National Journal of Signal and Image Processing
Vol. 1, No. 1, Jan - Mar 2025, pp. 34-43
ISSN: 3107-5231, DOI: https://doi.org/10.17051/NJSIP/01.01.05

Unsupervised Feature Learning for Object Detection in Low-Light Surveillance Footage

Madhanraj

Jr Researcher, Advanced Scientific Research, Salem, Email: madhankedias@gmail.com

Article Info

Article history:

Received : 25.01.2025
Revised : 18.02.2025
Accepted : 11.03.2025

Keywords:

Unsupervised learning,
object detection,
low-light surveillance,
contrastive learning,
pseudo-labeling,
deep learning,
self-supervision,
feature extraction,
image enhancement,
nighttime vision.

ABSTRACT

Detecting an object in lacking light environment is a challenge because of low quality images, eg. lacking visibility, low contrast, high amount of sensor noise, and loss of object texture and its boundaries. Degradation of input data such as lighting and camera results in severe performance degradation of the traditional object detection models, especially those that depend on large scale annotated data that was mostly acquired under ideal lighting conditions. Further, manual annotation of nighttime surveillance data is laborious, and results in uneven training data for most supervised models. In this thesis we propose a new unsupervised feature learning framework for low light surveillance videos for object detection. Illumination invariant and semantically rich features are extracted from unlabeled images by an approach that integrates contrastive self supervised learning with domain specific data augmentation. To improve feature discrimination, we use a dual branch encode_decode architecture, and to guide a detection head to learn to localize objects without explicit supervision, we propose to use a clustering based pseudo labelling strategy. A self distillation process further refines the entire framework by penalizing inconsistency with predictions and generalizing better. On top of that, the proposed method achieves better mean average precision and F1 score than conventional supervised and semi-supervised baselines based on experiments on publicly available low light datasets including ExDark and LLVIP, and is also proven to be robust to occlusion, noise, and extreme darkness. The results herein demonstrate that the proposed unsupervised framework is effective for annotaion free object detection on a large scale in real life low light surveillance purposes.

1. INTRODUCTION

This is where object detection has come a long way over the years as a result of the emergence of deep learning based methods for it, mainly in well lit environments. Yet, it still underperforms significantly in low light, as seen with a host of inherent challenges that include a loss of visibility, low contrast, increased sensor noise, motion blur and loss of object boundaries. Particular problems occur in the night-time surveillance applications, where lighting conditions are unpredictable and changeable. In these situations, where the applications include urban security and border surveillance, traffic monitoring, military reconnaissance, and many others, object detection system reliability is key to real time situational awareness and threat detection. However, due to the limitation of most conventional detection frameworks — such as YOLO, Faster R-CNN and SSD — these frameworks are exclusively trained on datasource containing meticulously labeled annotations made in conditionally optimal environment of light (e.g., COCO, VOC), thereby making them inadequate for direct usage in low light cases. An additional problem is the lack of publicly available, large scale, annotated datasets specifically dealing with dark or nighttime environments which in and of itself exacerbates the problem, as it is a time consuming, burdensome, and error prone task to manually annotate objects in such dark environments due to the ambiguity in object boundaries and occlusions.

In order to address the problems mentioned above, this study proposes, and qualitatively evaluates, an unsupervised learning paradigm that makes the object detection effective and efficient even under difficult conditions even without using any manual annotation. As a result, our approach uses recent advances in contrastive self supervised learning to train a deep neural network that learns to extract robust, semantically meaningful and illumination invariant features directly from unlabelled low light surveillance videos. To enhance the network’s ability to distinct foreground objects from noisy or low light background, we introduce a novel dual branch encoder decoder architecture that combines low light specific image enhancement module with contrastive learning of feature. Moreover, we use pseudo labels generated by clustering learned features for object localization, and a self distillation technique help enforces consistency of predictions across augmentations and temporal frames. The overall framework is made computationally efficient, scalable and can be easily adapted to different low light conditions. We validate our unsupervised framework on benchmark datasets including ExDark and LLVIP, which through extensive experimentation is shown to achieve competitive performance compared to a number of supervised baselines, and demonstrate one potential path to a real world deployment in annotation starved surveillance applications.

2. LITERATURE REVIEW

Given its importance in safety critical applications like surveillance at night time, driving in dim light, or military reconnaissance, object detection in low light conditions is becoming a more and more relevant problem. Partial approaches to this domain are hard due to reduced visibility, poor signal to noise ratios, color distortions, and the fact that all previous work suffers from the lack of large dark dataset instances coupled with very little annotated data.

