Évaluation de la qualité de nuages de points 3D sans référence en utilisant un transformer et la saillance visuelle

Salima BOURBIA[1,2] Ayoub KARINE[2] Aladine CHETOUANI[3] Mohammed EL HASSOUNI[1] Maher JRIDI[2]

[1]FLSH, FSR, Mohammed V University in Rabat, Morocco.

[2]L@bISEN, Vision-AD, ISEN Yncréa Ouest, 33 Quater Chemin du Champ de Manœuvre, 44470 Carquefou, France.

[3]Laboratoire PRISME, Université d'Orléans, France.

Résumé – Dans ce travail, nous proposons une approche basée sur l'apprentissage profond qui bénéficie de l'avantage du mécanisme d'auto-attention dans les Transformers pour prédire avec précision le score de qualité perceptuelle des nuages de points (3DPC) dégradés. De plus, nous avons introduit l'utilisation de cartes de saillance pour refléter le comportement du système visuel humain qui est attiré par certaines régions spécifiques par rapport à d'autres lors de l'évaluation. Pour ce faire, en utilisant une caméra virtuelle avec des angles pré-définis par rapport à l'objet, chaque 3DPC est projeté sous différentes vues 2D. Ensuite, nous pondérons les images projetées obtenues avec leurs cartes de saillance correspondantes. Après cela, nous éliminons la majorité des informations de fond en extrayant des patchs saillants. Ces derniers sont envoyés en entrée d'un modèle Vision Transformer (ViT) afin d'extraire les informations contextuelles globales et de prédire les scores de qualité des patchs. Enfin, nous calculons la moyenne des scores de tous les patchs saillants pour obtenir la qualité finale du 3DPC. Les performances de notre modèle ont été évaluées sur deux bases de données de l'évaluation de la qualité de nuages de points 3D : ICIP2020 et SJTU. Les résultats expérimentaux montrent que notre modèle atteint de bonnes performances par rapport aux méthodes de l'état de l'art.

Abstract – In this work, we propose a deep learning-based approach that benefits from the advantage of the self-attention mechanism in Transformers to accurately predict the perceptual quality score of degraded point clouds (3DPC). Additionally, we introduce the use of saliency maps to reflect the behavior of the human visual system, which is attracted to certain specific regions compared to others during evaluation. To do this, using a virtual camera with pre-defined angles relative to the object, each 3DPC is projected onto different 2D images (i.e. views). We then weight the projected images obtained with their corresponding saliency maps. After that, we remove majority of the background information by extracting salient patches. The latter are then introduced to the Vision Transformer (ViT) to extract global contextual information and predict patch quality scores. Finally, we calculate the average of scores of all salient patches to obtain the final quality of the 3DPC. The performance of our model has been evaluated on two 3D point cloud quality evaluation databases: ICIP2020 and SJTU. Experimental results show that our model achieves good performance compared to state-of-the-art methods.

1 Introduction

L'avancement rapide des technologies d'acquisition 3D, notamment les capteurs RGB-D et LiDAR, ont propulsé l'utilisation des nuages de points (3DPC) dans divers domaines tels que la robotique, la numérisation 3D et la réalité augmentée. Ces technologies fournissent des informations visuelles hautement réalistes et offrent des expériences immersives aux utilisateurs. Un 3DPC est une collection de points irréguliers représentant un objet ou une scène en 3D, chaque point contient des coordonnées géométriques qui indiquent sa position dans un système de coordonnées spécifique (x,y,z). De plus, des attributs associés optionnels tels que la couleur et les vecteurs normaux de surface décrivent l'apparence visuelle de chaque point. Dans le présent article, nous nous focalisons sur les objets 3D dont le traitement implique plusieurs étapes telles que l'acquisition, la compression, la transmission et le rendu. Chacune de ces étapes peuvent dégrader le rendu visuel du 3DPC. Par conséquent, il est important de disposer de méthodes aptes à prédire automatiquement la qualité perceptuelle des 3DPC dégradés. Dans la littérature, ces méthodes peuvent être classées en trois grandes familles en fonction de la disponibilité du 3DPC original ou de référence. Les méthodes avec référence (FR) et avec référence réduite (RR) nécessitent toutes ou une partie des informations du 3DPC de référence, respectivement, tandis que les méthodes sans référence (NR) n'utilisent que l'objet déformé pour évaluer sa qualité visuelle.

