Exploration de l'impact de l'estimation monoculaire de la profondeur et de la segmentation sémantique sur la tâche de détection de l'absence de mise au point et du flou de mouvement

Matthieu SERFATY, Tina NIKOUKHAH, Jérémy ANGER, Gabriele FACCIOLO, Jean-Michel MOREL

Université Paris-Saclay, CNRS, ENS Paris-Saclay, Centre Borelli, 91190 Gif-sur-Yvette, France
matthieu.serfaty@icloud.com, tina.nikoukhah@ens-paris-saclay.fr
angerj.dev@gmail.com, gabriele.facciolo@ens-paris-saclay.fr, moreljeanmichel@gmail.com

Résumé – Nous étudions l'impact des informations de haut niveau provenant de la segmentation sémantique et de l'estimation monoculaire de la profondeur sur la tâche de détection de flou de mouvement et de flou de mise au point. Nous montrons par une évaluation minutieuse sur plusieurs jeux de données que l'estimation de la profondeur monoculaire augmente la précision des détecteurs de flou et qu'une carte sémantique apporte également des informations précieuses à la tâche, mais avec un impact moindre sur le résultat.

Abstract – We study the impact of high level information coming from semantic segmentation and monocular depth estimation on the task of out of focus and motion blur detection. We show by careful evaluation on several datasets that monocular depth estimation boosts the accuracy of blur detectors and that a semantic map also brings valuable information to the task, but with less impact on the result.

1 Introduction

Le flou dans une image peut être le résultat d'un choix artistique, par exemple en sélectionnant une profondeur de champ restreinte ou une plus longue exposition. Cependant, elle peut aussi être le résultat d'un mouvement indésirable de la caméra ou d'un manque de netteté. Dans cet article, nous abordons le problème de la détection précise des régions floues d'une image et de l'identification du type de flou. L'utilisation de ces cartes de flou peut être très diverse, de la restauration locale du flou de mouvement [1] ou du flou de mise au point, à une simple évaluation de la qualité conduisant à écarter les images présentant un flou excessif [2].

Les réseaux de neurones convolutifs profonds (DCNN) ont permis des succès remarquables dans des tâches de traitement d'images telles que la détection d'objets, la segmentation sémantique [3], l'estimation de la profondeur monoculaire [4, 5] et la détection de la presence et type de flou [6, 7]. Ici, nous explorons ces avancées sur la tâche de bas niveau consistant en la détection et la classification du flou local.

De nombreuses méthodes efficaces de détection du flou basées sur des caractéristiques artisanales ont été proposées dans la littérature comme [8, 9]. Les DCNNs deviennent maintenant proéminents pour la détection du flou. L'une des méthodes de pointe proposées dans [6] propose un encodeur-décodeur pour segmenter l'image en régions floues, floues de mouvement et nettes. Grâce à l'encodeur-décodeur, ils parviennent à exploiter simultanément des caractéristiques de haut et de bas niveau. Ils atteignent des performances de pointe même dans les régions homogènes. Une autre étude [7] a proposé d'impliquer une estimation de la profondeur monoculaire dans l'entraînement d'un réseau dédié à la détection du flou. La profondeur et le flou de défocus sont en effet corrélés, ce qui explique amplement pourquoi l'association des deux peut être fructueuse.

Dans cet article, nous étendons les méthodes de [6] et de [7] en approfondissant le problème de la détection du flou dans les régions homogènes, qui est sans doute la partie la plus difficile de ce problème. Dans ce but, en plus de l'estimation de la profondeur, nous faisons intervenir une segmentation sémantique de l'image. La détection d'objets saillants dans la scène peut aider à prendre des décisions dans des régions homogènes mais sémantiquement identifiables, comme le cadre d'un vélo comme on peut le voir sur Figure 1.

