RUMPL : Transformeurs basés sur des rayons pour le lifting universel de pose humaine 2D vers 3D multi-vues

PO 15

Seyed Abolfazl Ghasemzadeh, Christophe De Vleeschouwer
ICTEAM UCLouvain, {seyed.ghasemzadeh, christophe.devleeschouwer}@uclouvain.be

La Problématique

- La capture multi-vues de l'activité (ex. Fig 2) est complexe (occultations, passage 2D vers 3D).
- Les modèles actuels peinent à généraliser ces poses du monde réel vers un squelette 3D fidèle.

L'Objectif

- Permettre une estimation de pose 3D universelle et robuste.
- Assurer un déploiement immédiat dans des environnements et scènes arbitraires.
- Garantir l'indépendance vis-à-vis de la calibration des caméras.

La Méthode

- **Représentation par Rayons 3D :** Conversion des points clés 2D en rayons dans l'espace mondial pour s'affranchir des paramètres de caméra.
- **Architecture Transformeur :** Utilisation d'un *View Fusion Transformer* (VFT) pour agréger les rayons de N vues via un jeton fusionné appris.
- **Générateur MHP :** Création de paires 2D-3D synthétiques à partir de maillages AMASS.

Figure 1 : Générateur de jeu de données 3D

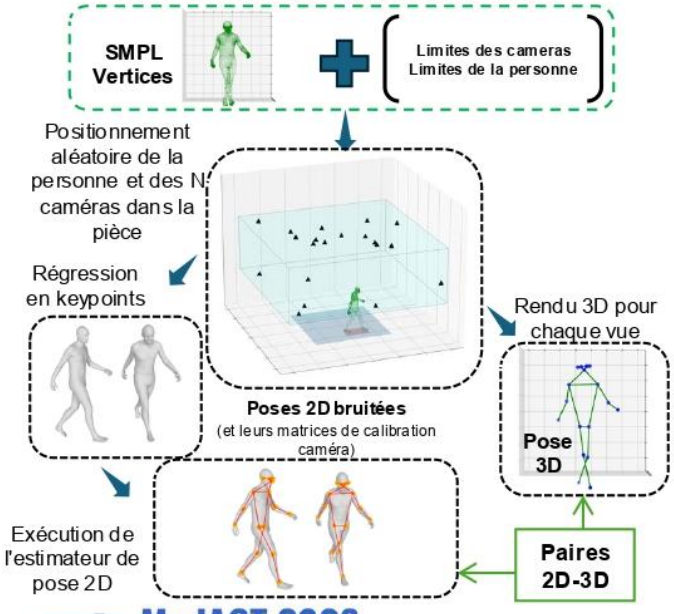

Figure 2 : Architecture de RUMPL

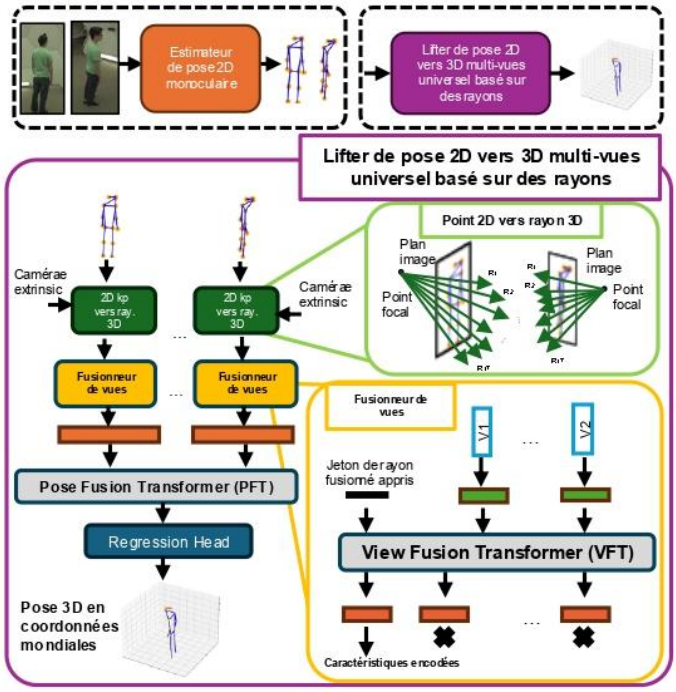

Tableau 1 : Résultats

En CMU [1] (2 vues)			
Method	MPL [3]	Triangulation	**RUMPL**
MPJPE (mm) ↓	55,7	44,0	**35,0**
En RICH [2] (2 vues)			
Method	MPL [3]	Triangulation	**RUMPL**
MPJPE (mm) ↓	75,3	54,5	**48,4**

Le MPJPE mesure l'écart moyen entre les articulations prédites et réelles.

