



Performance Analysis of Unsupervised Object Localization Using Autoencoder and K-Means Clustering

Ashiya^{1,*} and Dr. Kirti Jain¹

Sanjeev Agarwal Global Educational University Bhopal, Madhya Pradesh-462022, India
aasiyakhan577@gmail.com, kirti.j@sageuniversity.edu.in

Abstract. Object detection is a fundamental task in computer vision, enabling machines to identify and locate objects in images across applications such as surveillance, autonomous systems, and medical imaging. However, most existing approaches rely heavily on supervised learning and require large annotated datasets, which are expensive and time-consuming to obtain. To address this limitation, this paper proposes an unsupervised object detection framework that combines convolutional autoencoder-based feature learning with K-Means clustering. The model learns compact latent representations of input images and groups similar features to identify object regions without requiring labeled data. The proposed approach is evaluated on representative sample images, achieving reconstruction error values in the range of 0.015 to 0.018 and clustering variance between 0.22 and 0.25. The results demonstrate that the model can effectively capture meaningful visual structures and perform approximate object localization in annotation-scarce environments. This work highlights the potential of unsupervised learning as a practical alternative for object detection when labeled datasets are limited.

Keywords: Unsupervised Learning, Object Detection, Clustering, Autoencoder, Feature Learning, Image Localization, Computer Vision.

1 Introduction

Object detection focuses on the identification and location of objects within images. It plays an important role in many computer vision applications, such as autonomous systems, surveillance, medical imaging, and industrial inspection. By accurately detecting objects, systems can better understand visual scenes and make informed decisions.

Jing and Tian [15] stated that unsupervised learning offers a way to learn useful patterns directly from raw data without manual annotations. Vincent et al. [13] stated that autoencoder-based models are commonly used in this setting because they can learn compact feature representations by reconstructing input images. These learned features capture the overall structure of objects while reducing background noise. Jain [14] discussed advances in clustering techniques for unsupervised learning.

Lin et al. [12] and Jain [14] discussed that clustering techniques help group similar feature patterns, which can reveal prominent object regions in images. However, performing object localization without supervision remains challenging due to variations in object appearance, background complexity, and scale differences.

Zhou et al. [16] and van den Oord et al. [17] highlighted the importance of unsupervised representation learning for object detection tasks. In this work, we propose an unsupervised object detection framework that combines convolutional autoencoder-based feature learning with clustering techniques. The framework first learns latent representations from input images and then applies clustering to group similar features. This process enables approximate object localization without relying on labeled data. We evaluate the proposed approach through qualitative experiments, demonstrating its potential as a simple and effective proof-of-concept for unsupervised object detection.

Liu et al. [11] and Lin et al. [12] highlighted that most existing object detection approaches rely heavily on labeled datasets and complex architectures. There is a need for a simple and computationally efficient unsupervised framework that can perform object localization without requiring annotated data. This work aims to address this gap by proposing a lightweight and interpretable model.

2 Related Work

Simeoni et al. [6] surveyed recent advances in unsupervised object detection based on self-supervised representation learning. Longa and Henriques [2] introduced TokenCut, which leverages Vision Transformer (ViT) attention maps and a normalized

cut algorithm on the self-attention graph to perform object discovery without labels, showing significant improvements in localization performance.

Wang et al. [3] introduced CutLER, a "Cut-and-Learn" pipeline that uses MaskCut to generate initial masks and iteratively refines them through a self-training mechanism. While these methods achieve high Average Precision (AP), they often involve complex multi-stage training or depend on large-scale pre-trained models. Our work differs by focusing on a lightweight single-stage autoencoder-clustering pipeline that minimizes parameters while maintaining interpretability—a critical requirement for edge-computing applications and resource-constrained environments.

Unsupervised object detection and localization have gained significant attention in recent years as researchers aim to reduce dependency on large annotated datasets. Traditional object detection models rely heavily on supervised learning and bounding box annotations, which are expensive and time-consuming to obtain. Zunair and Hamza [1] and Longa and Henriques [2] explored unsupervised and self-supervised learning strategies to extract meaningful representations of objects without manual labeling. Autoencoder-based frameworks have been widely adopted to learn compact latent features that preserve structural and semantic information from input images. Wang et al. [3] proposed unsupervised attention-guided detection frameworks that encourage the network to focus on the salient regions of an image for object localization. Although such approaches demonstrate promising results, they often require carefully designed loss functions and additional hyperparameter tuning. In complex scenes with cluttered backgrounds, attention mechanisms may not consistently separate foreground objects from background regions.

