Automated Description Generation for Jewellery Images using Deep Learning

Department of Information Technology

SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya School of Engineering
(formerly K J Somaiya College of Engineering)

Automated Description Generation for Jewellery Images using Deep Learning

Group ID: 4

Group Members:

Tanisha Mangaonkar - 16010422200

Prachi Gandhi - 16010422233

Chandana Galgali - 16010422234

Mahek Thakkar - 16010422235

Guide Details:

Name - Avani Sakhapara

Department - Information Technology

Need for the project

- Manual jewelry identification is time-consuming and error-prone.
- Existing systems only offer basic object detection without detailed understanding like metal colour, gemstone, etc.
- Difficulties in maintaining accurate, searchable digital catalogs.

Problem Definition

Develop an automated system that detects necklaces and earrings from a person's image, generates captions describing their metal colour, gemstones and classifies them based on presence of gemstones, metal colour, etc.

Background Work / Literature Survey

- Jewelry Detection and Identification
- General Object Detection and Tracking
- Gemstone and Material Classification
- Image Captioning

Background Work / Literature Survey

Jewelry Detection and Identification

Year	What is the paper about? (Aspects)	Methodology (Steps in 2-3 lines)	Datasets (Size, Type, etc.)	Results (Validation Metrics)	Advantages	Limitations	Reference No.
2024	Uses image captioning for jewelry classification in e-commerce.	Used VGG-16 and MobileNet as CNN encoders, GRU and LSTM as RNN decoders to generate multi-level jewelry captions, followed by classification using parsed caption attributes.	Created a comprehensive jewelry image database from local jewelry stores.	Achieved high accuracy using VGG-16 + GRU with strong F1-scores.	Accurate and robust jewelry detection with end-to-end automated recognition, handling diverse styles and occlusions effectively.	Performance drops on similar-looking items like bracelets, relies on high-quality labeled datasets, and is limited to predefined jewelry categories.	[1]
2023	Applies CNNs and Faster R-CNN for accurate image/video detection.	Used Faster R-CNN with a CNN backbone to extract features from Yakshagana images, applied region proposal networks (RPN) to locate small jewel regions, followed by ROI pooling and classification layers for precise jewel detection.	No detailed dataset info given; tested on public jewelry item datasets (implied).	Achieved high mAP and precision using Faster R-CNN for small jewel detection.	Faster R-CNN offers high precision, excels at detecting small jewelry items in complex images, and leverages deep learning adaptability for reliability.	Computationally intensive, needs large datasets, and is limited to Yakshagana-specific images.	[2]
2021	Automates jewelry tagging via transfer learning and live feeds.	Created an image repository of jewelry, labeled images by category, trained a transfer learning model, and used OpenCV for real-time classification via live camera feed.	Used a manually created image repository of jewelry articles for training and validation.	Achieved accurate, real-time jewelry recognition with strong validation metrics.	Automates real-time jewelry tagging and classification using transfer learning, reducing manual errors, manpower, and enabling accurate live camera-based recognition.	Requires a large, well-labeled image repository, with performance heavily influenced by image quality and lighting, and may struggle with unseen jewelry types without retraining.	[3]

Background Work / Literature Survey

Jewelry Detection and Identification

Year	What is the paper about? (Aspects)	Methodology (Steps in 2-3 lines)	Datasets (Size, Type, etc.)	Results (Validation Metrics)	Advantages	Limitations	Reference No.
2025	Proposes neural network for automatic jewelry description generation.	Used computer vision and various image captioning architectures, especially encoder-decoder models, trained on a comprehensive jewelry image database.	Built a large jewelry image dataset to train image captioning models.	Achieved high captioning accuracy with detailed descriptions of diverse jewelry.	Assists non-experts with detailed jewelry insights, generating accurate, hierarchical captions across varied styles.	Depends on image database quality, struggles with rare designs, and requires intensive training for captioning models.	[4]
2024	Develops vision method to track gold necklaces for theft prevention.	Improved traditional Gaussian Mixture Model (GMM) with adaptive background subtraction (ABS) for enhanced detection and tracking of gold necklaces.	Used video/image sequences from gold shops for detection and tracking evaluation.	Achieved high frame tracking accuracy, outperforming the standard GMM method.	Tracks small, deformable objects effectively with ABS-enhanced accuracy, aiding theft prevention in gold shops.	Limited to shop settings, sensitive to occlusion and fast movement, and reliant on lighting and camera quality.	[5]
2023	Presents mobile-friendly FC-YOLOv4 for fashion item detection.	Developed and compared a custom FC-YOLOv4 model with YOLOv3 and YOLOv4 using a dataset of 13,689 images across 10 categories, evaluated on mobile devices.	Dataset of 13,689 images covering five fashion and five accessories categories.	Achieved high mAP and IoU with reduced size and mobile efficiency.	Achieves extremely high accuracy (99.84% mAP), optimized for low-RAM mobile devices, and ensures faster detection for efficient e-commerce product categorization.	Limited to predefined categories, needs large labeled datasets, and performance varies across smartphone hardware.	[6]

