.

Multimodal Learning for Skin Lesion Segmentation and Closed Visual Question Answering*

Notebook for the <Lab name> Lab at CLEF 2025

Bhagyashree Mallanaikar[1], Shradha Kekare[1], Padmashree Desai[2], Sujata C[2], Uma Mudenagudi[2], Ramesh Ashok Tabib[2], Anjali Savalkar[3] and Aishwarya S.H[3,*,†]

[1]*B.E(CS), India*
[2]*Professor at KLE Tech, India*
[3]*Student at KLE Tech, India*

Abstract

Accurate interpretation of medical images is essential for intelligent diagnostic systems. This project presents a unified deep learning framework that tackles two key challenges in medical image analysis: skin lesion segmentation and closed-ended visual question answering (VQA). For lesion segmentation, we introduce a model based on the Multi-Scale Feature Fusion Network (MSFNet), enhanced with boundary and reverse attention modules. This design improves the detection of irregular and low-contrast lesions. Tested on 314 dermatology images, the model achieved a mean Dice coefficient of 0.7021, a mean Jaccard index of 0.5410, and a maximum Dice score of 0.7512, supporting its effectiveness in aiding early melanoma detection. In parallel, our closed-ended VQA system combines visual feature extraction with language embeddings to answer structured questions—such as "yes/no," object types, and numeric values. On a set of 56 question-image pairs, it achieved 56.98% overall accuracy, with high scores in categories CQID012 (74.80%) and CQID035 (74.00%). Together, these results showcase the promise of deep learning in multi-modal medical image understanding. The integration of segmentation and VQA in a single pipeline highlights its potential for real-world applications, including clinical decision support, assistive tools, and automated medical interpretation.

Keywords

Skin Lesion Segmentation, Visual Question Answering (VQA), Deep Learning, MSFNet, Multi-modal Reasoning, Medical AI, Natural Language Processing CEUR-WS

1. Introduction

Skin lesion segmentation is crucial for the early detection of melanoma and other dermatological conditions [1]. Traditional methods using hand-crafted features often fail due to low contrast and irregular lesion boundaries. Deep learning models like U-Net [2], DeepLabV3+ [3], and Attention U-Net[4] have significantly improved performance but struggle with fine boundary refinement and semantic ambiguity.

To address these challenges, we adopt the Multi-Scale Feature Fusion Network (MSFNet), which offers an effective balance between precision and computational efficiency through its integration of attention modules and hierarchical feature fusion[5]. Our work evaluates this architecture on a challenging real-world dataset as part of the MEDIQA-MAGIC 2025 segmentation task [6].

The remainder of this paper is organized as follows: Section 2 reviews related work on skin lesion segmentation. Section 3 describes the proposed MSFNet architecture and methodology. Section **??** details the training strategy used, while Section **??** presents the evaluation metrics employed for performance

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
* CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
† All authors contributed equally to this work.
bhagyam1110@gmail.com (B. Mallanaikar); kenkreshraddha92@gmail.com (S. Kekare); padmashri@kletech.ac.in (P. Desai); sujata_c@kletech.ac.in (S. C); uma@kletech.ac.in (U. Mudenagudi); ramesh_t@kletech.ac.in (R. A. Tabib); 01fe22bcs196@kletech.ac.in (A. Savalkar); 01fe22bcs316@kletech.ac.in (A. S.H)


assessment. Section **??** discusses the experimental results. Finally, Sections 6 provide the conclusion and future scope.

Closed Visual Question Answering (Closed-VQA) is a vision-language task where systems answer image-based queries using a fixed set of responses [7], crucial for applications like medical diagnostics [8], autonomous systems [9], and HCI [10]. Unlike open-ended VQA, it ensures interpretability and control, vital for high-stakes domains [8]. Despite progress using CNNs and Transformers challenges remain in compositional reasoning and domain-specific contexts . To address this, we propose a model with cross-modal attention and fine-grained alignment for reliable, interpretable predictions in clinical and industrial applications.

2. Background Study

U-Net [2] established encoder-decoder networks with skip connections for biomedical segmentation. Subsequent works like DeepLabV3+ [3] introduced atrous convolution for multi-scale context, while Attention U-Net[4] incorporated attention gates to improve focus on lesion areas. MSFNet [5] combines parallel partial decoders (PPD), boundary attention (BA), and reverse attention (RA) modules to enhance edge sensitivity and semantic integration.