Les premières méthodes qui ont été proposées sont des méthodes FR basées sur les aspects de distorsion géométrique, notamment les méthodes point à point (P2P) [9, 4], point à plan (P2Pl) [12, 4], plan à plan (Pl2Pl) [1] et point à distribution (P2D-D) [6]. Ces méthodes calculent l'écart de distance entre les 3DPCs de référence et les 3DPCs déformés. Par la suite, Meynet *et al.* [10] ont proposé la métrique Point Cloud Quality Metric (PCQM) qui calcule un ensemble de caractéristiques dont les courbures et la couleur qui sont ensuite combinées par un modèle linéaire pour indiquer le niveau de dégradation de 3DPC. Alexiou *et al.* [2] se sont inspirés de la métrique de similarité structurelle (SSIM) 2D pour proposer la méthode PointSSIM. Les auteurs ont exploité les caractéristiques de géométrie, de couleur, de normale et de courbure des régions locales du 3DPC pour calculer le score de qualité en se basant sur une similarité entre le 3DPC de référence et le 3DPC déformé. Viola *et al.* [15] combinent linéairement la statistique de couleur et les caractéristiques géométriques extraites du 3DPC de référence et du 3DPC déformé pour estimer le score de qualité de l'objet déformé. Les mêmes auteurs ont

proposé une métrique RR [14] qui compare les caractéristiques extraites du côté du récepteur pour réduire la dépendance aux informations de référence. Dans les scénarios réels, les informations complètes ou réduites du 3DPC de référence ne sont pas toujours disponibles. Plusieurs méthodes NR ont été proposées pour prédire la qualité visuelle à partir de 3DPC déformés. Parmi ces méthodes, on peut citer la méthode de Yang *et al.* [16] qui représente le 3DPC avec six images de texture et de profondeur, puis agrège les caractéristiques extraites de ces cartes pour obtenir un indice de qualité du 3DPC. Liu *et al.* [8] ont proposé un réseau d'évaluation de la qualité des nuages de points (PQA-Net) qui se compose d'une étape d'extraction des caractéristiques des projections multi-vues de l'objet 3D, suivi par un réseau multitâche qui classe les types de distorsion et qui prédit la qualité perceptive du 3DPC dégradé avec un seul type de distorsion. Dans le même contexte, nous proposons une méthode NR basée sur la saillance visuelle et un modèle Transformer. La saillance permet de révéler les régions sur lesquelles l'œil humain se concentre et qui attirent son attention, tandis que le modèle Transformer a pour objectif d'extraire les informations contextuelles globales du 3DPC dégradé afin d'évaluer sa qualité perceptuelle. Le reste de ce document est organisé comme suit : La méthode proposée est décrite en détails dans la section 2. Les résultats expérimentaux sont présentés dans la section 3. La conclusion et les perspectives sont données dans la section 4.

2 Méthode proposée

La figure 1 présente le schéma global de la méthode proposée qui est basée sur deux étapes principales : le pré-traitement et l'estimation objective du score de qualité. Dans l'étape du pré-traitement, nous projetons le 3DPC dégradé en des vues 2D. Les patchs les plus saillants sont extraits à partir de ces projections en utilisant la saillance visuelle. L'étape suivante d'estimation objective de la qualité du 3DPC consiste à prédire le score de qualité de chaque patch extrait et ensuite d'en déduire le score de qualité global du 3DPC.

