Enhancing Object Detection Accuracy with Variational Autoencoders as a Filter in YOLO

Shubham Kumar Dubey[1], J. V. Satyanarayana[2] and C. Krishna Mohan[1]
[1]*Computer Science Department, Indian Institute of Technology Hyderabad, Hyderabad, India*
[2]*RCI-DRDO, India*

Keywords: Object Detection, YOLO, False Positive, Variational Autoencoders.

Abstract: Object detection is an important task in computer vision systems, encompassing a diverse spectrum of applications, including but not limited to autonomous vehicular navigation and surveillance. Despite considerable advancements in object detection models such as YOLO, the issue of false positive detections remain a prevalent concern, thereby causing misclassifications and diminishing the reliability of these systems. This research endeavors to present an innovative methodology designed to augment object detection accuracy by incorporating Variational Autoencoders (VAEs) as a filtration mechanism within the YOLO framework. This integration seeks to rectify the issue of false positive detections, ultimately fostering a marked enhancement in detection precision and strengthening the overall dependability of object detection systems.

1 INTRODUCTION

1.1 Background and Motivation

Object detection is a fundamental task in computer vision, and it plays a vital role in various applications such as autonomous vehicles, surveillance, healthcare and defence. The advent of deep learning and the availability of large-scale annotated datasets have propelled the field of object detection, with models like YOLO (You Only Look Once) (Redmon et al., 2016) achieving real-time performance. However, despite these advancements, false positive detections continue to challenge the reliability of these systems. False positives are instances where objects are incorrectly identified, leading to misclassifications, increased computational load, and even safety risks in applications like autonomous driving.

The motivation for this research stems from the need to reduce false positive detections in object detection systems, thereby improving their precision and reliability. By addressing this issue, the proposed approach aims to enhance the overall performance and safety of these systems.

1.2 Objective

The primary objective of this research is to enhance object detection accuracy by reducing false positive detections. This research proposes integrating Variational Autoencoders (VAEs) (An and Cho, 2015) into the YOLO framework to serve as a filtering mechanism. VAEs, renowned for their anomaly detection capabilities, aim to improve the precision and reliability of object detection systems.

2 LITERATURE SURVEY

There have been various approaches in the past for object detection. Detection methods like YOLO are widely used today.

2.1 Traditional Hand-Crafted Object Detection Methods

The Viola Jones (Viola and Jones, 2001) method uses a sliding window approach searching for haar wavelets as features in an image. HOG (Dalal and Triggs, 2005)used a dense pixel based grid called blocks where the gradients are given by the magnitude and direction change in the pixel intensity of the grid.

Deep convolutional neural networks performed much better for object detection due to their ability to learn detailed feature representations of an image.

Dubey, S., Satyanarayana, J. and Mohan, C.
Enhancing Object Detection Accuracy with Variational Autoencoders as a Filter in YOLO.
DOI: 10.5220/0012347700003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 4: VISAPP*, pages 270-277
ISBN: 978-989-758-679-8; ISSN: 2184-4321

2.2 Deep Learning Object Detection Methods

Generally two stage object detection methods like Faster RCNN (Ren et al., 2015) produce more accurate results compared to single stage detectors. On the other hand single stage detectors are much faster in terms of their computation time. With the advent of modern single stage detectors like YOLO we find detection accuracy to be on par with two staged detectors, while also being much faster than them.

The most widely used object detection methods today include CNN based methods. The state of the art methods include Faster RCNN, YOLO and SSD (single shot multi box detectors) (Liu et al., 2016). (Lin et al., 2017) emphasizes honing the model's skills on a limited set of challenging examples while simultaneously safeguarding against an inundation of numerous straightforward negatives that could otherwise overwhelm the training process.

(Ye et al., 2020) explored the use of YOLO along with VAE to detect and classify garbage from other objects. A trained gaussian curve representation of training samples is used for classifying new samples. It focuses on the classification task based on reconstruction and KL divergence losses along with the YOLO spatial information loss. Use of VAE along with YOLO could thus be further used to remove false positives while targeting the detection of objects of a single class like drones. The threshold can be increased or decreased by the factor (δ) to suit the specific detection task and scenario.

3 RESULTS BY YOLO

YOLO improves upon other object detection methods by re framing object detection as a regression task rather than a classification task. The working of YOLO starts by taking an image of dimensions H x W, where H represents the height and W represents the width of the image. Then we have the feature extractor module made of strong CNN networks like the VGG1 (Simonyan and Zisserman, 2014), ResNet-50 (He et al., 2016) etc. The next stage involves a single shot detector module using a grid layout on the image, where each grid cell is scanned for detecting an object of the required class.