Clustering-based methods have also been explored for unsupervised object discovery. Feature grouping strategies in latent space were investigated by Wu et al. [4] and Zhang et al. [5] to distinguish object regions from background patterns. Hierarchical and contrastive grouping approaches were further discussed by Simeoni et al. [6], where improved feature separation was achieved by similarity-based optimization. However, increased computational complexity and dependence on carefully designed contrastive losses were also observed, which may limit scalability across diverse datasets.

Motion-based and flow-driven optical methods were introduced by Zhu et al. [7] to improve unsupervised segmentation performance, particularly for video data. Although good performance was achieved in dynamic scenes, these methods were found to be less suitable for single-image object detection because of their reliance on temporal information. Representation disentanglement techniques were also explored by Charakorn et al. [8] to isolate object-specific features within learned embeddings. However, performance degradation was reported when significant variations in object scale, orientation, or appearance were present.

Recent survey studies by Zhang et al. [9] indicated that unsupervised object localization methods still face challenges related to robustness, background clutter, and computational efficiency. It was also observed that many existing frameworks rely on architectural complexity or auxiliary supervision signals to improve performance, which reduces the simplicity expected from unsupervised learning approaches.

To address these limitations, the proposed framework adopts a simple yet effective strategy by combining convolutional autoencoder-based feature extraction with K-Means clustering in the latent space. Instead of relying on complex attention modules or contrastive objectives, the method leverages compact latent representations and straightforward clustering to approximate object regions. This design reduces computational overhead while maintaining consistent performance across representative sample images. By emphasizing structural simplicity and interpretability, the proposed approach bridges the gap between lightweight unsupervised learning and practical object localization.

3 Proposed Methodology

This section presents the proposed unsupervised object detection framework. The approach consists of multiple stages including image preprocessing, feature extraction using a convolutional autoencoder, clustering in the latent space, and mapping clustered features back to the image domain to obtain object regions. The overall design aims to achieve object localization without requiring labeled data while maintaining computational simplicity.

The proposed unsupervised object detection framework processes input images, extracts meaningful representations, and identifies potential object regions without relying on labeled data. The workflow consists of preprocessing, feature extraction, latent-space clustering, and object-region mapping.

The framework begins with preprocessing the input images, where they are resized and normalized to ensure consistency and enhance the stability of downstream processing. Following preprocessing, the images are encoded through a convolutional autoencoder, which learns compact latent representations that capture essential visual patterns and structural information. The latent features are subsequently clustered to group similar patterns and isolate regions likely to correspond to objects. Finally, the clusters are mapped back to the original image space, producing precise object localization outputs. This high-level overview introduces the reader to the system's workflow, setting the stage for the following subsections where the autoencoder, latent feature processing, clustering strategy, and object region mapping are described in detail.

3.1 Autoencoder Architecture and Training

We used a convolutional autoencoder to learn useful features from unlabeled images. The autoencoder consists of two main components: an encoder and a decoder. The encoder gradually compresses the input image and learns important visual patterns

such as edges and shapes, while the decoder reconstructs the original image from the compressed representation. Related representation-learning approaches have been discussed in unsupervised object localization and feature-learning studies [1, 10, 13].

Let x represent an input image. The encoder converts the image into a latent feature representation z as:

$$z = f_{\phi}(x) \quad (1)$$

The decoder reconstructs the image from the latent features:

$$\hat{x} = g_{\phi}(z) \quad (2)$$

During training, the model minimizes the difference between the original image and the reconstructed image. We use the mean squared error as the re-construction loss:

$$L_{rec} = \|x - \hat{x}\|^2 \quad (3)$$

By minimizing this loss, the autoencoder learns to preserve important visual information while removing unnecessary background details.

3.2 Latent Feature Extraction

After training, we only use the encoder part of the autoencoder. For each input image, the encoder generates latent feature maps from the bottleneck layer. These latent features provide a compact and meaningful representation of the image.

Compared to raw pixel values, latent features have a higher resistance to noise, lighting changes, and background variations. Therefore, they are well suited for unsupervised object discovery.

3.3 Clustering Using K-Means

K-Means clustering is selected due to its simplicity and efficiency, while autoen-coders are used because of their ability to learn compact feature representations in an unsupervised manner. After learning compact feature representations using the convolutional autoencoder, K-Means clustering is applied to the latent features. K-Means groups similar feature vectors based on their distances in the latent space. The latent features obtained from the encoder are passed to the K-Means clustering module to separate foreground and background regions.

