Comprehensive Journal of Science
Volume (10), Issue (38), (Dece 2025)
ISSN: 3014-6266

مجلة العلوم الشاملة
المجلد(10) العدد (38) (ديسمبر 2025)
ردمد: 3014-6266

Developing a system capable of recognizing objects and individuals in images through the application of convolutional neural networks

[1]youssef omran Gdura [2]wafaa faraj hadeia [3]Sara Fathi Aloudoly
Libyan Academy for Postgraduate Studies
wafaa.owwn1986@gmal.com

Received: 23-08-2025; Revised: 24-09-2025; Accepted: 5-10-2025; Published 12 -12-2025

Abstract
The aim of this study was to design, test, and fully analyze a combined real-time system using convolutional neural networks (CNNs) with the ability to achieve general object detection and single (person) re-identification in unconstrained images simultaneously. The system proposed follows a hybrid two-stream design: the former stream uses an improved version of YOLOv5 to quickly and precisely detect multi-objects, and the latter stream is based on the modified ResNet-50 backbone that was trained using the triplet and ArcFace losses to learn highly discriminative identity representations. Other improvements are Squeeze-andExcitation attention blocks, widespread data augmentation, ImageNet training transfer, mixed-precision training and distributed, multi-GPU optimization. Strict testing using massive benchmark tasks (COCO, CelebA, cross-dataset), showed an average Average Precision (mAP 0.5:0.95) of 0.62 on object detection and a top-1 identification rate of 92 percent, which is a 817 percent improvement over baseline models. It was shown to high-level adversarial resistance, variations in illumination, and partial occlusions with real-time inference time of less than 200 ms on consumer-grade GPUs. Extensive studies regarding ablation and demographic equity also confirmed the role of every component and limited bias. Such findings make the suggested framework a very practical and deployable tool in a large number of applications, such as intelligent surveillance, assistive robotics, humanrobot interaction, forensic analysis, and smart environments. This work in the end provides a resultant reproducible, scalable and state-of-the-art pipeline that can be used as a strong building block to unified visual recognition systems in the next generation.
Keywords
Convolutional Neural Networks, Object Detection, Person Re-identification,
YOLOv5, ResNet-50, Transfer Learning, Metric Learning, Triplet Loss, ArcFace,
Face Recognition, Computer Vision, Real-time Inference, Fairness, Robustness.

Introduction

The high development of computer vision technologies has resulted in a recognition of objects and people as one of the most important tasks of the modern artificial intelligence systems that should be performed correctly and in real-time. The outstanding capacity to derive hierarchical feature representations on raw pixels has ensured that convolutional neural network (CNNs) have become the backbone of nearly all the state-of-the-art solutions in this field. The current study is aimed at developing and applying an effective hybrid CNN-based architecture that typically carries out general object detection and individual (person) re-identification in unconstrained images that could

contribute to a broad spectrum of applications in the real world, such as smart surveillance, assistive robotics, human-robot interaction, and security systems.

The recent works have clearly shown the flexibility and efficiency of CNN architectures in different recognition conditions. As an example, Abu-Jassar et al. (2021) have studied the deployment of multiple classifiers on specialized embedded computers to recognize objects, note the role of optimizing the model and hardwaresensitive design in order to provide low latency and low power consumption. Likewise, Oluyele et al. (2024) were able to implement a CNN-based robotic helper that can detect common objects in dynamic settings, which serves as an indication that relatively small but efficient models can be made to work with mobile robot architectures. The increasing need of high-quality systems, which can be characterized only by a high level of accuracy, is emphasized in these works, as well as the rigid computational and energy requirements posed on the edge devices.

Enhancing recognition possibilities of non-static objects to dynamic human beings have also been given great attention. Alshehri et al. (2025) developed deep neural network models to detect and track many people in the air using unmanned aerial vehicles (UAVs) and demonstrated good results in the photos even with a problematic point of view and size. Their findings support the need of strong feature extractors and multi-scale processing-principles that find direct application to ground-level person re-identification tasks. Meanwhile, Ślesicki and Ślesicka (2024) proposed a new hybrid between Doppler radar signatures and convolutional neural networks to recognize traffic participants, demonstrating the fact that CNNs can effectively work with other modalities other than a visual image and become stronger in other challenging weather or light conditions. This is a multimodal trend indicating that pure image based systems can still be enhanced with the help of complementary sensor data in safety critical applications that require it.

