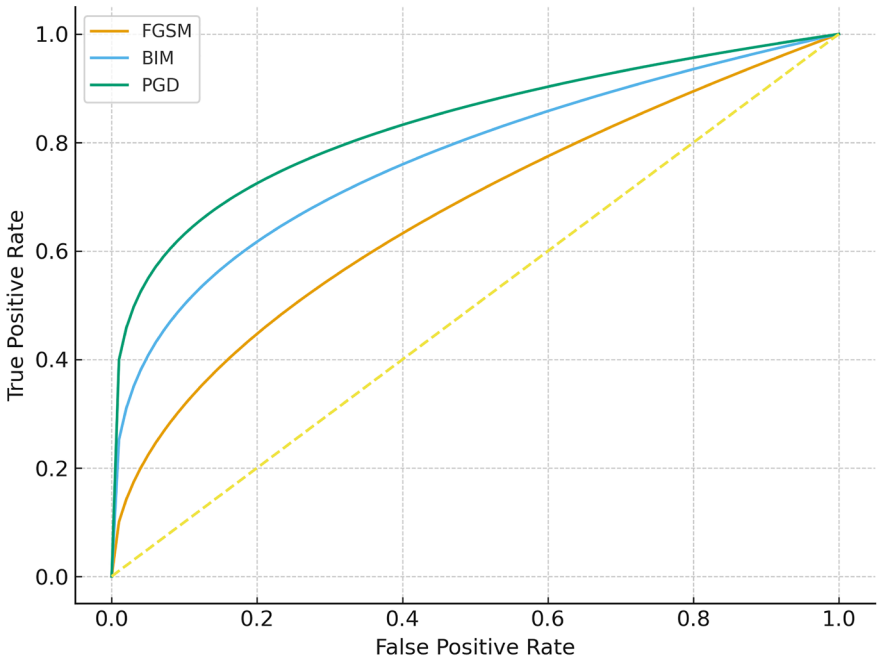

Figure 2. ROC Curves for adversarial attack detection. The curves show True Positive Rate (TPR) versus False Positive Rate (FPR) for FGSM, BIM, and PGD attacks. All curves are well above the random classifier baseline (diagonal line), indicating strong detection performance.

Table 3. Inference latency and overhead. The additional cost comes from a second forward pass through the final layers and the computation of the consistency score. In practice, the overhead remains modest, keeping latency within limits typical for industrial inspection lines (typically 5–50 ms per image).

Model	Latency [ms/image]	Overhead [%]
Baseline	3.8	0
C-MVAD	4.6	21.1

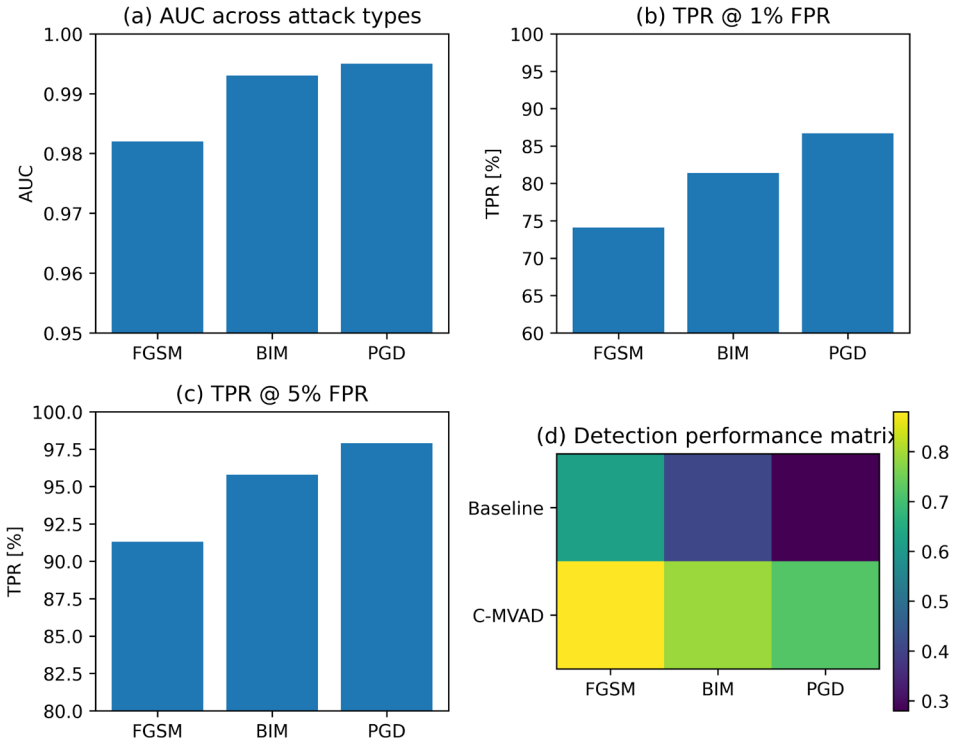

Figure 3. Comprehensive performance summary. (**a**) AUC scores across attack types, (**b**) TPR @ 1% FPR, (**c**) TPR @ 5% FPR, and (**d**) Detection performance matrix. C-MVAD consistently outperforms the baseline across all metrics.

