Attention Guidée par la Segmentation pour la Réponse Automatique à des Questions Visuelles à partir d'Images de Télédétection

Hichem Boussaid, Lucrezia Tosato, Flora Weissgerber, Laurent Wendling, Camille Kurtz, Sylvain Lobry

Attention Guidée par la Segmentation pour la Réponse Automatique à des Questions Visuelles à partir d'Images de Télédétection

Hichem Boussaid[*a] Lucrezia Tosato[*a,b] Flora Weissgerber[b]
Camille Kurtz[a] Laurent Wendling[a] Sylvain Lobry [a]

[a] LIPADE, Université Paris Cité, 75006 Paris, France
[b] DTIS, ONERA, Université Paris Saclay, FR-91123 Palaiseau, France

[*]Email : prenom.nom@u-paris.fr

Résumé

La Réponse automatique à des questions visuelles (Visual Question Answering, VQA) à partir d'images de télédétection (RSVQA) vise à permettre une extraction d'informations simplifiée à partir d'images de télédétection via l'utilisation du langage naturel. Nous proposons un mécanisme d'attention guidé par la segmentation pour améliorer l'extraction des caractéristiques visuelles en RSVQA. En exploitant la segmentation pour fournir un contexte, notre approche permet de mieux cibler les régions pertinentes de l'image. Pour l'évaluer, nous introduisons un nouveau jeu de données VQA composé d'orthophotos haute résolution, avec des paires question/réponse ainsi que des cartes de segmentations multi-canaux de 16 classes associées. Notre méthode surpasse l'approche classique, atteignant une précision globale supérieure de près de 10 % sur le jeu de données proposé.

Mots Clef

Réponse à des questions visuelles, attention, segmentation, traitement du langage naturel.

Abstract

Remote Sensing Visual Question Answering (RSVQA) is a task that aims to answer natural language questions about remote sensing images. We propose an attention mechanism guided by segmentation to enhance visual feature extraction in RSVQA. By leveraging segmentation to provide contextual understanding, our approach improves the focus on relevant image regions. To evaluate our contribution, we introduce a new VQA dataset with high-resolution RGB orthophotos, associated with multi-channels segmentation maps with 16 classes, and question/answer pairs. Our method outperforms a classical baseline approach, achieving nearly 10% higher overall accuracy on the proposed dataset.

Keywords

Visual Question Answering, Attention, Segmentation, Natural Language Processing.

1 Introduction

Le *Visual Question Answering* (VQA) vise à fournir des réponses en langage naturel à des questions sur le contenu d'une image [1]. Le *Remote Sensing Visual Question Answering* (RSVQA) applique le VQA aux images de télédétection et a été introduit dans [2]. Les avancées récentes en vision par ordinateur et en traitement du langage naturel ont amélioré les performances sur les jeux de données classiques de VQA et RSVQA. En particulier, les grands modèles de langage (*Large Language Models* LLM) permettent maintenant d'effectuer du VQA en intégrant des connaissances externes et du raisonnement [3]. Ils ont également été utilisés en RSVQA, montrant un fort potentiel pour encoder les questions et les fusionner avec les classes présentes dans les images satellitaires [4]. De plus, des *Transformers*, comme VisualBERT, ont été exploités pour fusionner l'image et le texte [5]. Les mécanismes d'attention ont aussi été utilisés en RSVQA pour améliorer l'extraction des caractéristiques en prenant en compte les alignements entre positions spatiales et mots [6].
En vision par ordinateur, les auteurs de [7] et [8] montrent que l'intégration des mécanismes d'attention améliore les performances par rapport à l'utilisation exclusive de l'image. Dans le cas des images naturelles, la segmentation sémantique a été exploitée pour guider le VQA vers l'objet d'intérêt [9] et pour orienter l'attention [10]. À notre connaissance, ces approches n'ont pas encore été appliquées au RSVQA.
Dans ce travail, nous proposons d'intégrer un mécanisme d'attention guidé par la segmentation dans une architecture dédiée au RSVQA. Pour calculer les poids d'attention, nous utilisons une segmentation sémantique multi-canaux, qui se distingue des cartes de segmentation sémantique

[*]Lucrezia Tosato et Hichem Boussaid ont contribué à parts égales. Ce travail est soutenu par l'Agence Nationale de la Recherche (ANR) dans le cadre du projet ANR-21-CE23-0011. Les expériences menées dans cette étude ont été réalisées en utilisant les ressources HPC/AI fournies par GENCI-IDRIS (2023-AD011012735R2).