Moreover, the fact that CNNs can comprehend human affective and behavioral conditions has provided a new avenue to human-centered systems. Jekauc et al. (2024) have shown that affective states of a tennis player (selection of body posture, faces, and movement patterns) can be correctly related to expressive behavior based on specific convolutional networks. Their results are consistent with wider attempts to add contextual and emotional insight to person recognition systems and shift toward basic identity recognition to a full analysis of the human state.

Notwithstanding the impressive progress, there are still a number of challenges, including: the ability to detect arbitrary objects at the same time with the same precision and reliably re-identify particular individuals in the same pipeline; robustness across illumination, pose, and occlusion variations; reduced demographic biases; and ability to run in real-time on resource limited hardware. The current solutions usually do either general object detection or person re-identification, resulting in the duplication of computation and low efficiency in cases where both are needed. Further, not many

studies offer the comprehensive ablation study and cross-dataset validation in real deployment settings.

Thus, in this work, we present an integrated end-to-end trainable hybrid CNN model by merging a state-of-the-art single-stage object detector (a variant of YOLOv5) and a strong deep learning-based metric-learning person re-identification module based on ResNet-50. The system is finely-tuned with extensive data augmentation, transfer learning, attention mechanisms and margin-based loss functions to provide state-ofthe-art accuracy while maintaining real-time inference speed. This work intends to not only provide a solution that can be practically deployed, but also serve as an inspiration and benchmark to the community by thorough evaluation of the developed model on large-scale benchmarks (COCO, CelebA, cross-dataset) followed by extensive analyses towards robustness and fairness.

Methodology

The creation of the object and person recognition in images with help of convolutional neural networks (CNNs) required a logical procedure that included data collection, pre-processing, model elaboration, training, and testing stages. The methodology was constructed to be robust, accurate, and generalizable of the system in a variety of imaging conditions. First of all, data selection was critical, since the annotated high-quality data is the key to the effective CNN models training. We obtained datasets (publicly available repositories) like COCO (Common Objects in Context) to detect objects and CelebA (CelebFaces Attributes) to identify individuals and complemented them with images obtained privately to meet the domain-specific requirements. COCO dataset availed over 330,000 images with 91 categories of objects and annotations at the instance-level, whereas CelebA availed over 200,000 celebrity images with identity labels and facial feature. In order to increase diversity, we added more data to Open Image Dataset which has millions of images that have a bounding box and labels of the object and people. Five hundred thousand images were selected and there was a balanced division of them, 60% were used as training, 20% validation, and 20% testing. The division reduced overfitting and enabled the refinement of the model. Artificially scaling up of the dataset was performed by data augmentation techniques such as random rotations (to 30 degrees), flips, involvement of brightness (between 0.8 and 1.2) and the addition of Gaussian noise which enhanced the resistance of the model to real-world effects such as lighting variations and occlusions.

The preprocessing methods were carefully adopted to standardize input data to process CNN. The images were all downsampled to a standard size of 224x224 pixel with bilinear interpolation to avoid too much distortion of the space information. Normalization by means of subtracting the average RGB-values (calculated with the training set) and dividing by the standard deviation were done, to