While results from YOLO are majorly precise, the limitations of YOLO show up when the objects in the image are small (Liu et al., 2021), or are of unusual aspect ratios. This can be seen in the example image 1 below.

4 VARIATIONAL ENCODERS (VAEs)

4.1 VAE Theory

Variational Autoencoders, or VAEs, are a class of generative models that merge neural networks with probabilistic modeling. VAEs extend traditional autoencoders, a type of neural network designed for data representation learning. In a VAE, data is encoded into a probability distribution in a lower-dimensional latent space, from which data samples can be generated. This probabilistic approach (Kingma and Welling, 2013) enables VAEs to model complex data distributions effectively.

At the core of VAEs is the idea of learning a probability distribution over the latent space, which allows for the generation of new data points. This is achieved through two main components: the encoder and the decoder. The encoder maps input data to a probability distribution in the latent space, while the decoder reconstructs data samples from this distribution.

4.2 VAEs for Anomaly Detection

VAEs excel in anomaly detection due to their inherent ability to model the distribution of normal data. Normal data points cluster densely in the latent space, whereas anomalies reside in less dense regions. As a result, anomalies yield higher reconstruction errors when decoded from the latent space, making them distinguishable from normal data (Li et al., 2019).

VAEs employ a loss function that measures the dissimilarity between input data and its reconstruction. In the context of anomaly detection, this loss function provides a quantifiable measure of how well a data point aligns with the model's understanding of normality. Anomalies exhibit significantly higher loss values, allowing for their identification.

4.3 Applications of VAEs

VAEs have found applications across diverse fields, including natural language processing, image generation, and healthcare. One of their most compelling uses is in anomaly detection. By utilizing the latent space learned by VAEs, anomalies in data can be identified based on their deviation from normal patterns.

In the realm of healthcare, VAEs have been applied to detect anomalies in medical images, such as X-rays and MRIs. Similarly, in finance, VAEs have been employed to detect fraudulent transactions by flagging deviations from typical spending patterns.

Figure 1: YOLO output shows how it detects bird(on top) as a drone thus giving a false positive in a video from the Drone vs. Bird dataset.

VAEs can also be useful in defence applications to give accurate target detection.

5 PROPOSED METHOD

We wish to apply the VAE-as a filter on YOLO method to reduce false positives for defence applications. In crucial on-field scenarios where we need to target drones accurately and discard any birds as false positives, our approach is implemented.

5.1 VAE Training and Architecture

To effectively harness VAEs for false positive reduction in object detection, a comprehensive training process is indispensable.

5.1.1 Data Collection

A crucial aspect of VAE training is the collection of a comprehensive dataset. This dataset should consist of normal, non-anomalous objects that are representative of real-world scenarios. To ensure the model's robustness, the dataset (Everingham et al., 2010) should encompass diverse environmental conditions and scenarios.

Data: YOLO object detection results $D = \{(b_i, c_i)\}$
Result: Filtered object detections $D_{filtered}$
initialization;
$D_{filtered} \leftarrow \emptyset$;
while *frame is captured* **do**
  Perform YOLO object detection to obtain D;
  foreach *detection* (b_i, c_i) *in* D **do**
    Compute reconstruction error R_i with VAE: $R_i = ||x_i - \hat{x}_i||^2$;
    if R_i *is below a predefined threshold* **then**
      Add (b_i, c_i) to $D_{filtered}$;
    end
  end
  Process $D_{filtered}$ for further use or display;
end

Algorithm 1: Integrating VAE as a Filter in YOLO Object Detection.

5.1.2 Drone vs Bird Dataset

The Drone-vs.-Bird dataset was released as a Detection Challenge in 2021. Seventy seven different video sequences were made available as training data. The Fraunhofer IOSB research institute, ALADDIN2 project and SafeShore jointly used the MPEG4-coded static cameras to record the dataset.

On average, the video sequences consist of 1,384 frames, while each frame contains 1.12 annotated drones. The video sequences are recorded with both

static cameras and moving cameras and the resolution varies between 720×576 and 3840×2160 pixels. In total, 8 different types of drones exist in the dataset, i.e. 3 with fixed wings and 5 rotary ones.

5.1.3 Training Procedure

The VAE is rigorously trained on this dataset to capture the distribution of normal objects effectively. We trained our VAE model on drone images from 45 videos of the Drone vs. Bird dataset with batch size 32 and for 100 epochs. The validation and testing was done on 16 videos each.

