Détection d'animaux camouflés par information temporelle et modèles Vision-Langage

Mounir Ammam
Université de Montréal
mounir.ammam@umontreal.ca

Byungsuk Min
Université de Montréal
byungsuk.min@umontreal.ca

Yanis Chikhar
Université de Montréal
yanis.chikhar@umontreal.ca

1 Revue de la littérature

La détection d'objets camouflés constitue un problème particulièrement difficile en vision par ordinateur, car les objets d'intérêt présentent une forte similarité visuelle avec leur environnement. Dans de nombreux cas, les textures, couleurs et contours de l'objet se confondent avec le fond, rendant l'identification ambiguë, même pour un observateur humain. Les approches basées uniquement sur des images statiques sont donc limitées lorsque les indices visuels sont faibles.

1.1 Limitations de la détection à partir d'images statiques

Li et al. (1) mettent clairement en évidence cette limitation dans le contexte des animaux camouflés. Les auteurs montrent que ces derniers sont presque indiscernables du fond dans des images isolées et deviennent détectables principalement lorsqu'ils sont en mouvement. Cette observation souligne que l'information contenue dans une seule image est souvent insuffisante pour résoudre le problème du camouflage, et motive l'exploration de sources d'information supplémentaires.

1.2 Exploitation du mouvement et de l'information temporelle

Afin de dépasser les limites des approches basées uniquement sur l'image, plusieurs travaux ont étudié l'exploitation du mouvement et de l'information temporelle dans les vidéos. Le mouvement constitue un indice discriminant important, car il permet de révéler des objets camouflés qui restent invisibles dans des images statiques.

Cheng et al. (2) s'intéressent à la détection d'objets camouflés dans des séquences vidéo et montrent que la prise en compte de l'information temporelle améliore significativement les performances de détection. Leur travail met en évidence que les variations temporelles, même subtiles, fournissent des signaux utiles pour distinguer les objets camouflés de leur arrière-plan. Ces résultats confirment que l'exploitation du mouvement est une direction prometteuse pour la détection d'animaux camouflés.

1.3 Vision–Language Models et connaissances sémantiques

En parallèle, les modèles Vision–Langage (VLM) ont récemment attiré beaucoup d'attention en raison de leur capacité à apprendre des représentations conjointes image–texte à grande échelle. Des modèles tels que CLIP, proposé par Radford et al. (3), sont entraînés sur de grandes collections de paires image–texte et apprennent à aligner des concepts visuels et linguistiques dans un espace de représentation commun. Grâce à cette supervision sémantique, ces modèles permettent une reconnaissance *zero-shot* et une bonne généralisation à des concepts non vus durant l'entraînement.

Bien que les VLM n'aient pas été conçus spécifiquement pour la détection d'objets camouflés, ils offrent une source de connaissances sémantiques complémentaires. Le langage peut fournir des indications de haut niveau sur la présence ou la nature d'un objet, ce qui peut être particulièrement utile lorsque les indices visuels sont ambigus ou peu informatifs.

1.4 Positionnement du projet

Dans ce contexte, ce projet adopte une approche exploratoire visant à combiner trois idées complémentaires issues de la littérature : les limites des approches basées sur des images statiques, l'apport du mouvement et de l'information temporelle dans les vidéos, et l'utilisation de connaissances sémantiques fournies par les modèles Vision–Langage. L'objectif du projet n'est pas de proposer une méthode à l'état de l'art, mais plutôt d'étudier comment ces différentes sources d'information peuvent être exploitées conjointement pour améliorer la détection d'animaux camouflés sur le jeu de données MoCA.

References

[1] Y. Li, Y. Zhang, Y. Wang, and C. Shen, "Moving camouflaged animals: A dataset and benchmark for motion-based object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. [Online]. Available: https://arxiv.org/abs/2011.11630v1

[2] X. Cheng, J. Fu, F. Zhu, and X. Li, "Implicit motion handling for video camouflaged object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. [Online]. Available: https://arxiv.org/abs/2203.07363

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020