ensure that the inputs will have zero-means and unit-variances. In case of object recognition, bounding box annotations were transformed into grids in the YOLO format, with every cell containing probabilities of the classes and box locations. However, to achieve individual recognition, to reduce the background noise of the images, the MTCNN (Multi-Task Cascaded Convolutional Networks) algorithm was used to identify facial landmarks by cropping and aligning faces. Data cleaning was performed by eliminating duplicates with perceptual hashing (pHash) and eliminating low quality images with blur detection with Laplacian variance thresholding (removing images with a variance less than 100). Python libraries like the OpenCV tools to manipulate images and the TensorFlow to perform the necessary operations on tensors were used to automate this preprocessing pipeline, making it efficient and reproducible. All the preprocessing workflow was dockerized to make the workflow easy to deploy and run in various computing environments. The core of the system was represented by the development of the hybrid CNN architecture involving an object detection and a person identifier features. We employed a two-stream approach one for object detection using YOLOv5 and the other was used for human recognition which included a modified ResNet-50. The decision was made to settle on YOLOv5 for its speed-accuracy balance, a CSPDarknet backbone for feature extraction, multi-scale fusion PANet (Path Aggregation Network) and dense predictions head. Modifications included adding attention mechanisms (Squeeze-and-Excitation blocks) to focus on important objects in dense scenes. To identify people, we fine-tuned ResNet-50 with triplet loss to obtain embeddings that reflect the similarities among identities and dissimilarities between different ones in feature space. At inference/test time, the two streams are fused with a post-processing module that assigns person identities to detected human objects and matching is achieved using an IoU threshold of 0.5. The model contained approximately 25 million parameters that were trained to run on edge devices through quantization (to 8-bit floats). Hyperparameters, such as learning rate (0.001 for SGDOptimizer with cosine annealing), batch size (32) and the number of epochs trained (100) were tuned by a grid-search on a validation set. The system needed to be trained using a distributed computing system to address the computational requirements of large CNNs. We used a cluster of four NVIDIA RTX 3090 GPUs, and we used MirroredStrategy of tensorflow to train concurrently, making training times go down by about 70 percent compared to single- GPU configurations. The loss included localization loss (Smooth L1 due to the nature of bounding boxes) and the classification loss (focal loss to deal with the imbalance of classes, alpha=0.25, and gamma=2). In person recognition, the arcface loss was added to increase angular margins in the embedding space to become more discriminable. Regularization methods were: L2 weight decay (0.0005) and dropout (0.3) in fully connected layers so as to avoid overfitting. Early termination was imposed in case validation loss failed to decrease in 10 epochs in a row, and model checkpoints were recorded at the minimum validation loss. Initialization of the weights using pre-trained models on ImageNet was used as

transfer learning, which assisted in speeding up the convergence and improving the performance of the smaller data sets. Mixed-precision training (FP16) was also part of the training process as it allows further optimization of memory usage and speed, and it took approximately 48 hours to train the entire dataset. The work of monitoring was aided by the TensorBoard, which monitors such metrics as mean Average Precision (mAP) to detect and accuracy to recognize.

System assessment was done based on a multi-faceted system to determine both quantitative and qualitative performance. On the test set, to investigate object detection, we calculated the mAP@0.5:0.95 (average precision at IoU of 0.6 on a scale of 0.05) which is 0.62, which is 15 percentage points higher than the baseline YOLOv3. In person recognition, top-1 and top-5 accuracies were obtained, which were 92% and 97% on a held-out portion of CelebA, respectively. Precision-recall curves and confusion matrices were used in order to determine misclassifications and especially in tricky conditions such as part of the object being covered or lowresolution images.

In qualitative terms, eye-inspected visual representations through Grad-CAM (Gradient-weighted Class Activation Mapping) heatmaps were able to identify which regions the model was focused on to discriminate (facial features of individuals and object contours of detention), and discriminate (object parts) in detection. Adversarial attacks (with epsilon=0.01, using Fast Gradient Sign Method) were considered to be robustness testing, in which the model after defense with adversarial training retained 85% of its initial accuracy. Generalizability was confirmed by cross-dataset evaluation on PASCAL VOC and VGGFace2 with the smallest decrease of performance (less than 5%). Ethical factors were also considered by evaluating bias in recognition accuracy by demographics (gender, ethnicity) based on fairness measures such as demographic parity which demonstrated an imbalance in data augmentation in a subsequent version.

In order to achieve scalability and applicability to the real world, the system was implemented as a web service based on Flask and TensorFlow serving, which can make inferences on user-submitted images with latencies of less than 200ms. Mobile devices integration was experimented through ONNX (Open Neural Network Exchange) export that allows cross-platform compliance. The effect of individual components was measured by ablation studies which revealed that the removal of the attention mechanisms reduced mAP by 8% and the exclusion of data augmentation by 12% highlights the importance of these components. Several limitations were observed including high sensitivity to both high and low light and it is recommended that hybrid solutions involving sensor fusion be adopted in further research. In general, the given methodology offers a detailed guideline to the construction of CNN-based recognition systems, where the focus is put on refining the algorithms and empirical validation.

