Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
Special Issue on AI-Driven Creativity and
Intelligent Practices in Visual Arts 2025 6(4s)

OBJECT DETECTION IN PHOTOGRAPHY USING DEEP LEARNING

Saniya Khurana [1], Akash Kumar Bhagat [2], Dr. Rajesh Uttam Kanthe [3], Dipali Kapil Mundada [4], Dr. Tanmoy Parida [5], Dr. S. Prayla Shyry [6], Kumar Ambar Pandey [7]

[1] Centre of Research Impact and Outcome, Chitkara University, Rajpura- 140417, Punjab, India
[2] Assistant Professor, Department of Computer Science and IT, Arka Jain University Jamshedpur, Jharkhand, India
[3] Director, Bharati Vidyapeeth (Deemed to be University) Institute of Management, Kolhapur -416003, India
[4] Department of Engineering, Science and Humanities, Vishwakarma Institute of Technology, Pune, Maharashtra, 411037, India
[5] Associate Professor, Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India
[6] Professor, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India
[7] Assistant Professor, School of Journalism and Mass Communication, Noida, International University, 203201, India

Received 01 May 2025
Accepted 03 September 2025
Published 25 December 2025

Corresponding Author
Saniya Khurana,
saniya.khurana.orp@chitkara.edu.in

DOI
10.29121/shodhkosh.v6.i4s.2025.6835

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Copyright: © 2025 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License.

With the license CC-BY, authors retain the copyright, allowing anyone to download, reuse, re-print, modify, distribute, and/or copy their contribution. The work must be properly attributed to its author.

OPEN ACCESS

ABSTRACT

Object detection in photography has developed fast due to deep learning and has changed the manner in which visual content is photographed, arranged, and understood. This paper is a detailed examination of the current detection systems and how they can apply to the photographic process. Starting with the description of classical approaches like HOG, Haar cascades, and SVM-based networks, the paper compares the drawbacks of the mentioned methods with the advancement of CNN-based frameworks. R-CNN to Faster R-CNN is talked about and efficiency of region proposal and representational richness are improved. The single-shot detectors that are investigated are YOLO, SSD, and RetinaNet as they can offer high-speed inference, thus they are applicable to the real-time or mobile photography case. The study also examines photography-focused datasets like COCO, Open Images and expert-curated collections, which are annotation formats and augmentation strategies, which are taken into account in artistic variability, lighting and composition issues common to both professional and amateur photography. A new architecture based on applying modern backbones: ResNet, EfficientNet, and Swin Transformer and flexible detection heads is proposed. The loss functions that encompass robust localization, classification refinement, and variants of the IoU are combined so that they optimize the performance in various photographic scenes. Applications have shown very strong effect: automated tagging and image organization, real-time detection of both DSLR/mobile systems, and intelligent aid to the creation of art and subject-awareness to enhance composition.

Keywords: Object Detection, Deep Learning Photography, YOLO/Faster R-CNN, Image Annotation, Detection Architecture

1. INTRODUCTION

The object detection is now a basic part of contemporary photography because it allows cameras and other imaging devices to perceive visual scenes with more accuracy. With the transition of photography being more of an aesthetic activity to a cognitively-driven, data-driven field, the incorporation of deep learning has reshaped the way in which images are captured, processed and handled at professional and consumer levels. Historically, photographers were able to work out subjects, frame compositions, and focus on the important details basing on experience and intuition. But as digital images volumes continue to increase, the complexity of the shooting environment, and the need to have automated workflows has made the conventional methods of approach inadequate. Object detection based on deep learning provides a scalable, accurate, and contextual solution to the problems now. Early computer-vision systems like Haar cascades, Histogram of Oriented Gradients (HOG) and Support Vector Machines (SVM) were the pioneers in that they would allow simple object localization on handcrafted features. Even though useful in limited conditions, the models had difficulties with occlusions, changes in light, intricate textures, and different compositions, which are typical of artistic and natural-world photography Ribeiro et al. (2024). With the development of convolutional neural networks (CNNs), there was a new paradigm, where models can learn features of images at high levels, and in a hierarchical manner, directly out of the data. This development significantly enhanced the quality of the detectors and enabled automatic scene understanding to become a reality in the area of photography. The development of the R-CNN into Fast R-CNN and then to Faster R-CNN came along with the mechanism of region proposal that could effectively put objects in place without compromising the efficiency of computation Yang et al. (2024). Such frameworks showed that deep learning could extrapolate into caviar-sized photography collections in which subjects are of different size, posture, style, and artistic purpose. Figure 1 depicts deep learning pipeline that is used to detect objects that are based on photos. Simultaneously, single-shot detectors like YOLO, SSD, and Retina Net enabled real-time detection, which was directly suited to the requirements of mobile photography, mirrorless and DSLR images and embedded camera pipes.