Our VAE is made of 7 convolutional layers, with batch normalization and ReLU activation, for both the encoder and decoder. Firstly, the VAE is completely trained on the 24,000 (approx. 60 % of total images) frame wise cropped images of drones, from the drone vs. bird dataset. Then, validation is done on 6000 images (approx. 20 % total images) of drones. Testing is done on the remaining 20 % of the images. This training process optimizes the VAE's parameters to minimize the reconstruction error between input data and its reconstructed counterpart. The objective is to create a latent space representation that accurately models the characteristics of normal objects.

5.2 Filtering in the Detection Pipeline

The core of the proposed approach is the integration of the VAE as a filtering mechanism within the YOLO-based object detection pipeline.

5.2.1 YOLO Object Detection

The YOLO (You Only Look Once) object detection system is a state-of-the-art model for real-time object detection. YOLO divides an image into a grid and assigns bounding boxes and class labels to objects within grid cells. Deep learning techniques, such as convolutional neural networks (CNNs), are used to achieve these detections.

5.2.2 VAE Filtering

In the proposed approach, YOLO generates a list of potential detections during the object detection process, denoted as $D = \{(b_i, c_i)\}$, where b_i represents the bounding box coordinates, and c_i represents the class label. These candidates are then passed through the trained Variational Autoencoder (VAE), which calculates the reconstruction error for each detection as:

$$R_i = ||x_i - \hat{x}_i||^2 \quad (1)$$

Here, x_i is the original detection, and $\hat{x}_i$ is the reconstructed detection obtained by passing b_i through the VAE. The reconstruction error, R_i, from equation 1 serves as a critical indicator of the detection's quality. A low reconstruction error indicates that the object is well-defined and easily recognizable ($R_i \approx 0$), while a high error suggests that the detection might be uncertain or noisy ($R_i \gg 0$). By using the VAE to assess the quality of each detection, the proposed approach effectively filters out false positives and focuses on the most reliable object candidates, ultimately improving the overall accuracy and robustness of object detection in computer vision applications.

5.2.3 Anomaly Classification

The VAE quantifies the dissimilarity between the original image patch and its VAE-reconstructed counterpart through the reconstruction error. Detections with reconstruction errors surpassing a predetermined threshold are identified as anomalies. This threshold can be adjusted to control the trade-off between sensitivity (recall) and specificity (precision).

5.2.4 Threshold Calculation

The threshold calculation method in this context involves utilizing the mean (μ) and standard deviation (σ) of reconstruction errors on a validation set to establish a threshold for anomaly detection in test data. By computing the pixel-wise mean squared errors between original and reconstructed images, the method captures the normal variability of the validation set.

$$Threshold = \mu + 2\sigma \quad (2)$$

In this work, assuming a normal distribution, our threshold is set as the mean error plus two times the standard deviation, providing a statistical measure to identify anomalies in the test data as shown in 2. This approach is advantageous as it adapts to the specific characteristics of the dataset, dynamically establishing a boundary for normalcy. It leverages statistical measures to discern anomalies, accommodating variations in image content and noise levels, making it a robust method for anomaly detection in the context of the Variational Autoencoder.

6 EXPERIMENTS AND RESULTS

6.1 Experiment Setup

To evaluate the effectiveness of the proposed approach, a series of experiments were conducted on

the drone vs bird dataset. The dataset encompasses a wide range of conditions, including different lighting, weather, and occlusion levels. The experiments aimed to assess the reduction in false positive detections and the impact on overall object detection precision.

For test evaluation, parameters $\lambda = 0$ and $A_{max} = 30$ frames were used. All our evaluation and testing was done on a machine with NVIDIA GeForce GTX 1050 Ti graphic card.

6.2 Experimental Results

The results of the experiments demonstrated a significant reduction in false positive detections when utilizing the VAE filtering mechanism. In particular, under challenging conditions such as distant, small targets and heavy occlusion, the approach exhibited a remarkable increase in precision. For the anomaly classification threshold, we choose to stay with the standard threshold as shown in equation 2.

6.2.1 Quantitative Results

The mAP scores compared for YOLO and YOLO with VAE filter at different IOU thresholds can be seen in figure 2 and table 1. Table 2 shows the percentage of false positive detections given by YOLO compared to YOLO-VAE. Table 3 compares the average execution time taken by YOLO and YOLO with VAE filter approaches.

