# Semantic Enrichment for Image Description Generation via Knowledge Graph and Contextual Inference

Le Vinh Thuan[1], Nguyen Minh Khoa[2], Nguyen Vinh Thanh[3], and Nguyen Thi Dinh[4,]

[{1,2,3}]Faculty of Information Technology, University of Science, VNU-HCM, Ho Chi Minh City, Vietnam,
[4]Ho Chi Minh City University of Industry and Trade, Ho Chi Minh City, Vietnam

[1]lvthuan23@apcs.fitus.edu.vn, [2]nmkhoa23@apcs.fitus.edu.vn,
[3]23120012@student.hcmus.edu.vn, [4]dinhnt@huit.edu.vn

**Abstract.** The abstract should briefly summarize the contents of the paper in 150–250 words.

**Keywords:** First keyword · Second keyword · Another keyword.

## 1 Introduction

The automatic generation of natural language descriptions from images has emerged as one of the most challenging and important tasks in computer vision and natural language processing [1, 14]. This interdisciplinary field, commonly known as image captioning or image description, requires machines to not only recognize objects and their spatial relationships within images but also to generate coherent, contextually appropriate textual descriptions that capture the semantic essence of visual content [1, 20].

Recent advances in deep learning have significantly improved the performance of image description systems, with approaches ranging from encoder-decoder architectures [14] to attention-based mechanisms [20] and transformer-based models [8]. However, existing methods often struggle with generating rich, contextually aware descriptions that go beyond simple object enumeration and spatial relationships [5, 18].

Knowledge graphs have shown tremendous potential in enhancing various computer vision tasks by providing structured semantic information and enabling reasoning capabilities [9, 19]. The integration of knowledge graphs with image understanding systems allows for more sophisticated semantic reasoning and can bridge the gap between visual perception and high-level conceptual understanding [22, 26]. Furthermore, the incorporation of external knowledge bases enables systems to generate more informative and contextually rich descriptions by leveraging world knowledge beyond what is directly observable in the image [?, 25].

This paper presents a comprehensive four-stage pipeline for generating detailed image descriptions that combines state-of-the-art object detection with knowledge graph construction and natural language inference. Our approach employs YOLOv11 [13] for robust object detection in the first stage, followed by knowledge graph construction and entity linking. Unlike traditional approaches that rely solely on visual features, our system incorporates enriched contextual data through knowledge graph reasoning to perform sophisticated contextual analysis and inference, ultimately generating more comprehensive and meaningful image descriptions.

The main contributions of this work are: (1) A novel four-stage architecture that seamlessly integrates computer vision, knowledge representation, and natural language processing for image description; (2) An enhanced knowledge graph-based reasoning approach that enriches visual understanding with external knowledge; (3) A streamlined contextual analysis module that focuses on enriched data interpretation; and (4) Comprehensive evaluation demonstrating the effectiveness of our approach in generating high-quality, contextually aware image descriptions.

## 2 Related Work

Image captioning has evolved significantly from simple encoder-decoder architectures [14] to sophisticated systems incorporating spatial awareness and external knowledge [1]. Our work builds upon three critical research directions: depth-aware image understanding, knowledge graph construction for vision-language tasks, and fine-tuning language models for spatial-aware captioning.

### 2.1 Object Detection and Spatial Feature Extraction

Traditional image captioning systems rely primarily on 2D visual features extracted from RGB images [1,20]. The YOLO family has evolved from YOLOv1's single-stage detection [?] through multiple generations [?, 2, 16] to more sophisticated architectures. Modern object detectors like YOLOv11 provide real-time performance with improved accuracy, particularly for small objects and crowded scenes. While object detection has advanced significantly, most systems still focus on 2D bounding boxes without incorporating depth information for true 3D spatial understanding.

### 2.2 Scene Understanding and Spatial Relationships

Scene graph generation has emerged as a method to capture object relationships explicitly [22, 23]. These approaches construct structured representations of visual scenes by identifying objects and their relationships [26]. However, most scene graphs focus on semantic relationships rather than precise metric spatial relationships. Our approach aims to incorporate depth estimation to enable more accurate spatial relationship modeling.

### 2.3 Knowledge-Enhanced Image Captioning

Knowledge graphs provide structured representations of visual scenes and external world knowledge, enabling richer semantic understanding. Visual Genome [?] pioneered large-scale visual knowledge graphs with detailed annotations of objects, attributes, and relationships. Flickr30k [?] and MS COCO [?] provide image-caption pairs widely used for training vision-language models.