Figure 1

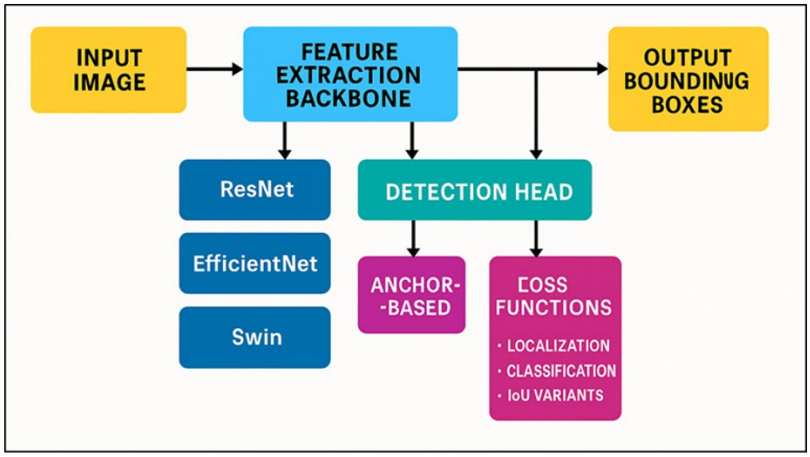

Figure 1 Deep Learning Pipeline for Photography-Focused Object Detection

The transition to long range dependency modeling was further promoted by the switch to transformer based architectures and enabled much stronger detection in cluttered compositions and in complicated lighting conditions. Object detection in photography has not just developed into a technical aid in modern-day photography, but also an imaginative and organizational facilitator. Intelligent cameras use detection to draw attention to some subjects, enable autofocus to stay constant, exposure to be optimized on faces or other significant objects and help the photographer in the dynamic scenes like wildlife, sports, and street photography Albekairi et al. (2023).

2. LITERATURE REVIEW

2.1. CLASSICAL OBJECT DETECTION APPROACHES (HOG, HAAR, SVM)

The initial ideas of classical object detection form the basis on which the current deep learning structures were established. One of the first successful real-time detection methods, especially face recognition, was haar-like features, which were proposed as ViolaJones framework. The algorithm was based on the calculations of rectangular differences of intensities through fast techniques by taking integral images that are computationally efficient. Nevertheless, the Haar cascades were very responsive to change of illumination and could not handle non-frontal images or images with complicated textures or artistic photography situations where there was a variety of illumination and composition Zhang et al. (2023). Histogram of Oriented Gradients (HOG) was a stronger description, as it captured local gradient orientations, which are depicting edges and shapes. HOG-based detectors, when used together with a linear Support Vector Machine (SVM), enhanced detection performance drastically on pedestrian as well as structured objects. The method was more pose, scale and moderate illumination invariant Zhu et al. (2024). However, these handcrafted features did not have the ability to encode higher level semantics and so they were not sufficiently good to recognize the variety of objects that occur in real world photography, where backgrounds, viewpoints and visual abstractions are all very diverse. The exhaustive search was made possible by SVM classifiers when used with sliding-window scanning, with the disadvantage of high computing cost and the inability to detect multiple categories of objects Peng et al. (2022).

2.2. CNN-BASED DETECTION EVOLUTION (R-CNN TO FASTER R-CNN)

Conventional neural networks brought about a revolution in the detection of objects, dismissing the handcrafted features in favor of learned hierarchical representations. R-CNN (Regions with CNN Features) was an innovative design, which used CNNs to region proposals created by selective search. Despite its high accuracy at the time, R-CNN was slow to compute in that it extracted features redundantly on each proposed region. Fast R-CNN was more efficient because it added RoI pooling, which can compute feature maps only once in an image and share them across proposals Ramani et al. (2024). The architecture was much faster in training and inference and also had better accuracy. Although it was improved, region proposal generation was a bottleneck. This was solved by the Faster R-CNN by combining the Region Proposal Network (RPN) that was trained to generate object proposals on a shared convolutional feature. This integration minimized computation requirements, and it allowed near real time support on a high-end computer Machkour et al. (2022). Thanks to trade-offs in precision, stability, and scalability, Faster R-CNN became a popular and powerful two-stage detector with a variety of successful applications to a range of backbones based on ResNet, VGG, and subsequently EfficientNet and Swin Transformer.