2 Travaux et état de l'art

Détection du flou avec une architecture encodeur-décodeur Le meilleur réseau pour détecter à la fois le mouvement et le flou de mise au point est sans doute celui proposé dans [6], qui est basé sur une architecture d'encodeur-décodeur bien établie inspirée par [10, 4, 5]. En effet, les tâches de prédiction d'images denses conduisent souvent à une architecture encodeur-décodeur. Dans [6], l'encodeur est entraîné sur une tâche de classification d'images sur le jeu de données ImageNet [11]. Le décodeur agrège les caractéristiques provenant de l'encodeur et les convertit en prédictions denses finales. La classification d'images et la détection de flou sont des tâches de vision différentes qui

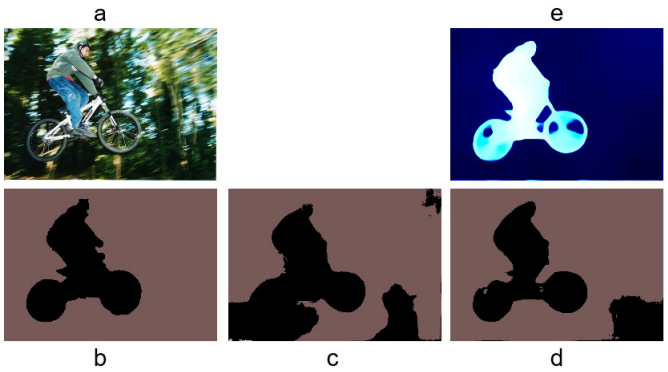

FIGURE 1 – Exemple d'une image partiellement brouillée par le mouvement provenant de l'ensemble de données CUHK (a), l'étiquette de vérité terrain (b), les résultats de détection (c, d) et la sortie MiDaS (e). Dans la carte de vérité terrain et les résultats de détection, le noir désigne une région nette et le marron une région floue. Nous obtenons (c) en entraînant le réseau avec des images RVB comme (a), (d) en entraînant le réseau sur la combinaison d'images RVB (a) avec leur sortie MiDaS (d).

font appel à des caractéristiques de bas niveau similaires, telles que les bords et les petites textures. Les caractéristiques de haut niveau utilisées pour la classification d'images sont donc susceptibles d'être utiles dans une tâche orientée sémantiquement telle que la détection de flou dans des régions homogènes.

Segmentation semantique avec DeepLabV3. DeepLabV3 [3] obtient des performances comparables à celles d'autres modèles de pointe sur le benchmark de segmentation sémantique d'images PASCAL VOC 2012. Ce modèle est constitué d'une convolution dilatée avec des filtres suréchantillonnés pour extraire des cartes de caractéristiques denses et capturer le contexte à longue portée. Les sorties de DeepLabV3 sont présentées dans la Figure 3.

Estimation de la profondeur monoculaire avec MiDaS. L'architecture de MiDaS [4, 5] est également un encodeur-décodeur. Cette architecture est combinée avec des *transformers*. L'image d'entrée est traitée comme un "sac de mots". Les patchs d'image qui sont individuellement intégrés prennent le rôle de "mots". Une succession de *transformers* extrait les caractéristiques de ces patchs. Enfin, le décodeur réassemble toutes les caractéristiques pour obtenir la sortie finale qui est une carte de profondeur monoculaire.

Distilation de la profondeur. À notre connaissance la méthode mise en place par [7] est la seule qui prospoe d'introduire des information de profondeur dans la détection du flou de défocus. Ils transfèrent les connaissances d'un réseau d'estimation de la profondeur monoculaire par distillation. Grâce à cela, leur réseau est capable de produire deux sorties : une carte d'estimation de flou et l'autre l'estimation de la profondeur monoculaire. Selon eux la profondeur monoculaire aide à la l'estimation du flou reciproquement le flou aide à l'estimation de la profondeur.

3 Méthode proposée

Dans cet article, nous voulons évaluer l'impact de la segmentation sémantique et de l'estimation monoculaire de la profondeur sur la tâche de détection de flou. Nous prenons comme modèle de base le modèle de [6]. Pour faire une comparaison équitable, nous avons réentraîné le réseau en suivant leur protocole mais avec nos données. Ainsi, le modèle de base est entraîné à détecter le flou de défocus, le flou de mouvement et les pixels non flous.