Conclusion

RUMPL introduit une représentation innovante par rayons 3D qui découple l'estimation de pose de la calibration des caméras. En apprenant à partir de données synthétiques randomisées, notre méthode atteint une véritable universalité, permettant une estimation robuste et précise dans des conditions réelles non contrôlées.

Références
[1] Ionescu et al., "Human3.6m …", IPAMI, 2014.
[2] Joo et al., "Panoptic Studio …", ICCV, 2015.
[3] Ghasemzadeh et al. "MPL …", ECCV, 2024.

ModACT 2026
20-22 Mai 2026
Paris

RUMPL : Transformeurs basés sur des rayons pour le lifting universel de pose humaine 2D vers 3D multi-vues

Seyed Abolfazl GHASEMZADEH, UCLouvain, ICTEAM/ELEN, Seyed.ghasemzadeh@uclouvain.be

Christophe DE VLEESCHOUWER, UCLouvain, ICTEAM/ELEN, Christophe.devleeschouwer@uclouvain.be

Mots-clefs. Estimation de pose humaine 3D, Multi-vues, Transformeurs, Représentation par rayons, Vision par ordinateur, Apprentissage profond.

Contexte et problématique

L'estimation de la pose humaine en 3D à partir d'images 2D est un problème fondamentalement mal posé en raison des occultations et de l'ambiguïté projective. Les approches multi-vues permettent d'atténuer ces difficultés en agrégeant les informations visuelles provenant de plusieurs caméras. Cependant, l'entraînement de ces modèles nécessite généralement des ensembles de données multi-vues annotés avec une vérité terrain 3D, qui sont rares et limités à des environnements de laboratoire très contraints. Par conséquent, la plupart des méthodes existantes peinent à se généraliser à de nouvelles scènes ou à de nouvelles configurations de caméras dans des conditions réelles.

Question de recherche

Comment concevoir un modèle d'estimation de pose 3D multi-vues qui soit véritablement universel, c'est-à-dire capable de fonctionner avec n'importe quel nombre de vues et n'importe quelle configuration de caméras, sans nécessiter de réentraînement ou d'ajustement (fine-tuning) pour chaque nouvelle scène ?

Méthodologie

Nous proposons RUMPL (Ray-based Universal Multi-view Pose Lifter), un réseau basé sur des Transformeurs. La méthode procède en deux étapes : des points clés 2D sont d'abord extraits indépendamment de toutes les images disponibles via un estimateur de pose standard, puis ils sont convertis en 3D. L'innovation majeure réside dans la représentation des points clés 2D sous forme de rayons 3D dans un système de coordonnées mondial. Cette formulation géométrique rend le modèle agnostique aux paramètres intrinsèques des caméras, bien qu'il s'appuie sur leur orientation dans un référentiel mondial. En pratique, les paramètres de calibration sont considérés comme des données d'entrée : la représentation par rayons encode nativement ces informations intrinsèques et extrinsèques dès l'entrée du réseau. Cela permet au système de traiter une géométrie universelle plutôt que des pixels dépendants d'un capteur spécifique. Un "View Fusion Transformer" agrège ensuite les informations le long de ces rayons pour gérer un nombre arbitraire de vues. Pour garantir cette universalité, le modèle est entraîné uniquement sur des paires 2D-3D synthétiques générées à partir de maillages 3D projetés sous des configurations de caméras totalement aléatoires.

Résultats

RUMPL a été évalué sur des jeux de données standards (Human3.6M, CMU Panoptic, et RICH) sans qu'aucune image de ces jeux n'ait été vue lors de l'entraînement. Notre approche réduit l'erreur moyenne par articulation (MPJPE) jusqu'à 53 % par rapport à la méthode classique de triangulation, et de plus de 60 % par rapport aux approches de référence basées sur des transformeurs utilisant des représentations d'images. Les expériences confirment la robustesse de RUMPL (Ghasemzadeh, Alahi, & De Vleeschouwer, 2025), sa flexibilité face à des configurations de caméras invisibles à l'entraînement, et sa scalabilité dans des scénarios complexes.

Références

Ghasemzadeh, S. A., Alahi, A., & De Vleeschouwer, C. (2025). RUMPL: Ray-Based Transformers for Universal Multi-View 2D to 3D Human Pose Lifting. *arXiv preprint* (pp. 1-12). arXiv, Ithaca.

Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., & Black, M. J. (2019). AMASS: Archive of motion capture as surface shapes. *International Conference on Computer Vision* (pp. 5442-5451). IEEE, Seoul.