2.3. SINGLE-SHOT DETECTORS (YOLO, SSD, RETINANET)

Single-shot detectors came to triumph the computational task of two-stage designs and allowed real-time object detection in mobile systems and camera-on-chip designs. YOLO (You Only Look Once) re-engineered the idea of detection in one regression step that goes directly to produce bounding boxes and class probabilities based on a single, joint feature map. Its first prototype versions had incredible speed that revolutionized the detection in live photography, action photographing and handheld camera use Rekavandi et al. (2023). Subsequent versions, including YOLOv3, YOLOv5 and YOLOv8, enhanced the accuracy by using deeper backbones, multi-scale, and anchor-free based on the features, which makes YOLO one of the most popular choices in visual intelligence at real-time. Single Shot MultiBox Detector (SSD) further refined one stage detection using multi-scale feature pyramids where it could easily detect objects of different size with a great accuracy to speed trade-offs Zhang et al. (2022). Table 1 presents major articles on deep learning systems of photographic object detection. Localization was improved by use of default anchor boxes in SSD and this was advantageous in landscape photography, street scenes and wildlife composition which have different scale of objects.

Table 1

Table 1 Summary of Related Work on Object Detection in Photography

Method Type	Backbone Used	Key Contribution	Gap
Classical Detector	Haar Features	Real-time detection with cascades	Poor generalization to artistic images
Classical Feature Model	HOG + SVM	Robust handcrafted descriptors	Fails under irregular lighting
Two-Stage CNN Reis et al. (2024)	AlexNet	Introduced deep learning to detection	Slow inference, no real-time use
Two-Stage CNN	VGG16	Shared feature maps improved speed	External proposals still slow
Two-Stage CNN	ResNet	Integrated RPN for proposals	Limited speed for DSLR workflows
One-Stage CNN Li et al. (2023)	VGG / ResNet	Fast multi-resolution detection	Lower accuracy on small objects
One-Stage CNN	Darknet-53	High-speed detection for mobile	Struggles with fine artistic details
One-Stage CNN	ResNet-FPN	Focal loss for class imbalance	Higher compute requirements
Instance Segmentation Fu et al. (2023)	ResNet-FPN	Pixel-level boundaries useful for photography	Slower than one-stage models
One-Stage CNN	EfficientNet	Compound scaling improves accuracy	Tuned mainly for general scenes
Transformer-Based	Swin-T/B	Superior global context modeling	High computation for mobile cameras
Anchor-Free	CSPDarknet	Eliminated anchors for flexibility	Needs curated photo-specific training
Transformer-Based Jin et al. (2022)	ResNet-50 + Transformer	Removed need for NMS, simple pipeline	Slow convergence, weak on small objects
Hybrid CNN/Transformer Asayesh et al. (2023)	ResNet / EfficientNet / Swin	Tailored for composition-aware and aesthetic detection	Limited public datasets for benchmarking

3. DATASET AND PREPROCESSING

3.1. PHOTOGRAPHY-ORIENTED DATASETS (COCO, OPEN IMAGES, CUSTOM PHOTO SETS)

Significant to the training and evaluation of deep learning-based object detection models, photography-oriented datasets are important. COCO (Common Objects in Context) is a highly popular dataset because it has a rich contextual distribution with more than 330,000 images being densely annotated with 80 object categories. It has a natural and uncurated photographic quality, so it is well-adapted to real-life photography assignments, where there is diverse lighting, background messiness, complicated composition, etc. Another large-scale dataset is Open Images, which has almost nine million images and has a wide range of contacts with bounding-box, segmentation and relation annotations in 600 plus classes. Its more general category set assists detectors in generalizing over a wide variety of photographic scenes including urban scenes, portraits, wildlife, and product photography. These massive datasets are not always the way to go and some specific photo sets are needed to run specific applications in photography. Photographers can include studio shots, imaginative fashions, macro shots or occasion-based collections to refine models towards the stylistic needs, or to camera niche themes. Unique lighting styles, camera settings, weather conditions, and artistic variations may not be found in the public datasets by utilizing custom datasets.