Pour évaluer l'impact de la ségmentation sémantique, nous calculons d'abord les cartes de segmentation avec un réseau DeepLabV3 pré-entraîné [3]. Ensuite, nous ajoutons les cartes de sortie de segmentation à leurs images RVB originales correspondantes sous la forme d'un $4^{\text{ème}}$ canal. Enfin, nous entraînons le réseau à partir de zéro avec ces nouvelles entrées. Nous appliquons le même protocole pour l'estimation de la profondeur monoculaire. Nous calculons toutes les cartes de sortie avec MiDaS [4, 5], nous les ajoutons en tant que $4^{\text{ème}}$ canal et nous entraînons le réseau. La Figure 2 illustre ce protocle.

Nous avons maintenant trois réseaux entraînés avec la même architecture mais entraînés sur plusieurs types d'images. Tout d'abord, nous voulons évaluer le modèle comme un simple détecteur de flou de défocus. Tous les réseaux sont entraînés pour détecter deux types de flou, mais nous ne le testons que sur des images partiellement floues. Ici, nous confirmons ce que les auteurs de [7] montrent, la profondeur monoculaire aide à détecter le flou de défocus.

Enfin, nous allons voir si l'estimation monoculaire de la profondeur peut aider à détecter le flou de mouvement. Nous testons les trois réseaux sur des images partiellement floues de mouvement ou de défocus et sur des images contenant les deux types de flou. Nous comparons tous nos résultats avec les méthodes de référence [6] et [7] pour montrer comment l'estimation de la profondeur monoculaire renforce la tâche de détection du flou.

Datasets. Nous utilisons 3 jeux de données différents pour nos expérimentations. Le jeux de données public **CUHK** [12], le **Synthetic** [6] et le **DUT** [13]. Ils contiennent respectivement 1000 images (704 flou de défocus et 296 de flous de mouvement), 8460 images contenant à la fois du flou de défocus et de mouvement et 500 images uniquement avec du flou de défocus.

Entraînement et évaluation. L'encodeur est initialisé avec les poids de VGG-19 [14] pré-entraînés et le décodeur avec l'iniatialisation de Xavier [15]. Nous entraînons le réseau sur 800 images (564 flous de défocus et 236 de mouvement) du CUHK. Les tests sont réalisée sur les jeux de données CHUK200 (140 flous de défocus et 60 de mouvements), Synthetic et DUT.

Nous avons évalué l'intersection moyenne des unions

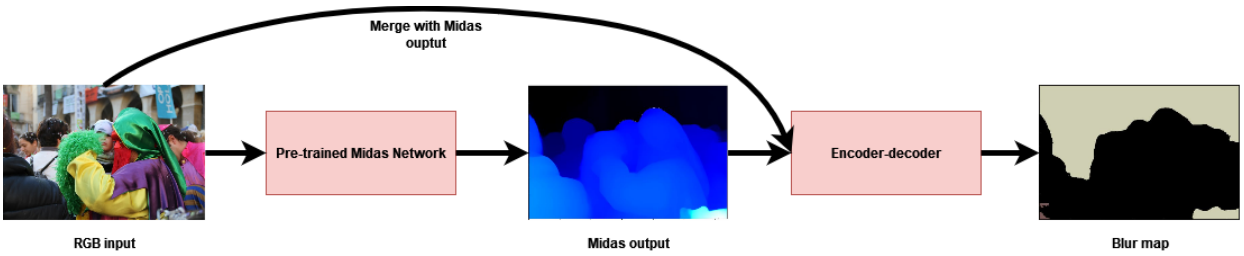

FIGURE 2 – À gauche, l'image RVB d'entrée, qui est partiellement floue. Son estimation de la carte de profondeur monoculaire par MiDaS (ou carte de segmentation de DeepLabV3) est ajoutée comme quatrième canal et utilisée pour entraîner le réseau encodeur-décodeur. À droite, la carte de flou estimée. Les pixels kaki sont flous et les pixels noirs sont nets.