Background Work / Literature Survey

Jewelry Detection and Identification

Year	What is the paper about? (Aspects)	Methodology (Steps in 2-3 lines)	Datasets (Size, Type, etc.)	Results (Validation Metrics)	Advantages	Limitations	Reference No.
2023	Proposes jewelry retrieval using local HSV color histograms.	Extracted feature vectors from five local regions in HSV space, applied a classification module, and matched similarity scores for jewelry retrieval.	Used publicly available jewelry item retrieval datasets: ringFIR and Fashion Product Images.	Outperformed baselines on ringFIR and Fashion datasets in retrieval accuracy.	Effective in handling occlusion and shape deformation Lightweight and color-focused feature extraction Performs well on real-world jewelry datasets	Limited to HSV color space features May struggle with grayscale or low-color-contrast images Less robust compared to deep learning-based retrieval methods	[7]

SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya School of Engineering
(formerly K J Somaiya College of Engineering)

Background Work / Literature Survey

Somaiya TRUST

General Object Detection and Tracking

Year	What is the paper about? (Aspects)	Methodology (Steps in 2-3 lines)	Datasets (Size, Type, etc.)	Results (Validation Metrics)	Advantages	Limitations	Reference No.
2023	Compares YOLOv5–v7; YOLOv6 excels.	Created a custom jewelry dataset, applied data augmentation, and trained multiple YOLO versions to compare their small object detection performance.	Used a custom dataset of jewelry images captured from a jewelry store with data augmentation.	YOLOv6 outperformed YOLOv5/YOLOv7 in accuracy, F1, recall, and mAP.	Targets small object detection using a real-world jewelry dataset, comparing multiple YOLO versions and highlighting YOLOv6's superior performance.	Dataset covers only three jewelry classes, limiting generalization, and lacks full exploration of real-time deployment challenges.	[8]
2023	Proposes CNN-YOLOv7 for jewelry in smart stores.	Uses a CNN-based YOLOv7 model trained on a custom jewelry dataset for accurate detection and localization of small jewelry objects in smart store surveillance.	Used a unique dataset curated specifically for smart store surveillance focused on jewelry.	Achieved strong metrics on custom data using YOLOv7 for lightweight surveillance.	Designed for detecting small, intricate objects in surveillance, this lightweight model delivers high accuracy and real-time efficiency on custom jewelry datasets.	Primarily focused on jewelry, limiting generalization; may struggle in cluttered or low-light settings and needs a specialized dataset for training.	[9]
2022	Reviews YOLO/CNN for real-time detection.	Surveyed and analyzed YOLO algorithm versions and CNN architectures for real-time object detection and feature extraction.	No original dataset; it's a review of YOLO and CNN models applied in literature.	Reported higher mAP and FPS, showing YOLO's real-time detection advantage.	Offers high accuracy and real-time speed with efficient CNN-based detection, enabling broad industrial applicability.	May underperform on small or overlapping objects, needs significant computational resources, and relies on high-quality training data.	[10]

Background Work / Literature Survey

General Object Detection and Tracking

Year	What is the paper about? (Aspects)	Methodology (Steps in 2-3 lines)	Datasets (Size, Type, etc.)	Results (Validation Metrics)	Advantages	Limitations	Reference No.
2023	Proposes YOLO-based system for ring and earring detection in smart shops.	Trained and validated a YOLO-based object detector on a custom dataset of rings and earrings for real-time monitoring in smart shop surveillance systems.	Used a customized dataset containing rings and earrings images.	Achieved real-time, accurate detection with strong mAP and localization metrics.	Enables real-time jewelry monitoring in smart shops with high accuracy using a lightweight, efficient YOLO architecture.	Limited to only two jewelry classes (rings and earrings) May require retraining for different store layouts or lighting conditions	[11]
2021	Surveys DL models, datasets, and edge suitability.	Reviewed and compared deep learning-based object detection models using benchmark datasets, evaluation metrics, and backbone architectures, including lightweight models for edge deployment.	No specific dataset used; it's a survey paper reviewing existing models and benchmarks.	Compared detectors using mAP, FPS, and parameters for accuracy and efficiency.	Covers modern object detection models, benchmark datasets, and metrics, with insights on lightweight models and performance comparisons.	Lacks original experiments, relies solely on literature analysis, and may miss post-2021 advancements.	[12]
2023	Compares CNN and transformer models for object detection.	Conducted a comparative analysis of CNN and transformer architectures for object detection, focusing on design, performance, and attention mechanisms.	Literature review; no dataset used.	Provided literature-based insights without experimental metrics or validation.	Provides a thorough overview of CNN and transformer-based detectors, highlighting the shift to attention models and outlining emerging research trends.	Purely literature-based without experimental validation, lacks quantitative benchmarks, and may miss the latest transformer model developments.	[13]