Other alternatives like Vision Transformers[11], MedT [12], and GAN-based methods[13] show promising results but with higher computational demands. Lightweight approaches such as SL-HarDNet [14] and hybrid optimization frameworks have emerged to address efficiency and generalization, especially in mobile or clinical settings.

Visual Question Answering (VQA) is a complex task requiring joint understanding of images and language, with Closed VQA framing it as a multi-class classification problem for consistent evaluation [7]. Early models used CNNs and LSTMs, but attention mechanisms like Bottom-Up and Top-Down Attention improved performance [15]. Transformer-based models such as ViLBERT [15], MCAN [16], BAN [10], and ViLT [17] enabled better cross-modal reasoning, building on Vaswani et al.'s architecture [8]. Despite progress, challenges remain in dense scenes and low-resource settings, addressed by models like LXMERT and Oscar.

3. Methodology

3.1. Segmentation

We employ the original MSFNet architecture[5], which integrates multi-scale feature extraction, boundary refinement, and attention-driven fusion. The network utilizes a deep CNN backbone with five convolutional blocks (Conv1–Conv5), capturing progressively abstract features from input images.

Boundary Attention (BA) modules operate on intermediate layers (Conv2, Conv3) to emphasize lesion edges. Parallel Partial Decoder (PPD) processes deep features from Conv4 and Conv5 to produce a coarse semantic prediction. Reverse Attention (RA) modules then iteratively refine this prediction by re-focusing on uncertain boundary regions.

Finally, the outputs of BA, RA, and PPD modules are fused using element-wise addition and convolution layers to generate the final lesion mask. A sigmoid activation produces a binary probability map for segmentation.

As this architecture is reused without structural modification, detailed formulations are omitted here and can be found in[5].

3.2. Closed VQA

The proposed VQA framework for the ImageCLEF 2025 challenge [18] employs a multimodal architecture that integrates clinical text queries and medical images. Clinical questions are encoded using a BERT-based model, while images from the DermaVQA-DAS dataset [6] undergo dual-path feature extraction via Vision Transformer (ViT) and Local Binary Patterns (LBP). The resulting visual and textual features

are fused through outer product-like matching to capture fine-grained cross-modal associations. This interaction enables accurate answer retrieval through similarity-based matching