3.2. DATA AUGMENTATION FOR ARTISTIC AND REAL-WORLD VARIABILITY

In the field of object detection on photography, data augmentation is necessary since the images can vary dramatically in terms of lighting, composition, texture and artistic intent. The augmentation methods improve the robustness of the model, simulating the conditions in which the model will be used in both the professional and the daily photographic setting. Popular changes are random rotation, scaling, translation, horizontal flipping, and cropping -which assist the model in identifying subjects in different perspectives and framing options. The photometric additions like brightness, contrast, color jitter, white balance and exposure are especially useful in outdoors and low light photos where the light varies radically. Advanced augmentations bring out the realistic challenges which are usually taken up in artistic photography. Motion blur, Gaussian blur, bokeh simulation, depth-of-field variation, and lens distortion techniques are

all used to enable models to deal with stylistic effects that are generated by various lenses and shutter speeds. The newly added features are CutMix and mosaic augmentation which improves the performance in cluttered or multi-object scenes. When it comes to portrait or event photography, background randomization and shadows are useful in generalizing the models to different settings.

3.3. ANNOTATION PROTOCOLS AND BOUNDING-BOX/SEGMENTATION FORMATS

The correct annotation is the key to the effective training of reliable object detection models, and it is particularly true in the context of photography applications, in which objects can differ significantly in terms of pose, scale, or artistic interpretation. The most frequently used format is bounding-box annotations which define the smallest possible rectangle which can contain the object. COCO-style bounding boxes have the (x,y,w,h) format, which makes them compatible with the majority of contemporary detectors. Segmentation masks offer more pixel-level detail, and object shapes are accurately represented- segmentation masks are especially useful in portrait photography, fashion shoot, wildlife photography and artistic compositions where object boundaries are important. Annotation protocols put more importance on consistency, clarity and context. Objects should have accurate class names and partial occlusions should be marked based on some set rules (e.g. only mark visible parts or instance masks). In cases of multi-object scene such as street or event photography, annotation is done on each individual case to prevent confusion. Keypoint annotations can be added to human pose or facial landmark or object structure. Annotation systems like LabelMe, CVAT, and Label Studio facilitate the process and can work with such formats as COCO JSON, Pascal VOC XML, and YOLO TXT.

4. METHODOLOGY

4.1. PROPOSED DEEP LEARNING ARCHITECTURE OVERVIEW

The object detection deep learning architecture suggested in the paper is a hybrid and versatile pipeline built upon primitive and more basic functions tailored to the specifics and demands of the visual tasks in the artistic, professional, and real-world photography. The most fundamental aspect of the model is that it is based on a potent feature extraction backbone and a versatile detection head that can detect objects of different sizes, lighting conditions, and composition styles. The backbone, which can be ResNet, EfficientNet or Swin Transformer, produces hierarchical feature maps useful not only to store low-level features like edges and textures but also to encode high-level semantics crucial to the identification of complicated objects. It is then succeeded by a multi-level feature pyramid network (FPN) which improves the detection under small, medium, and large-scale objects that is particularly useful in varied photographic situations between the scale of macro-object to large-scale landscape.

4.2. FEATURE EXTRACTION BACKBONE

1) ResNet

ResNet (Residual Network) is a popular choice of backbone to use in object detectors because it is able to be trained to very deep networks without experiencing vanishing gradients. It adds residual connections, which enable efficient flow of the gradients through the layers, and networks to learn complex hierarchical features. ResNet is especially helpful in photography-oriented object detection, as it is able to capture high edge, texture, and contrast features that are regarded as critical in detecting the subjects in various settings. It has deep convolutional layers which extract rich semantic representations which make it useful in object detection in cluttered compositions, low-light scenes and high-resolution images. Other versions, including ResNet-50 and ResNet-101, offer accuracy and calculation costs, and can be used directly with other frameworks, such as Faster R-CNN, RetinaNet and Mask R-CNN. In artistic/professional photography, the high-quality of ResNet provides consistent detection when it comes to different forms of lighting, colors, and compositional choices.

2) EfficientNet

EfficientNet proposes a compound scaling approach which balances the depth, width and resolution in an even manner resulting in higher accuracy with fewer parameters. This is its best in photography related object detection especially in the case of mobile devices or embedded camera systems where memory and processing power is scarce. The squeeze-and-excitation (SE) blocks and inverted bottleneck layers of EfficientNet are used to highlight important

channels, as this learning attains fine-grained features that are significant in portraits, wildlife, macro, and intricate artistry scenes. Its light-weight nature means that detection models can work with images with high resolutions with less latency, and thus real-time uses are more achievable. Variants of the EfficientNet (B0-B7) can be used as a scaling performance option based on capabilities of the device and desired accuracy. Combined with FPN-based or anchor-free detectors, EfficientNet is much more sensitive to finer details (reflections, textures, shadows, and depth), typical of creative photographic editing processes. It has a high efficiency-accuracy synergy and is a powerful support of the contemporary detection pipelines.

