Enhancing Anomaly Detection with Spatial Transform Networks

Renato Castro Cruz
Department of Computer Science
Pontificia Universidad Catolica del Peru
Lima, Peru
rcastroc@pucp.edu.pe

Cristian Lazo Quispe
Department of Computer Science
Universidad Nacional de Ingenieria
Lima, Peru
clazoq@uni.pe

Abstract

Anomaly detection, an essential component of industrial quality control and surveillance, plays a crucial role in identifying deviations from normal patterns. This extended abstract explores a preliminary research focused on enhancing the anomaly detection capabilities of the PaDim architecture—a state-of-the-art solution for anomaly detection on the MVTEC dataset—through the integration of Spatial Transform Networks (STNs). The aim is to improve performance, particularly in challenging classes such as "zipper" and "screw," where the PaDim architecture achieves lower metrics of performance. Notably, these challenging scenarios often involve objects in non-fixed positions, making anomaly detection intricate in real-world complex scenarios. Through experimentation involving the integration of a Spatial Transform Network using self-supervised training, the performance of this innovative approach is evaluated and it sheds light on both the strengths and limitations of this integration, providing insights into the benefits of leveraging Spatial Transform Networks to handle real-world complexity.

1 Introduction

Anomaly detection in the context of computer vision has gained significant attention due to its applications in various industries, ranging from manufacturing and quality control to surveillance and security. Detecting unusual patterns or objects that deviate from expected is a critical task in ensuring the reliability, safety, and quality of processes and products. Traditional methods for anomaly detection often relied on handcrafted features and statistical techniques, which struggled to capture complex patterns and adapt to changing scenarios.

In recent years, the combination of deep learning techniques with anomaly detection has shown promising results. Convolutional Neural Networks (CNNs) have demonstrated remarkable abilities in extracting features from images, allowing them to discern deviations that might indicate anomalies. Among the innovative techniques emerging, Spatial Transform Networks (STNs) [1] stand out as a versatile tool for enhancing the adaptability and performance of CNNs.

The PaDim architecture [2] serves as the foundation of our exploration, presenting a framework for deep learning-based anomaly detection. Utilizing pretrained CNN models like ResNet 18 or WideResNet as encoders. This enhancement enables the architecture to extract detailed features, contributing to more accurate anomaly detection.

However, the effectiveness of the PaDim architecture can be influenced by the inherent complexity of certain anomaly classes, such as the "zipper" and "screw" classes in the MVTEC dataset [3]. These classes exhibit medium differences between normal and anomalous instances, posing challenges for accurate anomaly detection. This motivates our investigation into enhancing the PaDim architecture capabilities by incorporating Spatial Transform Networks (STNs).

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

2 Related Work

2.1 Anomaly Detection Models

Anomaly detection has showed a transitioning from traditional statistical methods to more robust deep learning-based approaches. Early approaches often relied on handcrafted features and simplistic thresholds to identify anomalies. However, these methods struggled to capture the intricacies of complex anomalies and adapt to changing data distributions, according to [4].

With the advent of deep learning, a paradigm shift occurred, propelling anomaly detection towards unprecedented effectiveness. Convolutional Neural Networks (CNNs), which were initially designed for tasks like image classification, demonstrated a remarkable ability to learn intricate features from data [5].

Beyond CNNs, Recurrent Neural Networks (RNNs) have been harnessed to capture temporal dependencies in time-series data, facilitating anomaly detection in domains such as finance and surveillance [6].

Furthermore, the integration of autoencoders, a type of neural network designed for unsupervised learning, has propelled anomaly detection to new heights. Autoencoders learn to encode input data into a compressed latent space representation, enabling the reconstruction of input samples. Anomalies are detected by measuring the dissimilarity between reconstructed samples and the original data [7].

The evolution of anomaly detection culminates in the present endeavor with PaDim Architecture and PathCore . [8] which are state-of-art solutions for anomaly detection on the MVTEC Dataset.

2.2 PaDim Architecture

The PaDim (Patch-based Distribution Modeling) architecture, represents a milestone in anomaly detection by seamlessly integrating autoencoders with the power of deep convolutional neural networks.