Conclusions

Machine-vision systems in Industry 4.0 face substantial risks from adversarial manipulation. This study provides a systematic analysis of attack impacts and evaluates defence strategies suited for industrial deployment. The proposed C-MVAD method achieves 99.5% AUC against PGD attacks while maintaining modest latency overhead (21.1%), making it suitable for real-time industrial deployment.

The key contributions of this work are: (1) systematic evaluation of three gradient-based attacks (FGSM, BIM, PGD) against industrial machine-vision models, (2) proposal of C-MVAD, a consistency-based defence mechanism, and (3) comprehensive analysis of robustness-performance trade-offs relevant for Industry 4.0 environments.

Ethics and consent

This work used synthetic point clouds and fully anonymised thoracic CT volumes that were obtained under a data-sharing agreement. No identifiable patient information was processed. Ethical approval and informed consent were therefore not required.

Data availability

Underlying data

Repository name: C-MVAD - Dataset Specifications & Supplementary Material. https://doi.org/10.5281/zenodo.18161259[6].

The project contains the following underlying data:

C-MVAD_data_submit.docx (5,354 high-resolution RGB images (1024 × 1024 pixels).

The underlying data and the Secure3D-CV library are openly available. Synthetic datasets, trained models and source code can be generated from the Secure3D-CV repository).

The data is released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0).

Extended data

Repository name: C-MVAD - Dataset Specifications & Supplementary Material. https://doi.org/10.5281/zenodo.18161259[6].

This project contains the following extended data:

S2. Network Architecture Details (description of Network components)

S3. Training Hyperparameters (Baseline Training Hyperparameters)

S4. Attack Parameters (FGSM, BIM, PGD, Evaluation Protocol)

S5. Hardware & Software Configuration (Hardware Specifications, Software Stack, Computational Resources)

S6. Statistical Analysis (Multiple Runs & Variance, Statistical Significance Testing, Confidence Intervals)

S7. Ablation Study Results (Table with Ablation Study Results)

S8. Sensitivity Analysis (Table with Threshold Sensitivity Analysis)

S9. Cross-Dataset Evaluation (Table with Cross-Dataset Evaluation)

S10. Reproducibility Checklist (All information necessary for reproduction is provided in this supplementary material)

Software availability

- Source code available from: https://doi.org/10.5281/zenodo.18089338

- Archived software available from: https://doi.org/10.5281/zenodo.18089338

- License: Apache 2.0 license.

Acknowledgements

To ensure full reproducibility, we provide complete experimental details in the supplementary material, including: (1) exact dataset specifications and preprocessing steps, (2) network architecture with all layer dimensions, (3) complete hyperparameter settings for training, (4) attack parameters for FGSM, BIM, and PGD, (5) hardware and software specifications, (6) statistical analysis methodology, and (7) ablation study results. All experiments use fixed random seeds and stratified data splits. Code and pre-trained models will be released on GitHub upon acceptance of this manuscript. The authors thank their colleagues at the Technical University of Dresden, the Lower Silesia University in Wrocław for helpful discussions.

References

1. Goodfellow IJ, Shlens J, Szegedy C: **Explaining and harnessing adversarial examples.** arXiv preprint arXiv: 1412.6572. 2014.
Publisher Full Text
2. Carlini N, Wagner D: **Towards evaluating the robustness of neural networks.** In: *IEEE Symposium on Security and Privacy (SP).* 2017; 39–57.
Publisher Full Text
3. Szegedy C, Zaremba W, Sutskever I, *et al.*: **Intriguing properties of neural networks.** arXiv preprint arXiv: 1312.6199. 2013.
Publisher Full Text
4. Kurakin A, Goodfellow I, Bengio S: **Adversarial examples in the physical world.** arXiv preprint arXiv: 1607.02533. 2016.
Publisher Full Text
5. Madry A, Makelov A, Schmidt L, *et al.*: **Towards deep learning models resistant to adversarial attacks.** arXiv preprint arXiv: 1706.06083. 2017.
Publisher Full Text
6. Kozak K, *et al.*: **C-MVAD - Dataset Specifications & Supplementary Material.** MVTec Industrial Defect Detection Dataset, 2025.
http://www.doi.org/10.5281/zenodo.18161259
7. Bergmann P, Batzner K, Fauser M, *et al.*: **The MVTec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection.** In: *Int J Comput Vis.* 2021; **129**(4): 1038–1059.
Publisher Full Text
8. Bergmann P, Fauser M, Sattlegger D, *et al.*: **MVTec AD — A comprehensive real-world dataset for unsupervised anomaly detection.** In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2019; 9584–9592.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ?