Let z_i denote the latent characteristic vector obtained from the encoder. The objective function of K-Means clustering is given by:

$$\arg \min_{i=1}^{\Sigma} \|z_i - \mu_{c_i}\|^2 \quad (4)$$

where μ_{c_i} represents the centroid of the cluster assigned to z_i and where N represents the total number of feature vectors, and c_i denotes the cluster assigned to the i -th feature vector.

In this work, the number of clusters is set to $K = 2$, corresponding to the foreground and background regions. The resulting clusters provide a coarse separation of object-related regions from the background, which supports approximate object localization in an unsupervised manner.

3.4 Object Region Mapping

The clustered latent features are mapped back to their corresponding spatial locations in the original image. This mapping enables visualization of cluster distribution across the image.

Clusters forming compact and continuous regions are considered as potential object regions, while scattered clusters represent the background. Based on this distinction, approximate object locations are obtained in the form of masks or bounding box candidates.

3.5 Proposed Algorithm

1. **Preprocessing:** Resize and normalize input images.
2. **Feature Extraction:** Encode images with a convolutional autoencoder.
3. **Clustering:** Group latent features via K-Means.
4. **Mapping:** Map cluster labels back to image space for object localization.
5. **Output:** Generate object maps for evaluation.

4 Experimental Result

In this section, we present the experimental results of the proposed unsupervised object detection framework. The evaluation is carried out using two representative sample images containing a cat and a dog. The model is evaluated on sample images obtained from publicly available datasets. Training was performed using a learning rate of 0.001 for 30 epochs in a standard system configuration. These examples are used to demonstrate the behavior of the framework and to analyze each stage of the pipeline. The overall process includes autoencoder-based feature learning, clustering in the latent space, and mapping of clustered regions back to the image for object localization.

4.1 Dataset and Experimental Setup

The proposed model is evaluated on representative sample images to demonstrate the feasibility of the approach. The model is implemented using Python with deep learning libraries such as TensorFlow/Keras. Training is performed using a learning rate of 0.001 and a suitable batch size. The experiments were conducted on a system with standard computational resources. Although the size of the data set is limited, it is sufficient to demonstrate the working principle of the proposed framework. The pixel values were normalized before processing. An unsupervised convolutional autoencoder was trained to learn compact latent representations of the images, as discussed by Charakorn et al. [8]. The encoder compresses the input image into a low-dimensional feature space, while the decoder reconstructs the image from this representation.

After feature extraction, K-Means clustering was applied to the latent features to separate object regions from the background. The resulting cluster labels were then projected back onto the original image space, and rectangular bounding boxes were used to visualize approximate object locations.

4.2 Autoencoder Reconstruction Results

The autoencoder successfully captures the dominant visual structures present in the input images. Although the reconstructed images appear slightly blurred, they preserve the overall shape and spatial layout of the objects. This behavior indicates that the model focuses on essential features rather than fine pixel-level details, which is beneficial for subsequent clustering. The reconstructed outputs appear slightly smoother than the corresponding inputs while preserving the overall structure of the objects. This smoothing effect can help suppress some background variation and supports the subsequent separation of object-related regions.

4.3 Clustering and Object Localization

Clustering was performed on the latent representations learned by the encoder. This process groups regions with similar feature characteristics, allowing separation between foreground objects and background areas. By mapping the cluster labels back to the image space, the framework identifies the dominant object regions. The cluster-based localization results indicate that the detected regions can approximately cover the main objects without using labeled data. The reconstructed images preserve the overall structure of the objects, indicating effective feature learning. The clustering results show reasonable separation between the foreground and background regions. The consistency in reconstruction error and clustering variance indicates stable performance.

4.4 Quantitative Evaluation

In addition to qualitative evaluation, a basic quantitative analysis was performed using reconstruction error and compactness of the

clustering. The reconstruction error reflects how well the autoencoder preserves important visual information, while the cluster variance provides an indication of feature separation in the latent space. The two representative samples show similar reconstruction-error and cluster-variance ranges, providing a basic indication of consistent behavior. The experimental results indicate that the proposed model is capable of capturing essential object structures despite minor reconstruction blurring. The autoencoder effectively encodes important visual features, while clustering in the latent space enables an approximate separation of foreground objects from the background. The generated object regions demonstrate the feasibility of unsupervised object localization.