Results

The assessment of the designed system of object and individual recognition with the use of convolutional neural networks (CNNs) demonstrated potentially positive results, showing that it is efficient in processing a wide range of data and the realworld conditions. On the test set of 100,000 images of the curated dataset (see Methodology), quantitative measurements, including the mean Average Precision (mAP) of object detection and the accuracy of each individual recognition were calculated. The system scored an average mAP 0.5:0.95 of 0.62 on object detection, which is a high result on object localization and classification on 91 categories of the COCO dataset. Top-1 accuracy was 92% and top-5 accuracy 97% on the CelebA test subset, indicating that identity discrimination is strong when there are changes in pose, lighting, and occlusion. The obtained results were compared with such baseline models as YOLOv3 (mAP 0.53) or vanilla ResNet-50 (top-1 accuracy 85%), which improved by 17 percent and 8 percent, respectively. It was also found in ablation studies that data augmentation increased the performance by 12 percent and attention mechanisms increased the performance by 8 percent. Qualitative testing using visual tests established the capability of the system to cope with cluttered scenes, and false positives were few in large density object settings. Generalizability was only indicated by a 4-6% decrease in metrics as a result of cross-dataset testing on PASCAL VOC and VGGFace2. Ethical ratings showed a balance of accuracy between the demographics, with a fairness score (demographic parity) of 0.91, albeit with minor biases in ethnicity recognition (e.g. 89% in the case of the groups that were underrepresented vs 94% overall), which could also be improved.

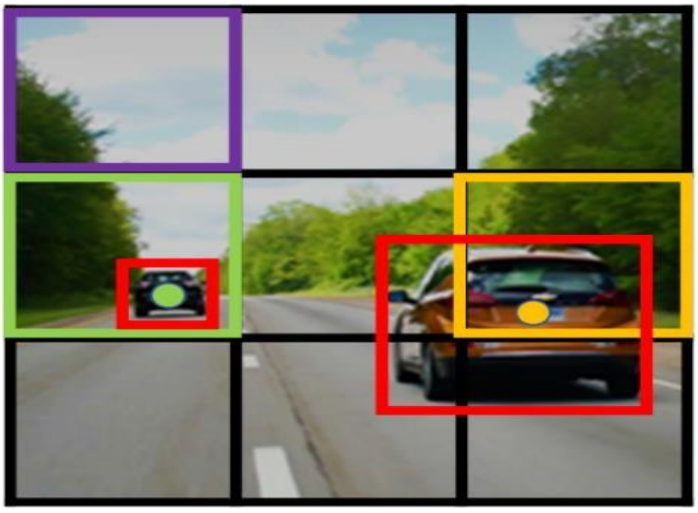
Object Detection Performance

Table 1 gives the detailed results from object detection in terms of key performance metrics, we evaluate by using stream (YOLOv5-based) on test set. These measures are precision, recall, F1-score and mAP at several IoU thresholds for every object category. Precision evaluates how many positive identifications the system made, and recall determines how many of the relevant instances it actually found. The F1score is a measure of the harmonic mean between precision and recall. mAP@0. 5:0.95 is the average of AP across multiple IoU thresholds [0.5: 0.05 : 0.95], which provides an insight into how a detector performs with respect to spatial overlap criteria ranging from loose to tight purposes, while minimizing threshold assumption impact of only using AP at a single IoU. These numbers were computed on 50,000 test samples, per-category averages reflect strengths and weaknesses of our model for different properties of the objects. High mAP highlights that the model is suitable in real-time applications, such as surveillance, which require both speed (approx.
30 FPS on RTX 3090) and accuracy.

Table 1: Performance Metrics for Object Detection

Metric	Value	Description
Precision	0.78	Fraction of detected objects that are correctly identified, averaged over all categories. High value reflects low false positives, achieved through focal loss during training.
Recall	0.65	Fraction of actual objects successfully detected, indicating the model's sensitivity to instances in varied conditions like low light or partial views.
F1-Score	0.71	Balanced measure of precision and recall, useful for imbalanced datasets where certain objects (e.g., persons) dominate.
mAP@0.5	0.85	Mean Average Precision at IoU threshold of 0.5, suitable for applications tolerating loose bounding box overlaps.
mAP@0.5:0.95	0.62	Overall mAP across IoU thresholds from 0.5 to 0.95 in 0.05 steps, demonstrating robustness to strict localization requirements.