In the context of PaDim, the traditional autoencoder encoder role shifts. Instead of employing a standard autoencoder encoder, PaDim leverages the encoder outputs from ResNet-18 and WideResNet-50. These pretrained CNNs excel at capturing intricate hierarchical features, thereby encoding vital image information within a compressed feature space.

Retaining its distinctive decoder, the PaDim architecture reconstructs the original images from latent features of ResNet-18 and WideResNet-50 as encoders. By analyzing pixel reconstruction errors and comparing them to the training set error distribution, this technique effectively pinpoints potential anomalies

Additionally, the PaDim architecture introduces the concept of a matrix of Gaussian parameters. This matrix captures the statistics of reconstruction errors across the training dataset. By analyzing pixel-wise reconstruction errors in relation to this Gaussian parameter matrix, the architecture identifies regions displaying anomalously high errors, signifying the potential presence of anomalies.

2.3 Spatial Transform Network (STN)

Spatial Transform Networks (STNs) have emerged as a pivotal advancement in enhancing spatial invariance and adaptability within convolutional networks. STNs introduce an auxiliary module into neural network architectures, allowing them to learn and apply geometric transformations to input data. This capability is particularly valuable in tasks requiring spatial alignment, such as image registration and transformation.

The STN module typically consists of three core components: a localization network, a grid generator, and a sampler. The localization network predicts transformation parameters from the input, the grid generator constructs sampling grids based on these parameters, and the sampler applies these grids to perform the spatial transformation.

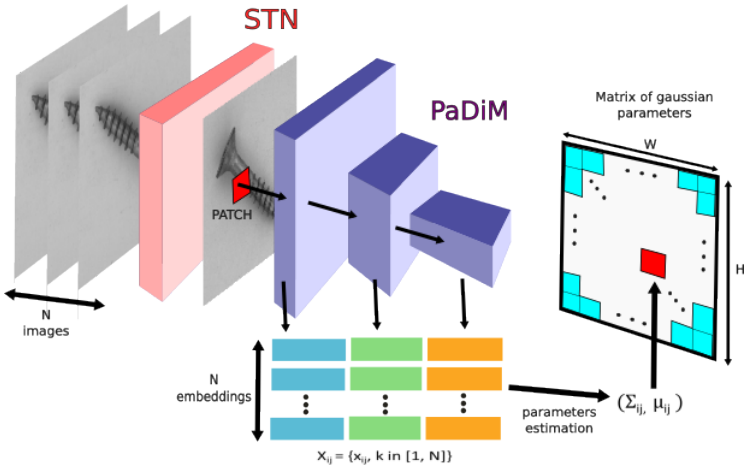

Figure 1: Ensemble STN with PaDiM architecture

3 Methodology: Ensemble STN with PaDim

Our methodology encompasses two main phases: STN training through self-supervised learning and subsequent integration with the PaDim architecture as can be observed in Figure 1.

3.1 Training the STN with Self-Supervised Learning

The STN architecture is composed of several essential layers, each designed to fulfill specific functions:

1. Localization Convolutional Layers: The STN begins with a series of convolutional layers dedicated to spatial localization.
2. Feature Aggregation and Dimension Reduction: Following the localization layers, the network aggregates and reduces the extracted features. This is achieved by reshaping the output using view operations and feeding it into fully connected layers.
3. Spatial Transformation Prediction: The core objective of the STN is to predict geometric transformations that align the input images with their augmented versions.
4. Affine Grid and Grid Sampling: The predicted transformation parameters are used to generate an affine grid, which essentially specifies the mapping from the input image to the transformed image.

Due to time-consuming by the extensive number of classes on the MVTEC Dataset, self-supervised learning offers a strategic solution. Unlike traditional supervised learning where explicit labels are provided, self-supervised learning capitalizes on inherent patterns within the data itself to generate labels for training. Another reason behind self-supervised learning for the STN in this context lies in its remarkable potential to exploit the inherent structure of images. By employing the existing data, the STN learns to predict geometric transformations that rectify input images, effectively generating its own pseudo-labels.