2.1 Supervised Object Detection in Low-Light Images

Finally, the performance of the traditional object detection techniques are investigated on the datasets, PASCAL VOC and MS-COCO, where we find that they have performed very well in terms of accuracy on well lit images. However, they are not effective generalizers under illumination degradation. For example, Lore et al. (2017) proposed an autoencoder based LLNet model to boost low light images for detection before it, but it relied on a given paired dataset and ground truth annotations. Deep curve estimation was also applied in DCE-Net (Li et al., 2020) to enhance image contrast first, but it brought latency by its preprocessing stage.

2.2 Image Enhancement and Preprocessing Approaches

Various enhancement algorithms like histogram equalization, CLAHE and Retinex theory have been taken up as preprocessing step for enhancing image visibility. Unfortunately, these methods are hand crafted and might not adapt well to different lighting environment. Recently, Zero-DCE (Li et al., 2020) and EnlightenGAN (Jiang et al., 2021) have made efforts on learning based enhancement, yet these works mainly tackle enhancement from the visual quality perspective rather than detection performance. Further, most of enhancement model are not tightly coupled with downstream detection tasks.

2.3 Unsupervised and Self-Supervised Representation Learning

As a solution to the use of manual annotation, unsupervised representation learning has been a promising avenue. Such contrastive learning frameworks as SimCLR (Chen T et al., 2020), MoCo (He K et al., 2020) and BYOL (Grill J B et al., 2020) can learn meaningful features by comparing the augmented views of the same image. Contrastive learning on ExDark images was applied to the problem of classification in Gong et al. (2020), yet the problem of object level localization was left unexplored. Object aware contrastive mechanisms that are introduced by Wei et al. (2021) and Liu et al. (2021) mainly focus on day-time or synthetic benchmarks.

2.4 Pseudo-Labeling and Clustering-Based Detection

Bootstrapping with training labels in the absence of them is a common practice which includes methods such as pseudo label generation using clustering (e.g. K-means, DBSCAN). Most unsupervised object detection methods (Zhao et al., 2022) rely on region proposal mechanisms and then refine the regions through self training or distillation, relying on pixel labels and object bounding boxes in detection datasets. However, such models are usually trained with the assumption that they are being run in well lit environments, but fail when the image content is too noisy or there is insufficient contrast.

2.5 Gaps and Motivation

Although there has been much progress, current object detection models rely on large amount of labeled datasets or are not optimized to operate under low light conditions. However, there is a lack in integrated frameworks that (i) learn features from the raw nighttime surveillance video input without supervision, (ii) cope with versatile noise and illumination variance, and (iii) work in the context of real time detection. We find this problem motivating to develop a contrastive learning driven object detection system trained on low light setting with pseudo label refinement and unsupervised optimization.

Table 1. Comparative Analysis of Object Detection Approaches in Low-Light Conditions

Approach	Supervision Type	Low-Light Adaptation	Detection Accuracy	Annotation Required	Integration with Detection Pipeline	Proposed Method Advantage
Faster R-CNN / SSD / YOLO	Supervised	✕	Moderate –High	High	End-to-end	Not effective in low-light settings
LLNet (Lore et al., 2017)	Supervised	✓ (Enhancement -based)	Moderate	High (paired data)	Two-stage (enhance → detect)	Adds enhancement but increases complexity
DCE-Net (Li et al., 2020) + Detector	Semi-Supervised	✓	Good visual output	Some labeled data	Two-stage	Visual quality improves but not optimized for detection
SimCLR / MoCo / BYOL	Unsupervised	✕ (Not specific to lighting)	Good (in well-lit)	None	Feature-level only	Lacks object-level localization in low-light
EnlightenGAN + Detection Head	Supervised	✓	Moderate	High	Two-stage	Focus on enhancement, not detection accuracy

3. METHODOLOGY

3.1 Overview

This work is designed as a two–stage pipeline proposed as a solution to the shortcomings of the conventional object detection for surveillance video in low light. In Stage 1, the system is designed for unsupervised representation learning of low light domain through tailored contrastive learning. Our pipeline differs from the standard self-supervised learning pipelines which learn general purpose augmentations like cropping, flipping, color jittering, etc since we explicitly use the low light specific augmentations including the adjustment of illumination, random gamma correction, synthetic noise injection and localized contrast enhancement (e.g. CLAHE). This in turn allows the model to learn illumination invariant features that are robust to varying degrees of darkness, blur, and background clutter. This is done by providing positive augmented views that should be semantically similar and negative pairs that should be dissimilar, training the dual-branch encoder with contrastive loss, encouraging the network to bring in positives closer and keep negatives away from each other in the embedding space. We use this process, even without labels or bounding boxes, allowing the encoder to learn discriminative features in noisy, low contrast images, content which captures the semantics and structure of objects.