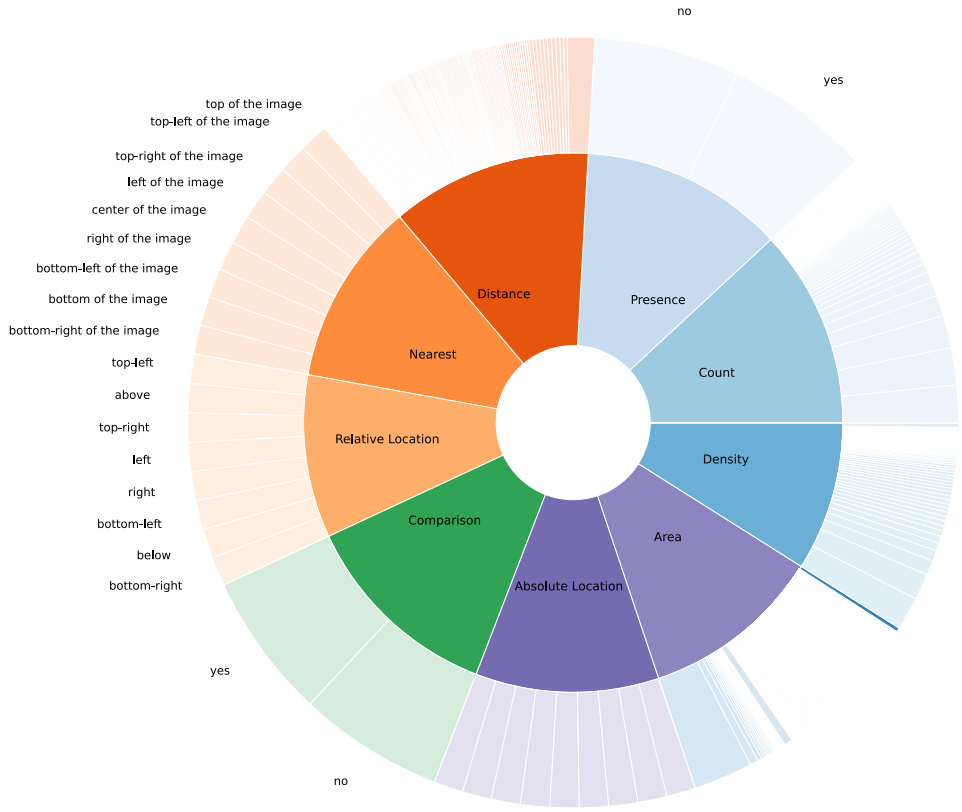

FIGURE 1 – Distribution des réponses par type de question. Nous omettons l'étiquetage des réponses numériques et nous les présentons dans l'ordre. Les valeurs numériques maximales sont 280 (questions de comptage), 40000m^2 (questions de superficie) et 273m (questions de distance).

usuelles en autorisant le chevauchement d'objets appartenant à différentes classes. Nous illustrons l'intérêt de cette approche sur un nouveau jeu de donnée de RSVQA centré sur la région Île-de-France. Ce jeu de données comprend des images aériennes à très haute résolution, des cartes de segmentation et des paires de question/réponse générées automatiquement.

La suite de cet article est organisée de la manière suivante : la section 2 présente le jeu de données utilisé, puis la section 3 introduit notre architecture RSVQA intégrant l'attention guidée par la segmentation. Enfin, la section 4 présente et discute les résultats obtenus.

