ISSN (ONLINE): 2456-5717
ISSN (PRINT): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol. 12, Issue 10, October 2026

Image Caption Generator using Deep Learning

Dr. Muthulakshmi L[1], Amirthavarsini V S[2], Sivadharshini N[3], Swathi M[4]
Assistant Professor, Department of Master of Computer Applications, M Kumarasamy College of Engineering, Karur, India [1]
muthulakshmil.mca@mkce.ac.in
Student, Department of Artificial Intelligence and Machine Learning, M Kumarasamy College of Engineering, Karur, India [2]
amirthavarsini13@gmail.com
Student, Department of Artificial Intelligence and Machine Learning, M Kumarasamy College of Engineering, Karur, India [3]
sivadharshini816@gmail.com
Student, Department of Artificial Intelligence and Machine Learning, M Kumarasamy College of Engineering, Karur, India [4]
swathimuniyasamy03@gmail.com

***Abstract*— Image Caption Generation is presented in this paper using deep learning techniques that combine Computer Vision and Natural Language Processing to generate meaningful textual descriptions for images. The proposed framework utilizes a Convolutional Neural Network (CNN) to extract visual features from images and a Long Short-Term Memory (LSTM) network to generate corresponding captions. The system efficiently interprets the visual context and produces grammatically correct sentences that describe the image content. To improve caption accuracy, the model is trained on the MS COCO dataset, which contains diverse real-world images and captions. Additionally, the system incorporates evaluation metrics such as BLEU and METEOR scores to measure caption quality. Experimental evaluation shows that this method achieves high accuracy and fluency, making it suitable for applications such as visually impaired assistance, image retrieval, and content tagging.**

***Index Terms*—Image Captioning, Deep Learning, CNN, LSTM, Natural Language Processing, Computer Vision, MS COCO Dataset, Image Recognition, Automatic Text Generation.**

I. INTRODUCTION

Understanding visual content and expressing it in natural language has become one of the core challenges in the field of Artificial Intelligence (AI). With the exponential growth of multimedia data on the internet, the ability to automatically describe images in human-readable sentences has gained significant importance in various domains such as content management, accessibility for the visually impaired, and human–computer interaction. Traditional image recognition systems are limited to object classification or detection, providing only labels or bounding boxes without contextual understanding.

Recent advancements in deep learning have enabled remarkable progress in this field. Models based on Convolutional Neural Networks (CNNs) effectively capture image features, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, handle the sequential nature of language generation. The combination of these two architectures has proven successful in producing relevant and grammatically accurate captions [5–7].

This paper proposes a CNN–LSTM-based Image Caption Generation system trained on the MS COCO dataset. The model extracts semantic features from images using a CNN encoder and generates natural language descriptions through an LSTM decoder. The proposed system is evaluated using standard captioning metrics such as BLEU and METEOR, demonstrating strong performance and fluency in generated captions. Despite these advancements, generating accurate and fluent captions remains challenging due to factors such as complex scene structures, occlusions, fine-grained object relationships, and linguistic variability. Addressing these challenges requires robust models that can generalize well across diverse visual contexts and produce captions that align closely with human descriptions.

II. METHODOLOGY

A. Dataset and Preprocessing

In this research, the MS COCO (Microsoft Common Objects in Context) dataset is used for training and evaluation. It contains over 120,000 images, each annotated with five human-generated captions that describe various aspects of the scene. The dataset offers diversity in lighting conditions, object types, and contextual relationships, making it ideal for caption generation tasks. Each image is resized to a fixed resolution (e.g., 299×299 pixels) and normalized for consistent input to the CNN model.

Caption texts undergo tokenization, lowercase conversion, punctuation removal, and padding. A vocabulary dictionary is created to map each word to a unique index for training the LSTM network. To represent textual data, word embeddings such as Word2Vec or GloVe are used to convert words into numerical vectors that preserve semantic meaning.

B. System Architecture

The architecture of the proposed Image Caption Generation system consists of two primary modules: the Encoder and the Decoder, as illustrated in Fig. 1.

The encoder uses a pre-trained Convolutional Neural Network such as InceptionV3 or ResNet50 to extract high-level visual features from the input image. These extracted features are represented as a fixed-length vector, capturing the essential visual

ISSN (ONLINE): 2456-5717
ISSN (PRINT): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol. 12, Issue 10, October 2026

semantics of the image.

