

Workflow of Image Captioning for Disaster Remote Sensing Images

Presenter: Sakura

Contact me: bili_sakura@zju.edu.cn

Date: August 19, 2024





Outline

- Problem Definition
- Image Captioning
- Modern Image Captioning Models
 - General VLM
 - Remote Sensing VLM
- Image Captioning for Disaster RSIs
 - Building Damage Assessment & Metrics



Problem Definition



Generate High Quality Descriptions for Disaster RSIs

- Imagen [4] showcases the effectiveness of frozen large pretrained language models as text encoders for the text-to-image generation using diffusion models. Inspired by this idea, DALL-E 3 [5] shows that prompt following abilities of text-to-image models can be substantially improved by training on highly descriptive generated image captions. Stable Diffusion 3 [6] follows previous findings, adopt 3 different text encoder for better image synthesis.
- In terms of remote sensing images (RSIs), high-quality descriptive captions for RSIs are scarce. Even for inspiring work as DiffusionSat [7], the caption of training dataset is quite short and less informative.
- Therefore, there is a strong need for **high-quality informative text-image RSIs dataset**.

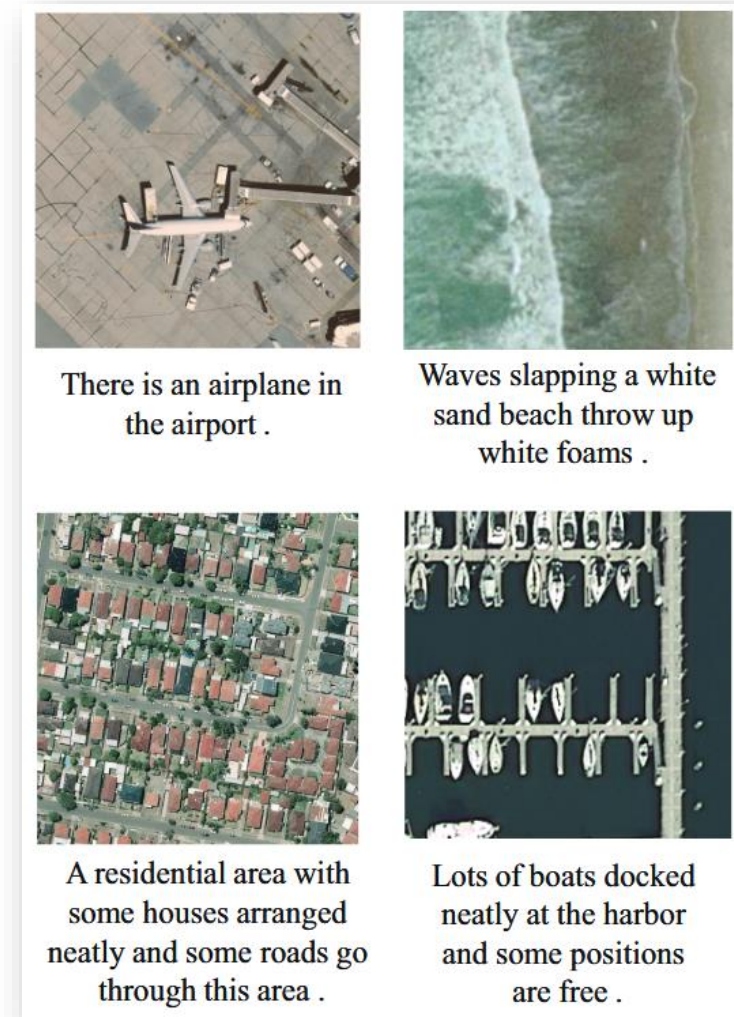
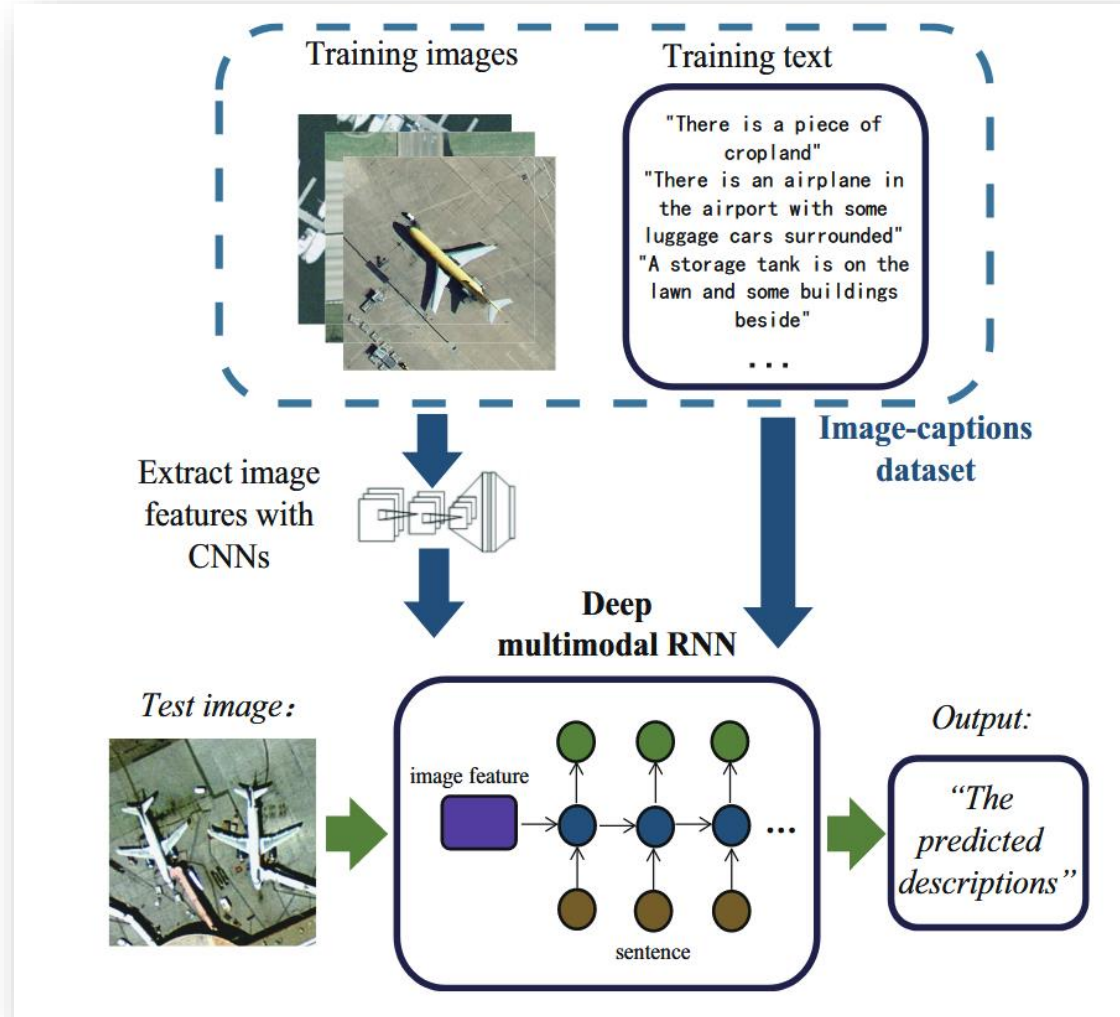
Dataset	Caption
fMoW	"a [fmow] satellite image [of a <object>] [in <country>]"
SpaceNet	"a [spacenet] satellite image [of <object>] [in <city>]"
Satlas	"a [satlas] satellite image [of <object>]"
Texas Housing	"a [satlas] satellite image [of houses] [built in <year_built>] [covering <num_acres> acres]"
xBD	"a [fmow] satellite image [<before/after>] being affected by a <disaster_type> natural disaster"

Captions created for each dataset type based on available label information from DiffusionSat [7].