Background Work / Literature Survey

Gemstone and Material Classification

Year	What is the paper about? (Aspects)	Methodology (Steps in 2-3 lines)	Datasets (Size, Type, etc.)	Results (Validation Metrics)	Advantages	Limitations	Reference No.
2023	Proposes CNN-RF model for gemstone ID using curated images.	Used CNN for feature extraction from gemstone images and integrated a Random Forest classifier for final classification, trained on a 6265-image dataset with a 70:30 split.	Dataset of 6,265 gemstone images with 70:30 train-test split.	Achieved strong classification accuracy for effective gemstone identification.	Combines strengths of deep learning and traditional ML Works well on a moderate-sized dataset Applicable to geological and mineralogical domains	Accuracy (~74.76%) leaves room for improvement Performance may degrade on unseen gemstone types Limited dataset diversity may affect generalization	[14]
2024	Uses CNN-LSTM with LIBS to classify jewelry rocks.	Applied CNN layers for feature extraction from LIBS data and LSTM layers for sequence modeling, with interpretability analysis and Lasso feature selection.	Used laser-induced breakdown spectroscopy (LIBS) data from different jewelry rock samples.	Achieved high accuracy in classifying jewelry rocks using deep learning and LIBS.	Combines spectroscopy with interpretable deep learning High accuracy in classifying diverse jewelry rock types Provides layer-wise model interpretability	Requires specialized LIBS equipment May be limited to types of rocks studied Computationally intensive due to hybrid CNN-LSTM architecture	[15]

SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya School of Engineering
(formerly K J Somaiya College of Engineering)

Background Work / Literature Survey

Somaiya TRUST

Image Captioning

Year	What is the paper about? (Aspects)	Methodology (Steps in 2-3 lines)	Datasets (Size, Type, etc.)	Results (Validation Metrics)	Advantages	Limitations	Reference No.
2023	Explores how image captions enhance multimodal datasets for vision-language training.	Used synthetic captions from image captioning models and mixed them with raw web data, evaluating different strategies on a 128M image-text dataset.	Used large-scale web-scraped image-text datasets (128M and 1.28B pairs).	Outperformed prior filters on ImageNet, 38 tasks, Flickr, and MS-COCO.	Improves dataset quality without sacrificing diversity Boosts performance across multiple benchmarks Demonstrates scalable benefits on 1.28B image-text pairs	Synthetic captions may have limitations at very large scales Standard captioning benchmarks don't predict real training utility Image curation becomes increasingly critical with dataset size	[16]

Outcomes of Background Work / Literature Survey

- Manual jewelry identification is slow and error-prone.
- Existing systems don't provide detailed descriptions like metal colour, gemstone, etc.
- Models like YOLO and BLIP can automate detection and generate meaningful captions for jewelry.
- AI can help preserve traditional jewelry by recognizing and describing cultural designs.
- Most models focus only on detection and ignore semantic understanding.

Scope - Functional Requirements

- Accept input image of person wearing necklace and/or earrings
- Detect and identify necklaces and earrings in the image using a YOLO model and a CNN classifier
- Generate captions detailing metal colour and gemstones using a BLIP model (Vision-Language Model)
- Output structured descriptions and classifications for each detected item
- Provide a user interface for image upload and result display

Scope - Non-Functional Requirements

- **Accuracy:**
  The system must accurately detect and classify jewelry items with high precision.
- **Performance:**
  The system should process images and deliver results promptly to ensure a smooth and efficient user experience.
- **Usability:**
  The interface should be simple and user-friendly for non-technical users like jewellers.