2 Jeu de données

2.1 Orthophotos à très haute résolution (BD ORTHO) et annotation de segmentation

La BD ORTHO est une base de données d'images optiques aériennes à une résolution de 20 cm obtenue auprès de l'Institut national de l'information géographique et forestière (IGN). Dans ce travail, nous utilisons des images RVB à très haute résolution (VHR) obtenues par la subdivision des tuiles BD ORTHO en images de 1000×1000 pixels (équivalent à 200 m $\times$ 200 m).

L'IGN fournit également une description vectorielle du territoire français dans la base de données BD TOPO. De cette dernière, nous extrayons 16 classes pour nos annotations de segmentation multi-classes : bâtiments (bâtiment, cimetière, terrain de sport, réservoir, pylône, construction surfacique), occupation du sol (zone d'estran, zone de végétation), plan d'eau, transport (aérodrome, équipement de transport, tronçon de route, tronçon de voie ferrée), zones réglementées (forêt publique, parc ou réserve), et services et activités (qui englobent les musées, les monuments, les écoles, etc).

Nous sélectionnons quatre départements de la région Île-de-France : Paris, les Hauts-de-Seine, le Val-de-Marne et la Seine-Saint-Denis. Au total, nous obtenons 16274 images de taille 1000×1000 pixels.

2.2 Paires de questions/réponses

En suivant une procédure similaire à [11], nous proposons une approche automatisée pour générer des ensembles de questions et de réponses liées à des images VHR. Cette méthode s'appuie sur la base de données BD TOPO, qui englobe à la fois des caractéristiques géographiques générales, telles que les bâtiments et les plans d'eau, et des entités plus spécifiques, telles que les musées et les lacs.

Pour une image VHR donné p_{VHR}, nous extrayons de la BD TOPO la collection d'objets géo-localisés qui sont pré-

sents dans l'emprise géographique de p_{VHR}. Les objets sont caractérisés par un élément présent dans la BD TOPO que nous appelons une classe, par exemple « route ». Nous définissons neuf types de questions, répartis en quatre catégories :

1. **Questions à propos d'une classe** qui ne considèrent qu'une seule classe d'objets à la fois. Les questions sont divisées en **présence** (a), **comptage** (b) et **densité** (c) ;
2. **Questions sur les objets** qui portent sur **l'emplacement absolu** dans l'image (a) ou sur **la surface** (b) d'un objet spécifique ;
3. **Questions portant sur deux classes** qui **comparent** le nombre d'objets de deux classes différentes ;
4. **Questions sur la relation entre les objets** qui prennent en compte **l'emplacement relatif** (a) et **la distance** (b) de deux objets spécifiques de deux classes différentes, ou l'emplacement absolu de l'objet **le plus proche** d'une classe par rapport à un objet pré-sélectionné d'une autre classe ou à la position d'un pixel (c).

L'un des défis de la construction stochastique d'un jeu de données VQA consiste à équilibrer le type de question et le type de réponse. Il s'agit d'une exigence pour réduire les biais linguistiques [12]. Pour équilibrer le type de question, nous commençons par générer aléatoirement 10 questions par type de question pour chaque image VHR p_{VHR}. Ensuite, pour équilibrer le type de réponse, nous définissons un nombre maximal de questions par réponse $N_{Q,A}$ pour chaque type de question. Seules les questions dont la réponse est moins présente que $N_{Q,A}$ sont conservées. Pour obtenir un nombre suffisant de questions, les images VHR sont parcourues deux fois. L'application de cette procédure permet d'obtenir un total de 146848 paires de question/réponse. Leur distribution est représentée graphiquement dans la figure 1.

2.3 Division du jeu de données

Le jeu de données est divisé en fonction des images VHR, de manière aléatoire en sous-ensembles d'entraînement, de validation et de test avec une proportion de 60 %, 20 % et 20 % respectivement. Nous utilisons le même découpage pour la tâche auxiliaire de segmentation et pour l'architecture du modèle RSVQA.

