Approches d'adaptation de domaine non supervisées pour la localisation de personnes dans la salle d'opération

Le candidat:	Vinkle Kumar Srivastav
Directeur de thése:	Prof. Nicolas Padoy, PhD
Co-directeur de thése:	Prof. Afshin Gangi, MD
Laboratoire:	CAMMA, AVR, ICube
École doctorale:	MSII, Université de Strasbourg, France

1 Motivation

Les approches de localisation de personne à haut niveau de granularité, telles que l'estimation de la pose humaine à partir de points clés (HPE) ou la segmentation d'instance sur les pixels peuvent fournir dans la salle d'opération (OR) [7, 8] des informations sémantiques de haut niveau sur les cliniciens et les patients. Les nouveaux systèmes d'assistance contextuelle pour la salle d'opération moderne peuvent utiliser ces modèles pour faciliter d'assistance applications telles que la surveillance

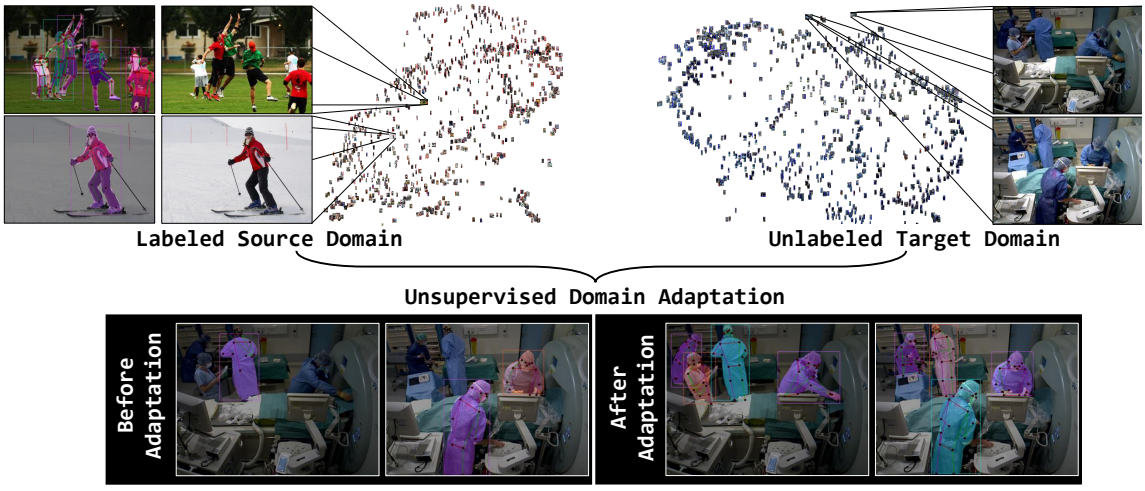

Figure 1: La figure montre les différences visuelles globales et au niveau de l'instance entre les images naturelles du *domaine source* et les images OR du *domaine cible*. Les approches UDA visent à généraliser un modèle à travers différents domaines avec accès aux données étiquetées du domaine source et aux données non étiquetées du domaine cible. Les clusters séparés du domaine source et des images du domaine cible sont obtenus en exécutant une technique de réduction de dimension : Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [6]. Les images des domaines source et cible sont respectivement un sous-ensemble des ensembles de données COCO et MVOR.

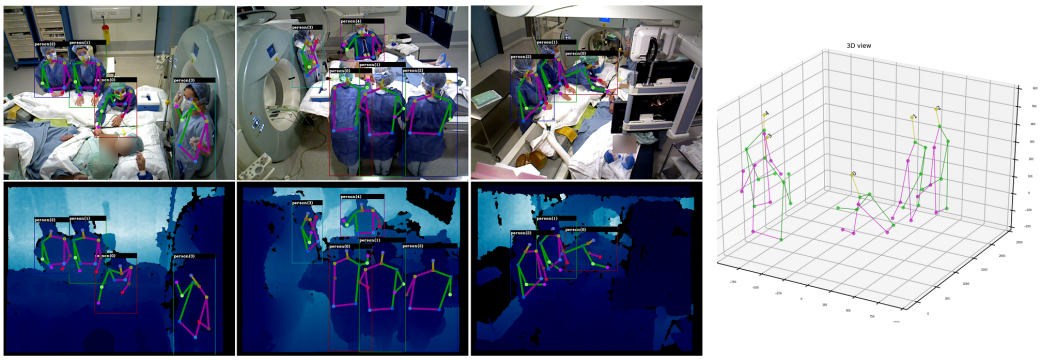

Figure 2: Une image multi-vues du jeu de données *MVOR+* [1] avec la visualisation de la vérité terrain 2D et 3D.