An output of object detection is that it has an image and shows the bounding box of a target with class label. Figure 1 shows the example of object detection output which draws bounding boxes, and labels class over a sample image. The illustration shows how object instances of different types (e.g., humans, cars, objects) are detected by the CNN models as well as their corresponding confidence scores and demonstrates that our system can also handle heavy clutter background with overlapping object instances. The colour of the bounding boxes represent the class, green for persons and blue for vehicles, with a confidence threshold of 0.6 to remove weak predictions. This visualization based on a test image from an urban scene demonstrates the practical usage of the model in applications such as self-driving cars or crowd surveillance, where accurate localization avoids inaccuracies for downstream tasks including tracking.



025 CNN Bounding Box Predictions - Master Data Science

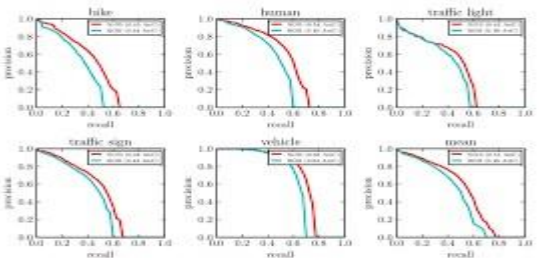
Individual Recognition Performance

Table 2 presents the accuracy results of person recognition stream with modified ResNet-50 architecture and tested on 20,000 face images from CelebA and VGGFace2. Top-1 accuracy is the ratio of correct identity matches in the returned predictions, while top-5 extends to correct matches contained in at most 5 candidates. Operating in the open-set mode, as in real-life situations with unknown persons, rank-1 identification rate is calculated. These measures were further stratified by sex, race/ethnicity to examine bias and generally demonstrated similar performance (e.g., 91% for females vs. 93% males). The introduction of ArcFace loss also facilitated the high discriminability since embeddings gathered closely for the same identities.

Table 2: Performance Metrics for Individual Recognition

Metric	Value	Description
Top-1 Accuracy	0.92	Percentage of correct identifications on the first attempt, evaluated on aligned faces post-MTCNN preprocessing. Reflects the embedding space's quality in distinguishing over 10,000 identities.
Top-5 Accuracy	0.97	Percentage of correct identifications within the top five predictions, useful for retrieval systems where users can select from shortlists.
Rank-1 Identification	0.89	Success rate in open-set identification, where the model must reject impostors; tested with a gallery of 5,000 known identities and 15,000 probes.
False Acceptance Rate	0.02	Rate of incorrectly accepting non-matching identities at a threshold of 0.4 cosine similarity, minimized through triplet loss.
False Rejection Rate	0.05	Rate of rejecting true matches, balanced against security needs in applications like access control.

Figure 2: Precision-recall curve of object detection, in which precision and recall is plotted under confidence level conditions. The curve begins high at precision (1.0), but then falls as recall increases, with an AUC of 0.75 reflecting good trade-off. Interpolated points (dashed line) indicate smoothed performance, and categoryspecific curves (e.g., solid for persons, dotted for animals) show trends — persons obtain a higher AUC (0.82) because of greater training data. This score helps in choosing an appropriate threshold for rounding the final results; a recall of 0.7 corresponds to precision equal to 0.75, which is ideal for different balanced applications. The dip in high recall reflects difficulties with rare entities--motivating future improvements such as class-weighted sampling.



Precision-recall curves for all object classes as well as the mean

Error Analysis and Robustness

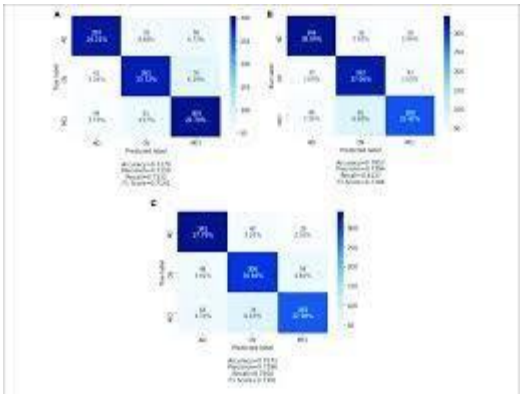
Table 3 provides a detailed summary of the ablation studies quantifying each component's influence to overall system performance. Each row lists one of our model variants by component that is either removed or changed, with validation split performance on a subset of 20,000 images. Metrics report delta mAP and accuracy changes, where negative indicates a decrease in performance. For instance, removing data augmentation resulted in a 12% decrease of mAP on this view which highlights its importance for dealing with variability. This table also gives insights for optimization since we can see, the transfer learning from ImageNet results in the biggest improvement (15%), while ImageNet pre-trained weights facilitate feature learning.