3 Méthode

Nous présentons une nouvelle approche de réponse aux questions visuelles de télédétection (RSVQA) en utilisant la segmentation pour orienter le mécanisme d'attention. L'architecture de notre modèle est illustrée dans la figure 2.

3.1 Extraction de caractéristiques

Pour extraire les caractéristiques visuelles des images VHR (notées f_{VHR}), nous utilisons un modèle ResNet-50 [13] pré-entraîné sur ImageNet dont nous retirons la dernière couche entièrement connectée.
Les caractéristiques textuelles f_{q} de la question q sont extraites avec un encodeur DistilBERT [14], entraîné sur le jeu de données BookCorpus [15]. Les poids de ces deux encodeurs ne sont pas mis à jour dans le reste de cette étude.

3.2 Attention guidée par la segmentation

Dans notre approche, nous incorporons une tâche auxiliaire de segmentation sémantique pour guider le calcul du poids de l'attention. Nous supposons que la tâche auxiliaire permet une meilleure initialisation du module d'attention en introduisant explicitement une première couche de sémantique dans l'espace des caractéristiques. Cela nous permet de faciliter le processus d'apprentissage du module d'attention.
Nous entraînons d'abord le module de segmentation en faisant un *finetuning* d'un modèle UPerNet avec un *Swin Transformer Backbone* introduit dans [16]. Ce modèle est pré-entraîné sur ADE20K [17] sur les 16 classes de segmentation. Techniquement, nous ajoutons une couche de convolution qui fait correspondre les cartes de caractéristiques f_{seg} aux 16 canaux de sortie. Cette sortie est finalement interpolée à la dimension de p_{VHR} pour obtenir le résultat de la segmentation. Nous obtenons les poids d'attention en appliquant une couche linéaire (avec un *dropout* de $d = 0,5$) sur f_{q} afin de le projeter dans un espace à 250 dimensions. De même, nous utilisons une convolution de 1×1 (avec un *dropout* de d) pour obtenir une représentation des 250 canaux de f_{seg}.
Nous concaténons les deux représentations à partir de la segmentation et du texte, et appliquons une fonction d'activation *ReLU* suivie d'une convolution pour obtenir l'information d'attention $\mathbf{a}$. Enfin, nous appliquons $\mathbf{a}$ aux caractéristiques visuelles f_{VHR}.

3.3 Prédiction

Nous concaténons la version de f_{VHR} après avoir appliqué l'attention avec f_{q} pour obtenir une représentation globale des entrées. Cette représentation est mise en correspondance avec une sortie à 1000 dimensions (correspondant aux 1000 réponses les plus fréquentes de jeu d'entraînement) à l'aide d'un perceptron à 2 couches. Concernant l'attention, nous utilisons la fonction d'activation *ReLU* et un *dropout* de d. Nous entraînons le modèle en utilisant une *Cross Entropy Loss*, optimisée avec Adam, avec un taux d'apprentissage de 10^{-6} et des lots de 4 échantillons.

3.4 Évaluation

Nous évaluons les résultats de la segmentation à l'aide de la précision globale, du rappel et du score F1 global.
Trois métriques sont utilisées pour évaluer les résultats de RSVQA : la précision par type de question, la précision globale (OA) et la précision moyenne (AA). La précision par type de question est définie comme étant le rapport entre le nombre de réponses correctes et le nombre total de questions pour l'un des neuf types de questions. L'OA