des activités, la compréhension de la dynamique d'équipe et la surveillance des risques radiologiques. La surveillance de l'exposition aux rayonnements [9], par exemple, pourrait utiliser de tels modèles pour comprendre quel pixel ou point clé des cliniciens est exposé à des rayonnements nocifs. Bien que l'émergence des techniques d'apprentissage profond ait conduit à des modèles très précis pour les images naturelles *in the wild*, la localisation des personnes dans la salle d'opération pose toujours des scénarios difficiles. Les données OR ont une complexité visuelle élevée en raison de l'occlusion et de l'encombrement de divers équipements, des vêtements amples portés par les cliniciens, de la proximité des personnes présentes sur la scène et du changement de distribution des données par rapport aux images naturelles, comme le montre la figure 2. Les méthodes d'apprentissage profond actuelles, qui fonctionnent remarquablement bien pour les images naturelles, montrent une précision réduite lorsqu'elles sont appliquées aux images OR en raison des différences de domaine visuel globales et au niveau de l'instance, comme le montre la figure 1. Les ensembles de données supervisés à grande échelle servent de cheval de bataille pour les modèles d'apprentissage profond hautement précis et déployables. Cependant, dans le cas de la salle d'opération, l'obtention d'une annotation de vérité terrain n'est pas triviale, et l'utilisation de techniques de crowdsourcing telles qu'Amazon Mechanical Turk est peu pratique en raison des problèmes de confidentialité des patients et des cliniciens. Les approches, par conséquent, qui peuvent adapter un modèle au domaine cible invisible et non étiqueté tel que la OR tout en préservant la confidentialité sont des poursuites prometteuses. Les approches d'adaptation de domaine non supervisé (UDA) visent à généraliser un modèle dans différents domaines avec accès aux données étiquetées du domaine source et aux données non étiquetées du domaine cible, voir l'exemple dans la figure 1. Cette thése explore plusieurs approches d'adaptation de domaine non supervisée (UDA) pour l'estimation HPE et la segmentation d'instances de personnes, comme indiqué ci-dessous.

2 Contributions

Nous commençons par la génération d'un nouvel ensemble de données de salle d'opération multi-vues (*MVOR+* [1]). Il se compose de 732 images multi-vues échantillonnées à partir de quatre jours d'enregistrement en salle d'opération, et les annotations manuelles de vérité terrain consistent en 4699 boîtes d'encombrement délimitant les personnes, 2926 poses 2D du haut du corps et 1061

poses 3D du haut du corps. Un exemple de image multi-vues avec des annotations de vérité terrain est montré dans la figure 2. L'ensemble de données *MVOR+* est destiné à servir de banc d'essai pour évaluer les algorithmes de détection de personnes et HPE. Nous avons également évalué nos approches sur le jeu de données *TUM-OR-test* [8] contenant 2400 images avec 11611 instances de personnes.

Comme le jeu de données *MVOR+* est petit et inadapté à l'entraînement de modèles de deep learning, nous exploitons également un jeu de données non étiqueté, *MVOR-unlabelled*, provenant de la même OR mais capturé à des jours différents de MVOR. Il se compose de 160k images d'images de couleur et de profondeur synchronisées (80k couleur, 80k profondeur). Nous abordons plusieurs scénarios pour HPE au bloc opératoire et proposons à chaque fois des méthodes pour exploiter les données non annotées, comme décrit ci-dessous.

HPE sur les images de profondeur basse résolution préservant la confidentialité [2]

Nous proposons une architecture HPE multi-personnes pour des images de profondeur basse résolution afin de répondre aux préoccupations de confidentialité des patients et des cliniciens. Le modèle proposé a une précision comparable à celle du modèle conçu pour les images à haute résolution. Nous proposons également un moyen automatisé de générer des pseudo-étiquettes sur les images de profondeur de *MVOR-unlabelled* à toutes les résolutions en utilisant un estimateur de pose humaine à plusieurs étages très précis sur les images couleur haute résolution de *MVOR-unlabelled* .

