**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 and 4). 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 \cite{bergmann2022mvtec} 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 \cite{bergmann2022mvtec}.*

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

*\begin{itemize}*

*\item \textit{Natural Variations:} Bagel, carrot, cookie, peach, and potato exhibit significant natural variations in shape, size, and texture.*

*\item \textit{Deformable Objects:} Foam, rope, and tire have standardized appearances but can easily deform.*

*\item \textit{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.*

*\end{itemize}*

*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 incorporated them into the 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 section 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, such as token normalization, attention computation, and graph-based refinement, 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. These revisions aim to establish a stronger technical foundation for the proposed method and provide the basis for further analysis, such as ablation studies on attention mechanisms and feature refinement.

We acknowledge the need for additional experimental validation and an in-depth analysis of the Transformer-based Visual-to-Geometric feature reconstruction. We will enhance the technical explanation by providing a detailed mathematical formulation of the reconstruction process and further experiments, such as ablation studies, to evaluate its contribution to anomaly detection performance.

**Author action:**

* Expanded the methodology section with a more detailed explanation of the Visual-to-Geometric reconstruction network, including equations and implementation specifics.
* Added an ablation study evaluating the impact of components such as non-local attention and GCN on the reconstruction quality.

**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:** We agree that reproducibility is critical and will provide more detailed implementation specifics for the Transformer-based reconstruction module, including training parameters, hardware specifications, and hyperparameter tuning. Additionally, we will conduct more simulation tests to demonstrate the method’s robustness across various scenarios.

**Author action:**

* Detailed the implementation of the Transformer-based module:

*“The Visual-to-Geometric Feature Reconstruction module is designed to map the visual features ($224 \times 224 \times 768$) to the geometric feature map ($224 \times 224 \times 1152$). This module begins by refining the visual features using the visual feature enhancement module $\mathcal{M}\_{FE}$, which combines non-local attention and graph convolutional networks (GCNs) to enhance both global and local feature interactions. Specifically:*

*\begin{itemize}*

*\item \textbf{Non-Local Attention:} Visual tokens are passed through linear projection layers to compute query, key, and value matrices with dimensions of $384$ each, i.e., $d/2$ where $d = 768$. Attention weights are computed as described in Equation (4) of the paper, and the refined tokens are aggregated using adaptive average pooling with a kernel size of $2 \times 2$. This ensures efficient computation while preserving salient features.*

*\item \textbf{Graph Convolutional Networks (GCNs):} The adjacency matrix for the GCN is constructed dynamically based on cosine similarity between tokens, with a threshold of 0.6 to retain meaningful connections. A single-layer graph convolution propagates contextual information across connected nodes, using a ReLU activation for non-linearity. The GCN weight matrix $w\_g$ has dimensions $8 \times 8$, corresponding to the patch size.*

*\end{itemize}*

*The refined visual features are reshaped into patch-based feature maps and bilinearly upsampled to match the original spatial resolution ($224 \times 224$). These features are then passed through a lightweight Multi-Layer Perceptron (MLP) comprising three fully connected layers with GeLU activation functions. The dimensions of the MLP layers are $768$, $960$, and $1152$, respectively. Dropout with a rate of $0.1$ is applied after each layer to prevent overfitting, and the final output is a predicted geometric feature map $\hat{\mathbf{F}}\_{geo}$ aligned with the extracted geometric features $\mathbf{F}\_{geo}$.”*

* Included additional simulation results using variations in dataset size and defect characteristics.

**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 \cite{roth2022towards}. We have clarified this in the revised manuscript. Regarding the threshold, we empirically determined that a fixed threshold of 0.5 provides the best balance between precision and recall when evaluated on both the Simulation and MVTec 3D-AD datasets. In future work, we plan to explore adaptive thresholding methods that can dynamically adjust the threshold.