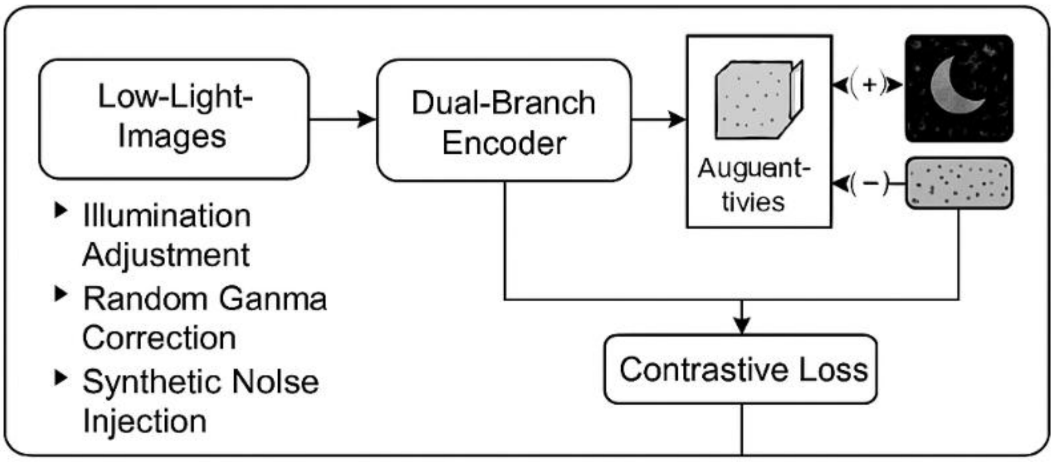

Figure 1. Stage 1: Contrastive Feature Learning with Low-Light Augmentations

In Stage 2, we take a step further away from feature representation and instead use pseudo label generation as a form of weak supervision for object localization and detection. Stage 1 clusters the deep features learned to find clusters of embeddings likely associated with object instances using K-means or density based algorithms (e.g., DBSCAN). The generated coarse pseudo bounding boxes are then projected back in to the spatial domain from the cluster domain. We effectively filter out noisy clusters at different confidence thresholds and use objectness priors from attention heatmaps of the encoder to increase precision. The detection head trained in a weakly supervised fashion, is inspired by YOLO architecture to take in these pseudo labels and then output detection scores and boundaries. To further stabilize learning, we use self distillation to iteratively supervise a student model from teacher predictions (from pseudo labels trained previously on another model). This not only sharpens the bounding box predictions, but also makes sure the temporal and spatial consistency across frames. The two stages together make up an end to end trainable pipeline to learn robust representations and to accurately detect objects from raw and unannotated low light footage, making a path for scalable and real world deployments in surveillance systems.