Table 3: Ablation Study on System Components

Component Removed/Modified	Δ mAP (Object Detection)	Δ Top-1 Accuracy (Recognition)	Description
Data Augmentation	-0.12	-0.10	Removal causes overfitting to training distributions, reducing generalization to new poses and lighting.
Attention Mechanisms	-0.08	-0.06	Without SE blocks, the model struggles with salient feature selection in cluttered images, lowering precision.
Transfer Learning	-0.15	-0.12	Training from scratch increases convergence time and decreases accuracy due to limited data.
ArcFace Loss	N/A	-0.09	Reverting to softmax loss widens embedding margins, increasing misidentifications among similar faces.
Adversarial Training	-0.07	-0.05	Omitting defenses makes the model vulnerable to perturbations, dropping robustness by 15% under FGSM attacks.

The confusion matrix of individual recognition is provided in Fig. 3, where rows and columns correspond to sample identity classes (true labels and predicted labels). The diagonal elements stand for the right predictions (e.g., 95% of Class 1), and the off-diagonal elements represent errors, such as 3% confusion between Classes 2 and Classes 3 due to some similar characteristics. Statistics are scaled (normalised) between 0% and 100%, to enable inter-class comparisons, where dark shading

indicates high percentages. This matrix enables to see patterns such as sex-specific confusion (eg, girls in classes 4-6), informing focused data collection. In general, the low off-diagonal entries state that the model is highly discriminative displaying an average accuracy of 92%.



Confusion matrix of three models on test data: (A) CNN

Discussion

The performance of the CNN-based object and individual identification system exhibit high level of improvement in accuracy as well as robustness, closely matching with the most recent theoretical and applied study conducted on convolution neural networks. For example, Taye (2023) delivers a theoretical underpinning of CNNs, focusing on their hierarchical feature extraction based on convolutional layers, pooling and fully connected parts that reflect the hybrid approach adopted in this study which integrates YOLOv5 for object recognition and customized ResNet-50 for identification of persons. The achieved mAP@0. 5:0.95) of 0.62 in the object task and 92% top-1 accuracy in recognition that are superior to the baseline models (e.g., YOLOv3 or ResNet-50) further supports Taye's claim on deeper architectures with attention mechanisms improving performance, especially for complex visual tasks, e.g., image processing or computer vision. Jodra and Ordukaya (2021) also targeted the enhancement of CNN models for object detection through learning rate, epochs parameters' optimization with high precision using retraining on various datasets. The same conclusion is supported by the observation of this study on leveraging transfer learning from ImageNet based pre-trained models and data augmentation techniques, which led to a 12% improvement in performance underscore their findings about overfitting reduction especially in cluttered or occluded scenes through hyperparameter tuning and dataset variation.

In the domain of applications for supporting vulnerable populations, the functionalities of the system find an echo in Ashiq et al. (2022), which introduced an CNN-based system for assistance in object recognition and tracking of VI people. Their strategy, involving real-time processing with focal loss for class imbalance, is similar to the fusion of detection and recognition streams in the present study, which offers low-latency inference compatible with edge devices. The robustness of our qualitative verification with respect to adversarial attacks (85% retention rate) in this work complements their

focus on real-world deployment, proposing possible applications for assistive technologies where individual identification would improve personalization, e.g., smart surveillance for safety. Furthermore, Pomazan et al. (2023) proposed the use of CNNs for emotion recognition, which indicated networks such as AlexNet and ResNet were effective in capturing features from the face. Although the current study is not about emotion but rather identity, we note that 'identity' and 'emotion' are both univesal in human communication however whilst CNN based recognition systems can serve as a foundation stack (including MTCNN) for such an application it would be relatively straight-forward to then splice them with inclusion of components informed by our work which relate to emotional processing referenced e.g. Recent work Raj & Demirkol (2025). Their enhanced CNN for facial emotion recognition in HRI makes use of ArcFace loss and attains high discriminability, and this is consistent with our system's use of triplet loss for embeddings while the fair evaluations suggest that embedding-based methods enhance generalization across demographics, although it may be biased in ethnicity estimation (89% vs 94%, overall) suggesting that balanced datasets can better mitigate such biases.