Image Captioning

Classic Image Captioning Method



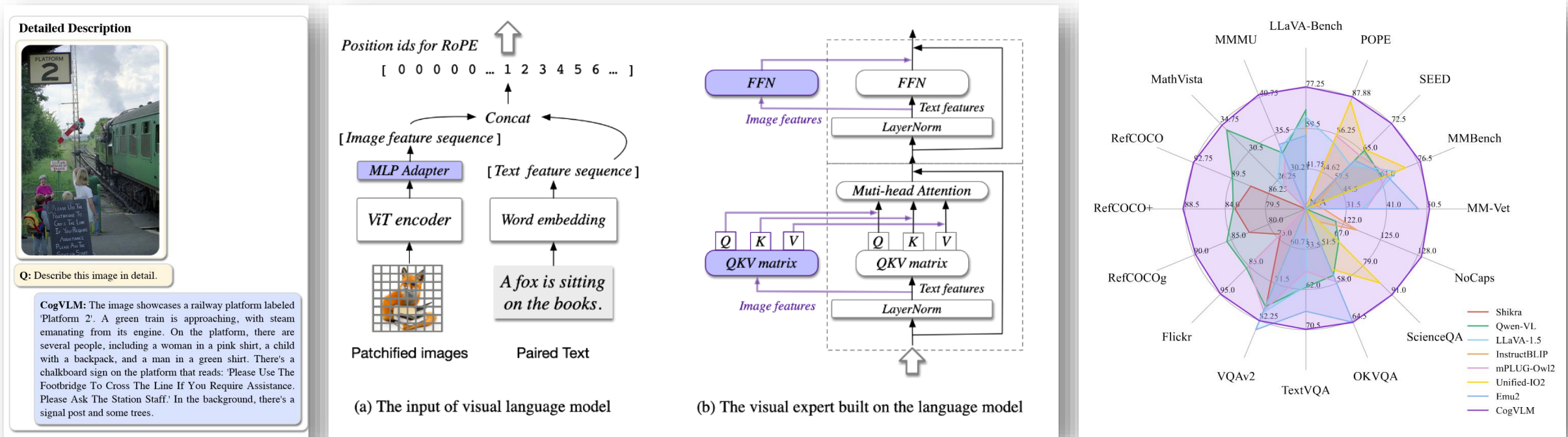
Left: Overview of model proposed in [8]. Right: The result of HSR image caption generation.



Image Captioning Models

CogVLM 2


- Developed by Tsinghua University and Zhipu AI
- Checkpoint accessed [here](#)



Left: CogVLM image captioning example. Mid: Architecture of CogVLM. Right: Evaluation of CogVLM.

MiniCPM-V 2.6

- Developed by OpenBMB
- Checkpoint accessed [here](#)

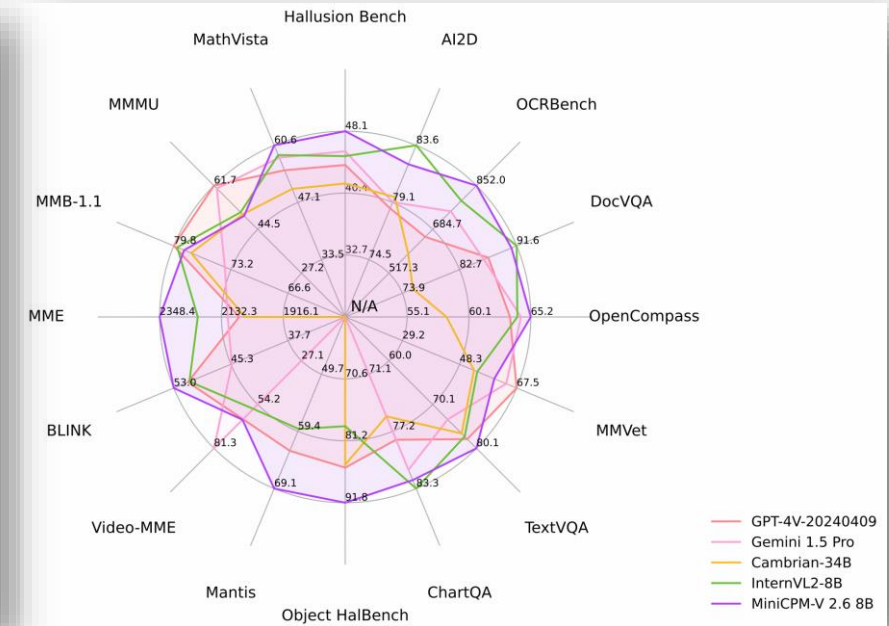


Veuillez analyser et déduire où la photo a été prise
 (English Translation)
 Please analyze and deduce where the photo was taken.