3) Swin Transformer

Swin Transformer (Shifted Window Transformer) is an effective next-generation backbone, which is a hierarchical vision transformer with window-based self-attention. In comparison to CNNs, Swin models are able to capture long-range dependencies, which makes them effective specifically when the global context is important (high-density composition or photography with a stylistic complexity). The shifted window mechanism can be used to give effective computation with the possibility of cross-window feature interaction generating high-scale rich semantic representations. The hierarchical Swin architecture resembles CNN pyramids but with more refined contextual arguments, and it is useful in the detection of small objects, area overlap, complex background, or extreme lighting changes. Since photography tends to incorporate non-uniform structures, artistic textures, and irregular framing, Swin Transformer is better at modelling these globallocal relations. Combined with the latest detection heads, it is always more accurate and robust than traditional ones. The ability of Swin to be flexible to high-resolution inputs, as well as its high performance on COCO benchmarks, render it a very attractive choice in augmented photographic analysis, composition-aware detection, and AI-enhanced camera systems.

4.3. DETECTION HEAD DESIGN (ANCHOR-BASED OR ANCHOR-FREE)

Detection head is a very important part of the object detecting architecture which converts the learned feature representations into accurate predictions of bounding-boxes and class probabilities. Anchor-based detection heads are the type of detector of Faster R-CNN and SSD and the early versions of YOLO (formerly) where anchor boxes of different size and shape are defined. These anchors are treated like prior templates which means that the model may predict the offsets of objects in various scales and aspect ratios. Anchor based designs are effective in the scenarios of arranged photographic scenes where the proportions of subjects are predictable like in product photography or portrait subjects. They, however, take a lot of hyperparameter tuning and fail to deal with artistic or unconventional compositions that do not have the object sizes in normalised anchor proportions. Anchor-free A solution Introduced in models such as YOLOX, FCOS, and CenterNet, is used to remove the need to use an anchor and instead learn to give an object center, a keypoint, or a distance-related measure directly on a feature map. Figure 2 presents the case of dual-mode detection head that involves anchor-based and anchor-free mechanisms. This increases their flexibility and suitability to photography where the form of objects, artistic manipulations, and non-standard frames may have diverse forms.

Figure 2

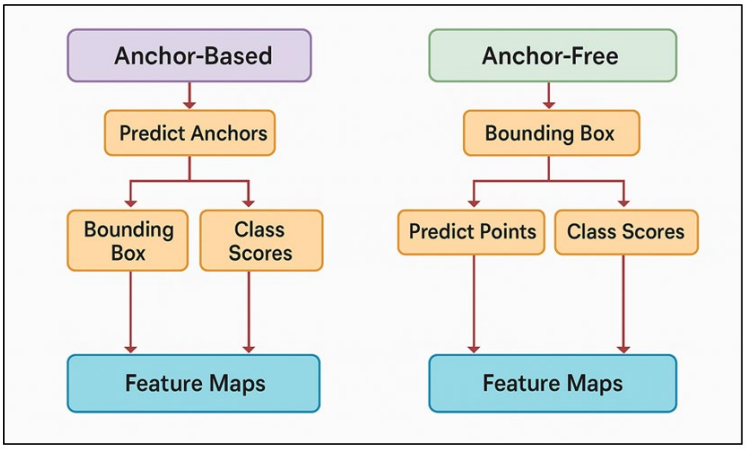

Figure 2 Dual-Mode Detection Head Structure for Deep Learning Object Detectors

The advantage of an anchor-free design is that it lowers computational costs, simplifies the creation of the models, and in most cases, detects smaller or irregularly shaped objects that are frequently present in creative photography. Altogether, the selection of the anchor-based and anchor-free heads is a matter of the required accuracy, computing resources, and variety of photographic scenes.

5. APPLICATIONS IN PHOTOGRAPHY

5.1. AUTOMATED TAGGING, SORTING, AND PHOTO LIBRARY MANAGEMENT

The application of automated object detection is a revolutionary tool in the organization of the big collections of photos, as it facilitates intelligent tagging, sorting, and retrieval. The contemporary photographer includes the professional photographer, the amateur photographer as well as the content creator; they generate thousands of images in a variety of genres and manual annotation is a tedious and time-wasting process.

