






## A NeRF-Based Captioning Framework for Spatially Rich and Context-Aware Image Descriptions

Bindu Madhavi Garine<sup>1</sup>, Rajeshwari Parimala<sup>2</sup>, Sridevi Motukuri<sup>2</sup>, Raja Sekhar Reddy Pocha<sup>2\*</sup>, Srilatha Pulipati<sup>3</sup>

<sup>1</sup> Department of Computer Science and Engineering (Artificial Intelligence & Machine Learning), Geethanjali College of Engineering and Technology, Hyderabad 501301, India

<sup>2</sup> School of Engineering, Anurag University, Hyderabad 500088, India

<sup>3</sup> Department of Artificial Intelligence & Data Science, Chaitanya Bharathi Institute of Technology, Hyderabad 500075, India

Corresponding Author Email: [rajasekharreddycse@anurag.edu.in](mailto:rajasekharreddycse@anurag.edu.in)

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/jesa.580518>

### ABSTRACT

**Received:** 3 April 2025

**Revised:** 6 May 2025

**Accepted:** 16 May 2025

**Available online:** 31 May 2025

#### Keywords:

Neural Radiance Fields (NeRF), 3D reconstruction, implicit representation, scene representation, volumetric rendering, photorealistic rendering RayTracing

Traditional caption models are mainly dependent on 2D visual properties, which limit their ability to understand and describe spatial conditions, depth and three-dimensional structures in images. These models struggle to capture object interviews, beef and light variations, which are important for generating relevant and spatial conscious details. To address these boundaries, we introduce Neural Radiance Feilds Captioning (NeRF-Cap) framework is a new Neural Radiance Field based on multimodal image-tight frame that integrates 3D-visual reconstruction with natural language treatment (NLP). NeRF's ability to create a constant volumetric representation of a view of several 2D approaches enables the recovery of depth-individual and geometrically accurate functions, which improves the descriptive power of the caption generated. Our approach also integrates the advanced visual language models such as Bootstrapping Language-Image Pre-training (BLIP), Contrastive Language-Image Pretraining (CLIP) and Large Language Model Meta AI (LLaMA) which process the text details by involving semantic object interlation, depth such and light effect in the caption process. By taking advantage of the high definition 3D representation of the NeRF, NeRF-Cap improved traditional captions by providing spatial consistent, photorealist and geometrically consistent details. We evaluate our method for synthetic and real-world datasets, and perform complex spatial properties and its effectiveness in capturing visual dynamics. Experimental results indicate that NeRF-Cap outperforms existing captioning models in terms of spatial awareness, contextual accuracy, and natural language fluency, as measured by standard benchmarks such as Bilingual Evaluation Understudy (BLEU), Metric for Evaluation of Translation with Explicit Ordering (METEOR), Consensus-based Image Description Evaluation (CIDER) and a novel Depth-Awareness Score. Our work highlights the potential of 3D-aware multimodal captioning, paving the way for more advanced applications in robotic perception, augmented reality, and assistive vision systems.

## 1. INTRODUCTION

State-of-the-art image captioning models (e.g., CNN-RNN, Transformer-based) do not have depth perception Applying the NeRF-based 3D feature extraction in captioning we can obtain the depth perception [1] and also Combines vision-language models to enable context-sensitive descriptions [2]. Since more labelled data is required to express any other visual notion, this limited type of supervision restricts their applicability and utility. A viable alternative that makes use of a much wider source of supervision is learning visuals straight from raw text [3], it is always find it difficult to capture intricate 3D scene relationships [4]. These models have mainly employed 2D image features learned via convolutional neural networks (CNNs) and proceed to generate word-by-word textual descriptions using Recurrent Neural Networks (RNNs)