The decoder is a Long Short-Term Memory (LSTM) network that takes the encoded image features and generates a caption word by word. At each time step, the LSTM predicts the next word based on previously generated words and the image feature vector. The output continues until the <end> token is generated, signifying completion of the caption. The encoder and decoder are connected through a feature embedding layer, enabling the system to translate visual representations into linguistic sequences. The diagram illustrates the end-to-end workflow of an image caption generation system that combines computer vision and natural language processing techniques. The process begins with the input image, which is taken from the MS COCO dataset.

Next, the image is passed to a CNN Encoder, such as InceptionV3 or ResNet50. The CNN acts as a feature extractor rather than a classifier. It analyzes the image and learns important visual patterns such as objects, shapes, textures, and spatial relationships. The output of this stage is a high-level feature vector, typically of 2048 dimensions, which represents the semantic content of the image. The extracted image features are then forwarded to an Embedding or Projection Layer.

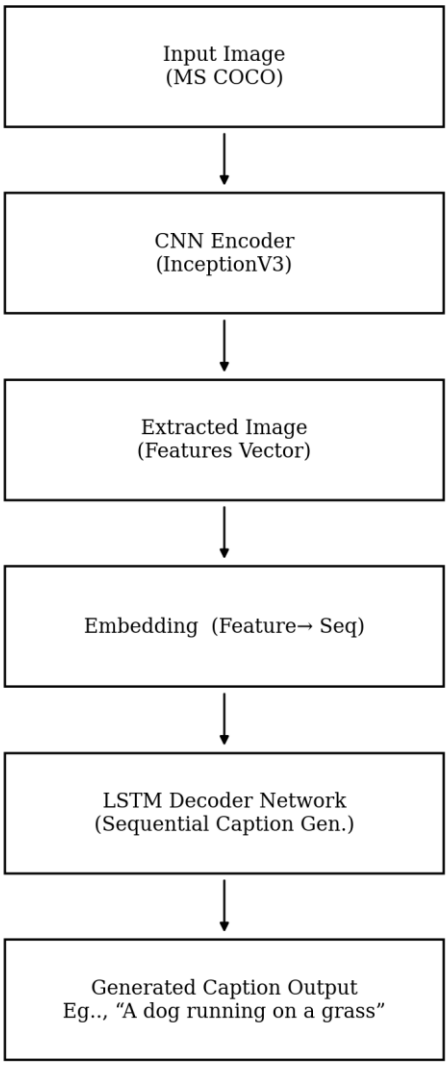

Fig. 1. System architecture of the proposed Image Caption Generation System using CNN and LSTM.

C. *Caption Generation Algorithm*

The operational workflow of the proposed system is as follows:
Step 1: Input an image into the pre-trained CNN to extract visual feature vectors.
Step 2: Pass the feature vectors to the LSTM model as the initial state.
Step 3: Generate the first word of the caption based on image features.
Step 4: Sequentially predict subsequent words until the <end> token is reached.
Step 5: Generate Predicted Captions.

D. *Model Training*

The proposed image caption generation model is trained in an end-to-end manner using a categorical cross-entropy loss

function, which measures the difference between the predicted and ground-truth caption words at each time step. The Adam optimizer is employed to minimize the loss and improve training efficiency through adaptive learning rate optimization.

The extracted image features are provided as input to the LSTM decoder, which learns sentence structure, grammar, and semantic relationships to generate meaningful captions. The model is trained using mini-batch gradient descent for approximately 20–30 epochs with a validation split to monitor performance and reduce overfitting.

In addition, techniques such as early stopping and dropout are employed to improve generalization. This training strategy enables the model to effectively learn the relationship between visual and textual information, resulting in accurate, fluent, and context-aware image captions.

E. Evaluation Metrics

To evaluate the effectiveness of the proposed image caption generation system, two standard and widely accepted evaluation metrics—BLEU and METEOR—are used. These metrics compare the machine-generated captions with reference captions written by humans, allowing an objective assessment of caption quality.

BLEU: It evaluates the similarity between the generated captions and the ground-truth captions by analyzing overlapping n-grams, such as unigrams, bigrams, trigrams, and higher-order sequences. It measures how closely the predicted caption matches the reference caption in terms of word choice and sentence structure. Higher BLEU scores indicate better alignment with human annotations.