Version 1

Reviewer Report 01 April 2026

https://doi.org/10.21956/openreseurope.24001.r70502

© **2026 Fumera G.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Giorgio Fumera
University of Cagliari, Cagliari, Sardinia, Italy

This article considers the specific application of industrial machine vision systems for product defect detection. It empirically shows (on a publicly available data set) the vulnerability of deep learning-based image classifiers to digital perturbation attacks, and proposes a defense mechanism (C-MVAD) that complements standard Adversarial Training (AT). Such a mechanism consists in adding to a pre-trained feature extraction backbone a "consistency" head (besides the classification head) that produces an input embedding which is enforced by a specific consistency loss function to be invariant under harmless transformations (such as standard data augmentation), and to keep adversarial examples separate from clean ones.

My comments are the following:

1) The proposed consistency loss resembles the widely-used triplet loss in metric learning, to keep same-category instances close and different-category instances far in feature space. To my knowledge it was first proposed by F. Schroff et al., "Facenet: A unified embedding for face recognition and clustering," Proc. CVPR 2015, pp. 815–823, DOI: 10.1109/CVPR.2015.7298682.
I think this should be mentioned.

2) It is not clear to me how C-MVAD works in practice.
On page 6 it is stated that C-MVAD "can be applied to existing deployed systems without retraining," so I assume that the pre-trained backbone remains frozen during the training stage of the classification and consistency heads. If so, it seems to me that the two heads cannot influence each other during training, i.e., the classification head is affected only by L_clean and L_adv, as in standard AT, whereas the consistency head is affected only by L_cons (in this case, the values of the weights \alpha and \beta would have no effect).
Moreover, it is also not clear whether and how the embedding computed by the consistency head is used during inference/testing. On page 6 it is stated that "each testing sample is associated with a consistency score s(x). Clean samples tend to have low scores, adversarial samples higher ones." However it is not specified how such a score is computed, nor how it is used. I guess a detection threshold, mentioned in table 2 and in the supplementary material, is used to this aim. If so, how

are input instances handled, if they are detected as adversarial? And how is the corresponding robust accuracy computed?

3) There are several other unclear points and missing details:

- The ratio of defective vs non-defective instances in the MVTec AD data set is not reported.

- On page 5 it is stated that "PGD or BIM" is used during C-MVAD training, but in the supplementary material (S3.3) only PGD is mentioned.

- How is the consistency loss L_cons computed from L_inv and L_cons? From the supplementary material (table S4) it seems that L_cons = L_inv + 0.5L_cons.

- It is not clear what the AUC and TPR/FPR values in table 2 refer to. Do they refer to attack detection performance? However, in the text in the same page an improvement over the AUC of the baseline is mentioned, whereas, if I correctly understood, the baseline model does not implement any detection mechanism.
By the way, in table 2 the standard deviation is not reported, as stated in the text.

- The AUC values reported in figure 3(a) and on table 2, and mentioned on page 6 as "exceeding 0.98 for all attack types," do not seem to correspond to the ROC curves in figure 2, whose AUC values are clearly lower than 0.98.

3) The proposed defense mechanism could also be evaluated against stronger attacks, such as the well-known AutoAttack [a], and possibly against an adaptive attack to be specifically devised against C-MVAD, as suggested by N. Carlini et al. [b].

[a] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," Proc. ICML 2020, pp. 2206-2216, http://proceedings.mlr.press/v119/croce20b.html

[b] N. Carlini et al., "On Evaluating Adversarial Robustness," arXiv:1902.06705v2, 2019, https://arxiv.org/abs/1902.06705.

4) To compare C-MVAD against the baseline, a paired t-test is used. Since a long time the paired t-test is however not recommended to compare two predictors on a single data set [a]. McNemar test, 5x2 cross-validation [b], or a corrected version of the t-test [c] are recommended, instead.

[a] J. Demsar, "Statistical Comparisons of Classifiers over Multiple Data Sets," Journal of Machine Learning Research 7, 2006, pp. 1–30, https://jmlr.org/papers/v7/demsar06a.html

[b] T.G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Computation 10, 1998, pp. 1895-1923, DOI: 10.1162/089976698300017197.

[c] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," Journal of Machine Learning Research 5, 2004, pp. 1089–1105, https://jmlr.org/papers/volume5/grandvalet04a/grandvalet04a.pdf.

5) Bibliographic references on adversarial machine learning are pertinent, but relatively old.

6) Minor remarks:
- Figure 1 is not informative as a schematic representaiton of C-MVAD.
- The "Ethics and consent" section contains unrelated text, referring to medical data.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistical pattern recognition, adversarial machine learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.