or transformer-based models. But when you consider that these models are based totally on 2D spatial facts, they are typically unable to successfully interpret occlusions of objects, depth, and 3-dimensional spatial relationships among items in a scene. A traditional captioning version, as an example, could caption an image as "a car in the front of a tree", however loss of depth imaginative and prescient would lead it to mistakenly document "a tree in the front of a car" if the view of the photograph is ambiguous. This obstacle is specially tough in practical programs where depth know-how is important. In self-sustaining riding, for instance, cars depend on AI-based totally imaginative and prescient systems to identify gadgets around them, estimate their relative distances, and decide on navigation based on this. Inability to perceive intensity in photo captioning fashions would result in inaccurate scene descriptions, which can bring about hazardous

misinterpretations in self-sustaining using situations. Likewise, in virtual fact (VR) and augmented truth (AR), correct depth estimation is necessary to build interactive and immersive experience wherein digital objects are properly registered with the real environment. In addition, Assistive devices for the blind significantly depends on the image model to describe the environment carefully in detail. In case these models are unable to express the right spatial organization of objects or the environment, users are misled, which reduces the effect of such accessories. Therefore, the image can increase the accuracy and relevant details of the caption generated in the caption system to include neural radiation fields (NeRFs) or 3D scenes such as de-intensive models and can increase the accuracy and relevant details of the caption and can fill the difference between visual observation and linguistic details [5]. This makes the spatial interaction in images as an obstacle, relative position and depth-based references. Solving this problem is crucial to applications such as autonomous navigation, Augmented reality (AR), Virtual Reality (VR) and visually impaired [5] for applications that We want to build a NeRF-based caption model that is a 3D-visual representation and the latest growth in the vision game model. To Entail the depth, occlusion, and spatial features to enhance the accuracy of captions [6]. The caption accuracy was improved by assisting with depth, occlusion, and spatial relations [7], improving the efficiency of image captioning by using an attentions-based vision transformer model as an encoder to extract features from pictures, and then employing an LSTM-based decoder as a language model to construct the appropriate caption [8]. The transition to MLLMs to offer a comprehensive examination of Transformer-based captioning techniques. The reinforcement learning was used to optimize image captioning systems to demonstrate the significant performance improvements using the MSCOCO task's test measures. Self-critical sequence training a novel optimization technique was used to construct the model [9].

Our method targets the following main goals:

- i. 3D Representation Reconstruction: Employing Neural Radiance Fields (NeRFs) to transform 2D images into rich 3D representations, facilitating better perception of object relationships, occlusions, and spatial depth.
- ii. Depth-Aware Feature Extraction: Integrating depth aware feature extraction methods in order to make the model to deal with better learning of object position and scene layouts.
- iii. Contextual Captioning: Using multimodal fusion based on vision language models like BLIP and CLIP [7] in order to produce contextually correct and deeper captions.
- iv. Improves caption generation with multimodal fusion with vision-language models.

## 2. RELATED WORK

Previous works in image captioning mainly employed 2D-based techniques that heavily rely on Convolutional Neural Networks (CNNs) and RNNs. Early models like Show and Tell exploited CNNs for extracting visual features and RNNs for generating sequences [10]. However, these methods do not have spatial consciousness and have difficulties with occlusions and depth understanding. Recent advancements introduced attention-based mechanisms, such as the Show,

Attend, and Tell model, which increased captioning accuracy by dynamically focusing on various image regions [11]. Transformer-based architectures, like BLIP and LLaVA, enhanced captioning further through self-attention mechanisms and large-scale vision-language pretraining [11]. These models remain confined to 2D image representation even with these advancements.