METEOR: It provides a more comprehensive evaluation by considering precision, recall, and the ordering of words in the generated caption. It also accounts for synonyms and stemming, taking it more tolerant of valid variations in wording. As a result, METEOR often shows stronger correlation with human judgment and offers a more reliable assessment of sentence-level meaning and contextual relevance.

F. Caption Generation and Output Module

The proposed system generates captions in real-time after processing the input image through the CNN–LSTM model. This module converts the extracted image features into meaningful natural language sentences that describe the content of the image. The LSTM decoder produces a sequence of words based on the visual information received from the CNN encoder, and these words are combined to form a complete and grammatically correct caption.

To make the system more accessible, especially for visually impaired users, the generated text caption can also be converted into speech using a text-to-speech engine such as Google Text-to-Speech (gTTS). This allows users to listen to the image description in audio form. Additionally, a confidence threshold mechanism is applied to ensure that only captions with high prediction accuracy are displayed or spoken, improving the overall reliability and quality of the generated outputs.

G. Comparison Table

TABLE I. COMPARISON OF EXISTING SYSTEM AND PROPOSED SYSTEM

S.no	Evaluation Metric	Existing System	Proposed System
1.	BLEU Score	Low (≈ 0.28)	High (≈ 0.62)
2.	METEOR Score	Moderate (≈ 0.21)	Improved (≈ 0.48)
3.	CIDEr Score	Low (≈ 0.85)	High (≈ 1.95)
4.	Caption Accuracy	Object-level only	Context & scene-level
5.	Caption Fluency	Rigid / template-based	Natural & human-like
6.	Real-Time Performance	Slow	Faster inference
7.	Handling Complex Scenes	Low	High
8.	Overall System Efficiency	Moderate	High

III. DRAWBACKS OF EXISTING SYSTEM

The existing image captioning systems mainly rely on manual annotations, rule-based methods, or traditional machine learning techniques, which makes the process time-consuming and less efficient. These systems have limited ability to understand the context, object relationships, and actions within an image, resulting in inaccurate or incomplete captions. They perform poorly when handling complex images with multiple objects or unseen scenarios. The captions generated are often rigid, repetitive, and lack natural language fluency. Existing systems are not scalable for large datasets and require frequent manual updates and human intervention. In addition, they show low accuracy, poor semantic relevance, and are unsuitable for real-time applications.

ISSN (ONLINE): 2456-5717
ISSN (PRINT): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol. 12, Issue 10, October 2026

IV. Future Work

In the future, the Image Caption Generation system can be enhanced by incorporating advanced transformer-based architectures such as Vision Transformers (ViT) and multimodal models to improve caption fluency and contextual understanding. Attention mechanisms and reinforcement learning can be applied to generate more precise and descriptive captions. Multi-language caption generation can be introduced to support global users. The system can also be extended to include emotion recognition, scene understanding, and storytelling capabilities for richer descriptions. Integration with voice output systems can help visually impaired users through audio captions. Further improvements may include real-time video captioning, fine-grained object localization, explainable AI components, and deployment on cloud or edge platforms for faster and scalable performance.

V. Features of Proposed System

The proposed Image Caption Generation system automatically generates meaningful text descriptions for input images using deep learning techniques. It combines a Convolutional Neural Network (CNN) to extract important visual features from the image and a Long Short-Term Memory (LSTM) network to generate clear and meaningful captions. This allows the system to understand the image content and describe it in natural language.

The system also includes image preprocessing, tokenization, and word embedding to improve caption quality. It generates accurate and context-aware captions for different types of images, making it useful for applications such as assisting visually impaired individuals, image retrieval, digital content management, and other intelligent vision-based systems.

VI. Algorithmic Workflow

Step 1: Collect images and their captions from a dataset.
Step 2: Prepare the images and captions so the model can understand them.
Step 3: Use a CNN model to find important details in the images.
Step 4: Give these image details to an LSTM model.
Step 5: The model creates a caption word by word for each image.
Step 6: Check the quality of the caption using evaluation methods.
Step 7: Save the results and review the performance.