Several works have explored incorporating external knowledge into image captioning. Knowledge graphs have been used for visual question answering [17] and image classification [9]. Entity linking techniques [4] help map visual detections to knowledge base entities. Knowledge-based approaches [25, 27] have shown improvements by incorporating common-sense and factual knowledge.

Most existing approaches integrate knowledge at the generation stage [19]. Our pipeline differs by constructing a comprehensive knowledge graph that fuses visual observations from datasets like Visual Genome and Flickr with external knowledge sources before the generation phase.

### 2.4 Natural Language Generation for Image Captioning

Natural language generation is the final component of image captioning systems. Early encoder-decoder models [?,14] used RNNs to generate captions from visual features. Attention mechanisms [1,20] improved performance by allowing models to focus on relevant image regions during generation.

More recent approaches employ transformer-based architectures [5,8] for better long-range dependency modeling. The T5 model [10] treats all NLP tasks as text-to-text problems, offering a unified framework adaptable to various generation tasks including image captioning.

Our approach leverages text-to-text transformers for generating descriptions from structured semantic representations. By fine-tuning on data that explicitly encodes spatial relationships and depth information, we aim to generate descriptions that accurately reflect 3D spatial configurations.

### 2.5 Research Gaps and Our Contributions

Existing image captioning systems face several limitations:

1. **Limited Spatial Awareness:** Most systems [14, 20] rely on 2D visual features without depth information for true 3D spatial understanding.
2. **Knowledge Integration:** While some works integrate external knowledge [9, 25], comprehensive frameworks that systematically combine visual observations with external knowledge graphs remain limited.
3. **Spatial Description Quality:** Generated descriptions often lack precise spatial relationships and depth-aware positioning information.

Our proposed three-system pipeline addresses these gaps:

1. **System 1 - Spatial Feature Extraction:** We combine YOLOv11 for object detection with Depth Anything V2 for monocular depth estimation, enabling extraction of objects, their relationships, and depth information.
2. **System 2 - Knowledge Graph Framework:** We construct knowledge graphs based on Visual Genome [?] and Flickr [?] datasets, enriched with external knowledge from ConceptNet and DBpedia.
3. **System 3 - Depth-Aware Caption Generation:** We fine-tune language models to generate image descriptions with emphasis on spatial and depth features, producing captions that accurately reflect 3D spatial configurations.

The key contribution is the systematic integration of depth information throughout the pipeline, from detection through knowledge graph construction to caption generation, emphasizing spatial accuracy in generated descriptions.

## 3 Theoretical Basis

This section presents the theoretical foundations underlying our four-stage architecture, providing formal definitions and mathematical models that support the integration of computer vision, knowledge graph reasoning, and natural language generation.

### 3.1 Object Detection and Spatial Analysis Framework

**YOLOv11 Detection Model** The YOLOv11 object detection framework can be formalized as a function $f_{YOLO} : \mathbb{R}^{H\times W\times 3} \to \mathcal{D}$, where the input image $I \in \mathbb{R}^{H\times W\times 3}$ is mapped to a detection set $\mathcal{D} = \{d_1, d_2, ..., d_n\}$. Each detection $d_i$ is represented as:

$$d_i = (bbox_i, c_i, s_i) \tag{1}$$

where $bbox_i = (x_i, y_i, w_i, h_i)$ represents the bounding box coordinates, $c_i \in \mathcal{C}$ is the class label from the predefined class set $\mathcal{C}$, and $s_i \in [0, 1]$ is the confidence score.

**Depth Estimation and Spatial Reasoning** We incorporate the Depth-Anything-V2 model [?] to estimate depth information, formally defined as:

$$f_{depth} : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H\times W} \tag{2}$$

The depth map $D = f_{depth}(I)$ provides spatial context that enhances object relationship understanding. Combined with the Shapely geometry library [6], we compute spatial relationships between detected objects using geometric operations:

$$R_{spatial}(d_i, d_j) = \{distance(d_i, d_j), overlap(d_i, d_j), relative\_position(d_i, d_j)\} \tag{3}$$

### 3.2 Knowledge Graph Construction Theory

**Entity Linking and Knowledge Fusion** Given the detection set $\mathcal{D}$ from Stage I, we define the entity linking function as:

$$f_{link} : \mathcal{D} \times \mathcal{KB} \rightarrow \mathcal{E}_{linked} \tag{4}$$

where $\mathcal{KB}$ represents the external knowledge base and $\mathcal{E}_{linked}$ is the set of linked entities. The linking process employs semantic similarity scoring:

$$sim(d_i, e_k) = \alpha \cdot sim_{text}(c_i, label(e_k)) + \beta \cdot sim_{context}(context(d_i), context(e_k)) \tag{5}$$

where $\alpha$ and $\beta$ are weighting parameters, and $sim_{text}$ and $sim_{context}$ represent textual and contextual similarity measures respectively.