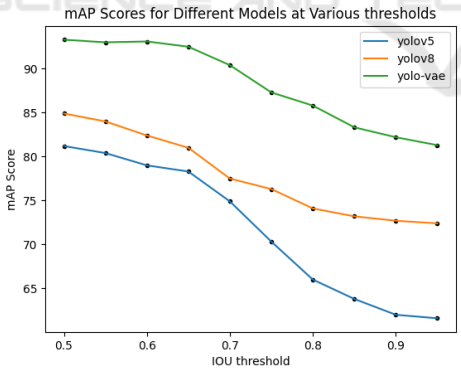

Figure 2: Comparing mAP at different thresholds.

Table 1: Comparing mAP at different IOU thresholds.

IOU	YOLOv5	YOLOv8	YOLO-VAE
0.5	81.2	84.9	93.3
0.95	61.6	72.4	81.3

Table 2: Comparing percentage of false positive detections by YOLO and YOLO with VAE filter.

YOLOv5	YOLOv8	YOLO-VAE
33.6 %	22.3 %	15.9 %

Table 3: Comparing execution times of YOLO and YOLO with VAE filter.

Model	Exec.time
YOLO	0.016s
YOLO with VAE	0.021s

6.2.2 Qualitative Results

Figure 3 below shows sample of how YOLO with VAE compares to the results by YOLO on the Drone vs. Bird dataset. We can observe that the birds falsely detected as drones by YOLO (on left), have been clearly rectified and only true drones were detected by our work (on right). In the first image (a), the two small black birds on the top, are detected as drones by YOLO, and the small white drone below is not detected at all, whereas YOLO with VAE detects only the white drone correctly. In the second image (b), the white bird is detected as a drone by YOLO, but YOLO with VAE correctly discards it as a false positive. In the third comparison, we see how only a bird's image has been detected as a drone by YOLO, but our work does not detect it as a drone.

7 BENEFITS AND IMPLICATIONS

7.1 Reduced False Positives

One of the primary benefits of the proposed approach is a significant reduction in false positive detections. By leveraging VAEs' anomaly detection capabilities, the system is better equipped to distinguish anomalies from normal objects, contributing to a more reliable object detection process.

The proposed approach markedly improves object detection precision. Even in complex and dynamic real-world scenarios, the system maintains high accuracy, minimizing the chances of misclassification and mislabeling.

7.2 Application in Safety-Critical Scenarios

The application of this approach is pivotal in safety-critical fields. For instance, in autonomous vehicles, where precise object detection is essential, the reduction of false positives significantly contributes to system safety. This has the potential to save lives and reduce accidents.

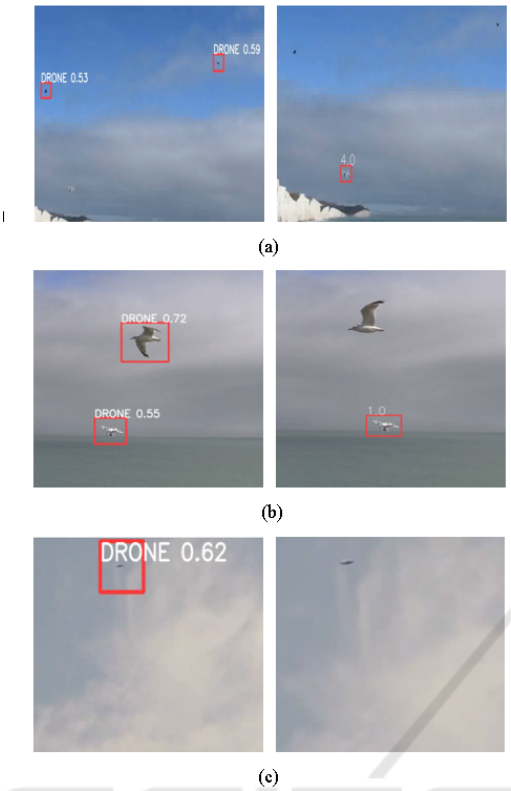

Figure 3: Results by YOLO (left) vs. Results by YOLO with VAE filter (right). Image (a) has 2 small black birds on top and 1 small white drone below. Image (b) has 1 white drone and 1 white bird. Image(c) has a single small black bird on top.

7.3 Threshold Adaptability

The classification threshold for anomaly detection can be adjusted to meet specific application requirements. This adaptability allows users to balance precision and recall based on the desired performance characteristics. This flexibility makes the approach applicable to a wide range of use cases.

7.4 Potential for Real-Time Applications

The proposed approach is amenable to real-time applications, making it suitable for scenarios where timely decision-making is crucial, such as targeting a drone.