MiniCPM-V 2.6

Cette image a été prise à Paris, en France. La Tour Eiffel, l'un des monuments les plus reconnaissables du monde, est clairement visible dans le fond, et la Seine, le fleuve qui traverse la ville, peut être vu coulant sous le pont. Les autres éléments de l'image, tels que la présence d'arbres, de bannières et de bâtiments, renforcent l'identité française de la scène.

(English Translation) This image was taken in Paris, France. The Eiffel Tower, one of the most recognizable monuments in the world, is clearly visible in the background, and the Seine, the river that flows through the city, can be seen flowing under the bridge. Other elements in the image, such as the presence of trees, banners, and buildings, enhance the French identity of the scene.



Left: MiniCPM-V 2.6 image captioning example. Mid: Architecture of MiniCPM-V. Right: Evaluation of MiniCPM-V 2.6.

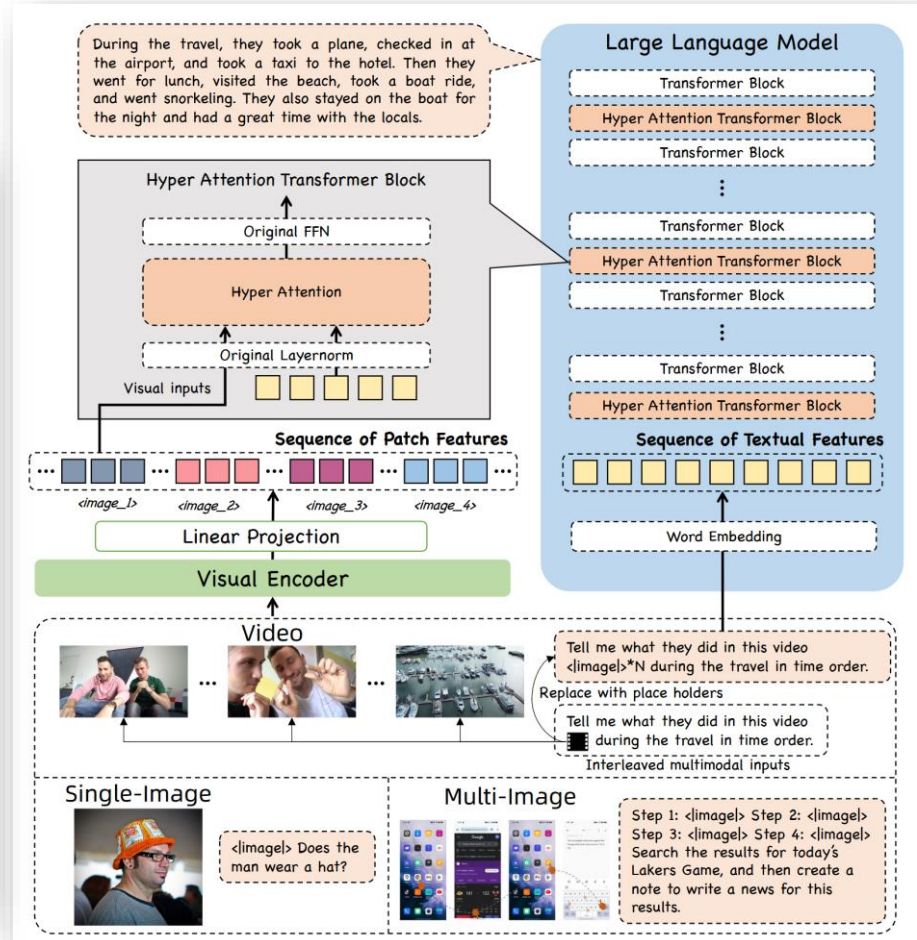
mPLUG-Owl3

- Developed by Alibaba Group
- Checkpoint accessed [here](#)



Which country can be identified in the clothing styles in each image? Tel me one by one