**Knowledge Graph Formalization** The constructed knowledge graph is formally defined as a directed graph $KG = (V, E, R, A)$ where $V = \{v_1, v_2, ..., v_m\}$ is the set of entity vertices, $E \subseteq V \times V$ represents the edges between entities, $R$ is the set of relation types, and $A : V \rightarrow \mathcal{P}(\mathcal{A})$ maps entities to their attribute sets.

### 3.3 T5-based Natural Language Generation Framework

**Text-to-Text Transfer Learning** Our approach leverages the T5 (Text-to-Text Transfer Transformer) model [10] for generating natural language descriptions from structured semantic representations. The T5 framework treats all NLP tasks as text-to-text problems:

$$f_{T5} : \mathcal{S}_{semantic} \rightarrow \mathcal{T}_{text} \tag{6}$$

where $\mathcal{S}_{semantic}$ represents the semantic input derived from the knowledge graph and $\mathcal{T}_{text}$ is the generated natural language output.

**Semantic Representation Encoding** The knowledge graph entities and relationships are encoded into a structured semantic representation $S_{semantic}$ that serves as input to the T5 model:

$$S_{semantic} = encode(KG, R_{spatial}, C_{context}) \tag{7}$$

where $C_{context}$ represents the contextual information derived from the enriched data analysis.

**Attention Mechanism for Semantic Focus** The T5 model employs multi-head self-attention to focus on relevant semantic elements:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

where $Q$, $K$, and $V$ represent query, key, and value matrices respectively, derived from the semantic representation.

### 3.4 Contextual Reasoning and Inference Theory

**Enriched Data Analysis** Our approach performs contextual reasoning using enriched data from the knowledge graph without explicit relationship modeling. The contextual inference function is defined as:

$$f_{inference} : \mathcal{E}_{enriched} \rightarrow \mathcal{I}_{context} \tag{9}$$

where $\mathcal{E}_{enriched}$ represents the enriched entity set and $\mathcal{I}_{context}$ denotes the inferred contextual information.

**Semantic Coherence Optimization** To ensure semantic coherence in the generated descriptions, we optimize the following objective function:

$$\mathcal{L}_{total} = \mathcal{L}_{generation} + \lambda_1 \mathcal{L}_{coherence} + \lambda_2 \mathcal{L}_{semantic} \tag{10}$$

where $\mathcal{L}_{generation}$ is the standard language modeling loss, $\mathcal{L}_{coherence}$ measures textual coherence, $\mathcal{L}_{semantic}$ ensures semantic consistency with the input knowledge graph, and $\lambda_1$, $\lambda_2$ are regularization parameters.

### 3.5 Multi-Modal Integration Theory

**Cross-Modal Alignment** The integration of visual features, spatial information, and semantic knowledge requires cross-modal alignment. We define the alignment function as:

$$f_{align} : \mathcal{F}_{visual} \times \mathcal{F}_{spatial} \times \mathcal{F}_{semantic} \rightarrow \mathcal{F}_{unified} \tag{11}$$

where $\mathcal{F}_{visual}$, $\mathcal{F}_{spatial}$, and $\mathcal{F}_{semantic}$ represent visual, spatial, and semantic feature spaces respectively, and $\mathcal{F}_{unified}$ is the unified representation space.

**Information Fusion Strategy** The multi-modal information fusion employs weighted combination of feature representations:

$$\mathcal{F}_{unified} = \sum_{i=1}^{3} w_i \cdot \phi_i(\mathcal{F}_i) \tag{12}$$

where $\phi_i$ represents the transformation function for each modality, and $w_i$ are learned weights that adapt to the relative importance of each information source.

This theoretical framework provides the mathematical foundation for our four-stage architecture, ensuring principled integration of computer vision, knowledge representation, and natural language generation components.

## 4 Proposed Approach

This section presents our comprehensive four-stage architecture for knowledge graph-enhanced image description generation. Our approach integrates advanced computer vision techniques with knowledge representation and natural language processing to generate rich, contextually aware image descriptions.