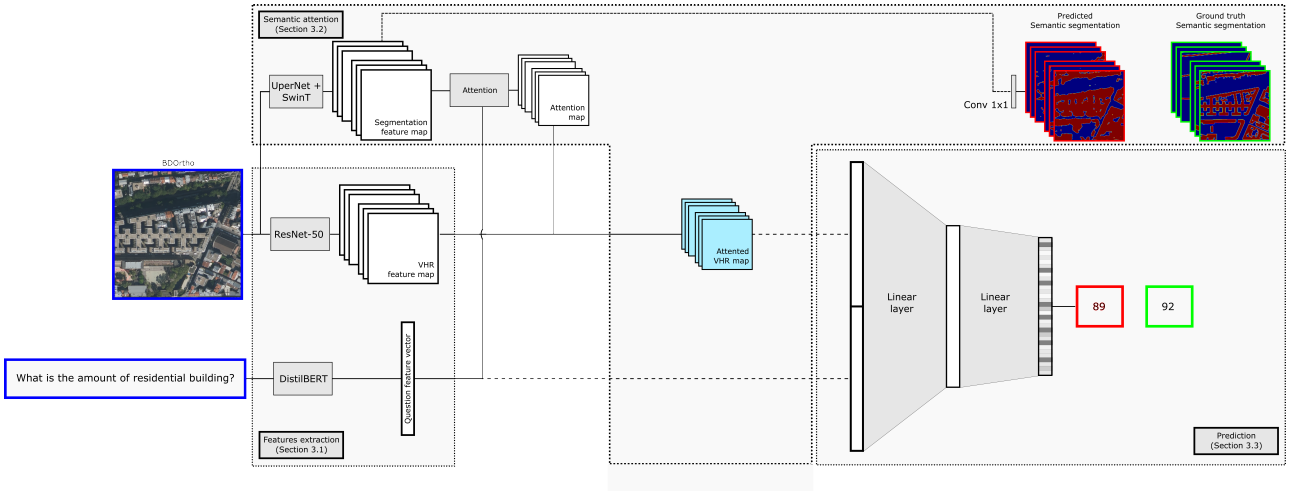

FIGURE 2 – Schéma de l'architecture proposée. Les entrées (images de télédétection à très haute résolution et questions en anglais) sont représentées dans des cadres bleus, les sorties dans des cadres rouges (réponse et segmentation) et les vérités de terrain (réponse et segmentation) dans des cadres verts.

est le rapport entre les réponses correctes et le nombre total de questions dans le jeu de données. Enfin, l'AA est la moyenne des précisions par type de question.

4 Résultats et discussion

4.1 Tâche auxiliaire de segmentation

Notre modèle a été entraîné à l'aide d'un GPU NVIDIA GeForce RTX 4090 24G. Ce modèle atteint une précision et un rappel globaux de 53,6% et 73,2% respectivement, ainsi qu'un score F1 global de 60,1%. Ces résultats sont obtenus en utilisant une courbe de précision-rappel avec 20 seuils différents, afin d'identifier sur le jeu de validation la valeur optimale de seuil pour chaque classe.

4.2 Tâche VQA

Notre modèle a été entraîné pendant 100 heures, en utilisant un GPU NVIDIA V100-16G. Les résultats globaux pour la tâche VQA sont présentés dans la Table 1. Nous observons un AA de 43,24% et un OA de 45,44% en utilisant notre approche d'attention (i.e. Proposé).
Pour étudier l'impact de l'attention guidée par la segmentation, nous concevons deux études d'ablation. La première (i.e. A1) utilise un mécanisme d'attention basique (c'est-à-dire non guidé par la segmentation) et la seconde (i.e. A2) n'utilise pas de mécanisme d'attention. Ces études d'ablation démontrent que l'utilisation de l'attention guidée par la segmentation améliore considérablement les performances, avec un gain de 10% en précision globale et de 5% par rapport au mécanisme d'attention basique.
À partir de la Table 1, on observe une amélioration de la performance en utilisant l'attention guidée par la segmentation pour chaque type de question. Dans presque toutes les catégories, l'amélioration est progressive en passant du modèle A2 à A1 puis vers la méthode proposée. Cela démontre que le mécanisme d'attention permet au réseau neuronal de se concentrer spécifiquement sur des parties sélectionnées de l'entrée, aidant ainsi la tâche RSVQA. La tâche auxiliaire de segmentation quant à elle augmente le niveau de supervision du processus, ce qui permet d'obtenir de meilleurs résultats.