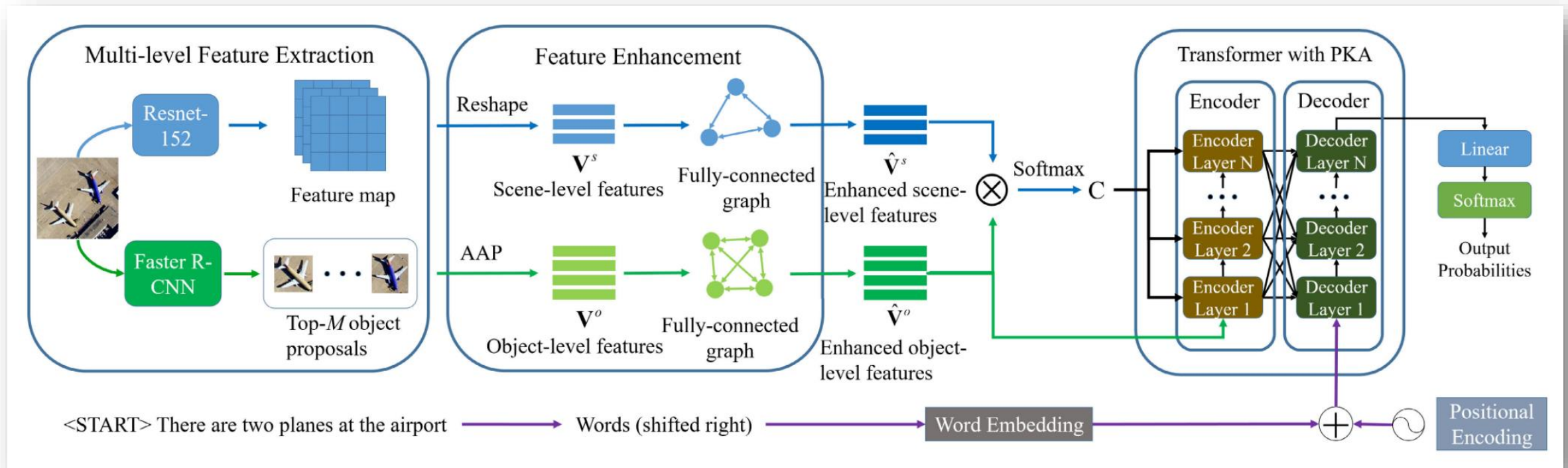
The clothing styles in each image suggest different countries. The first image features a woman wearing a cheongsam, which is a traditional Chinese dress, indicating that the country is likely China. The second image shows a man in traditional Mexican attire, including a sombrero and poncho, which are commonly associated with Mexican culture. The third image depicts two individuals in traditional Japanese kimonos, which are traditional Japanese garments, suggesting that the country is Japan.



Left: mPLUG-Owl3 multi-images understanding example. Right: Architecture of mPLUG-Owl3.