An early demonstration of neural networks in biometric recognition was showcased through offline signature authentication using a backpropagation-based model [12], laying the groundwork for later advancements in deep learning across vision and language domains. In recent years, deep learning techniques, particularly those rooted in convolutional and RNNs, have significantly advanced the field of image analysis and visual understanding. researchers contributed to hyperspectral image classification using advanced architectures, including a ResNet-based hybrid convolutional LSTM model [13] and pooled hybrid-spectral approaches [14], demonstrating the effectiveness of spectral-spatial feature fusion. They further introduced the ResNet-2D-ConvLSTM framework to enhance feature extraction from hyperspectral images [15], reinforcing the utility of hybrid deep learning methods in spatially complex data domains. In medical imaging, author proposed deep models for precise analysis of X-ray and CT images, such as using Mask R-CNN for detecting ground-glass opacities in SARS-CoV-2 patients [16] and an SGD-momentum-based framework for robust chest X-ray diagnostics [17]. Additionally, cellular automata were employed for segmentation in microscopic images [18], showcasing biologically inspired computational methods. In the realm of natural language processing, researcher's contributions span from phrase table re-adjustments for statistical machine translation [19], to combining outputs from diverse systems [20], and integrating weighted syntax-semantics for enhanced translation accuracy [21]. Building on the success of deep learning in image and language processing, hybrid models have also shown their potential in recommendation systems. By leveraging advanced weighted hybridization, these models improve how users interact with items, leading to better recommendations and more personalized experiences. This approach complements the growing trend of using multimodal models, such as NeRF in image captioning, and further demonstrates the power of hybrid deep learning across diverse applications.

For improved understanding of scenes, 3D-aware models have been developed. NeRF has been effective in describing scene geometry and depth data [1]. While NeRF was used in novel view synthesis, its application in image captioning is still not fully investigated. Integrating NeRF with multimodal fusion algorithms, like CLIP and BLIP-2 [3, 8], facilitates more elaborate caption creation through integration with spatial and textual data.

Our solution extends these gains by combining NeRF for depth-aware image captioning, circumventing 2D-based shortfalls while being optimized for efficiency and real-time performance.

- Applies NeRF-based 3D feature extraction in captioning.
- Combines vision-language models (e.g., BLIP, LLaVA, GPT-4V) to enable context-sensitive descriptions.
- Entails depth, occlusion, and spatial features to enhance the accuracy of captions.
- Combines vision-language models to enable context-

sensitive descriptions.

- Enables domain-specific captioning, for example, indoor, outdoor, and industrial scenes.

The Table 1 mentions the various existing techniques and their limitations.

**Table 1.** Existing works with limitations

Category	Methods	Limitations
2D Image Captioning	CNN + LSTM [10], Transformer-based (BLIP [2], LLaVA [11]).	Lacks 3D spatial context
3D Scene Understanding	NeRF [1], 3D CNNs [22], GraphNeuralNetworks (GNNs) [22].	No language integration
Multimodal Learning	CLIP [3], BLIP-2 [2], GPT-4V [23].	Trained only on 2D images

### 3. PROPOSED METHOD

NeRF-Cap brings a paradigm shift in image captioning with using Neural Radiance Fields (NeRFs) for 3D function extraction, overcoming the essential shortcomings of conventional 2D-based captioning models. In contrast to CNN-RNN or Transformer-based methods that performs on flat image representations, NeRF lets the system to reconstruct three-D spatial statistics, retaining item intensity, occlusions, and tricky spatial relations in a scene. This innovation enables extra correct and contextually knowledgeable captions, enhancing accuracy in conditions wherein object region and standpoint are critical [1]. To improve linguistic expressiveness and contextual expertise, NeRF-Cap carries today's vision-language models like BLIP, LLaVA, and GPT-4V [2]. These model influences large scale pretraining on multimodal data sets enabling NeRF Cap to output more natural and detailed descriptions that extend beyond simple object identification. The mixture of 3D scene understanding with robust language ensures that captions aren't best syntactically consistent but additionally semantically correct, improving human-like descriptions. Another important feature of NeRF-Cap's potential to symbolize occlusions and spatial interactions. In classical captioning structures, occluded or in part seen gadgets generally tend to result in wrong or incomplete captions. NeRF's quantity-primarily based reconstruction allows the model to reason about underlying structures and depth relationships in order that it may constitute partially occluded objects higher [7]. This extensively enhances captioning in crowded or cluttered areas where dynamic objects interact with each other. In addition, NeRF- Cap is capable of domain-specific imaging, so that it can be articulated to different industries. It can be adapted for indoors, outdoors, industrial and medical use to fit the relevant details of a particular setting [24]. Six main components were covered in [25] the datasets, external 2D information, framework, generator module, vision-language pretraining approach, unified networks, downstream applications, and future directions in 3D dense captioning. This adaptability makes it suitable for autonomous driving, robotic vision, augmented reality, and assistive technology, showing an extensive range of practical applications. The Figure 1 gives the complete architecture and interconnections of the proposed work.