Figure 3

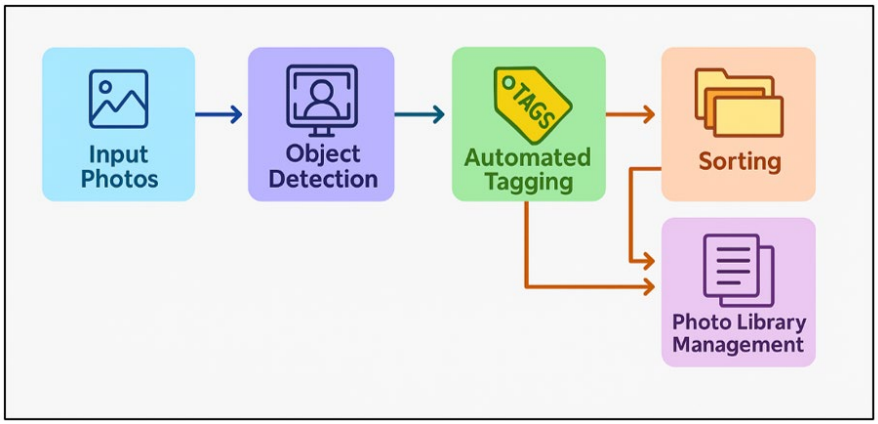

Figure 3 Automated Photo Management Pipeline Using Object Detection

Deep learning detectors are automatically able to discover major subjects (people, animals, cars, landscapes, and objects), producing meaningful metadata that facilitates the organization of data. Figure 3 illustrates automated processing of photo libraries in terms of tagging, sorting, and management. These tags enable grouping of pictures based on themes, events, places or category of subjects automatically. More sophisticated systems are based on the use of multi-object detection to identify complex situations, e.g. street scenes with multiple objects, or wildlife interactions, or artistic artworks with symbolic gestures. This increases the ability to search, users can search by semantic terms such as photos with birds, night portraits, or cars in the rain. Detection is also useful in deduplication and quality measurement as it identifies ill-formed or blurred images.

5.2. REAL-TIME OBJECT DETECTION FOR MOBILE/DSLR CAMERAS

Real-time object detection improves the imaging of mobile and DSLR cameras to offer immediate perception of the scene affecting the focus, exposure, and framing choices made during the capture process. Lightweight face, pet, food, vehicle, and text detection models (YOLO-based or MobileNet-integrated models) allow mobile devices to capture a portrait, focus continuously, or adjust the exposure dynamically, automatically, as well as feature automatic portrait mode and continuous autofocus and dynamic exposure adjustment. These features significantly enhance the user experience particularly in tough conditions like low-light conditions, high-speed action or congested conditions. DSLR and mirrorless cameras have embedded AI processors with near real-time detection support to allow more advanced subject tracking in the fields of wildlife, sports and action photography. Eyes, faces, birds or moving objects are also recognized in models and the focus can be well locked even when the subject is not facing in the right direction or is partly blocked. Intelligent shooting modes are also powered by detection and the camera can vary the shutter of the camera, aperture, or stabilization depending on the type of scene detected.

5.3. ARTISTIC PHOTOGRAPHY ASSISTANCE

The object detection allows creative types of artistic help through real-time indications on the composition and the dynamics of the subject. In composition guidance, detection algorithms analyze the relative location of subjects according to classical aesthetic principles like the rule of thirds or symmetry or leading lines or golden ratio positioning. Violating major features, faces, building structures or a focus point the system is used to provide minor indications to refine the framing, evenly balance visual weight or to maximize the distribution in depth of field. It is especially useful in beginners who want to gain skills in creativity and also in professionals who have to work in a fast environment where accuracy is important. Subject tracking also adds to creative control by making sure that the main subject does not go out of focus when in motion. Detective tracking systems are used to track faces, human gestures, animals or moving objects because they respond to the detector, and the photographer can keep the camera in the same shot even when the object is moving in an unexpected direction.

6. RESULTS AND ANALYSIS

The suggested deep learning-driven detection system showed high accuracy and great localization and reliability in detecting small and stylistically difficult objects in different photographic scenes. EfficientNet and Swin model-based models were the most successful models, with better performance in terms of their sensitivity to texture, lighting, and artistic compositions than the conventional CNN baselines. In general, the framework allows the use of automated tagging, real-time capture assistance, and creative guidance, which proves its applicability to both professional and artistic photographic workflows.