Si l'on s'intéresse davantage aux résultats par catégorie, nous observons une plus faible précision de prédiction pour des questions spécifiques (2a, 4a, 4c) qui impliquent de déterminer les positions relatives et absolues des objets. Cette tâche difficile contribue probablement à la baisse des performances dans ces cas. Il est en effet difficile de reconnaître simultanément des objets de télédétection et leur relation spatiale de bout en bout en s'appuyant uniquement sur les réseaux profonds actuels [18]. Pour augmenter la précision de ces classes, en disposant de cartes de segmentation, une perspective porte ici sur l'emploi de descripteurs de positions relatives comme les histogrammes de forces pour modéliser les relations spatiales directionnelles entre les objets géo-localisés, comme proposé dans [19].

Nous pouvons observer que dans les questions de densité de classe (1c), nous avons une précision légèrement inférieure, ce qui est un résultat intéressant lorsqu'on le compare aux performances des questions de surface d'objets (2b). Ces deux tâches, bien que similaires, présentent une différence substantielle : la prédiction de surface est en effet faite par des valeurs numériques supérieures à zéro et entières, alors que les prédictions de densité sont toutes des valeurs de 0 à 1. Nous pensons que la différence d'échelle des résultats impose une plus grande sensibilité et une plus grande précision dans les prédictions de densité. Une autre raison pourrait être que le nombre de questions sur la densité est inférieur à celui des autres questions. Enfin, alors que les questions de surface portent sur un seul objet, les questions de densité portent sur une classe d'objets, ce qui complexifie la tâche du modèle.

Modèle	Paramétrage		Précision par catégorie										
	Att.	Seg.	1.(a)	1.(b)	1.(c)	2.(a)	2.(b)	3.	4.(a)	4.(b)	4.(c)	AA	OA
Proposé	✓	✓	**89.36**	**26.36**	**23.42**	**14.61**	**56.98**	**93.08**	**13.76**	**74.19**	**17.21**	**43.24**	**45.44**
A1	✓		85.79	22.98	20.99	13.82	44.69	82.94	11.91	74.19	15.60	39.40	41.43
A2			80.94	10.50	20.38	12.96	40.94	74.00	13.41	59.80	13.44	34.91	36.26

TABLE 1 – Résultats du modèle proposé et des études d'ablation. Tous les résultats sont des pourcentages de précision.

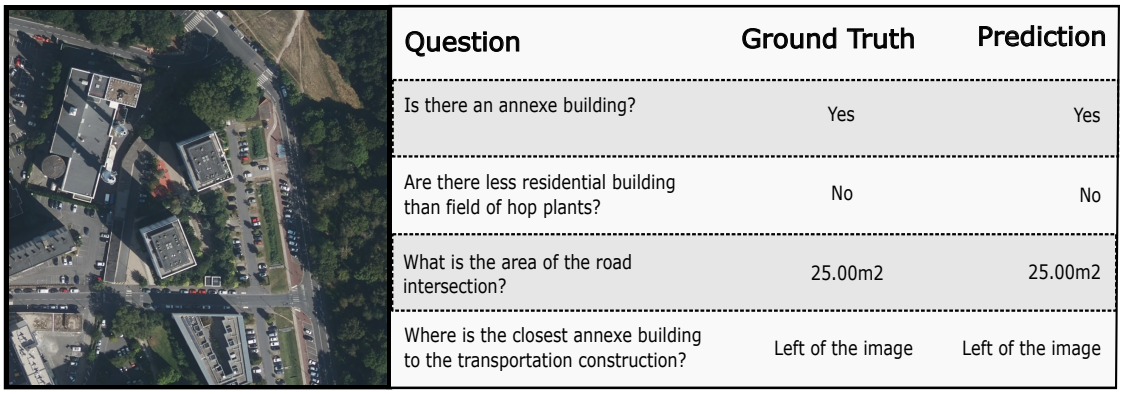

FIGURE 3 – Exemple d'une image dans le département des Hauts-de-Seine (92) avec des questions, les étiquettes et les prédictions associées.