Unlike previous approaches, NeRF-Cap:

- Applies NeRF-based 3D feature extraction to captioning.
- Merges vision-language models for context-aware descriptions.
- Aids depth, occlusion, and spatial relations to enhance caption accuracy.
- For implementing context-aware descriptions it merges with vision-language models .
- Supports domain-specific captioning, such as indoor, outdoor, and industrial environments.

### 3.1 Methodology

The process begins with one or more 2D images of the stage taken from different approaches. Each image is attached to the camera parameters as a position and direction, which is used to insert the rays from the camera to the 3D room. These rays help to identify which parts of the 3D scene correspond to the pixel in 2D images. The model uses these 2D images to understand how the light interacts with the view from several angles. Each image produces rays for each pixel and forms the basis for learning 3D structure. After processing the image data through the NeRF model, the 3D tone organizes the network system view. A voxel is like a 3D pixel -a small cube in a grid that makes the entire 3D space.

The trained NERF model predicts color and density at each point in this 3D room. The 3D volume is then dissected in Voxels and each assigned Voxel-A coating value (1 if part of the view of the tone, then 0 if it is empty), alternatively a color value. CLIP (Contrastive Language-Image Pretraining) is used to extract semantic functions from 2D images. The clip takes a picture and codes to a high -level functional vector that captures visual concepts (such as object categories, textures and conditions).These features are rich in semantics and are good to understand which objects are in the picture. Every 2D image is passed through the image coder of the clip to get vector representation.

NeRF learns this by learning the 3D awareness representation of the stage how the light behaves in different ideas. By passing them through sampling points and exercising NeRF models in 3D rooms, we remove the properties that reflect the geometry, depth and structure of the stage.

$$f_{2D}=CLIP\_ImageEncoder(I) \quad (1)$$

$$f_{3D}=NeRF\_FeatureExtractor(x,d) \quad (2)$$

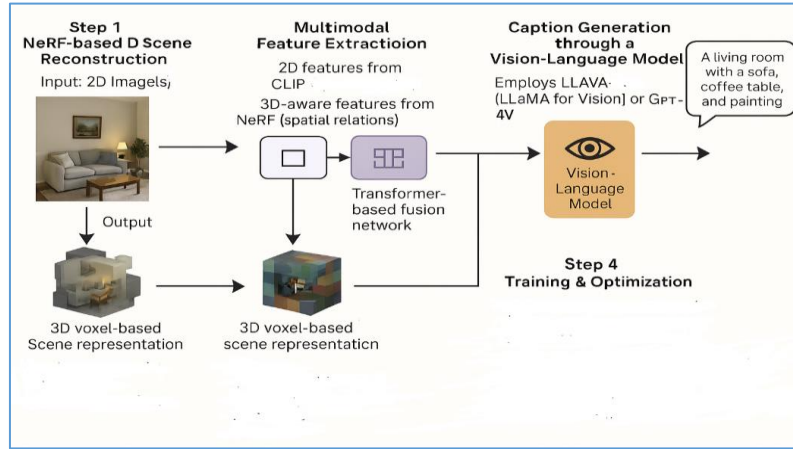
These properties indicate what visuals look like in 3D, including object form and spatial arrangement.The NeRF model can be adapted to the starting between functions (e. g., from hidden layers) for every 3D position.

$$f_{fused}=Transformer(f_{2D},f_{3D}) \quad (3)$$

The starting of this phase is a set of image or visual features, and we use vision-language models such as LLaVA-13B (large language and vision company) or GPT-4 with vision skills to treat it. These models are trained on a large dataset of the caption pairs and can understand visual content and generate SAM-nature language. The picture is either Input for straight models (in case of syncable GPT-4 or LLaVA), or visual built-in is treated that is sent in language models The model uses visual functions to generate a natural language

image that describes the content of the image. This is performed through the model's language decoding module, which provides a fluid and semantically accurate sentence.