8 CHALLENGES AND CONSIDERATIONS

8.1 Dataset Bias

One significant challenge is dataset bias. The performance of the VAE as a filter heavily depends on the quality and representativeness of the training dataset. A biased or incomplete dataset may lead to unintended filtering outcomes.

8.2 Threshold Tuning

Selecting an appropriate reconstruction error threshold for anomaly detection is a non-trivial task. It requires a balance between false positives and false negatives, and the optimal threshold may vary across applications.

8.3 Computational Overhead

The introduction of VAE filtering adds a computational overhead to the object detection pipeline. Ensuring real-time performance in resource-constrained environments is a critical consideration.

8.4 Ethical and Privacy Concerns

The use of object detection systems in surveillance and other applications raises ethical and privacy concerns. Enhanced object detection should be paired with appropriate ethical frameworks to address these issues.

8.5 Adversarial Attacks

Adversarial attacks against VAE-based filtering systems pose a significant threat, as attackers may manipulate input data to deceive the filtering mechanism and bypass security measures (Xu et al., 2020). Research efforts should focus on enhancing the robustness of VAE-based systems to defend against such attacks, ensuring the reliability and integrity of these systems, particularly in critical applications like autonomous vehicles, surveillance, and industrial automation.

9 CONCLUSION

In conclusion, the integration of Variational Autoencoders as a filtering mechanism within the YOLO

architecture holds great promise for enhancing object detection precision. By harnessing the VAE's anomaly detection capabilities, a substantial reduction in false positives can be achieved, thereby improving the reliability of object detection systems.

This approach is particularly pertinent in safety-critical applications, and further research and experimentation will be essential to fine-tune the system for optimal performance in diverse and dynamic real-world scenarios.

10 FUTURE WORK

The proposed approach opens the door to various avenues for future research and development:

10.1 Robustness Testing

To assess the robustness of the VAE filtering mechanism, a comprehensive testing plan should cover various environmental conditions and scenarios. This includes evaluating performance under different lighting, temperature, humidity, indoor and outdoor settings, static and dynamic scenarios, crowded or sparse environments, and adverse conditions like rain, fog, and sensor interference. The VAE should also be tested with various sensor types, calibrations, and occlusions. Assessing its adaptability to temporal changes and real-world applications is crucial. Quantitative metrics and qualitative user feedback should be used to evaluate performance, and an iterative testing process should be employed for continuous improvement.

10.2 Integration with Multi-Modal Data

Extending the approach to accommodate multi-modal data, such as the fusion of images and lidar data in autonomous driving, holds significant promise. Combining these data modalities can enhance the perception capabilities of autonomous vehicles, enabling them to better understand their surroundings and make more informed decisions. The synergy between image and lidar data can provide depth information, object detection, and contextual awareness, which is crucial for safe and efficient navigation. Research in this direction has the potential to unlock advanced solutions for autonomous systems, improving their reliability and safety in complex real-world environments.

10.3 Real-World Deployment

Real-world deployment and testing in safety-critical applications, such as autonomous vehicles, will provide valuable insights into the practicality and effectiveness of the approach.

10.4 Ethical Frameworks

The development of ethical frameworks and guidelines for the use of object detection systems enhanced with Variational Autoencoder (VAE) filters is imperative to tackle privacy and fairness concerns. VAE filters have the potential to significantly impact data privacy by filtering sensitive or unnecessary information, yet their implementation can raise ethical questions about what information is filtered and retained. Furthermore, fairness concerns arise when decisions made based on filtered data disproportionately affect certain groups or individuals. Robust ethical frameworks (Diakopoulos, 2016) are essential to establish guidelines for responsible use, data handling, transparency, and accountability, ensuring that VAE-enhanced object detection systems operate ethically, respecting privacy and promoting fairness in their decision-making processes.

REFERENCES

An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2(1):1–18.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee.

Diakopoulos, N. (2016). Algorithmic accountability: A primer. *Data Society Research Institute*.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Li, Y., Huang, X., Li, J., Du, M., and Zou, N. (2019). Specae: Spectral autoencoder for anomaly detection in attributed networks. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2233–2236.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In

Proceedings of the IEEE international conference on computer vision, pages 2980–2988.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer.

Liu, Y., Sun, P., Wergeles, N., and Shang, Y. (2021). A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems with Applications*, 172:114602.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee.

Xu, H., Ma, Y., Liu, H.-C., Deb, D., Liu, H., Tang, J.-L., and Jain, A. K. (2020). Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17:151–178.

Ye, A., Pang, B., Jin, Y., and Cui, J. (2020). A yolo-based neural network with vae for intelligent garbage detection and classification. In *Proceedings of the 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 1–7.