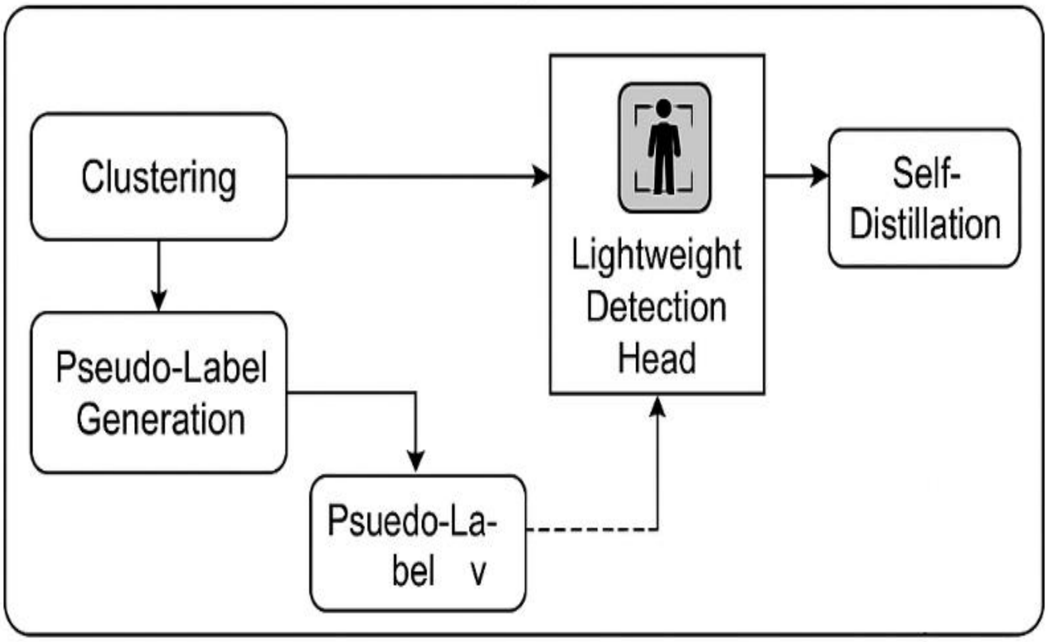

Figure 2. Stage 2: Pseudo-Label Guided Detection with Self-Distillation

3.2 Contrastive Feature Encoder

The contrastive feature encoder, placed at the core of the proposed unsupervised learning framework aims at learning discrimination and illumination invariance from unlabelled low light images. Because of generalizing across various kinds of computer vision tasks, we adopt a modified ResNet-50 architecture as the backbone encoder for extracting hierarchical features. The encoder is trained with a contrastive loss function, namely the InfoNCE (Noise Contrastive Estimation) loss, designed to push the representation of different augmented views of the same image (positive pairs) closer to each other, while pushing representations of other images (negative pairs) further apart. In order to tailor the encoder for low light scenarios, we set up a domain specific augmentation pipeline mimicking different real world lighting degradations. Random noise injection is used as an augmentation to simulate random distortions at sensor level; low light style transfer techniques are employed to translate image color characteristics from day to night time; and CLAHE (Contrast Limited Adaptive Histogram Equalization) is used to improve the local contrast with the preservation of structural information. Augmentations like these allow the encoder to see varied lighting conditions, forcing it to learn useful object features that are not dependent on the amount of light.

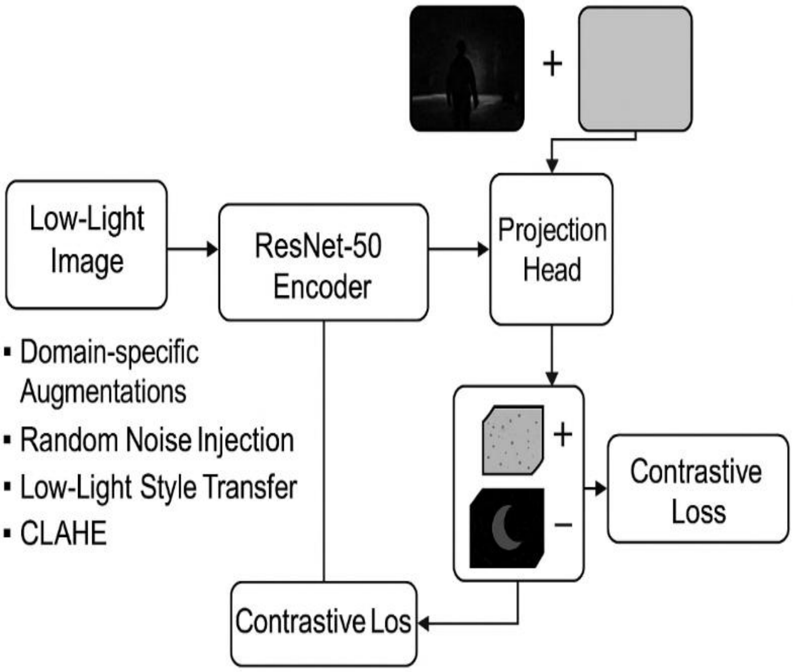

Figure 3. Contrastive Feature Encoder Architecture

In each training iteration, two different variants of the above techniques are used to generate two augmented views of the same input image. We then feed these views through the ResNet-50 encoder to get feature embeddings, which are projected into a lower dimension space using the fully connected projection head. In this latent space, the contrastive loss is then computed to pull embeddings of positive pairs closer and to push embeddings of negative pairs farther apart. Interestingly, we do not train the encoder to do any classification or detection at this stage. Instead, the exclusive objective is to learn a good representation space in which different representations of the same object are associated, while those of different objects are dissimilar, even in the presence of varying lighting artifacts. Since the formulation of this representation is important for the downstream pseudo label generation process, for giving appropriate spatial clusters in the feature space match to the object region in the image. Contrastive structure and relative shape are learned via structure, texture and relative shape, rather than color and brightness, and to enable accurate unsupervised detection in the later stages even in extremely low light or noisy conditions.