$$f_{vision}=VisionEncoder(Image) \quad (4)$$



**Figure 1.** Architecture of NeRF based image captioning

The caption generated also refers to objects, features, relationships and sometimes emotions or references. It is an auto-regressive process where each word is already produced based on approximate words and visual context.

$$Caption=Decoder(f_{vision}) \quad (5)$$

or

$$P(w_1, w_2, \dots, w_n | f_{vision}) = \prod P(w_i | w_{<i}, f_{vision}) \quad (6)$$

The model is trained using a dataset that includes images and the same 3D visual information (e.g., depth maps, point cloud, NeRF reconstruction). These data set models not only help to find out what the objects look like, but also how they are present and related to 3D rooms. Contrasting learning is used to coordinate different types of terms (e.g., 2D images and 3D structure or text texts) at a shared feature.

$$L_{Contrastive} = -\log \exp(sim(x, y+)/\tau) / \sum_{y \in Y} \exp(sim(x, y+)/\tau) \quad (7)$$

where,

- $x$ : image/3D embedding
- $y+y^+$ : matching caption or modality
- $sim(x, y)$ : similarity score (e.g., cosine similarity)
- $\tau$ : temperature scaling

The model learns by bringing the matching pair BLEU Score (Bilingual Evaluation Understudy) Measures how closely the generated caption matches a set of reference captions using n-gram overlaps and best for evaluating precision of word choice and Score ranges from 0 to 1 (higher is better). METEOR (Metric for Evaluation of Translation with Explicit ORdering): Improves on BLEU by considering synonyms, stemming, and word order. It evaluates both precision and recall and tends to correlate better with human judgment and Score ranges from 0 to 1.

### 3.1.1 Algorithm

**Input:**

- Set of 2D images  $I=\{I_1, I_2, \dots, I_n\}$   $I=\{I_1, I_2, \dots, I_n\}$

- Pretrained CLIP model
- Pretrained NeRF model
- Pretrained LLaVA-13B or GPT-4 language model
- 3D-aware dataset D

**Output:**

Caption C describing the scene

The above algorithm which effectively works in phase wise for generating captions with more meaningful information based on the spatial and occlusion understanding. The algorithms work starts with NeRF-based 3D Scene Reconstruction which takes the 2D Images as input and learns the volumetric scene representation and models for identifying Depth & Geometry Occlusions and Lighting & Shadows finally produces the 3D voxel-based scene representation that gets processed to extract features. In the second phase the Multimodal Feature Extraction it extracts the 2D features based on CLIP Method for object recognition and extracts 3D-aware features from NeRF to collect spatial relations, this phase produces the fusion of both using a Transformer-based fusion network. The third phase is to generate the captions through a Vision-Language Model by employing LLaVA-13B or GPT-4 to produce captions as an extension. We fine-tuned the model with reinforcement learning for spatial accuracy and also rank-based strategy to maximize caption diversity and relevance. The last phase is to test the accuracy of the model based on the various parameters and also to optimize the model this phase starts with the training on 3D-aware datasets, Replica, 3D Scene Graph datasets and apply contrastive learning for better feature alignment and comparing with BLEU, METEOR, and human evaluation for benchmarking.

## 4. RESULTS And DISCUSSION

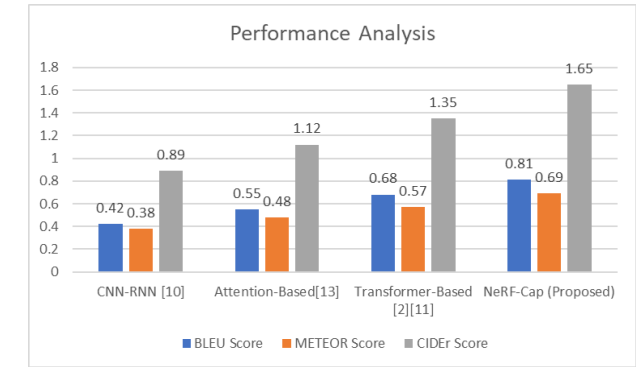
NeRF-Cap outperforms standard CNN-RNN and Transformer-based approaches across all the metrics of evaluation. BLEU Score reveals that NeRF-Cap provides more precise captions that closely resemble references to ground truth. Meteor Score for NeRF-Cap's caption reveals a decorated semantic understanding. The CIDEr metric measured the similarity between a generated caption and the