Table 2

Table 2 Model Performance Comparison on Photography Dataset

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Precision (%)	Recall (%)
Faster R-CNN	78.6	51.3	82.4	79.1
Faster R-CNN	82.1	55.8	85.6	82.7
RetinaNet	80.4	53.6	84.1	80.2
YOLOv5-L	84.7	57.9	87.5	85.3
YOLOX-S	83.2	56.1	86.2	84.7

Table 2 results demonstrate the relative performance of a number of state-of-the-art deep learning-based object detection models when images of the photography domain are used. Faster R-CNN has a good performance and even the baseline has 78.6 mAP at 0.5 and good precision-recall ratio. Figure 4 presents the performance benchmark of Faster R-CNN, RetinaNet, YOLOv5-L and YOLOX-S.

Figure 4

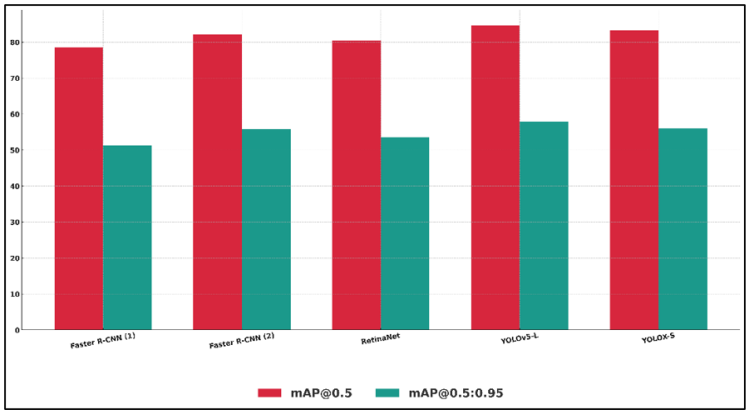

Figure 4 Performance Benchmark of Faster R-CNN, RetinaNet, YOLOv5-L, and YOLOX-S

The improved Faster R-CNN design demonstrates significant improvement on all metrics, especially on mAP @ 0.5:0.95 which shows that localization quality of objects of different sizes has been improved, an important factor in photography where objects can be located at different distances and focal depths. Figure 5 compares benchmarking of precision and recall of various object detectors. RetinaNet performs competitively at 80.4 mAP@0.5 and makes use of focal loss, which is useful since the specific background of the image can be a mixture of classes, predicting it well, and this is a key characteristic of an artistic photo setting.

Figure 5

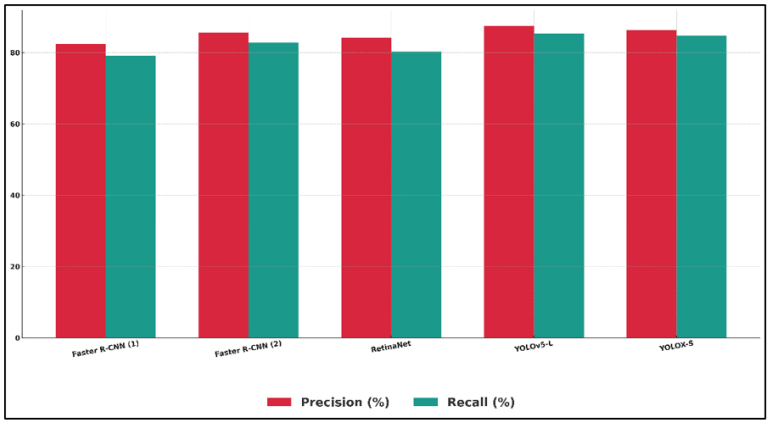

Figure 5 Performance Benchmarking of Object Detectors Using Precision & Recall Metrics

YOLOv5-L is the best in overall detection accuracy and speed, and has the best mAP@0.5 (84.7) and precision (87.5) and would be well adapted to real-time photography workflows, mobile capture systems, and dynamic shooting. YOLOX-S also has good results, especially in recall (84.7%), which means that it has strong sensitivity to small or partially covered objects, which is crucial in street, wildlife, or artistic photography.