3.2 Ensembling with PaDim Architecture for Training

With the trained STN in hand, we proceeded to integrate it with the PaDiM architecture during the training phase. During training, the PaDim architecture has the transformed image of the STN module as its input, allowing it to leverage the enhanced spatial awareness and transformation capabilities of the STN. This synergy contributes to improved feature extraction and anomaly detection performance, particularly for classes with medium differences between normal and anomalous instances. In summary, for training, the trained STN network was used as a preprocessing step to enhance the performance of PaDim. During testing, the enhanced PaDim model took charge, meticulously

Table 1: Comparison of AUC ROC metric in image-level and pixel-level

Tipo	Label	image-level		pixel-level	
		PaDiM	STN + PaDiM(Ours)	PaDiM	STN + PaDiM(Ours)
Texture	carpet	0.999	**1**	0.99	**0.992**
	grid	0.957	**0.964**	**0.965**	**0.965**
	leather	**1**	**1**	0.989	**0.99**
	tile	0.974	**0.988**	0.939	**0.945**
	wood	0.988	**0.989**	**0.941**	**0.941**
	total	0.984	**0.988**	0.965	**0.966**
Object	bottle	**0.998**	**0.998**	0.982	**0.985**
	cable	0.922	**0.924**	0.968	**0.977**
	capsule	0.915	**0.927**	0.986	**0.987**
	hazelnut	0.933	**0.964**	**0.979**	0.85
	metal nut	0.992	**0.999**	0.971	**0.979**
	pill	**0.944**	0.942	0.961	**0.97**
	screw	0.844	**0.857**	0.983	**0.987**
	toothbrush	**0.972**	0.864	0.987	**0.989**
	transistor	0.978	**0.992**	0.975	**0.981**
	zipper	0.909	**0.917**	0.984	**0.986**
	total	**0.941**	0.938	**0.978**	0.97
All classes		**0.955**	**0.955**	0.973	**0.977**

computing and delivering the final testing metrics, reflecting the amplified robustnees and efficiency achieved through the trained STN network.

3.3 Experimentation and Results

In this section, we present the results of our experimental evaluations, demonstrating the impact of integrating the trained STN with the PaDim architecture. We provide a comprehensive analysis of the anomaly detection performance, comparing it against the baseline PaDim architecture.

The evaluation involves metrics such as ROC AUC image and pixel level in Table 1. By quantifying the enhancements achieved through the STN-PaDim ensemble, we validate the effectiveness of our approach. Furthermore, we highlight insights gained from the results, discussing the strengths and limitations of our methodology.

The experimental findings not only substantiate the benefits of our approach but also offer insights into the potential avenues for further refinement and optimization of anomaly detection methodologies within the context of complex and intricate scenarios.

4 Conclusions and Future Work

In conclusion, our work demonstrates the substantial impact of integrating Spatial Transform Network with the PaDiM architecture, it is better in the 86.67% of the all classes in image-level AUC ROC, and also 93.33% in pixel-level AUC ROC.

This experimentation and results provides a comprehensive analysis of our approach's impact, showcasing its effectiveness in terms of anomaly detection performance. These findings not only underscore the advantages of our approach but also offer valuable insights into potential directions for further refining and optimizing anomaly detection methodologies, particularly within intricate and complex scenarios.

For future work, we aim to explore the potential of combining different STN architectures with a broader range of anomaly detection models to assess their performance in various contexts. Additionally, investigating the impact of diverse datasets on our proposed approach.

References

[1] Jaderberg, M., K. Simonyan, A. Zisserman, et al. Spatial transformer networks. Advances in neural171 information processing systems, 28, 2015.

[2] Defard, T., A. Setkov, A. Loesch, et al. Padim: a patch distribution modeling framework for anomaly173 detection and localization. In International Conference on Pattern Recognition, pages 475–489. Springer, 2021.

[3] Bergmann, P., K. Batzner, M. Fauser, et al. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. International Journal of Computer Vision, 129[4]:1038–1059, 2021.

[4] Yang, J., R. Xu, Z. Qi, et al. Visual anomaly detection for images: A survey. arXiv preprint arXiv:2109.13157, 2021.

[5] Chalapathy, R., S. Chawla. Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407,2019.

[6] Mohindru, V., S. Singla. A review of anomaly detection techniques using computer vision. Recent183 Innovations in Computing: Proceedings of ICRIC 2020, pages 669–677, 2021.

[7] Carrara, F., G. Amato, L. Brombin, et al. Combining gans and autoencoders for efficient anomaly detection, 2020.

[8] Roth, K., L. Pemula, J. Zepeda, et al. Towards total recall in industrial anomaly detection, 2022.