2.1 Pré-traitement

La première étape de la méthode proposée consiste à projeter chaque objet 3DPC déformé sous forme de vues 2D prises à différents angles. Étant donné un 3DPC, nous générons N vues 2D en utilisant une projection perspective qui incorpore les distorsions de géométrie et de couleur, tout en imitant le système visuel humain lors de l'évaluation de la qualité du 3DPC à partir de différents points de vue. Des caméras virtuelles sont fixées à différents angles pour entourer le 3DPC. Il est important de mentionner que le centre de chaque 3DPC est dans la même position que le système de coordonnées sphériques d'origine de la caméra virtuelle. Les coordonnées des caméras virtuelles $(r, \theta_{el}, \phi_{az})$ sont obtenues en faisant varier l'angle d'azimut $\phi_{az} \in [0, 2\pi]$ par $\frac{\pi}{24}$ et en fixant l'angle d'élévation θ_{el} à zéro. En plus de ces vues, nous ajoutons également les vues supérieures et inférieures de l'objet. La distance r entre la caméra et le 3DPC varie en fonction de la taille de l'objet. En conséquence, nous avons obtenu, pour chaque 3DPC, 50 vues de projection de 512×512 pixels. Un exemple de projection d'un objet 3D est illustré dans la figure 2.

À partir de chaque projection résultante, nous calculons la carte de saillance correspondante. Cette dernière reflète les régions qui attirent le plus l'attention humaine. Selon tong *et al.* [13], l'inconfort visuel causé par les distorsions du 3DPC dépend fortement de ces zones spécifiques contenant des informations de saillance. Dans notre travail, nous nous appuyons sur l'utilisation de la saillance visuelle en pondérant les images projetées à partir du 3DPC avec leurs cartes de saillance correspondantes par une opération de multiplication élément par élément. Un exemple de ces étapes est donné à la Figue 3. Dans ce contexte, les cartes de saillance sont calculées en utilisant la méthode Kroner *et al.* [7], qui exploite la capacité de l'encodeur d'images VGG16 à extraire des caractéristiques pertinentes à partir d'images brutes et à les décrire en une distribution de saillance dans des scènes arbitraires, en attribuant une forte saillance aux régions contenant des informations sémantiques.

2.2 Estimation du score de qualité objective

Le défi de cette étape est de permettre à un modèle profond d'extraire, dès ses premières couches, les informations contextuelles globales des patchs saillants d'entrée qu'on a obtenu à partir de l'étape précédente. Pour ce faire, nous avons utilisé le modèle de Vision Transformer (ViT) [3] que nous avons adapté pour estimer la qualité visuelle du 3DPC. Tout d'abord, les patchs saillants obtenus d'une taille de 224×224 pixels sont projetés en encodage de patchs grâce à une couche cachée de convolution. Nous concaténons les encodages de patchs obtenus $[P_1, ..., P_n] \in R^{nxD}$ avec les encodages de position apprenables $[Pos_1, ..., Pos_n] \in R^{nxD}$ afin de préserver l'information de position de chaque patch dans l'image d'entrée. De plus, nous utilisons un token d'encodage apprenable supplémentaire P_{QE}, qui est concaténé avec un encodage positionnel Pos_0. L'utilisation de P_{QE} permet d'agréger la représentation visuelle globale de la qualité d'image perçue indépendamment de la qualité de chaque patch spécifique individuellement. Ensuite, la séquence de tokens est utilisée en tant qu'entrée de l'encodeur Transformer qui est composé de L couches identiques empilées. Chaque couche de l'encodeur est une combinaison d'une sous-couche d'auto-attention multi-têtes et d'une sous-couche à propagation avant. La normalisation et les connexions résiduelles sont appliquées entre chaque sous-couche. Enfin, le premier token dans la sortie de l'encodeur Transformer qui contient l'information contextuelle globale de la qualité perceptive du patch saillant est utilisé comme entré d'une couche entièrement connectée pour prédire le score de la qualité visuelle du patch saillant d'entrée. Nous soulignons que nous entraînons le modèle Transformer en utilisant la fonction de perte L_1, définie comme suit :

$$\mathcal{L}oss = \frac{1}{K}\sum_{k=1}^{K}|PMOS_k - MOS_k| \quad (1)$$

où $PMOS$ et MOS représente les scores prédits et les scores de vérités terrain des patchs saillants respectivement. Pour obtenir le score global de l'objet 3D, la moyenne des scores de qualité des patchs saillants est calculée.