### 4.1 Overall Architecture

Figure **??** illustrates the complete pipeline of our proposed system. The architecture consists of four interconnected stages: (1) Computer Vision Processing using YOLOv11 and depth estimation, (2) Knowledge Graph Construction with entity linking, (3) Contextual Analysis and Inference using enriched data, and (4) Natural Language Generation powered by T5 transformer. Each stage contributes essential components that collectively enable sophisticated image understanding and description generation.

### 4.2 Stage I: Enhanced Computer Vision Processing

The first stage performs comprehensive visual analysis using multiple computer vision techniques to extract rich visual information from input images.

**Multi-Scale Object Detection** Our system employs YOLOv11 [13] for robust object detection, which provides real-time object detection with high accuracy, multi-scale feature extraction for objects of varying sizes, and confidence-based filtering to ensure reliable detections.

The detection process generates a comprehensive object set $O = \{o_1, o_2, ..., o_n\}$, where each object $o_i$ contains class information, spatial coordinates, and confidence scores.

**Depth-Based Spatial Understanding** We integrate Depth-Anything-V2-Base [**?**] to obtain detailed depth information, enabling 3D spatial relationship understanding between objects, distance estimation for relative positioning, and depth-aware scene composition analysis.

The depth estimation provides spatial context that enhances the understanding of object interactions and scene layout.

**Geometric Relationship Analysis** Using the Shapely geometry library [6], we compute precise geometric relationships including spatial overlaps and intersections between object regions, relative positioning (above, below, left, right, inside, outside), distance calculations for proximity analysis, and containment relationships for hierarchical object understanding.

**Optical Character Recognition** For images containing textual elements, we incorporate OCR capabilities to extract textual information from signs, labels, and documents, integrate text as additional contextual entities, and enhance semantic understanding through textual cues.

### 4.3 Stage II: Knowledge Graph Construction and Entity Linking

The second stage transforms visual detections into a structured knowledge representation that enables semantic reasoning and contextual understanding.

**Entity Enrichment and Aggregation** Detected objects are enhanced with additional semantic information including visual attributes (color, size, texture, orientation), spatial properties derived from depth analysis, contextual tags based on scene understanding, and confidence-weighted importance scoring.

**External Knowledge Base Integration** Our system performs entity linking with multiple knowledge sources: **ConceptNet** for common-sense relationships and properties, **WordNet** for semantic hierarchies and synonyms, **YAGO/DBpedia** for factual information and entity properties, and **Visual Genome** for visual relationship patterns.

The linking process employs semantic similarity matching:

$$similarity(e_{visual}, e_{kb}) = \alpha \cdot sim_{text}(label(e_{visual}), label(e_{kb})) + \beta \cdot sim_{context}(context(e_{visual}), context(e_{kb})) \tag{13}$$

**Dynamic Knowledge Graph Construction** The system constructs a scene-specific knowledge graph $KG = (V, E, R, A)$ where $V$ contains both detected visual entities and linked knowledge entities, $E$ represents relationships derived from spatial analysis and knowledge linking, $R$ includes spatial, semantic, and functional relationship types, and $A$ maps entities to their enriched attribute sets.

### 4.4 Stage III: Contextual Analysis and Enriched Data Processing

The third stage focuses on sophisticated reasoning using the enriched knowledge graph without explicit relationship modeling.

**Semantic Enrichment Strategy** Unlike traditional approaches that explicitly model all relationships, our system focuses on enriching entities with contextual information. This includes **Semantic Context** for inferring implicit meanings from entity combinations, **Functional Context** for understanding purposes and activities, **Temporal Context** for inferring time-related aspects from visual cues, and **Causal Context** for identifying cause-effect relationships.

**Multi-Level Inference Engine** Our inference engine operates at multiple abstraction levels: (1) **Object-Level Inference** for direct properties and attributes, (2) **Scene-Level Inference** for overall scene understanding and context, (3) **Activity-Level Inference** for actions and events happening in the scene, and (4) **Conceptual-Level Inference** for high-level concepts and themes.

**Context Propagation Mechanism** The system employs a context propagation algorithm that spreads semantic information through the knowledge graph, weights context based on spatial proximity and semantic similarity, resolves ambiguities through multi-source evidence combination, and maintains uncertainty estimates for probabilistic reasoning.