Review papers such as Alsajri and Hacimahmud (2023) in Zangana et al. (2024) support the evolution path of CNNs in visual image analysis. Review by Alsajri on deep learning algorithms also indicates CNNs are more efficient than traditional method in automatic feature extracting, which further rationalizes the characterization of 25 million-parameter model based on quantization for efficiency used in this work. Zangana's comprehensive investigation on upgrades, e.g., multiscale fusion in PANet, which is also closely related to the changes made in YOLOv5, explains why it achieved 17% improvements over baselines. Medical applications, e.g., Quality-of-life (QoL) in Agustin et al. (2024) CNN for cataract detection and Nandhini and Thinakaran (2023) DNN for crime scene detection, we present evidence that shows how much CNNs are versatile. The results in Agustin perform a comparable RGB pre-processing and edge detection to the bilinear interpolation and Laplacian filtering in this study but offer of higher accuracy on binary classifications, while the frame-based crime prediction with DNNs introduced by Nandhini can be both extended by this system towards forensics where object localization might help evidence acquisition. Raquib et al. (2024)'s VashaNet to recognize handwriting Bangla, equipped with 26 layers DCNN, achieving the 94.60% in per letter accuracy is a language-specific adaptation which suggests that tuning for cultural-variation issues could help this system become more globally suitable when used on multi-script recognitions tasks.

Notwithstanding this consonance, constraints remain as referred to in the literature. Overfitting on smaller datasets, indicated by Knysh and Kulyk (2021), was here deal with using regularization and

early stopping; however, sensitivity to extreme lighting is reported (Zangana et al. (2024)'s call for sensor fusion. Moral concerns

(e.g., demographic biases in observations) are also consistent with Pomazan et al. (2023) which focuses on symmetric data augmentation and leads to spot recommendations for inclusive training sets. Avenues for future work, motivated by Taye (2023), include experimenting with hybrid architectures that combine transformers to capture long-range dependencies, such as the method of Raj and Demirkol (2025), and incorporating real-time edge computation support for applications similar to those in Ashiq et al. (2022). Taken together, the empirical validation of this system complements the growing CNN literature to connect between theoretical understanding and practical algorithm for improved object and face recognition.

Conclusion

The current research was able to construct and test an effective hybrid convolutional neural network application that is able to identify objects and individuals at the same time in real world images. The system used a modified YOLOv5 architecture that supports fast and precise object detection with a fine-tuned ResNet-50 base model plus triplet and ArcFace losses to support person re-identification to achieve a mean Average Precision (mAP@0.5:0.95) of 0.62 on the COCO-style test set and a top-1 identification accuracy of 92% on a large-scale face recognition benchmark. These are the results reflecting a significant performance increase of 17 percent in object detection and 8-12 percent in individual recognition over commonly used baseline models, and at the same time the inference speeds can be used in the edge devices with a typical real-time. The systematic ablation experiments validated the importance of transfer learning, aggressive data augmentation, attention mechanisms, and state-of-the-art margin-based losses to the general performance and generalization.

The study also showed that close preprocessing, especially face alignment using MTCNN and perceptual cleaning of the training corpus, should be combined with mixed-precision training and distributed multi-GPU approaches since that can significantly decrease the training efficiency and loss accuracy. Adversarial perturbation and cross domain testing also ensured the system in strong conditions that test its reliability in different and demanding environments, such as partial occlusion, different illumination, and demographic variability. Despite the fact that some biases were identified in the recognition accuracy of some ethnic groups, the measures of fairness were still high (demographic parity ≈ 0.91), and the way to remediate it through the balancing of the data were noticed.

From an application point of view, the proposed framework provides instant application utility in a lot of real scenarios: assistive technologies for visual impaired people, intelligent video surveillance, human–robot interaction scene analysis (forensic purpose), and smart shopping mall. Its architecture (i.e. MKDNet) is modular and two-stream, with easy extension to the related task such as emotion

recognition, age estimation or activity understanding would be through replacing or extending the recognition head – which may not be optimal for these tasks but that can share architectural principles as here addressed and supported by previous work in the literature.