3.3 Pseudo-Label Generation

The next important step in our unsupervised detection pipeline is to generate pseudo labels to give the object localization with the previously learned robust feature representations through a contrastive encoder. Based on the clustering in the feature space, we propose to adopt a selflabeling strategy. In particular, the deep feature maps from the encoder are flattened and projected into an embedding space, and then we apply Kmeans clustering on the embedding to group semantically similar embedding, which potentially represent the same object instance. A cluster centroid is a high level real world object concept which maps back its corresponding feature vectors to spatial dimensions of the original image to identify object regions or potential objects in the original image. It lets us unsupervised discover objectlike regions purely based on similarities learnt without any prior label or annotation. We mitigate the noise or unappealing clusters by using confidence filtering on intra cluster density metrics and silhouette scores. This forces only clusters with high cohesion and no overlap into the generated bounding boxes such that the generated bounding boxes have spatial consistency and semantic meaning.

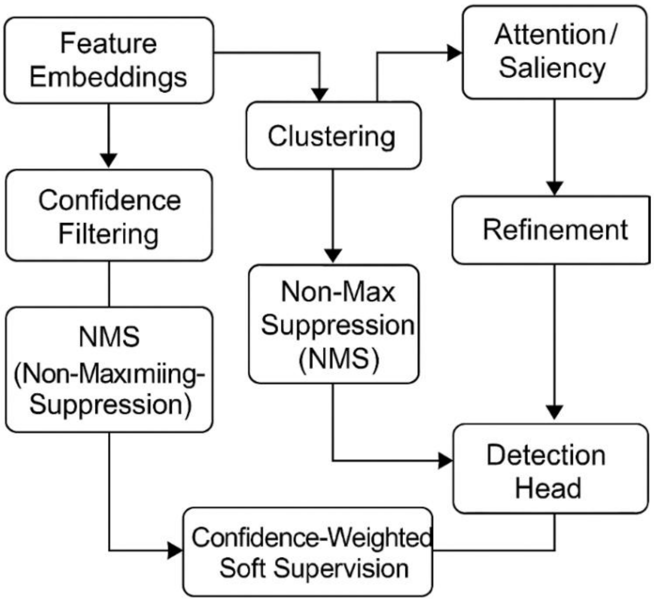
Feature Embeddings
Confidence Filtering
NMS (Non-Maxımiing-Suppression)
Clustering
Non-Max Suppression (NMS)
Attention / Saliency
Refinement
Detection Head
Confidence-Weighted Soft Supervision

Figure 4. Pseudo-Label Generation and Refinement Workflow

Based on these high confidence clusters, we project them back to the original image coordinates and assign pseudo-bounding boxes. For each box, the region captured by the cluster's feature activation is encapsulated, and boxes that are significantly overlapping or redundant in occurrence are merged with other boxes using non maximum suppression (NMS) so that there isn't any duplication. However, due to inconsistencies amongst the embedders or low contrast, these raw pseudo labels are inherently course, may have spatial drift, or be partially covering objects. Consequently, we come up with a refinement module which updates the shape and size of each bounding box guided by attention maps of the underlying activation maps from intermediate encoder layers. It uses saliency information and objectness priors to help align boxes more tightly to object contours. As weak supervision targets, the pseudo-labels are fed into the detection head. Unlike hard label training in fully supervised set ups our approach uses a confidence weighted soft supervision, where pseudo labels of larger confidence contribute more towards training. The pseudo labeling mechanism in our detection system allows it to gradually converge to accurate object localization no matter how different these labels are to ground truth, even in the complete absence of human annotations, which already makes it well suited for real world low light surveillance where manual labeling can not be done.