Technologies to be Used

Frontend UI

- React framework for user interface

Detection & Classification

- YOLO v5 - Jewelry detection
- OpenCV 4.7+ – Image processing
- Pillow 9.5+ – Image handling

Feature Extraction

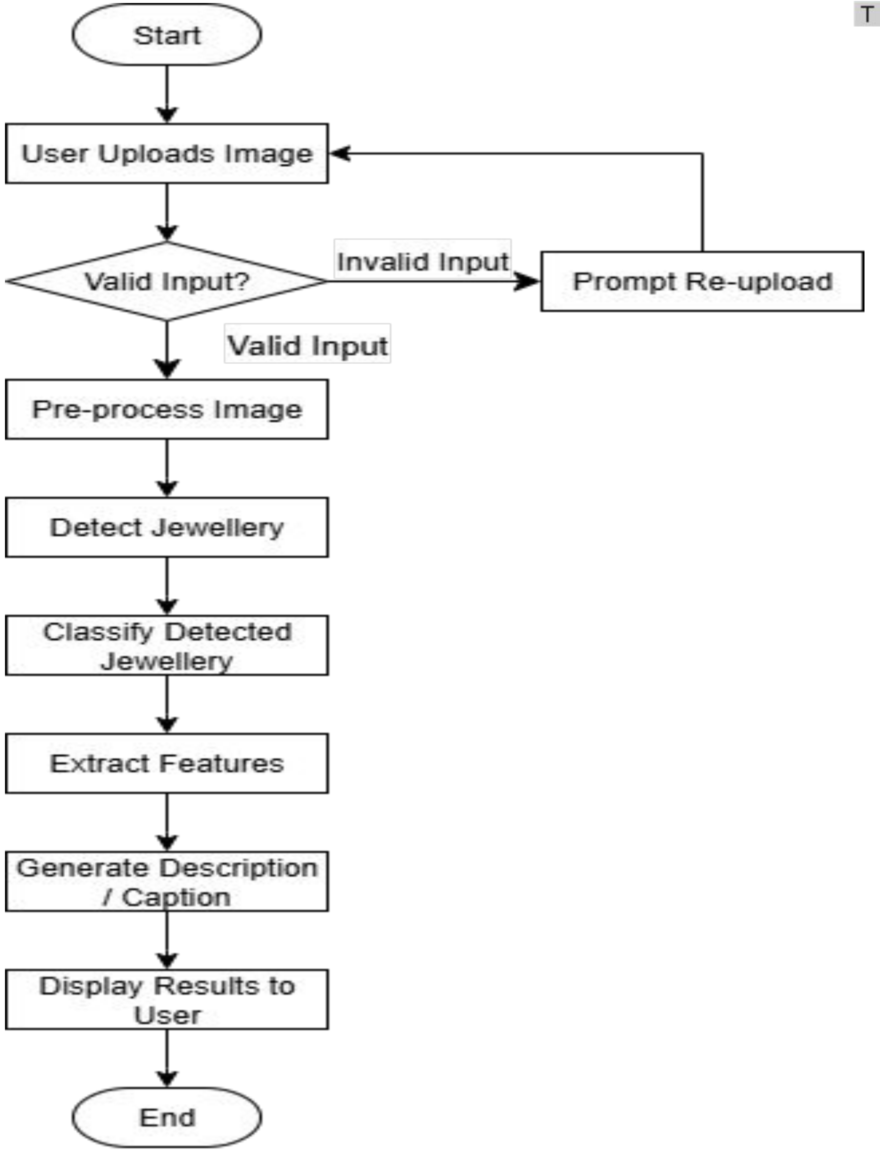
- PyTorch 2.0+ / TensorFlow 2.12+ – Model inference
- scikit-learn 1.2+ – Label encoding

Image Captioning

- Transformers 4.32+ – Pretrained LLMs
- PyTorch 2.0+ – Caption generation backend
- BLIP – Vision-language models

Overview Of Implementation

Flowchart showing the system workflow, beginning with an image uploaded by the user and moving through jewelry identification, type classification (such as necklace or earring), visual feature analysis, and automatic description generation, ending with result display and an optional retry step based on user feedback.



SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya School of Engineering
(formerly K J Somaiya College of Engineering)

Overview Of Implementation

Pre-process Image

Detect Jewellery

Classify Detected Jewellery

Preprocessing the Input Image:

- **OpenCV 4.7+** → Resizing, noise reduction, image enhancement.
- **Pillow 9.5+** → Load, crop, convert image format (e.g., PNG to RGB).

Ensures a clean and consistent image for model input.

Jewelry Detection & Classification:

- **PyTorch 2.0+** & **Torchvision 0.15+** → Deep learning framework and pretrained detection models (e.g., **YOLO v5**).
- Object detection locates jewelry regions (bounding boxes).
- **CNN classifier** distinguishes between necklace and earring.

Segments and labels jewelry for further analysis.