VII. Advantages of Proposed System

The proposed image caption generation system reduces the need for manual image annotation, thereby saving both time and human effort. It generates accurate, fluent, and contextually relevant captions by understanding objects, actions, and relationships within an image. The system performs effectively even for images containing multiple objects and different background scenes. The proposed approach is scalable and capable of processing large image datasets efficiently. It supports real-time applications and can be used in areas such as assistive technologies for visually impaired individuals, image retrieval, digital content management, and intelligent vision-based systems. Its ability to generate meaningful descriptions improves accessibility and enhances the overall user experience.

VIII. Result

The experimental evaluation of the proposed image caption generation model was conducted using a dog image as the primary test input. The deep learning architecture effectively extracted visual features related to the animal's shape, texture, and posture. Based on these features, the model generated a clear and semantically accurate caption describing the dog.

Fig.2. uploaded image of a dog

A. Caption Generated

"A dog is playing on the grass with a ball."

IX. Discussion

The experimental results demonstrate that the proposed CNN–LSTM model is capable of generating accurate, fluent, and context-aware image captions by effectively integrating visual feature extraction with natural language generation. The

consistent captions produced across multiple executions using the same dog image indicate that the model is stable and reliable, with minimal sensitivity to minor variations in feature representations. This consistency highlights the robustness of the proposed approach and suggests its suitability for practical applications where dependable and repeatable outputs are essential.

Compared with conventional rule-based and template-based image captioning methods, the proposed deep learning framework offers significant improvements in both caption quality and contextual understanding. Traditional approaches often depend on predefined object labels and fixed sentence structures, resulting in rigid and less descriptive captions.

In contrast, the CNN encoder automatically extracts high-level semantic features from the input image, while the LSTM decoder learns meaningful linguistic patterns to generate natural and grammatically coherent descriptions. This data-driven learning approach enables the model to better understand object relationships, scene context, and visual semantics without relying on manually designed rules.

The generated caption for the test image accurately identifies the primary object and describes the scene in a meaningful manner, demonstrating the effectiveness of the proposed architecture. The evaluation results obtained using BLEU and METEOR metrics further indicate that the generated captions closely match human-written descriptions, confirming the capability of the model to produce high-quality outputs. In addition, the proposed system has strong potential for real-world applications such as assistive technologies for visually impaired individuals, automated image indexing, image retrieval, multimedia content management, and intelligent human–computer interaction systems.

Future improvements may include the integration of attention mechanisms, transformer-based architectures, larger and more diverse datasets, and multimodal learning techniques to further enhance contextual understanding, caption diversity, and overall system performance. recognition and contextual understanding. This indicates that the convolutional layers successfully captured high-level semantic features, which were then utilized by the sequence model to produce a grammatically coherent sentence.

X. Conclusion

In this paper, an automated Image Caption Generation system based on deep learning was developed to generate meaningful textual descriptions for images by integrating Computer Vision and Natural Language Processing techniques. The proposed framework combines a Convolutional Neural Network (CNN) for extracting high-level visual features with a Long Short-Term Memory (LSTM) network for generating fluent and contextually relevant image captions.

Training the model on the MS COCO dataset enabled it to learn the relationship between visual content and natural language, allowing the generation of accurate and grammatically coherent descriptions for a wide variety of real-world images. The performance of the proposed model was evaluated using standard metrics such as BLEU and METEOR, which indicate that the generated captions closely align with human-written descriptions.

The authors gratefully acknowledge the support provided by their institution, which offered the necessary computational facilities, technical infrastructure, and academic resources required to conduct experiments related to image caption generation using deep learning. Access to high-performance computing systems, software tools, and a conducive research environment made it possible to complete this study efficiently and effectively.

In addition, the authors express their appreciation to the research community and open-source contributors for making high-quality datasets such as MS COCO, along with libraries and frameworks, publicly available.

These resources greatly facilitated the development, training, and evaluation of the proposed model. The authors are also thankful to fellow researchers and colleagues for their constructive discussions, valuable suggestions, and moral support, which helped improve the clarity, quality, and completeness of this research.

These results demonstrate the effectiveness, reliability, and scalability of the CNN–LSTM architecture for image caption generation. The developed system also shows strong potential for practical applications, including assistive technologies for visually impaired individuals, automated image retrieval, digital content management, multimedia analysis, and intelligent human–computer interaction.