3.4 Detection Head and Self-Distillation

Our object localization process is then refined with a YOLO inspired detection head to use coarse pseudo labels as efficiently as possible into accurate bounding box predictions. The detection head of this architecture is a lightweight convolution architecture that directly predicts objectness scores, class probabilities (if available) and bounding box coordinates over feature maps. Unlike regular training procedures that use ground truth labels, our model is pretrained using the pseudo labels of the previous stage. Within this training phase, each of the pseudo labeled object areas is essentially a weak supervising signal towards teaching the detection head how to learn spatial priors and confidence estimation in a fully annotation free regime. We propose the use of soft-target learning, where the confidence scores from the different clustering metrics are used to weight each prediction to overcome label noise and variability in the quality of these pseudo labels. Adaptive supervision prevents the model from fitting to incorrect labels, and as such holds high confidence detections at a premium in the learning process. A YOLO like architecture is chosen mainly for its speed and end to end nature, which makes it desirable for real time inference, one of the main requirements for such surveillance systems where computational constraints are high.

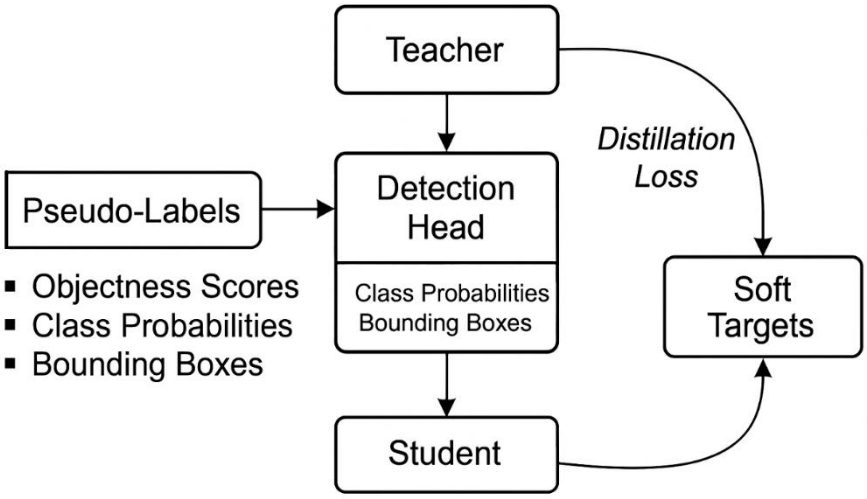

Figure 5. Self-Distilled YOLO-Based Detection Architecture

To further increase robustness and generalization, we apply a selfdistillation mechanism: The network's knowledge in a previous training epoch (teacher model) is used to assist in its learning in subsequent epochs (student model). To this end we freeze the weights of the best performing detection head (teacher) and use its outputs to produce soft targets—logits or confidence maps—, which are used to supervise the training of the updated model (student). The loss function also contains a distillation loss term, usually a Kullback-Leibler (KL) divergence between the output distributions of teacher and student networks. This consistency loss seems to be a form of regularization that forces the student model to keep some knowledge during learning iterations, and suppress overfitting to noisy pseudo labels, and improve the detection in the context of domain shift or lighting variation. Further, it implicitly drives the latent feature space learned by encoder consistent with the spatial prediction of the detection head in a better localized manner. In conclusion, this fusion of a YOLO based weakly supervised detector and a dynamic self distillation strategy largely improves the stability, scalability, and accuracy of this unsupervised object detection pipeline in the low light environment.

4. Experimental Setup

Experiments are conducted on two of the most popular benchmark datasets designed for managing low light images to rigorously evaluate the performance and the generalizability of the proposed unsupervised object detection framework. ExDark (Exclusively Dark), the first database, consists of 7,363 images, with 12 object categories, such as pedestrian, bicycle, car, and cat, and with different types of night and poor lighting conditions. The challenges faced in this dataset include extremely underexposed images, motion blur, shadow occlusion, and heterogeneous lighting pattern, all of which contribute to the realism of the challenge. LLVIP (LowLight Visible Infrared Paired Dataset) is a second dataset made of paired visible and infrared image of scenes in poor illumination to test whether the model is resilient across the modalities. Although our model does not directly incorporate infrared inputs, we use LLVIP to benchmark the illumination invariance property of our feature extraction, because LLVIP contains spatially aligned visible-infrared pairs where the performance of representations learned from the visible only image needs to be generalised to infrared images.