For Sample Image 1, the reconstruction error (MSE) is 0.015 and the cluster variance is 0.22. For Sample Image 2, the reconstruction error is 0.018 and the cluster variance is 0.25. The relatively close values across the two samples indicate consistent behavior for these representative examples.

4.5 Discussion and Future Work

The experimental results demonstrate that the proposed unsupervised framework is capable of learning meaningful feature representations and performing approximate object localization without relying on labeled data. The use of a convolutional autoencoder enables effective feature extraction, while K-Means clustering helps in identifying object-like regions by grouping similar latent representations.

Although the proposed approach produces consistent and stable results, its localization accuracy is lower compared to fully supervised methods. This limitation is expected due to the absence of annotated training data and the simplicity of the clustering strategy. Additionally, the evaluation is conducted on a limited number of sample images, which may not fully capture the variability present in large-scale datasets. Future work can focus on improving the robustness and accuracy of the model by incorporating more diverse datasets and advanced clustering techniques. The integration of self-supervised or semi-supervised learning methods can further enhance feature representation. Moreover, extending the framework to support multi-scale feature extraction and more precise object boundary detection can significantly improve localization performance.

5 Conclusion

We presented an unsupervised framework for object detection using autoencoder-based feature extraction and clustering. The autoencoder successfully captured key features from representative sample images, while clustering allowed approximate object localization without any labeled data. The reconstructed images highlighted essential structures, making the clustering more accurate and stable.

Quantitative analysis using reconstruction error and cluster variance showed that the framework performs consistently across different images. Low reconstruction errors confirm that the autoencoder retains important visual information and the cluster variance values indicate effective separation of object regions from the background. These results demonstrate that unsupervised representation learning can provide meaningful object localization without manually annotated training labels.

The proposed method demonstrates the potential of unsupervised learning for object detection in scenarios with limited labeled data. Although the results are promising, the study is limited by a small dataset and the lack of a comprehensive evaluation. Future work will focus on large-scale validation, inclusion of standard evaluation metrics, and comparison with existing methods.

References

1. Zunair, H., Hamza, A.B.: PEEKABOO: Hiding parts of an image for unsupervised object localization. arXiv preprint arXiv:2407.17628 (2024).
2. Longa, M., Henriques, J.F.: Unsupervised Object Detection with Theoretical Guarantees. In: Proceedings of the Neural Information Processing Systems (NeurIPS) (2024).
3. Wang, X., Girdhar, R., Misra, I.: Cut and Learn for Unsupervised Object Detection and Instance Segmentation. arXiv preprint arXiv:2301.11320 (2023).
4. Wu, Z., Moemeni, A., Caleb-Solly, P.: Ensemble Foreground Management for Unsupervised Object Discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025).
5. Zhang, L., et al.: Towards Unsupervised Object Detection from LiDAR Point Clouds. arXiv preprint arXiv:2311.02007 (2023).
6. Simeoni, O., et al.: How far can we go without manual annotation? A survey of unsupervised object localization methods. arXiv preprint arXiv:2310.12904 (2023).
7. Zhu, V., et al.: GMAIR: Unsupervised Object Detection Based on Spatial Attention and Gaussian Mixture. arXiv preprint arXiv:2106.01722 (2021).
8. Charakorn, R., et al.: SPLIT: Representation Disentanglement for Deep Clustering and Unsupervised Object Detection. arXiv preprint arXiv:2001.09597 (2020).
9. Zhang, X., et al.: HEAP: Unsupervised Object Discovery and Localization with Contrastive Grouping. arXiv preprint arXiv:2312.17492 (2023).
10. Mei, G., Saltori, C., Ricci, E.: Unsupervised Point Cloud Representation Learning by Clustering and Neural Rendering. International Journal of Computer Vision (2024).
11. Liu, W., et al.: SSD: Single Shot MultiBox Detector. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 21–37 (2016).
12. Lin, T.Y., et al.: Focal Loss for Dense Object Detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017).
13. Vincent, P., et al.: Extracting and Composing Robust Features with Denoising Autoencoders. In: Proceedings of the International

Conference on Machine Learning (ICML), pp. 1096–1103 (2008).

14. Jain, A.K.: Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters 31(8), 651–666 (2010).

15. Jing, L., Tian, Y.: Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 43(11), 4037–4058 (2020).

16. Zhou, Z., et al.: Unsupervised Object Detection: A Survey. arXiv preprint (2021).

17. Van den Oord, A., et al.: Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv:1807.03748 (2020).