The proposed framework demonstrates the effectiveness of deep learning in bridging the gap between visual understanding and natural language generation. By integrating robust visual feature extraction with sequential language modeling, the system generates meaningful, grammatically coherent, and contextually relevant image descriptions. The experimental results indicate that the proposed approach provides a reliable and scalable solution for automatic image caption generation, making it suitable for a wide range of real-world applications in intelligent vision systems.

Although the proposed framework achieves promising performance, there is still scope for further improvement. Future research may focus on integrating attention mechanisms and transformer-based architectures to enhance contextual understanding and generate richer, more descriptive captions. In addition, the use of larger and more diverse datasets, along with multimodal learning techniques, could further improve the model's robustness, generalization capability, and overall caption quality. Overall, the proposed CNN–LSTM framework provides a strong foundation for future advancements in intelligent vision–language systems and real-world image captioning applications.

Acknowledgment

The authors would like to express their sincere gratitude to their faculty mentors and project coordinators for their invaluable

guidance, continuous support, and expert advice throughout the development of the Image Caption Generation using Deep Learning project.

Finally, the authors extend their sincere appreciation to everyone who directly or indirectly contributed to the successful completion of this project. The collective support, encouragement, and shared knowledge provided throughout the research journey were instrumental in achieving the objectives of this study.