Auto-supervision sur images OR non étiquetées pour HPE 2D/3D [3]

Nous proposons un modèle léger en une seule étape pour HPE 2D/3D conjoint sur les images couleur en utilisant la distillation des connaissances dans un cadre enseignant/étudiant pour exploiter les connaissances présentes dans un ensemble de données non étiqueté à grande échelle - *MVOR-unlabelled* - et un réseau d'enseignants précis mais complexe à plusieurs niveaux. Nous proposons également un schéma d'entraînement simple et efficace pour adapter le modèle sur des images couleur OR basse résolution préservant la vie privée. Les résultats sur l'ensemble des données *MVOR+* montrent que le réseau d'étudiants proposé fonctionne à égalité avec un réseau d'enseignants complexe, bien qu'il s'agisse d'un réseau léger et à une seule étape. En outre, il peut également donner des résultats précis sur des images à faible résolution, au besoin pour garantir la confidentialité, même en utilisant un taux de sous-échantillonnage de 12x.

UDA pour l'estimation de la pose d'une personne et la segmentation des instances [4]

Nous abordons la tâche conjointe d'estimation de pose de personne et de segmentation d'instance et proposons une approche UDA pour adapter un modèle entraîné sur le domaine source étiqueté au domaine cible non étiqueté. Nous utilisons un cadre d'*auto-entraînement* avec des augmentations de données avancées pour générer des pseudo-étiquettes pour le domaine cible non étiqueté. L'efficacité de haute qualité des pseudo-étiquettes est assurée en imposant des contraintes géométriques explicites aux différentes augmentations sur l'image d'entrée non étiquetée. Nous utilisons également des couches de normalisation séparées pour le domaine source et cible et le paradigme *enseignant-étudiant* pour stabiliser l'entraînement. L'évaluation de la méthode sur les deux ensembles de données du domaine cible, *MVOR+* et *TUM-OR-test*, avec des études d'ablation approfondies, suggère l'efficacité de notre approche. Nous démontrons en outre que l'approche proposée peut être

adaptée efficacement aux images basse résolution du domaine cible, au besoin pour assurer la confidentialité, même jusqu'à un facteur de sous-échantillonnage de 12x. La figure 1 montre un résultat qualitatif comparatif avant et après l'adaptation du domaine. Les résultats nettement meilleurs sur les images à basse résolution sont particulièrement encourageants à utiliser de tels modèles pour l'environnement OR sensible à la confidentialité.

Détection de visage au bloc opératoire [5]

Nous étudions également les performances de modèles de détection de visage de pointe sur des images OR et améliorons la précision de l'un des détecteurs de visage de pointe par une approche auto-supervisée itérative.

Publications

[1] V. Srivastav, T. Issenhuth, K. Abdolrahim, M. de Mathelin, A. Gangi, and N. Padoy, "Mvor: A multi-view rgb-d operating room dataset for 2d and 3d human pose estimation," in *MICCAI-LABELS workshop*, 2018.

[2] V. Srivastav, A. Gangi, and N. Padoy, "Human pose estimation on privacy-preserving low-resolution depth images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2019, pp. 583–591.

[3] V. Srivastav, A. Gangi, and N. Padoy, "Self-supervision on unlabelled or data for multi-person 2d/3d human pose estimation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2020, pp. 761–771.

[4] V. Srivastav, A. Gangi, and N. Padoy, "Unsupervised domain adaptation for clinician pose estimation and instance segmentation in the or," in *submission to Medical Image Analysis*, 2021.

[5] T. Issenhuth, V. Srivastav, A. Gangi, and N. Padoy, "Face detection in the operating room: Comparison of state-of-the-art methods and a self-supervised approach," *IJCARS*, vol. 14, no. 6, pp. 1049–1058, 2019, (Runner Up award in the Bench-to-Bedside category at IPCAI 2019).

Autres références

[6] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[7] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin, and N. Padoy, "A multi-view rgb-d approach for human pose estimation in operating rooms," in *WACV*, IEEE, 2017, pp. 363–372.

[8] V. Belagiannis, X. Wang, H. B. B. Shitrit, K. Hashimoto, R. Stauder, Y. Aoki, M. Kranzfelder, A. Schneider, P. Fua, S. Ilic, *et al.*, "Parsing human skeletons in an operating room," *MVA*, vol. 27, no. 7, pp. 1035–1046, 2016.

[9] N. L. Rodas, F. Barrera, and N. Padoy, "See it with your own eyes: Markerless mobile augmented reality for radiation awareness in the hybrid room," *IEEE-TBE*, vol. 64, no. 2, pp. 429–440, 2017.