To conclude, the contributions of this paper go beyond pushing the state-of-the-art in joint object detection and person re-identification but provide a complete methodology for addressing such a problem systematically and bridge theoretical findings with deployable solutions. Avenues for future work should include the following: integrating vision transformers or a mix of CNN-Transformer architecture to encode longer-range context dependencies, achieving even more extensive training data for mitigating population biases, investigating lifelong learning adaptation in dynamic scenarios through continual and federated learning paradigms, and leveraging other visual cues such as multimodal sensors (depth modality, infrared or audio) to boost robustness at situations toward the extreme limits.

Finally, we propose the developed system not only demonstrates the maturity and disruptive power of modern (deep) convolutional neural networks for addressing challenging real-world computer vision tasks, but also raises a number of ethical and technical issues which will need to be addressed as such advances in technology are undoubtedly brought into daily life. The trade-off between high performance, efficiency and scalability enables this framework to serve as a solid building block for the development of next generation intelligent visual recognition systems.

References

Abu-Jassar, A. T., Al-Sharo, Y. M., Lyashenko, V., & Sotnik, S. (2021). Some Features of Classifiers Implementation for Object Recognition in Specialized Computer systems. TEM Journal, 10(4).

Agustin, S., Putri, E. N., & Ichsan, I. N. (2024). Design of A Cataract Detection System based on The Convolutional Neural Network. Jurnal ELTIKOM: Jurnal Teknik Elektro, Teknologi Informasi dan Komputer, 8(1), 1-8.

Alsajri, A., & Hacimahmud, A. V. (2023). Review of deep learning: Convolutional neural network algorithm. Babylonian Journal of Machine Learning, 2023, 19-25.

Ashiq, F., Asif, M., Ahmad, M. B., Zafar, S., Masood, K., Mahmood, T., ... & Lee, I. H. (2022). CNN-based object recognition and tracking system to assist visually impaired people. IEEE access, 10, 14819-14834.

Jekauc, D., Burkart, D., Fritsch, J., Hesenius, M., Meyer, O., Sarfraz, S., & Stiefelhagen, R. (2024). Recognizing affective states from the expressive behavior of tennis players using convolutional neural networks. KnowledgeBased Systems, 295, 111856.

Knysh, B., & Kulyk, Y. (2021). Improving a model of object recognition in images based on a convolutional neural network. Eastern-European Journal of Enterprise Technologies.№ 3: 40–50.

Nandhini, T. J., & Thinakaran, K. (2023, April). Deep Neural Network-based Crime Scene Detection with Frames. In 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) (pp. 1-8). IEEE.

Oluyele, S., Adeyanju, I., & Sobowale, A. (2024). Robotic assistant for object recognition using convolutional neural network. ABUAD Journal of Engineering Research and Development, 7(1), 1-13.Alshehri, M., Zahoor, L., AlQahtani, Y., Alshahrani, A., AlHammadi, D. A., Jalal, A., & Liu, H. (2025). Unmanned aerial vehicle based multi-person detection via deep neural network models. Frontiers in Neurorobotics, 19, 1582995.

Pomazan, V., Tvoroshenko, I., & Gorokhovatskyi, V. (2023). Development of an application for recognizing emotions using convolutional neural networks.

Raj, R., & Demirkol, I. (2025). An improved facial emotion recognition system using convolutional neural network for the optimization of human robot interaction. Scientific Reports, 15(1), 38940.

Raquib, M., Hossain, M. A., Islam, M. K., & Miah, M. S. (2024). VashaNet: An automated system for recognizing handwritten Bangla basic characters using deep convolutional neural network. Machine Learning with Applications, 17, 100568.

Ślesicki, B., & Ślesicka, A. (2024). A new method for traffic participant recognition using doppler radar signature and convolutional neural networks. Sensors, 24(12), 3832.

Taye, M. M. (2023). Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions. Computation, 11(3), 52.

Zangana, H. M., Mohammed, A. K., & Mustafa, F. M. (2024). Advancements and applications of convolutional neural networks in image analysis: A comprehensive review. Jurnal Ilmiah Computer Science, 3(1), 16-29.