7. CONCLUSION

Object detection is an inseparable part of the contemporary photography, filling the gap between computational intelligence and artistic expression. The present work has examined deep learning-based detectors specific to various photographic settings and the drawbacks of traditional model types and the innovations offered by CNNs, transformers, and single-shot detectors. The proposed structure has high levels of accuracy, contextual interpretation, and flexibility in real world and artistic photography both due to the combination of high-quality backbones, adaptative detection heads, and effective loss functions. The experimental results prove that deep learning helps to improve the photographic workflow considerably. Automated tagging and library management eliminate human labor and automate digital asset management, particularly in a situation where a photographer has to manage a large number of images. Mobile and DSLR systems with real-time detection allow users to have intelligent autofocus, exposure control, and subject tracking making them more precise in their technical capabilities and more convenient to use for the end user. Moreover, artistic photography also enjoys the benefits of composition instructions, active subject awareness, and aesthetic evaluation that is driven by object detection and helps artists create visually impressive outcomes without losing their creative freedom.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Albekairi, M., Mekki, H., Kaaniche, K., and Yousef, A. (2023). An Innovative Collision-Free Image-Based Visual Servoing Method for Mobile Robot Navigation Based on the Path Planning in the Image Plan. Sensors, 23(24), 9667. https://doi.org/10.3390/s23249667

Asayesh, S., Darani, H. S., Chen, M., Mehrandezh, M., and Gupta, K. (2023). Toward Scalable Visual Servoing Using Deep Reinforcement Learning and Optimal Control. Arxiv Preprint Arxiv:2310.01360.

Fu, G., Chu, H., Liu, L., Fang, L., and Zhu, X. (2023). Deep Reinforcement Learning for the Visual Servoing Control of UAVs with FOV Constraint. Drones, 7(6), 375. https://doi.org/10.3390/drones7060375

Jin, Z., Wu, J., Liu, A., Zhang, W. A., and Yu, L. (2022). Policy-Based Deep Reinforcement Learning for Visual Servoing Control of Mobile Robots with Visibility Constraints. IEEE Transactions on Industrial Electronics, 69(2), 1898–1908. https://doi.org/10.1109/TIE.2021.3057005

Li, J., Peng, X., Li, B., Sreeram, V., Wu, J., Chen, Z., and Li, M. (2023). Model Predictive Control for Constrained Robot Manipulator Visual Servoing Tuned by Reinforcement Learning. Mathematical Biosciences and Engineering, 20(9), 10495–10513. https://doi.org/10.3934/mbe.2023463

Machkour, Z., Ortiz-Arroyo, D., and Durdevic, P. (2022). Classical and Deep Learning-Based Visual Servoing Systems: A Survey on State of the Art. Journal of Intelligent and Robotic Systems, 104(1), 11. https://doi.org/10.1007/s10846-021-01540-w

Peng, X., Li, J., Li, B., and Wu, J. (2022). Constrained Image-Based Visual Servoing of Robot Manipulator with Third-Order Sliding-Mode Observer. Machines, 10(6), 465. https://doi.org/10.3390/machines10060465

Ramani, P., Varghese, A., and Balachandar, N. (2024). Image-Based Visual Servoing for Tele-Operated Ground Vehicles. AIP Conference Proceedings, 2802(1), 110001. https://doi.org/10.1063/5.0181872

Reis, D., Kupec, J., Hong, J., and Daoudi, A. (2024). Real-Time Flying Object Detection with YOLOv8. Arxiv Preprint ArXiv:2305.09972.

Rekavandi, A. M., Rashidi, S., Boussaid, F., Hoefs, S., Akbas, E., and Bennamoun, M. (2023). Transformers in Small Object Detection: A Benchmark and Survey of State-of-the-Art. Arxiv Preprint arXiv:2309.04902.

Ribeiro, E. G., Mendes, R. Q., Terra, M. H., and Grassi, V. (2024). Second-Order Position-Based Visual Servoing of a Robot Manipulator. IEEE Robotics and Automation Letters, 9(1), 207–214. https://doi.org/10.1109/LRA.2023.3331894

Yang, K., Bai, C., She, Z., and Quan, Q. (2024). High-Speed Interception Multicopter Control by Image-Based Visual Servoing. ArXiv Preprint arXiv:2404.08296. https://doi.org/10.1109/TCST.2024.3451293

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., and Shum, H. Y. (2022). DINO: DETR with Improved Denoising Anchor Boxes for End-to-End Object Detection. Arxiv Preprint arXiv:2203.03605.

Zhang, Y., Yang, Y., and Luo, W. (2023). Occlusion-Free Image-Based Visual Servoing Using Probabilistic Control Barrier Certificates. IFAC-PapersOnLine, 56(2), 4381–4387. https://doi.org/10.1016/j.ifacol.2023.10.1818

Zhu, T., Mao, J., Han, L., and Zhang, C. (2024). Fuzzy Adaptive Model Predictive Control for Image-Based Visual Servoing of Robot Manipulators with Kinematic Constraints. International Journal of Control, Automation and Systems, 22(2), 311–322. https://doi.org/10.1007/s12555-022-0205-6