### 4.5 Stage IV: T5-Based Natural Language Generation

The final stage employs the T5 (Text-to-Text Transfer Transformer) model [10] to generate natural language descriptions from the enriched semantic representation.

**Semantic-to-Text Encoding** The enriched knowledge graph is converted into a structured semantic representation featuring entity-centric encoding highlighting important objects and their properties, relationship-aware structuring preserving spatial and semantic connections, context-enriched formatting including inferred information, and hierarchical organization from concrete objects to abstract concepts.

**T5 Model Adaptation** We fine-tune the T5 model specifically for our image description task with **Input Format** as structured semantic representations derived from knowledge graphs, **Output Format** as natural language descriptions with varying levels of detail, **Training Strategy** using multi-task learning with description generation and semantic consistency, and **Attention Mechanism** enhanced attention over semantic structures.

**Description Generation Logic** The text generation process follows a structured approach: (1) **Content Planning** for organizing semantic information into narrative structure, (2) **Sentence Structure Logic** for constructing grammatically correct and coherent sentences, (3) **Style Adaptation** for adjusting linguistic style based on content type and context, and (4) **Coherence Optimization** for ensuring logical flow and narrative consistency.

**Multi-Level Description Generation** Our system generates descriptions at multiple levels of detail: **Basic Level** for simple object enumeration and spatial relationships, **Detailed Level** for rich descriptions including attributes, activities, and context, **Narrative Level** for story-like descriptions with inferred activities and emotions, and **Technical Level** for precise descriptions suitable for accessibility applications.

### 4.6 Integration and Optimization

**End-to-End Training Strategy** While individual components are pre-trained separately, the system employs end-to-end fine-tuning through joint optimization of knowledge graph construction and text generation, reinforcement learning for description quality improvement, and multi-objective optimization balancing accuracy, fluency, and informativeness.

**Quality Assurance Mechanisms** The system incorporates several quality control measures including semantic consistency checking between visual content and generated text, factual accuracy verification against knowledge bases, linguistic quality assessment using automated metrics, and diversity promotion to avoid repetitive descriptions.

This comprehensive approach ensures that our system generates high-quality, contextually rich, and semantically accurate image descriptions that surpass traditional methods in both information content and linguistic quality.

## 5 Evaluation

### 5.1 Environment and Experiment Data

## 6 Future Work

While our current framework demonstrates promising results in knowledge graph-enhanced image description generation, several avenues for future research and improvement remain to be explored.

### 6.1 Scalability and Real-time Optimization

Future work will focus on optimizing the computational efficiency of our four-stage pipeline to enable real-time applications. This includes developing lightweight versions of the knowledge graph construction module for mobile and edge computing environments, implementing parallel processing techniques to reduce inference time across all four stages, particularly in the entity linking and contextual reasoning phases, exploring model compression and quantization techniques for the T5 transformer without significant performance degradation, and investigating efficient caching mechanisms for frequently accessed knowledge graph entities.

### 6.2 Enhanced Knowledge Graph Integration

We plan to expand and improve the knowledge graph component through integration of domain-specific knowledge bases (medical, scientific, cultural) for specialized image description tasks, development of dynamic knowledge graph updating mechanisms that can incorporate new entities and relationships during inference, investigation of multimodal knowledge graphs that combine visual, textual, and temporal information for richer semantic representation, and exploration of federated knowledge graph architectures to leverage multiple distributed knowledge sources.

### 6.3 Advanced Contextual Reasoning

Future research will explore more sophisticated inference mechanisms including implementation of causal reasoning models to better understand cause-effect relationships in complex scenes, development of temporal reasoning capabilities for video description generation and dynamic scene understanding, investigation of emotional and sentiment inference from visual cues to generate more nuanced descriptions, and enhancement of the context propagation mechanism with graph neural networks for better semantic diffusion.

### 6.4 Multimodal and Cross-domain Applications

We aim to extend our approach to broader applications through adaptation for video description generation by incorporating temporal dynamics and motion analysis, extension to medical image analysis for generating diagnostic descriptions with clinical knowledge integration, development of cross-lingual description generation capabilities using multilingual knowledge bases, integration with accessibility technologies for visually impaired users, including audio description generation, and application to augmented reality scenarios for real-time scene understanding and description.

### 6.5 Evaluation and Benchmarking

Future work will include comprehensive evaluation strategies involving development of new evaluation metrics that better capture semantic richness, factual accuracy, and contextual appropriateness, creation of specialized benchmarks for knowledge-enhanced image description tasks with ground truth knowledge annotations, human evaluation studies to assess the quality and usefulness of generated descriptions in real-world applications, and comparative analysis with human-generated descriptions to identify areas for improvement.