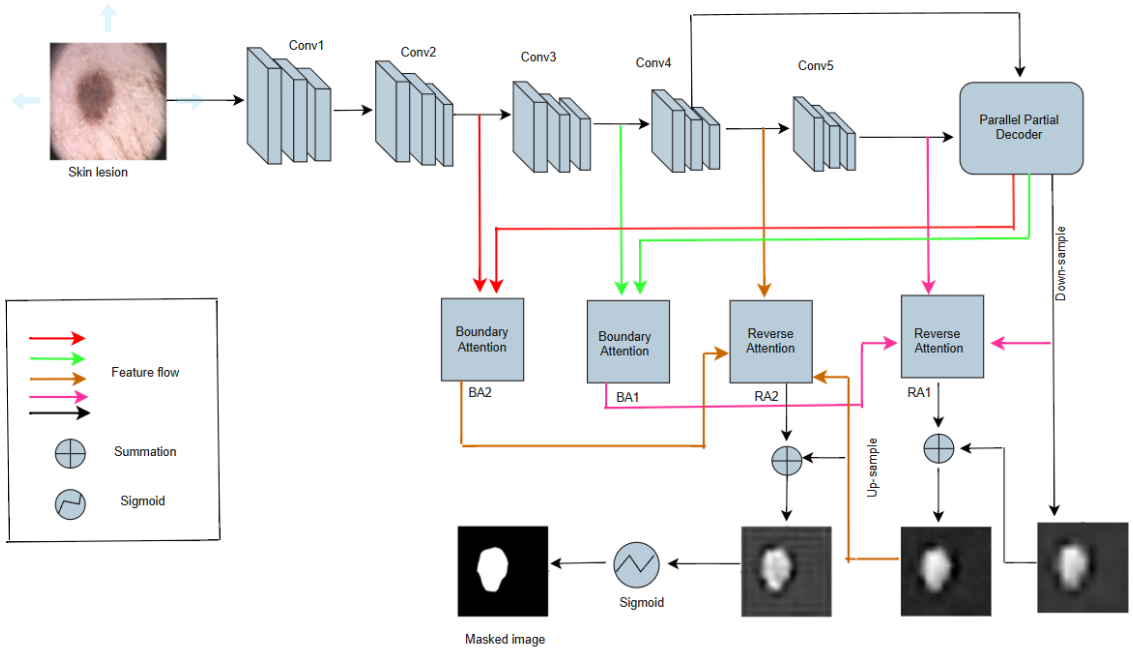

Figure 1: MSFNet architecture overview [5], illustrating feature flow through BA, RA, and PPD modules.

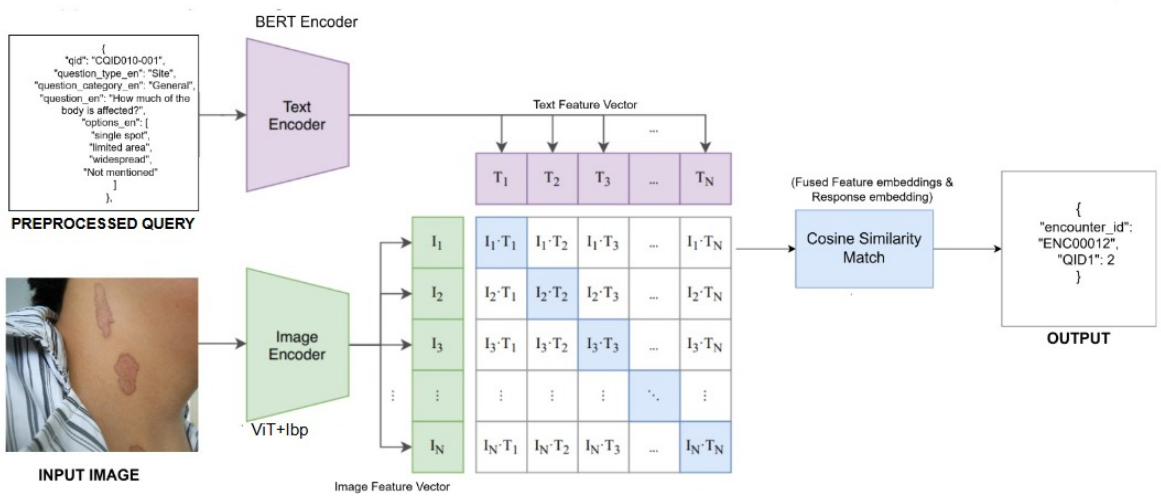

Figure 2: Schematic overview of the proposed multimodal VQA framework leveraging text and image encoders with cosine similarity-based response matching.

4. Training and Evaluation Summary

For segmentation, all training was conducted using the gold-standard masks provided in the MEDIQA-MAGIC 2025 dataset [19], comprising 314 pixel-annotated dermatology images. To maintain consistency and reduce computational load, both input images and corresponding ground truth masks were resized to 352×352 pixels. The model was trained using a hybrid loss function designed to balance pixel-level accuracy and region-level overlap:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{WBCE}} + \beta \cdot \mathcal{L}_{\text{IoU}} \tag{1}$$

Here, $\mathcal{L}_{\text{WBCE}}$ denotes the Weighted Binary Cross-Entropy Loss, which mitigates class imbalance by assigning greater importance to underrepresented lesion pixels, while $\mathcal{L}_{\text{IoU}}$ is the Intersection-over-Union Loss that directly promotes region overlap between prediction and ground truth. The coefficients were empirically set as $\alpha = 1$ and $\beta = 1$ in all experiments.

To enhance generalization and prevent overfitting, standard data augmentation techniques such as flipping, rotation, and contrast adjustment were applied during training. The Adam optimizer was employed with an initial learning rate of 1×10^{-4}. A learning rate decay strategy was incorporated, reducing the learning rate by a factor of 0.1 if the validation loss did not improve for 10 consecutive epochs. Training was conducted over a maximum of 100 epochs using a batch size of 8, with early stopping enabled based on validation loss. During inference, predicted masks were upsampled to the original image resolution using bilinear interpolation.

