Original Manuscript ID: Access-2024-41876

Original Article Title: **“**Unsupervised Geometric-guided Industrial Anomaly Detection”

To: IEEE Access Editor

Re: Response to reviewers

Dear Editor,

Thank you for allowing a resubmission of our manuscript, with an opportunity to address the reviewers’ comments.

We are uploading (a) our point-by-point response to the comments (below) (response to reviewers, under “Author’s Response Files”), (b) an updated manuscript with yellow highlighting indicating changes (as “Highlighted PDF”), and (c) a clean updated manuscript without highlights (“Main Manuscript”*)*.

Best regards,

Dinh-Cuong Hoang, et al.

Reviewer#1, Concern # 1: If the authors could include more qualitative figures across various types and provide more detail on the datasets used, it would enhance the paper further. Additionally, citing peer-reviewed publications rather than preprints from arXiv would improve the reliability and credibility of the references.

Author response: Thank you for this suggestion. In the revised manuscript, we have added more qualitative figures (please see Figure 3, 4, and 5). We have also expanded the dataset description section to include detailed information. Furthermore, wherever possible, we have replaced arXiv preprints with peer-reviewed publications to enhance the credibility of the references.

Author action: We expanded the dataset description:

"Most existing industrial anomaly detection datasets provide only 2D RGB images without corresponding 3D point cloud data. In contrast, 3D industrial anomaly detection is still in its early stages. The MVTec 3D-AD dataset [60] is the first dataset designed specifically for industrial anomaly detection using 3D data. It consists of 4147 scans acquired by a high-resolution industrial 3D sensor (Zivid One+ Medium) under conditions similar to real-world inspection setups. The dataset includes 10 categories of industrial objects, comprising 2656 training samples, 294 validation samples, and 1197 test samples. The training and validation sets contain only anomaly-free scans, while the test set includes both anomaly-free and anomalous samples.

The anomalies in the dataset include a wide range of real-world defect types, such as scratches, dents, contaminations, cracks, holes, and deformations. These defects were devised and fabricated to closely simulate actual defects encountered in industrial settings. For example, the bagel and cookie categories feature cracks, while the carrot exhibits a hole, and the peach and rope contain contaminations. Prototypical examples of anomalies from the dataset's 41 distinct defect types are shown in [60].

The dataset's 10 object categories can be grouped based on their properties:

- *Natural Variations: Bagel, carrot, cookie, peach, and potato exhibit significant natural variations in shape, size, and texture.*
- *Deformable Objects: Foam, rope, and tire have standardized appearances but can easily deform.*
- *Rigid Objects: Cable gland and dowel are rigid and could, in principle, be inspected using CAD models, but the dataset is designed to test unsupervised methods that can handle all object types.*

Each scan includes a pixel-aligned RGB image and a point cloud containing x, y, and z coordinates, which enables a one-to-one mapping between the 2D image and the 3D geometric data. The scans were cropped to a fixed rectangular domain to minimize background pixels while retaining sufficient margins for data augmentation techniques such as cropping, translation, and rotation. The preprocessing ensures consistency with real-world scenarios, where objects are typically positioned in predefined locations with controlled lighting. The acquisition setup features an indirect and diffuse light source, with the sensor statically mounted to maintain a consistent view for each object category. Calibration of internal camera parameters ensures precise projection of 3D points into their corresponding 2D pixel coordinates. This configuration not only aligns with real-world practices but also simplifies data augmentation and preprocessing."

Reviewer#1, Concern # 2: The authors may consider citing additional state-of-the-art works that use GANs in the literature review, particularly from IEEE Access, relevant to this research area. Suggested references include:

T. Ganokratanaa, S. Aramvith, and N. Sebe, "Unsupervised Anomaly Detection and Localization Based on Deep Spatiotemporal Translation Network," IEEE Access, vol. 8, pp. 50312-50329, 2020. doi: 10.1109/ACCESS.2020.2979869.

Thittaporn Ganokratanaa, Supavadee Aramvith, Nicu Sebe, "Video anomaly detection using deep residual-spatiotemporal translation network," Pattern Recognition Letters, vol. 155, pp. 143-150, 2022. ISSN 0167-8655. https://doi.org/10.1016/j.patrec.2021.11.001.

Author response: We have reviewed the suggested papers and cited them in literature review to provide a more comprehensive overview of related GAN-based anomaly detection techniques.

Reviewer#2, Concern # 1: The title should be refined to provide precise information about the focus of the research study by using "image-based anomaly detection" or "Industrial quality inspection" instead of "Industrial Anomaly Detection", which is actually a very broad topic dealing with anomaly detection using various sensors embedded in machinery and equipment to capture detailed information about the process, equipment's sensors and/or the quality of products.

Author response: We agree that the title could be refined to better reflect the specific focus on image-based industrial anomaly detection. We have revised the title to emphasize the use of 2D and 3D data in quality inspection rather than the broader field of anomaly detection.

Author action: Revised the title to: "Unsupervised Visual-to-Geometric Feature Reconstruction for Vision-Based Industrial Anomaly Detection"

Reviewer#2, Concern # 2: The technical aspects of the proposed method are not strong enough and lack pertinent experimental tests and an in-depth analysis, in particular the Transformer-based Visual-to-Geometric feature reconstruction.