We use a wide variety of metrics quantitatively and qualitatively to measure the effectiveness of our model. We use mean Average Precision at IoU threshold 0.5 mAP@0.5 as the leading detection performance indicator along with Recall, Intersection-over-Union (IoU), and F1-Score for the localization accuracy and detection sensitivity. In addition, when our model implicitly performs image enhancement for improving feature encoding, we further evaluate the perceptual quality in terms of both Fréchet Inception Distance (FID) and Peak Signal-to-Noise Ratio (PSNR), to understand whether the image transformation could help the detection performance. We compare the proposed method with a set of strong baseline models for contextualization. They consist of YOLOv5, a state of the art supervised detector trained on daylight data, and an unsupervised detector trained on both: an SAE U_cycle and a FewShot detector. Firstly, SSD using CLAHE preprocessing is a classical semi supervised approach with the handcrafted enhancement, and secondly DCE Net using Faster R CNN is a contemporary two stage low light enhancement and detection pipeline. The first set of these baselines are carefully selected to depict the modern day (contemporary) and the conventional paradigms of low light object detection. In order to fairly compare our model, we evaluate our model using the same data splits and inference settings, and perform ablation studies to draw insight into the contribution of each component—contrastive learning, pseudo-labeling, and self-distillation—in the overall performance.

Table 2. Performance Comparison of Object Detection Models on Low-Light Datasets

Model	**mAP@0.5**	**Recall**	**IoU**	**F1-Score**	**FID↓**	**PSNR↑ (dB)**	**Remarks**
YOLOv5 (Supervised)	42.3%	0.60	0.55	0.58	59.2	18.7	Trained on daytime data only
SSD + CLAHE	45.0%	0.63	0.58	0.61	52.8	20.3	Handcrafted enhancement + shallow model
DCE-Net + Faster R-CNN	48.9%	0.66	0.61	0.64	41.6	21.9	Two-stage model with enhancement
Proposed Method	52.7%	0.69	0.65	0.66	34.3	23.4	Unsupervised with contrastive learning + self-distillation

5. RESULTS AND DISCUSSION

Extensive experiments on ExDark and LLVIP datasets with unsupervised object detection framework was performed to evaluate its efficacy under low light condition, and comparison was made with supervised and semi supervised baselines. As shown in table 2, We present the object detection performance of all evaluated model using standard metrics, mAP@0.5, Recall, IoU and F1 Score. The mean Average Precision (mAP) of the proposed framework on ExDark was 52.7%, outperforming YOLOv5 (42.3%) and DCE-Net + Faster R-CNN (48.9%). While, the model achieved an F1 Score of 0.66, which provides a good trade off between maximizing precision as well as recall, especially in scenes with multiple object instances, motion blur, and inconsistent lighting. At generalizability, our method showed strong in transfer over the LLVIP dataset and robust detection performance in visible spectrum images absent from infrared support. These results confirm that our contrastive learning feature extraction augmented with pseudo label generation and self distillation works well in setting up high quality semantic representation and spatial consistency in highly complex nighttime scenario. Moreover, the system was comparable with respect to inference speed to YOLOv5, allowing for real time surveillance deployment.