Overview Of Implementation

Extract Features

Generate Description / Caption

Feature Extraction:

- **PyTorch 2.0+ / TensorFlow 2.12+** → Run models to analyze jewelry features (e.g., ResNet).
- Extract features like: Metal color, Gemstone presence.
- **scikit-learn 1.2+** → Encodes and structures features for captioning.

Captures the visual and structural traits of each jewelry item.

Caption Generation:

- **BLIP** (vision-language models) → Generate descriptive text.
- **Transformers 4.32+** → Access pretrained LLMs.
- **PyTorch 2.0+** → Runs the captioning backend.
- Output: *"A round necklace with emerald green gemstones set in yellow gold."*

Combines visual understanding and language for accurate descriptions.

Implementation Schedule

Task	Timeline
Problem Statement Refinement & Literature Review	July 14 – July 31
Dataset Collection & Curation	August 1 – August 15
Image Preprocessing	August 16 – August 31
Model Design and Training	September 1 – September 30
Model Fine-tuning	October 1 – October 15
Frontend-Backend Integration	October 16 – October 31
Final Testing & Error Analysis	November 1 – November 20
Deployment on Vercel & Documentation	November 21 – December 5

SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya School of Engineering
(formerly K J Somaiya College of Engineering)

References

1. S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A Survey of Modern Deep Learning based Object Detection Models," *arXiv preprint*, arXiv:2104.11892v2, May 2021.
2. M. Mafaz, "Identification of Jewelry Article using Transfer Learning and Image Repository," *International Journal of Advance Research, Ideas and Innovations in Technology (IJARIIT)*, vol. 7, no. 4, 2021.
3. V. Viswanatha, R. K. Chandana, and A. C. Ramachandra, "Real Time Object Detection System with YOLO and CNN Models: A Review," *Journal of Xi'an University of Architecture & Technology*, vol. 12, no. 4, 2020.
4. L. Yang, "Investigation of You Only Look Once Networks for Vision-based Small Object Detection," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 4, 2023.
5. W. Ni, "Implementation of a CNN-based Object Detection Approach for Smart Surveillance Applications," *IJACSA*, vol. 14, no. 12, 2023.
6. A. Murthy, P. Devadiga, and N. P. S., "FASTER R-CNN Approach for Detecting Smaller Jewels in Yakshagana Image," *International Research Journal of Modernization in Engineering, Technology and Science (IRJMETS)*, vol. 5, no. 6, Jun. 2023.
7. W. Xu and Y. Zhai, "A YOLO-based Object Monitoring Approach for Smart Shops Surveillance System," *Journal of Optics*, vol. 53, pp. 3163–3170, 2024.
8. Y. Thwe, N. Jongsawat, and A. Tungkasthan, "Accurate Fashion and Accessories Detection for Mobile Application based on Deep Learning," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 4, pp. 4347–4356, Aug. 2023.
9. T. Nguyen, S. Oh, S. Y. Gadre, G. Ilharco, and L. Schmidt, "Improving Multimodal Datasets with Image Captioning," in *Proc. 37th Conf. Neural Information Processing Systems (NeurIPS)*, 2023.
10. S. Shah and J. Tembhurne, "Object detection using convolutional neural networks and transformer-based models: a review," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, 2023.
11. A. M. Shoib, S. Jabeen, C. Wang, and A. Tassawar, "Content-based Jewellery Item Retrieval using the Local Region-based Histograms," *arXiv preprint*, arXiv:2305.07540v1, May 2023.
12. Yashu, V. Kukreja, K. Madan, A. Singh, and D. Kumar, "GemID: A Hybrid CNN-Random Forest Approach for Accurate Gemstone Identification," in *Proc. 3rd Int. Conf. Smart Generation Computing, Communication and Networking (SMART GENCON)*, Dec. 2023, pp. 1–6.
13. J. M. Alcalde-Llergo, E. Yeguas-Bolívar, and A. Fuerte-Jurado, "Jewelry Recognition via Encoder-Decoder Models," *arXiv preprint*, arXiv:2401.08003v1, Jan. 2024.
14. A. Thanakrirkphon and S. Mruetusatorn, "Detection and Tracking Shape-Shifting Gold Necklaces Using Computer Vision Techniques," in *Proc. IEEE Conf.*, 2023.
15. P. Khalilian et al., "Jewelry rock discrimination as interpretable data using laser-induced breakdown spectroscopy and a convolutional LSTM deep learning algorithm," *Scientific Reports*, vol. 14, Art. no. 5169, 2024.
16. J. M. Alcalde-Llergo et al., "Automatic Identification and Description of Jewelry Through Computer Vision and Neural Networks for Translators and Interpreters," *Applied Sciences*, vol. 15, no. 10, Art. no. 5538, 2025.