PKG-Transformer & MG-Transformer

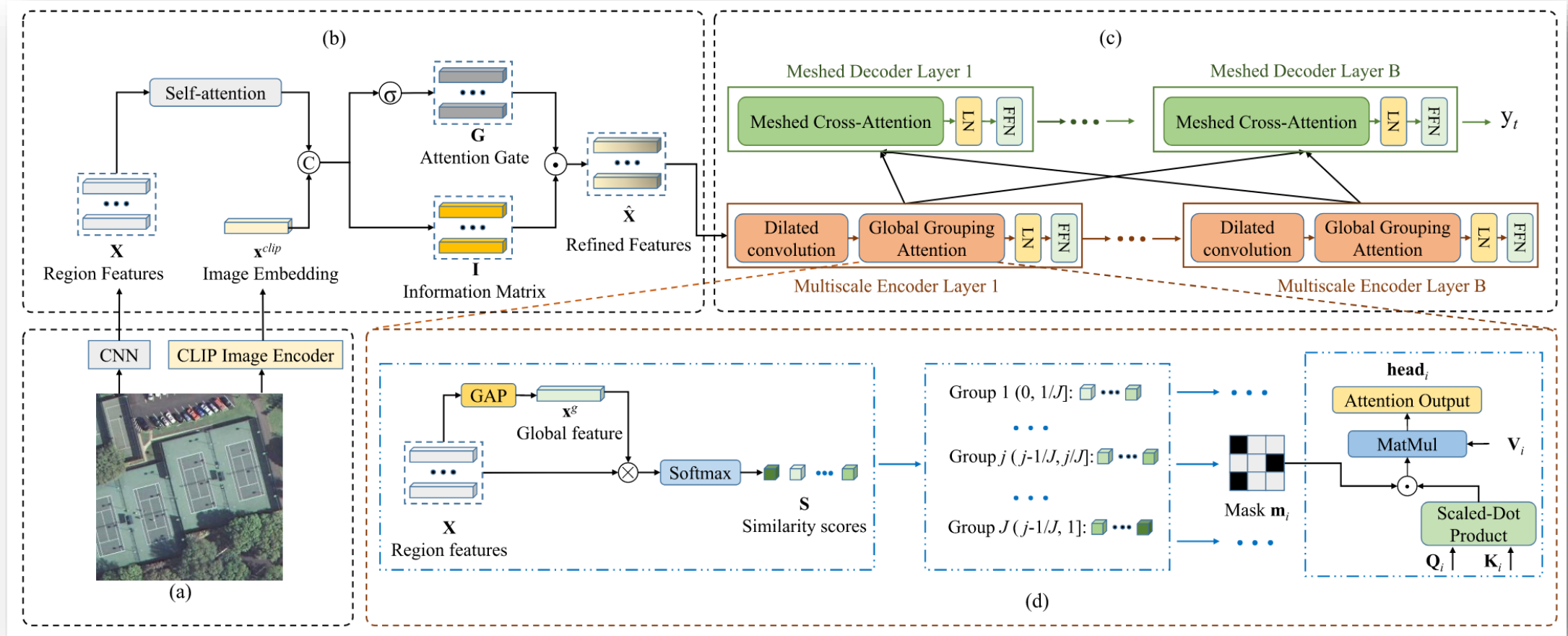
- PKG-Transformer (TGRS 2023) checkpoint accessed [here](#)
- MG-Transformer (TGRS 2024) checkpoint accessed [here](#)



Framework of PKG-Transformer [9].

PKG-Transformer & MG-Transformer

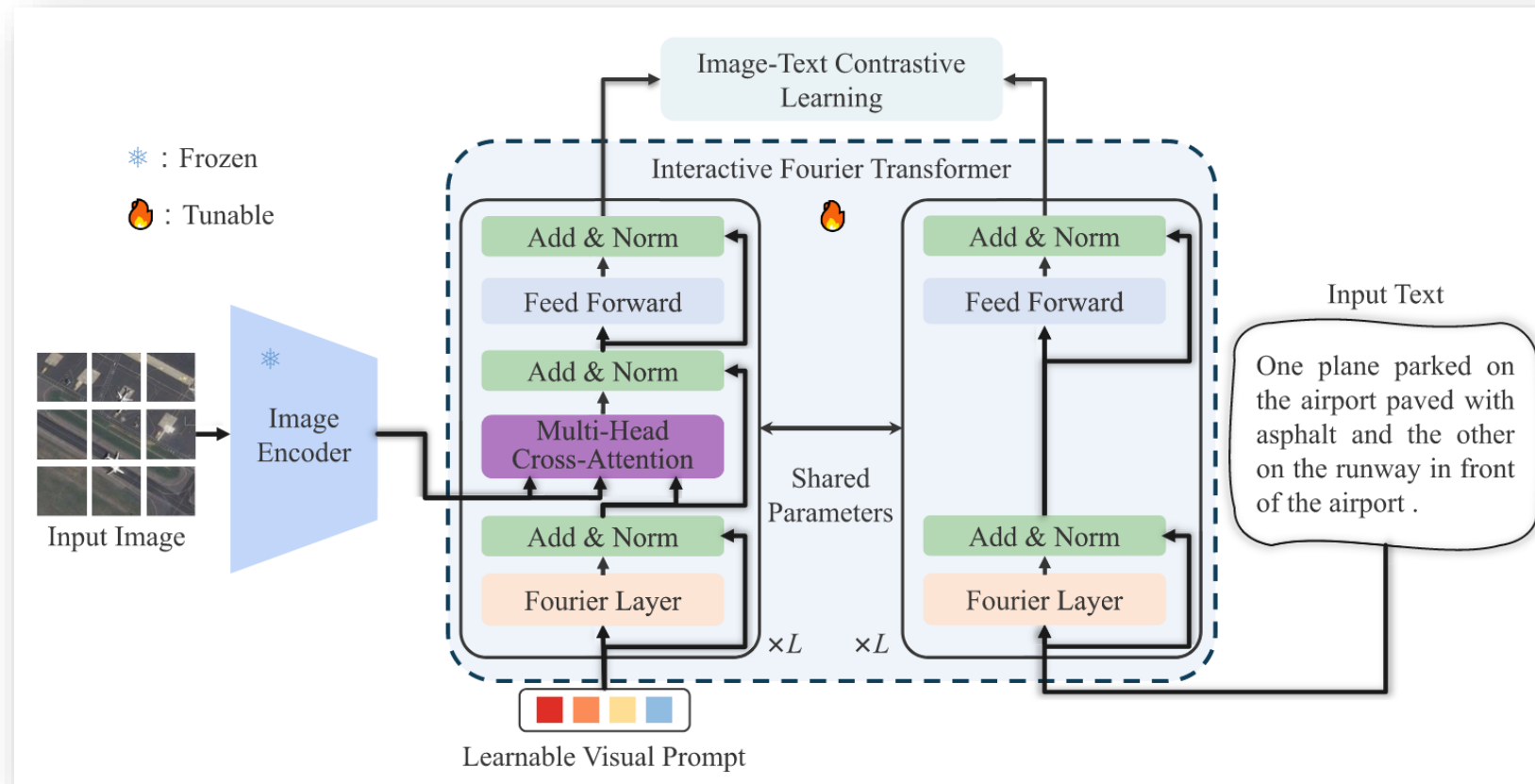
- PKG-Transformer (TGRS 2023) checkpoint accessed [here](#)
- MG-Transformer (TGRS 2024) checkpoint accessed [here](#)