### 6.6 Robustness and Generalization

We plan to investigate the robustness of our approach through evaluation on adversarial examples and out-of-domain images to test system resilience, development of uncertainty estimation mechanisms for confidence-aware description

generation, investigation of few-shot learning capabilities for adapting to new domains with limited training data, and analysis of failure modes and development of error detection and correction mechanisms.

### 6.7 Integration with Emerging Technologies

Future directions will also explore integration with cutting-edge technologies including integration with large language models (LLMs) for enhanced natural language generation capabilities, exploration of diffusion models for generating visual explanations alongside textual descriptions, investigation of neural-symbolic approaches for more interpretable reasoning processes, and development of interactive description generation systems that can respond to user queries and preferences.

These future directions will contribute to advancing the field of knowledge-enhanced image understanding and enable the deployment of our framework in real-world applications requiring detailed, accurate, and contextually appropriate image descriptions.

## Acknowledgments

The authors would like to thank the Faculty of Information Technology, Ho Chi Minh City University of Industry and Trade, and University of Science, VNU-HCM, which are sponsors of this research. We also thank anonymous reviewers for their helpful comments on this paper.

## References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6077–6086 (2018)
2. Cao, C., Song, Y., He, J.: An overview of yolo series: From yolov1 to yolov8. Journal of Physics: Conference Series **2685**(1), 012002 (2023)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (ECCV). pp. 213–229 (2020)
4. Chen, Y., Ma, L., Huang, Y., Zhang, S., Qian, Y.: A survey of knowledge graph-based entity linking from text to knowledge graph. IEEE Access **9**, 13000–13020 (2021)
5. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10578–10587 (2020)
6. Gillies, S.: Shapely: manipulation and analysis of planar geometric objects (2007), python package, available at https://pypi.org/project/Shapely/
7. Kulkarni, M., Kulkarni, G.S., Krishna, V.S.R., Singh, S.K.: Generating image descriptions using template and semantic composition. In: European Conference on Computer Vision (ECCV). pp. 126–140 (2014)

8. Li, G., Zhu, L., Liu, P., Yang, Y.: Entangled transformer for image captioning. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8928–8937 (2019)
9. Marino, K., Salakhutdinov, R., Gupta, A.: The more you know: Using knowledge graphs for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2673–2681 (2017)
10. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020)
11. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016)
12. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **39**(6), 1137–1149 (2017)
13. Ultralytics: Yolov8: A new state-of-the-art computer vision model. https://github.com/ultralytics/ultralytics (2023), accessed: 2023
14. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3156–3164 (2015)
15. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **39**(10), 1974–1985 (2017)
16. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies for real-time object detection. arXiv preprint arXiv:2207.02696 (2022)
17. Wang, P., Wu, Q., Shen, C., van den Hengel, A.: Fvqa: Fact-based visual question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **40**(10), 2413–2427 (2018)
18. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Z.: Controllable image captioning with part-of-speech guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8689–8698 (2019)
19. Wu, Q., Shen, C., Wang, P., Dick, A., van den Hengel, A.: Image captioning and visual question answering based on attributes and external knowledge. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(6), 1367–1381 (2017)
20. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 2048–2057 (2015)
21. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning (ICML). pp. 2048–2057 (2015)
22. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: European Conference on Computer Vision (ECCV). pp. 684–699 (2018)
23. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph generation with global context. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5831–5840 (2018)
24. Zhao, S., Sun, Z., Li, X., Gong, M., Tao, D., Wei, W., Zhang, X., Li, X.: Attentional relational reasoning for image captioning. In: AAAI Conference on Artificial Intelligence (AAAI). pp. 9463–9470 (2019)

25. Zhao, W., Wu, B., Ma, S.: Knowledge enhanced fine-grained image captioning. In: ACM International Conference on Multimedia (MM). pp. 1631–1640 (2021)
26. Zhong, Y., Wang, L., Chen, J., Yu, D., Li, Y.: Comprehensive image captioning via scene graph decomposition. In: European Conference on Computer Vision (ECCV). pp. 211–229 (2020)
27. Zhu, Y., Yang, Z., Salakhutdinov, R., Xing, E.P.: Incorporating commonsense knowledge into image captioning via graph convolutional networks. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9008–9017 (2019)