TABLE 1 – Comparaison de mIoU entre les méthodes de pointe et la nôtre. Ici, mIoU est calculé sur la détection du flou de défocus.

Dataset	Baseline [6]	Depth [7]	w/MiDaS	w/DeepLab	Combiné
DUT500	0.84788	0,66324	**0.88468**	0.85918	0.88123

(mIoU), définie comme

$$mIoU = \frac{1}{N_c}\sum_c \left(\frac{B^c \cap G^c}{B^c \cup G^c}\right), \quad (1)$$

où B est une carte de flou estimée, G est la carte de vérité terrain et c est le type de flou et N_c est le nombre de types de flou considérés, il peut être égale à deux ou un.

4 Résultats et discussion

Le Tableau 1 compare les résultats obtenus pour la détection d'absence de mise au point. Il confirme que l'ajout d'informations de haut niveau provenant de la segmentation sémantique et de l'estimation de la profondeur monoculaire améliore la tâche de détection de la mise au point. La Figure 1 illustre cet impact. La prédiction du vélo est beaucoup plus précise lorsqu'on utilise ces informations de haut niveau.

Nous notons également que l'estimation de la profondeur monoculaire est plus efficace qu'une segmentation sémantique pour cette tâche. En effet, l'estimation de la profondeur monoculaire implique une sorte de segmentation sémantique de la scène, mais sans étiquettes de classe. Par rapport à une carte sémantique, une carte de profondeur est également plus directement liée au flou hors champs. De plus, tous les objets ne sont pas appris dans un réseau de segmentation sémantique. La Figure 3 illustre comment des objets sont manqués. En ce sens, l'estimation monoculaire de la profondeur est plus fiable car elle repose sur des caractéristiques de bas niveau.

Dans le Tableau 2, nous considérons les deux types de flou. L'estimation monoculaire de la profondeur surpasse légèrement les méthodes de segmentation baseline et sémantique sur CUHK200 et DUT. Ceci est conforme à ce que nous avons observé avec les résultats dans le Tableau 1. Les résultats obtenus sur l'ensemble de données synthétiques soulignent le fait que la segmentation sémantique ne donne des informations que sur les objets sur lesquels elle a été entraînée.

TABLE 2 – Comparaison mIoU de la détection du flou pour chaque ensemble de données. nous détectons les pixels flous de defocus, flous de mouvement et nets.

Dataset	Baseline [6]	w/MiDaS	w/DeepLab	Combiné
CUHK200	0.82898	**0.83134**	0.81245	0.82523
DUT500	0.84788	**0.88468**	0.85918	0.88123
Synthetic	**0.91248**	0.90877	0.64896	0.88333

Dans de nombreux cas d'échec, nous ne détectons aucun objet dans la scène. La méthode de base est également légèrement meilleure que l'estimation de la profondeur monoculaire sur le jeu de données synthétique. L'estimation monoculaire de la profondeur ne fournit pas d'informations de haut niveau permettant de détecter le flou de mouvement comme c'est le cas pour le flou de défocus. L'estimation de la profondeur monoculaire ne parvient pas à distinguer le mouvement du flou de défocus dans les cas ambigus. Dans ce jeu de données, l'état de l'art est légèrement meilleure que l'estimation de la profondeur monoculaire.

La détection du flou est une tâche difficile, rendue encore plus ardue par le manque de grands ensembles de données de bonne qualité disponibles publiquement avec une vérité de terrain fiable. Comme les ensembles de données CUHK et DUT disponibles ne contiennent pas suffisamment d'images pour entraîner un réseau complet, nous avons dû utiliser des réseaux pré-entraînés. En outre, nous avons observé de nombreuses erreurs d'annotation, notamment pour la distinction entre le flou de mise au point et le flou de mouvement.

Le Tableau 1 montre que l'ajout de cartes de profondeur améliore la prédiction. De plus, on voit également que la méthode [7] est bien en dessous des autres, ce qui nous amène à penser qu'elle a du mal à généraliser sur un autre dataset.