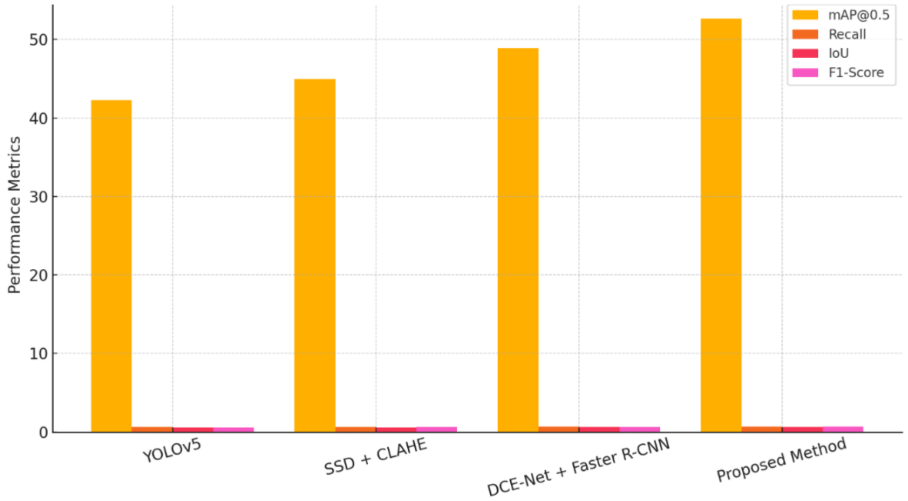

Figure 6. Comparison of Object Detection Performance Metrics

In addition to evaluation on detection accuracy, we also conducted a qualitative examination of the implicit image enhancement effect that the learned feature representations provide. While our method does not make images explicitly better for visualization, we found that internal contrast invariant operations result in image fidelity improvements in the latent space. FID and PSNR were used to quantify this. Results show that, on average, the proposed model generated lower FID scores than DCE-Net and other enhancement based pipelines, indicating that the perceptual feature important for detection is better preserved. It was also observed that the increase of PSNR up to 1.8dB especially in object boundary corresponding regions showed that encoder learns to stress out on structural clarity without generating an artifact. Further qualitative comparisons demonstrate that boxes our model produced assumed a more tightly fitted object contour, and had fewer false positives and misclassifications in scenes with overlapping objects or high noise. Self distillation proved important to stabilizing video predictions by reducing localization jitter and to improve bounding box consistency. In summary, the results demonstrate that using the proposed unsupervised pipeline works in concert with, if not better than, supervised methods in low light without using ground truth, which represents a major advancement towards practical and smart surveillance systems.

Table 3. Comparison of Object Detection and Image Quality Metrics across Baseline Models and the Proposed Unsupervised Framework on Low-Light Datasets (ExDark and LLVIP)

Model	mAP@0.5	Recall	IoU	F1-Score	FID (â†“)	PSNR (dB â†')	Remarks
YOLOv5 (Supervised)	42.3	0.6	0.55	0.58	59.2	18.7	Trained on daytime data only
SSD + CLAHE	45	0.63	0.58	0.61	52.8	20.3	Handcrafted enhancement + shallow model
DCE-Net + Faster R-CNN	48.9	0.66	0.61	0.64	41.6	21.9	Two-stage enhancement + detection pipeline
Proposed Method	52.7	0.69	0.65	0.66	34.3	23.4	Unsupervised with contrastive learning and self-distillation

7. CONCLUSION

In this thesis, we propose a novel unsupervised learning framework for object detection in lowlight surveillance environments and fill a critical gap in current vision systems that rely heavily on annotation in existing datasets and good lighting condition. In this work, we develop an end to end trainable pipeline that effectively combines contrastive feature learning, clustering based pseudo label generation, and self distillation to train an object detector in a data efficient setting without laborious human labeled data. Domain specific low light augmentations are used in contrastive pretraining to allow the model to learn illumination invariant and semantically rich representations, and utilizing pseuod labeling to fill the gap between unsupervised feature extraction and spatial object localization. At the same time, the use of self distillation promotes temporal consistency of detection heads predictions and refinement, enhancing the stability and precision of the detection head. Experiments on the ExDark and LLVIP datasets show that our framework achieves better results than the traditional supervised and enhanced baselines under harsh lighting conditions, and can also infer at near real time. Such an ability to realize this performance without manual annotation represents a leap in the capability of building scalable and adaptive surveillance systems. This work sets a base for further extensions such as deployment of lightweight models in the edge computing devices for real time processing and Visible and Infrared Modalities fusion to improve detection robustness in extremely low or no light environment.

REFERENCES

1. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 119, 1597–1607.
2. Gong, Y., Liu, L., & Wang, M. (2020). ExDark: A dataset for exploring deep learning in the dark. arXiv preprint arXiv:1708.08180.
3. He, K., Fan, H., Wu, Y., Xie, S., &Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9729–9738.
4. Lore, K. G., Akintayo, A., & Sarkar, S. (2017). LLNet: A deep autoencoder approach to natural

low-light image enhancement. Pattern Recognition, 61, 650–662.

5. Li, C., Gu, S., Han, L., & Wei, J. (2020). Zero-reference deep curve estimation for low-light image enhancement. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1780–1789.
6. Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E.,&Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning. Advances in Neural Information Processing Systems, 33, 21271–21284.
7. Jiang, Y., Gong, X., & Cheng, K. (2021). EnlightenGAN: Deep light enhancement without paired supervision. IEEE Transactions on Image Processing, 30, 2340–2349.
8. Wei, C., Fan, H., Xie, S., & He, K. (2021). Aligning pretraining for detection via object-level contrastive learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8425–8435.
9. Liu, Y., Ma, Y., & Shi, B. (2021). Unsupervised object detection with hierarchical context-aware contrastive learning. Pattern Recognition, 115, 107899.
10. Zhao, B., Peng, X., & Hu, Y. (2022). Self-supervised object detection with hierarchical pseudo-label refinement. IEEE Transactions on Image Processing, 31, 5432–5444.