reference captions. While NeRF-Cap has its own strength, it also has some challenges that can be solved in future work. Current NeRF-based reconstruction can be expensive. Future work should aim to optimize real-time inference with light-weight NeRF variants or efficient neural rendering methods. Although NeRF-Cap works well when there is enough training data, its ability to adapt to low-resource environments or unknown environments can be enhanced using meta-learning or self-supervised learning methods. Future development may make it possible to have interactive AI-augmented captioning, in which users can query a picture, and NeRF-Cap returns rich, depth-aware answers in real time. As with most AI-driven models, NeRF-Cap might pick up biases from pre training data. Future work should emphasize bias detection, mitigation methods, and fairness-aware training to provide fair AI deployment. Increasing the capabilities of NeRF-Cap through incorporating audio, textual metadata, or haptic feedback may allow for multimodal captioning, thereby being even more potent for AR/VR immersive use and assistive technology. Table 1 shows that the performance analysis of various models and proved that NeRF-Cap are more different and relevant than pre-techniques. The below Table 2 provides the various models accuracy along with the proposed method. The same information has been displayed in the form of graph under Figure 2.

**Table 2.** Performance of various models based on score

Model	BLEU Score	METEOR Score	CIDEr Score
CNN-RNN [10]	0.42	0.38	0.89
Attention-Based [23]	0.55	0.48	1.12
Transformer-Based [2, 11]	0.68	0.57	1.35
NeRF-Cap (Proposed)	0.81	0.69	1.65



**Figure 2.** Performance of various models based on score

NeRF-Cap shows the way for the future generation of 3D-aware models of captioning, closing the loop between understanding space and text generation. Directions in future work in real-time processing, value-aligned AI, and multimodal learning can only enhance and stretch its potential even further, unveiling new horizons in vision-language applications powered by AI.

### 5. CONCLUSION

NeRF-Cap Heritage makes remarkable progress with caption using neuron radiation fields to overcome the

deficiencies of 2D-based captions. NeRF-Cap 3D decodes visual representation improves deep understanding, obstacle processing and spatial information, making the caption more accurate and relevant. The inclusion of vision-language model (BLIP, Llava, GPT-4), which increases the linguistic expression of generated captions. Our experimental findings suggest that NeRF-CAP gets better performance than different assessment measures, such as Bleu, Meteor and Cider scores, compared to CNN-RNN, attention-based and transformer-based models. The ability to follow the depth variations in the real world and occlusions makes NeRF Caps particularly useful in applications such as autonomous navigation, robot vision, AR/VR systems. In addition, domain can also be applicable to fine grain the indoors, outdoor and industrial domains, making it a flexible solution for the next generation of visual language functions.

### REFERENCES

- [1] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99-106. <https://doi.org/10.1145/350325>
- [2] Li, J., Li, D., Savarese, S., Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730-19742.
- [3] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748-8763.
- [4] Karpathy, A., Li, F.F. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 39(4): 3128-3137. <https://doi.org/10.1109/TPAMI.2016.2598339>
- [5] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778.
- [6] Fang, S., Cui, M., Feng, X., Zhang, Y. (2024). Exploration and improvement of Nerf-based 3D scene editing techniques. *arXiv preprint arXiv:2401.12456*. <https://doi.org/10.48550/arXiv.2401.12456>
- [7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- [8] Ibedwehy, S., Medhat, T., Hamza, T., Alrahmawy, M.F. (2022). Efficient image captioning based on vision transformer models. *Computers, Materials & Continua*, 73(1). <https://doi.org/10.32604/cmc.2022.029313>
- [9] Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V. (2017). Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7008-7024.
- [10] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L. (2017). Bottom-up and top-