Enfin nous avons entraîné un dernier réseau en combinant à la fois les cartes d'estimation de la profondeur et de ségmentation sémantique. Les résultats correspondant aux colones "combiné" des Tableaux 2 et 1 sont proche de ceux obtenu uniquement avec la carte d'estimation de la profondeur. Ce qui renforce l'idée que la ségmentation sémantique n'apporte rien de plus.

FIGURE 3 – Résultats de la détection du flou pour les images de chaque ensemble de données. Dans l'ordre lexicographique, nous montrons : l'image d'entrée, la carte de profondeur correspondante [5] et la carte de segmentation [3]. Puis les prédictions obtenues en utilisant uniquement l'image RVB, en incorporant la carte de profondeur, et en incorporant la carte de segmentation. Le noir indique l'absence de flou, le marron le flou de mouvement et le kaki le flou de défocus.

5 Conclusion

Les informations de haut niveau provenant de l'estimation monoculaire de la profondeur constituent un outil puissant pour la détection du flou de défocus. Elle n'est pas d'une grande aide pour détecter le flou de mouvement, car ce flou n'est pas caractérisé par les mêmes caractéristiques de haut et bas niveau que le flou de défocus. Le plus grand défi à la détection du flou vient du manque de jeux de données, combiné à l'inévitable ambiguïté de la vérité terrain. En effet, seule la sémantique peut permettre de prendre une décision sur des régions homogènes, qui, prises isolément, pourraient être indifféremment floues ou nettes. Même pour un humain, il est difficile de trancher dans certains cas. Il pourrait donc être intéressant de demander au réseau une carte supplémentaire de "confiance", ou même de créer une étiquette "je ne sais pas". On notera également qu'avec notre méthode la complexité et le temps d'exécution augmente grandement par rapport aux autres méthodes. Une possibilité serai de remplacer l'encoder par un réseaux pré-entraîné pour la profondeur monoculaire en choisissant les meilleures caractéristiques à extraire.

Remerciements Ce travail a été partiellement financé par la subvention N00014-17-1-2552 de l'Office of Naval research, et le MENRT. Ce travail a bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2022-AD011012472R1 attribuée par GENCI.

Références

[1] J. Anger, G. Facciolo, and M. Delbracio, "Estimating an Image's Blur Kernel Using Natural Image Statistics, and Deblurring it : An Analysis of the Goldstein-Fattal Method," *IPOL*, vol. 8, pp. 282–304, 2018.

[2] J. Anger, C. de Franchis, and G. Facciolo, "Assessing the sharpness of satellite images : Study of the planetscope constellation," in *IGARSS*. IEEE, 2019, pp. 389–392.

[3] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017.

[4] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation : Mixing datasets for zero-shot cross-dataset transfer," *TPAMI*, 2020.

[5] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," *ArXiv preprint*, 2021.

[6] K. Beomseok, H. Son, S.-J. Park, S. Cho, and S. Lee, "Defocus and motion blur detection with deep contextual features," *Computer Graphics Forum*, vol. 37, pp. 277–288, 10 2018.

[7] X. Cun and C.-M. Pun, "Defocus blur detection via depth distillation," in *ECCV*. Springer, 2020, pp. 747–763.

[8] R. Liu, Z. Li, and J. Jia, "Image partial blur detection and classification," *CVPR*, pp. 1–8, 07 2008.

[9] A. Chakrabarti, T. Zickler, and W. T. Freeman, "Analyzing spatially-varying blur," in *CVPR*. IEEE, 2010, pp. 2512–2519.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab : Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *TPAMI*, vol. 40, no. 4, pp. 834–848, 2017.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet : A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[12] J. Shi, L. Xu, and J. Jia, "Discriminative blur detection features," in *CVPR*, 2014, pp. 2965–2972.

[13] W. Zhao, F. Zhao, D. Wang, and H. Lu, "Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network," in *CVPR*, 2018, pp. 3080–3088.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv :1409.1556*, 2014.

[15] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS*. JMLR, 2010, pp. 249–256.