For evaluation, segmentation performance was quantified using the Dice coefficient and mean Intersection over Union (mIoU). The Dice score is computed as:

$$\text{Dice}(P, G) = \frac{2TP}{2TP + FP + FN} \tag{2}$$

where TP, FP, and FN represent the true positives, false positives, and false negatives, respectively. The mIoU metric averages IoU across all instances:

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{3}$$

For Visual Question Answering (VQA), the model was trained in a supervised manner to classify inputs into one of 12 answer categories. The input comprises a fused feature vector of dimension 1024, created by concatenating a 512-dimensional text embedding (extracted using CLIP) and a 512-dimensional visual embedding (obtained from a Vision Transformer and Local Binary Patterns).

The classifier outputs logits over the answer space, which are converted into class probabilities using the softmax function:

$$P(y = c|\mathbf{x}) = \frac{e^{\hat{y}_c}}{\sum_{j=1}^{C} e^{\hat{y}_j}}, \quad c \in \{1, 2, \ldots, C\} \tag{4}$$

The model was trained to minimize the categorical cross-entropy loss:

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log P(y_i|\mathbf{x}_i) \tag{5}$$

where N is the number of training samples, y_i is the ground truth label, and $P(y_i|\mathbf{x}_i)$ is the predicted probability for the true class. Optimization was performed using the AdamW optimizer with a fixed learning rate of 1×10^{-4}, and training was run for 1000 epochs with a batch size of 32. Mixed precision training on CUDA-enabled GPUs was used to accelerate computation and reduce memory usage.

Evaluation was conducted according to the ImageCLEF VQA-Med 2024 protocol. The performance was measured using macro-averaged accuracy across grouped question types. Let $\mathcal{E}$ represent the set of encounter IDs and $\mathcal{Q}$ the set of grouped question IDs. For each encounter-question pair (e, q), with gold-standard answers $G_{e,q}$ and predictions $P_{e,q}$, instance-level accuracy is defined as:

$$\text{Accuracy}(e, q) = \frac{|G_{e,q} \cap P_{e,q}|}{\max(|G_{e,q}|, |P_{e,q}|)} \tag{6}$$

Group-level accuracy is computed as:

$$\text{Accuracy}_q = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \text{Accuracy}(e, q) \tag{7}$$

And the final macro-averaged accuracy across all question types is given by:

$$\text{Accuracy}_{\text{overall}} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \text{Accuracy}_q \tag{8}$$

The evaluation process involved parsing both gold and predicted JSON files, grouping responses by question and encounter ID, and computing the aforementioned accuracy metrics. Predictions with missing instances were assigned an accuracy of zero to ensure consistency and fairness across all model submissions.

5. Results

We evaluated the segmentation model on the MEDIQA-MAGIC 2025 DermaVQA-DAS dataset [6, 19], comprising 314 annotated dermatology images. Performance was measured using Dice and Jaccard indices, including both mean-of-mean and mean-of-maximum variants. The model achieved a mean Dice coefficient of 0.7021 and a mean Jaccard index of 0.5410, indicating strong lesion overlap accuracy. Best-case alignment was reflected by Dice (mean of max) at 0.7512 and Jaccard (mean of max) at 0.6377, while Dice (mean of mean) and Jaccard (mean of mean) were 0.6711 and 0.5538, respectively, demonstrating stable performance across the dataset. A visual example of segmentation outputs, including original images, ground truth, and model predictions, is shown in Figure 3.

Table 1
Skin lesion segmentation performance metrics.

Metric	Score
Dice coefficient (mean)	0.7021
Jaccard index (mean)	0.5410
Dice (mean of max)	0.7512
Dice (mean of mean)	0.6711
Jaccard (mean of max)	0.6377
Jaccard (mean of mean)	0.5538
Segmentation instances	314

We also evaluated our closed-ended Visual Question Answering (VQA) model on a test dataset containing 56 image-question pairs as part of the ImageCLEF VQA-Med 2024 task. Submitted under the team name KLE1 (Rank 12), the model was assessed based on per-question-type accuracy and overall performance. As shown in Table 2, the model achieved an overall accuracy of 56.98%. It performed particularly well in CQID012 (74.80%) and CQID035 (74.00%), indicating strength in those question categories. However, it showed lower performance in CQID034 (39.00%) and CQID036 (35.00%), suggesting scope for improvement in those areas. These results confirm the model's effectiveness in handling diverse closed-form visual questions while identifying areas for future refinement.