Framework of MG-Transformer [10].

Bootstrapping Interactive Image–Text Alignment for Remote Sensing Image Captioning

- BITA checkpoint accessed [here](#)



Framework of BITA [11].



References and Related Works

- [1] W. Wang et al., ‘CogVLM: Visual Expert for Pretrained Language Models’, Feb. 04, 2024, arXiv: arXiv:2311.03079. doi: 10.48550/arXiv.2311.03079.
- [2] Y. Yao et al., ‘MiniCPM-V: A GPT-4V Level MLLM on Your Phone’, Aug. 03, 2024, arXiv: arXiv:2408.01800. doi: 10.48550/arXiv.2408.01800.
- [3] J. Ye et al., ‘mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models’, Aug. 13, 2024, arXiv: arXiv:2408.04840.
- [4] C. Saharia et al., ‘Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding’, in Advances in Neural Information Processing Systems, Dec. 2022, pp. 36479–36494.
- [5] J. Betker et al., ‘Improving Image Generation with Better Captions’, 2023.
- [6] P. Esser et al., ‘Scaling Rectified Flow Transformers for High-Resolution Image Synthesis’, presented at the Forty-first International Conference on Machine Learning, Jun. 2024.
- [7] S. Khanna et al., ‘DiffusionSat: A Generative Foundation Model for Satellite Imagery’, presented at the The Twelfth International Conference on Learning Representations, Oct. 2023.
- [8] B. Qu, X. Li, D. Tao, and X. Lu, ‘Deep semantic understanding of high resolution remote sensing image’, in 2016 International Conference on Computer, Information and Telecommunication Systems (CITS), Jul. 2016, pp. 1–5. doi: 10.1109/CITS.2016.7546397.
- [9] L. Meng, J. Wang, Y. Yang, and L. Xiao, ‘Prior Knowledge-Guided Transformer for Remote Sensing Image Captioning’, IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–13, Aug. 2023, doi: 10.1109/TGRS.2023.3328181.
- [10] L. Meng, J. Wang, R. Meng, Y. Yang, and L. Xiao, ‘A Multiscale Grouping Transformer With CLIP Latents for Remote Sensing Image Captioning’, IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–15, Apr. 2024, doi: 10.1109/TGRS.2024.3385500.