Un exemple visuel de résultats de notre modèle est présenté dans la figure 3.

5 Conclusion

Dans cette étude, nous appliquons un modèle d'attention guidé par la segmentation dans le contexte de RSVQA sur un nouveau jeu de données construit à partir d'images à haute résolution BD ORTHO et BD TOPO pour l'annotation de la segmentation, ainsi que les paires questions/réponses.
par le biais d'une étude expérimentale préliminaire, nous observons que la segmentation parvient à diriger l'attention plus efficacement que l'attention seule. Nous pensons qu'en utilisant la segmentation en 16 canaux, l'attention identifie les canaux liés à un mot spécifique de la question et qu'il est donc plus aisé de localiser le bon objet dans l'image finale. D'autres expériences avec un jeu de données plus complet seront nécessaires pour vérifier que les résultats peuvent être généralisés à d'autres zones géographiques et à des questions plus variées.

Références

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA : Visual Question Answering," in *IEEE/CVF ICCV*, pp. 2425–2433, 2015.

[2] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA : Visual Question Answering for Remote Sensing Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555–8566, 2020.

[3] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, and J. Luo, "PROMPTCAP : Prompt-guided image captioning for VQA with GPT-3," in *IEEE/CVF ICCV*, pp. 2963–2975, 2023.

[4] C. Chappuis, V. Zermatten, S. Lobry, B. Le Saux, and D. Tuia, "Prompt-RSVQA : Prompting visual context to a language model for remote sensing visual question answering," in *IEEE/CVF CVPR*, pp. 1372–1381, 2022.

[5] T. Siebert, K. N. Clasen, M. Ravanbakhsh, and B. Demir, "Multi-modal fusion transformer for visual question answering in remote sensing," in *SPIE Remote Sensing*, pp. 162–170, 2022.

[6] X. Zheng, B. Wang, X. Du, and X. Lu, "Mutual attention inception network for remote sensing visual question answering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[7] Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3536–3545, 2020.

[8] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10941–10950, 2020.

[9] J. Wu and R. J. Mooney, "Faithful multimodal explanation for visual question answering," *arXiv preprint arXiv :1809.02805*, 2018.

[10] C. Gan, Y. Li, H. Li, C. Sun, and B. Gong, "VQS : Linking segmentations to questions and answers for supervised attention in VQA and question-focused semantic segmentation," in *IEEE/CVF ICCV*, 2017.

[11] S. Lobry, J. Murray, D. Marcos, and D. Tuia, "Visual question answering from remote sensing images," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4951–4954, IEEE, 2019.

[12] C. Chappuis, E. Walt, V. Mendez, S. Lobry, B. L. Saux, and D. Tuia, "The curse of language biases in remote sensing VQA : the role of spatial attributes, language diversity, and the need for clear evaluation," *arXiv preprint arXiv :2311.16782*, 2023.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF CVPR*, pp. 770–778, 2016.

[14] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT : smaller, faster, cheaper and lighter," *arXiv preprint arXiv :1910.01108*, 2019.

[15] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies : Towards story-like visual explanations by watching movies and reading books," in *IEEE/CVF ICCV*, 2015.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer : Hierarchical vision transformer using shifted windows," in *IEEE/CVF ICCV*, 2021.

[17] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ADE20K dataset," *International Journal of Computer Vision*, vol. 127, pp. 302–321, 2019.

[18] W. Cui, F. Wang, X. He, D. Zhang, X. Xu, M. Yao, Z. Wang, and J. Huang, "Multi-scale semantic segmentation and spatial relationship recognition of remote sensing images based on an attention model," *Remote Sensing*, vol. 11, no. 9, p. 1044, 2019.

[19] M. Faure, S. Lobry, C. Kurtz, and L. Wendling, "Embedding spatial relations in visual question answering for remote sensing," in *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 310–316, IEEE, 2022.