6. Conclusion and Future Work

Our MSFNet-based segmentation framework demonstrated strong performance on the MEDIQA-MAGIC 2025 dataset, achieving a mean Dice coefficient of 0.7021 and Jaccard index of 0.5410 across 314 skin lesion instances. The hybrid loss formulation, combining Weighted Binary Cross-Entropy and IoU losses, was effective in handling class imbalance and enhancing boundary delineation. These results highlight the model's robustness and generalization capability in binary lesion segmentation. For

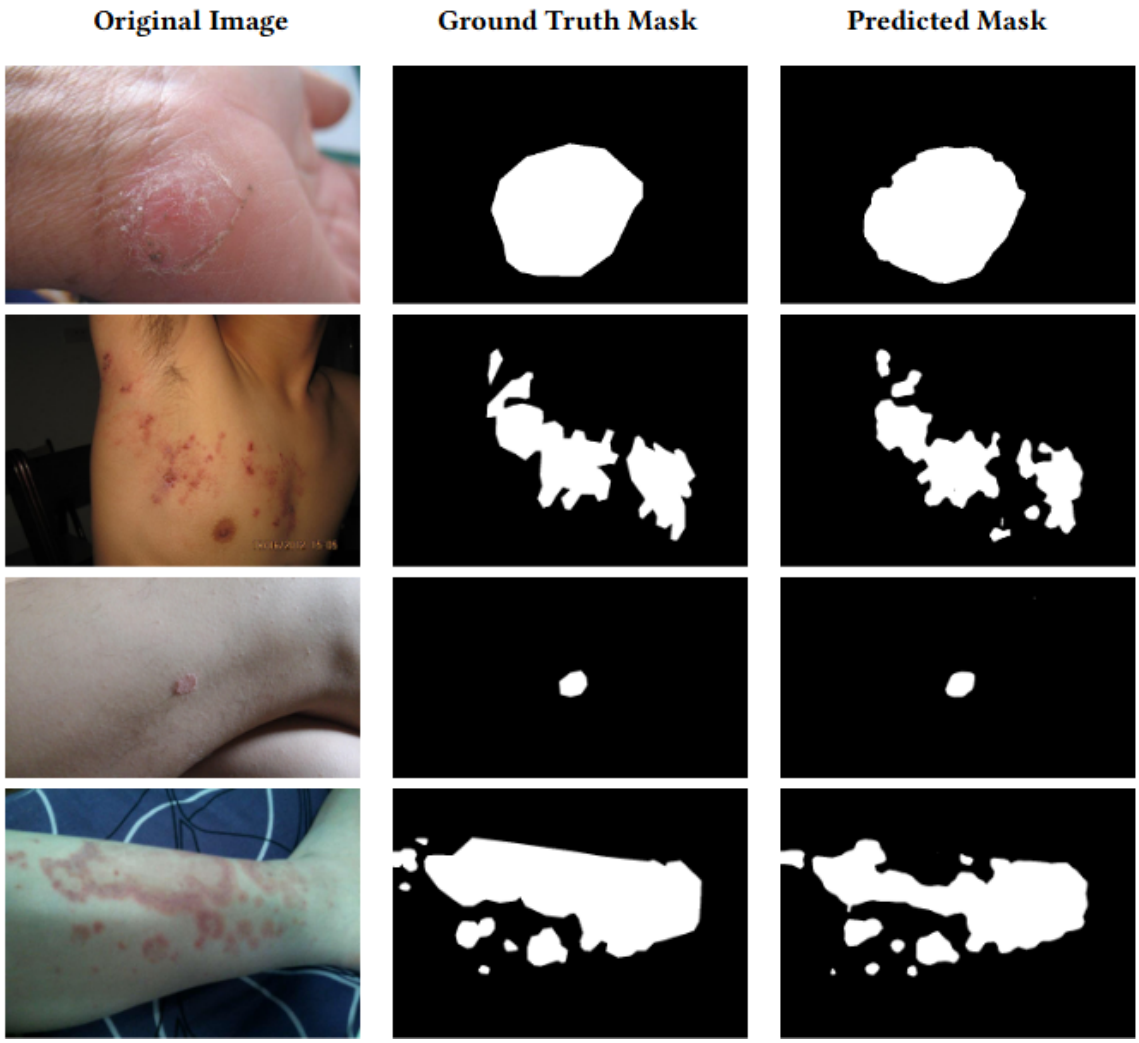

Figure 3: Visualization of the model's segmentation output. Each row shows the original image, the ground truth mask, and the corresponding prediction generated by the model.