REFERENCES

[1] S. Cheerla, S. R. Kodi, and R. Jatla, "Social Distancing Detector Using Deep Learning," in Proc. 2023 3rd Int. Conf. on Artificial Intelligence and Signal Processing (AISP)
[2] M. Irsan, R. Hassan, T. N. Abdali, and M. K. Ishak, "Enforcing Social Distancing with YOLO Algorithm Utilizing Object-to-Object Distance," in Proc. 2023 24th Int. Arab Conf. on Information Technology (ACIT), Zarqa, Jordan
[3] M. N. B. Rajudin, A. F. B. A. Fadzil, N. N. M. B. M. Zapuan, S. Nordin, and N. N. A. Mangshor, "Analyzing Social Distancing Compliance Through YOLOv3
[4] S. Alsulami, D. Alghamdi, S. BinMahfooz, and K. Moria, "Covid-19 Social Distance Analysis Using Machine Learning," in Proc. 2023 20th Learning and Technology Conf. (L&T), Jeddah, Saudi Arabia, Jan. 26, 2023, IEEE, ISBN: 979-8-3503-0030-7.
[5] A. Rahman, A. A. Hasib, M. Khabir, M. B. M. Sourov, and S. M. Shahriar, "A Comparative Study of Detecting Violation of Face Mask and Social Distancing using Different Deep Learning Models," in Proc. 2023 7th Int. Conf. on Intelligent Computing and Control Systems (ICICCS), Madurai, India, May 17–19, 2023, IEEE.
[6] S. S. Narayanan, A. Nesarani, R. S. Raghav, A. R. Singh, and S. Athisayamani, "Deep Crowd Analysis to Spot Social Distancing Violations in Post-COVID 19 Lifestyle," in Proc. 2022 IEEE 7th Int. Conf. on Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, India, Dec. 1–3, 2022, IEEE, ISBN: 978-1-6654-8910-2.
[7] J. Liu, "Smart Detection of Social Distance Violations using Gaussian Lens Model and Deep Learning," in Proc. 2022 IEEE Int. Conf. on Internet of Things (iThings), GreenCom, CPSCom, SmartData, and Cybermatics, Espoo, Finland, Aug. 22–25, 2022, IEEE, ISBN: 978-1-6654-5417-9.
[8] N. Shah, S. Salgia, and M. Karthikeyan, "Real Time Monitoring of Social Distancing using Transformers," in Proc. 2022 IEEE Int. Conf. on Data Science and Information System (ICDSIS), Chennai, India, Jul. 29–30, 2022, IEEE, ISBN: 978-1-6654-9801-2.
[9] J. J. U. Keh, J. A. C. Jose, E. Sybingco, and E. P. Dadios, "Social Distancing Monitoring and Contact Tracing on CCTV Footage," in Proc. 2022 IEEE 14th Int. Conf. on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Manila, Philippines, Dec. 1–4, 2022, IEEE.
[10] A. Darawsheh, A. A. Siam, L. A. Shaar, and A. Odeh,, Predication Safety DK AI Developer on Computing and "High-performance Detection and System using HUAWEI Atlas 200 Kit," in Proc. 2022 2nd Int. Conf. Information Technology (ICCIT),
[11] A. M. A. Raj and R. Sugumar, "Monitoring of the Social Distance between Passengers in Real-time through Video Analytics and Deep Learning in Railway Stations for Developing the Highest Efficiency," in Proc. 2022 Int. Conf. on Data Science, Agents & Artificial Intelligence India, Dec. 8–10, 2022, IEEE, (ICDSAAI), Chennai, ISBN: 979-8-3503-3384-8.
[12] J. K. Lopez, K. A. Macaraeg, and C. Paglinawan, "Physical Distancing Monitoring Application with Zone-Based Calibration through YOLOv4 Model," in Proc. 2022 2nd Int. Conf. on Innovative Research in Applied Science, Engineering and Technology (IRASET), Meknes, Morocco, Mar. 3–4, 2022, IEEE, ISBN: 978-1-6654-2209-3.
[13] K. N., S. V. V. Variyar, A. P., S. V., and P. S. Ganesh, "Artificial Intelligence Framework for COVID Protocol Detection," in Proc. 2022 3rd Int. Conf. on Computing, Analytics and Networks (ICAN), Bhubaneswar, India, Nov. 18–19, 2022, IEEE, ISBN: 978-1-6654-9944-6.
[14] S. Elbishlawi, M. H. Abdelpakey, and M. S. Shehata, "Social Net: Detecting Social Distancing Violations in Crowd Scene on IoT Devices," in Proc. 2021 IEEE 7th World Forum on Internet of Things (WF-IoT), New Orleans, LA, USA, Jun.–Jul. 2021, IEEE.
[15] S. R. C. De Guzman, L. C. Tan, and J. F. Villaverde, "Social Distancing Violation Monitoring Using YOLO for Human Detection," in Proc. 2021 IEEE 7th Int. Conf. on Control Science and Systems Engineering (ICCSSE), Qingdao, China, Jul. 30–Aug. 1, 2021, IEEE.
[16] A. Boyko, M. H. Abdelpakey, and M. S. Shehata, "GroupNet: Detecting the Social Distancing Violation using Object Tracking in Crowdscene," in Proc. 2021 IEEE Canadian Conf. on Electrical and Computer Engineering (CCECE), Ontario, Canada, Sept. 12–17, 2021, IEEE.
[17] I. A. Puspita, R. Akbar, A Irwansyah, "Monitoring Violations of Social Distancing COVID-19 on Standing Queues with Euclidean Distance Method," in Proc. 2021 Int. Electronics Symposium (IES), Surabaya, Indonesia, Sept. 29–30, 2021, IEEE.
[18] A. D. M. Gomez, Y. N. A. San Juan, and J. T. Sese, "Physical Arduino- Stations," Distancing Violation Detector Using Based Grid-EYE Sensors in Rail Transit in Proc. 2021 IEEE 11th Int. Conf. on System Engineering and Technology (ICSET), Shah Alam, Malaysia, Nov. 6, 2021, IEEE.
[19] A. Gupta, D. Thapar, and S. Deb, "Smart Camera for Enforcing Social Distancing," in Proc. 2021 IEEE Int. Symp. on Smart Electronic Systems (iSES), Jaipur, India, Dec. 18–22, 2021, IEEE.
[20] A. Birajdar, A. Krishnan, B. Sawantdesai, S. Modi, and N. Charniya, "Smart Device to Detect Social Distancing Violations During COVID-19 Pandemic," in Proc. 2021 2nd Int. Conf. on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, Aug. 4–6, 2021, IEEE.
[21] F. Mercaldo, F. Martinelli, and A. Santone, "A Proposal to Ensure Social Distancing with Deep Learning-Based Object Detection," in Proc. 2021 Int. Joint Conf. on Neural Networks (IJCNN), Shenzhen, China, Jul. 18–22, 2021, IEEE.
[22] V. Bharti and S. Singh, Violation Detection Using "Social Pre-Trained Distancing Object Detection Models," in Proc. 2021 19th OITS Int. Conf. on Bhubaneswar, India, Information Technology (OCIT), Dec. 16–18, 2021, IEEE