Author response: We thank the reviewer for insightful comments regarding the technical aspects of the proposed Transformer-based Visual-to-Geometric Feature Reconstruction and the need for additional experimental and analytical depth. We have addressed this concern by significantly revising the relevant sections to enhance its clarity, technical rigor, and depth of explanation.

Specifically, we have restructured the section to provide a more detailed and systematic description of the Visual-to-Geometric Feature Reconstruction process. The updated text now explicitly describes the role of non-local attention and graph convolutional networks (GCNs) in capturing global and local feature interactions, respectively. We included equations for each critical step to highlight the mathematical foundations of the method. Additionally, we expanded on the rationale for combining global context (via non-local attention) and local feature refinement (via GCNs), explaining how this approach effectively captures subtle correlations between visual and geometric features. In addition, we clarified how spatial alignment between visual and geometric features is maintained through bilinear upsampling, ensuring consistency in feature reconstruction. We also elaborated on the role of the L2 loss function in learning correlations between 2D appearance and 3D structure, demonstrating its critical contribution to anomaly detection. Please see section "VISUAL-TO-GEOMETRIC FEATURE RECONSTRUCTION".

Moreover, we have expanded additional experimental validation and provided an in-depth analysis of the Transformer-based Visual-to-Geometric feature reconstruction. We conducted an ablation study to analyze the contributions of individual components in the proposed Visual-to-Geometric feature reconstruction module. This study evaluated four configurations: (1) without the visual feature enhancement module (M_{FE})*, (2) without the graph convolutional network (GCN), (3) without the non-local attention mechanism (NLA), and (4) the full module (Proposed).* The results show that each component contributes uniquely to the overall performance, with the full module achieving the highest accuracy and precision. Removing M_{FE} resulted in the most significant drop in performance, highlighting its critical role in integrating global and local feature refinements. Similarly, the exclusion of GCN or NLA reduced performance, underscoring the importance of capturing both local spatial relationships and global contextual dependencies. Please see section "ADDITIONAL EXPERIMENTS AND ABLATION STUDY".

Reviewer#2, Concern # 3: The claim of a substantial improvement of the proposed method requires more evidence to be credible. To get reproducible results, authors should expand simulation tests and provide a better description of the implementation of the Transformer-based Visual-to-Geometric feature reconstruction, which is the core of this research study.

Author response: To strengthen our experimental evidence, We further expanded simulation tests to demonstrate the effectiveness and robustness of the proposed Transformer-based Visual-to-Geometric feature reconstruction module. Following [67], we generated a synthetic dataset using the Blender framework, a popular 3D modeling software that provides seamless interoperability with Python via the BlenderProc package. BlenderProc is a procedural synthetic data generation tool that allows for controlled and reproducible creation of 3D environments. Using this framework, we simulated industrial scenarios with various object shapes, textures, lighting conditions, and geometric anomalies. Figure 5 shows examples of the generated samples. The dataset consists of 10,000 images for training and 1,000 images for testing, with corresponding 3D point cloud data and ground-truth anomaly maps. The results of these simulations, now included in the manuscript, demonstrate the consistent superiority of our method over baselines, particularly in high-precision metrics such as AUPRO(0.05), which is critical for industrial anomaly detection applications.

To address the reviewer's concern about reproducibility, we have provided a more detailed description of the implementation of the Transformer-based Visual-to-Geometric feature reconstruction module. This includes the design of the visual feature enhancement module M_{FE}, which integrates non-local attention and GCNs, the geometric feature reconstruction pipeline, and specific training configurations (e.g., learning rates, batch sizes, optimizer settings). Hardware specifications, including the use of an NVIDIA GeForce RTX 4090 GPU, and regularization techniques for stable training have also been detailed. Furthermore, we have clarified our use of metrics such as AUPRO(0.05), which provides a stringent and realistic evaluation for high-precision industrial applications.

To ensure complete transparency and reproducibility, we have made our codebase, synthetic dataset generation scripts, and detailed training instructions publicly available at https://github.com/hoangcuongbk80/GeoAD.

We believe that these enhancements address the reviewer's concerns by providing stronger experimental evidence, comprehensive implementation details, and accessible resources for reproducibility. Thank you for highlighting these critical aspects of our research.

Reviewer#2, Concern # 4: Why the global anomaly score for each sample is based on the maximum value in the anomaly map? Apparently this decision is sensitive to noise, increasing false alarm rates. How to determine the best threshold to avoid a lot of false positives?

Author response: Actually, we computed the global anomaly score for each sample using the maximum value from the smoothed anomaly map. The predicted anomaly map is smoothed using a Gaussian kernel with σ=4, following [47]. We have clarified this in the revised manuscript. Regarding the threshold, we employed a well-defined way from [23, 60] to estimate the threshold using a set of randomly selected validation images, which are excluded from the training set. For each category, we define a minimum defect area that a connected component in the thresholded anomaly map must have to be classified as a defective region. Using this criterion, we iteratively segment the anomaly maps of the anomaly-free validation set with increasing thresholds. The process stops when the area of the largest anomalous region on the validation set is just below the user-defined area, and the corresponding threshold is used for further evaluation. We have also clarified this in the revised manuscript.