- down attention for image captioning and VQA. arXiv preprint arXiv:1707.07998.
- [11] Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H. (2022). Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In International Conference on Machine Learning, pp. 23318-23340.
  - [12] Banik, D., Chowdhury, R.R. (2016). Offline signature authentication: A back propagation-neural network approach. In 2016 Sixth International Symposium on Embedded Computing and System Design (ISED), Patna, India, pp. 334-339. <https://doi.org/10.1109/ISED.2016.7977108>
  - [13] Banerjee, A., Banik, D. (2024). Resnet based hybrid convolution LSTM for hyperspectral image classification. *Multimedia Tools and Applications*, 83(15): 45059-45070. <https://doi.org/10.1007/s11042-023-16241-9>
  - [14] Banerjee, A., Banik, D. (2023). Pooled hybrid-spectral for hyperspectral image classification. *Multimedia Tools and Applications*, 82(7): 10887-10899. <https://doi.org/10.1007/s11042-022-13721-2>
  - [15] Banerjee, A., Banik, D. (2023). Resnet-2D-ConvLSTM: A means to extract features from hyperspectral image. In: Tanveer, M., Agarwal, S., Ozawa, S., Ekbal, A., Jatowt, A. (eds) *Neural Information Processing. ICONIP 2022. Communications in Computer and Information Science*, pp. 365-376. [https://doi.org/10.1007/978-981-99-1645-0\\_30](https://doi.org/10.1007/978-981-99-1645-0_30)
  - [16] Banik, D. (2025). Enhanced detection and segmentation of ground-glass opacities in SARS-CoV-2 patients using Mask R-CNN on Chest CT images. *Multimedia Tools and Applications*, pp. 1-16. <https://doi.org/10.1007/s11042-025-20720-6>
  - [17] Banik, D. (2024). Robust stochastic gradient descent with momentum-based framework for enhanced chest X-ray image diagnosis. *Multimedia Tools and Applications*, 84: 18687-18710. <https://doi.org/10.1007/s11042-024-19721-8>
  - [18] Ayach, T., Behera, S., Padhi, S., Naskar, N., Banik, D. (2023). Segmentation in microscopic images using cellular automata. In 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, pp. 1-5. <https://doi.org/10.1109/INCET57972.2023.10170641>
  - [19] Banik, D. (2021). Phrase table re-adjustment for statistical machine translation. *International Journal of Speech Technology*, 24: 903-911. <https://doi.org/10.1007/s10772-020-09676-0>
  - [20] Banik, D., Ekbal, A., Bhattacharyya, P., Bhattacharyya, S., Platos, J. (2019). Statistical-based system combination approach to gain advantages over different machine translation systems. *Heliyon*, 5(9): e02504. <https://doi.org/10.1016/j.heliyon.2019.e02504>
  - [21] Banik, D., Ekbal, A., Bhattacharyya, P. (2020). Statistical machine translation based on weighted syntax- semantics. *Sādhanā*, 45: 191. <https://doi.org/10.1007/s12046-020-01427-w>
  - [22] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
  - [23] Fei-Fei, L., Fergus, R., Perona, P. (2004). Learning generative visual models from few examples: An unsupervised approach. In *Proceedings of the Conference on Computer Vision Pattern Recognit. Workshop*, pp. 1789-1798.
  - [24] Zhang, S., Li, J., Yang, L. (2023). Survey on controllable image synthesis with deep learning. arXiv preprint arXiv:2307.10275. <https://doi.org/10.48550/arXiv.2307.10275>
  - [25] Yu, T., Lin, X., Wang, S., Sheng, W., Huang, Q., Yu, J. (2023). A comprehensive survey of 3D dense captioning: Localizing and describing objects in 3D scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3): 1322-1338. <https://doi.org/10.1109/TCSVT.2023.3296889>