Table 2

Closed VQA performance across different question categories.

Question ID	Accuracy (%)
CQID010	51.00
CQID011	63.10
CQID012	74.80
CQID015	57.00
CQID020	62.90
CQID025	56.00
CQID034	39.00
CQID035	74.00
CQID036	35.00
Overall Accuracy	**56.98**

future enhancements, we aim to extend the framework to multi-class segmentation, integrate advanced attention modules, explore alternative loss strategies, and optimize the model for real-time clinical or mobile deployment.

In parallel, our closed-ended Visual Question Answering (VQA) model attained an overall accuracy of 56.98% on a 56-pair test set, performing notably well in specific categories such as CQID012 and CQID035.

This reflects the model's potential in accurately interpreting visual inputs and providing consistent answers to domain-specific questions. However, lower performance in certain categories also reveals areas for improvement. Future work will focus on expanding training data diversity, incorporating larger-scale multimodal datasets, and experimenting with transformer-based and attention-driven fusion architectures. These efforts aim to boost model generalization and accuracy across broader medical VQA tasks.

Declaration on Generative AI

During the preparation of this work, we have used generative AI tool (ChatGPT) for tasks such as grammar checking and paraphrasing. All AI-generated content was reviewed and edited by the authors, who take full responsibility for the final version of the manuscript.

References

[1] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, nature 542 (2017) 115–118.

[2] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, Springer, 2015, pp. 234–241.

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.

[4] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al., Attention u-net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999 (2018).

[5] H. Basak, R. Kundu, R. Sarkar, Mfsnet: A multi focus segmentation network for skin lesion segmentation, Pattern Recognition 128 (2022) 108673.

[6] W. Yim, A. Ben Abacha, N. Codella, R. A. Novoa, J. Malvehy, Overview of the mediqa-magic task at imageclef 2025: Multimodal and generative telemedicine in dermatology, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[7] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, Vqa: Visual question answering, in: ICCV, 2015.

[8] X. Li, X. Yin, C. Li, S. Li, Medical visual question answering: A survey, Medical Image Analysis (2021).

[9] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollar, C. L. Zitnick, Microsoft coco captions: Data collection and evaluation server, in: arXiv preprint arXiv:1504.00325, 2015.

[10] D. Kiela, F. Faghri, I. Vulić, S. Clark, Supervised multimodal bitransformers for classifying images and text, EMNLP (2020).

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[12] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, V. M. Patel, Medical transformer: Gated axial-attention for medical image segmentation, in: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24, Springer, 2021, pp. 36–46.

[13] K. Innani, et al., Efficient-gan: Adversarial learning framework with morphology-aware loss for skin lesion segmentation, arXiv preprint arXiv:2305.18164 (2023).

[14] Y. Chao, et al., Sl-hardnet: A lightweight network for skin lesion segmentation with boundary enhancement, Frontiers in Bioengineering and Biotechnology 10 (2022) 1028690.

[15] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: NeurIPS, 2019.
[16] Z. Yu, J. Yu, Y. Cui, D. Tao, Deep modular co-attention networks for visual question answering, in: CVPR, 2019.
[17] J.-H. Kim, J. Jun, B.-T. Zhang, Bilinear attention networks, in: NeurIPS, 2018.
[18] A. Singh, V. Natarjan, X. Jiang, D. Hudson, M. Rohrbach, D. Batra, D. Parikh, Towards vqa models that can read, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019) 8317–8326.
[19] W. Yim, Y. Fu, A. Ben Abacha, M. Yetisgen, N. Codella, R. A. Novoa, J. Malvehy, Dermavqa-das: Dermatology assessment schema (das) and datasets for closed-ended question answering and segmentation in patient-generated dermatology images, CoRR (2025).

A. Online Resources

The sources for the ceur-art style are available via

- GitHub,
- Overleaf template.