3 Résultats expérimentaux

3.1 Base de données et protocole de validation

Pour évaluer la performance de la méthode proposée et la comparer avec les métriques de l'état de l'art, nous utilisons deux jeux de bases de données :

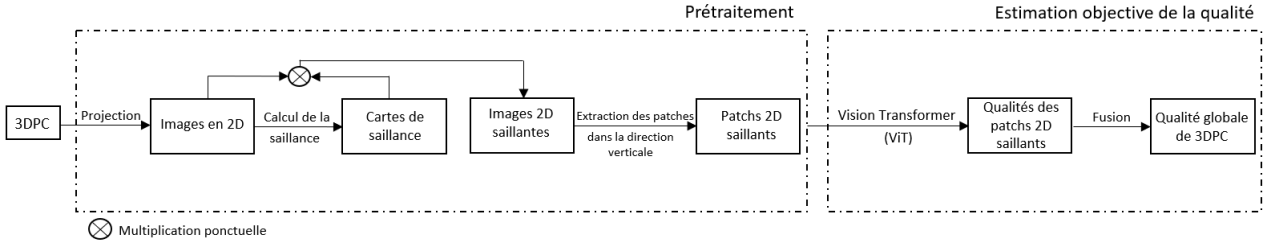

FIGURE 1 : Architecture globale de la méthode proposée.

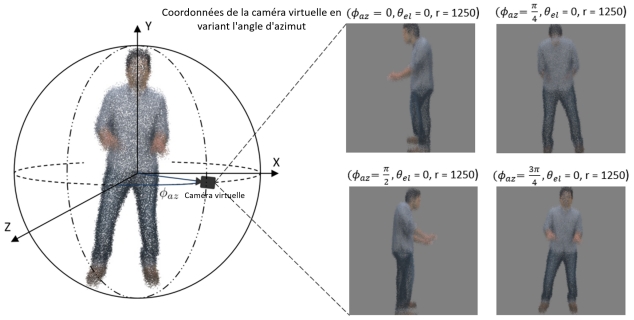

FIGURE 2 : Exemple de vues projetées à partir d'un 3DPC dégradé de la base de données SJTU. Les coordonnées de la caméra virtuelle sont r, θ_{el} et ϕ_{az}. r est le rayon, θ_{el} représente l'angle d'élévation et ϕ_{az} est l'angle d'azimut

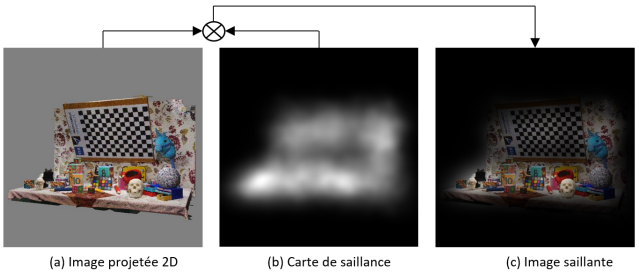

FIGURE 3 : Exemple d'une image saillante résultante (c) d'une opération de multiplication élément par élément entre une image projetée de l'objet ULB Unicorn de la base de données SJTU (a) et sa carte saillante correspondante (b).

- **ICIP2020** [11] contient 6 objets 3DPC de référence et 90 versions dégradées par 3 types de distorsion de compression : G-PCC Octree, G-PCC Trisoup et V-PCC avec 5 niveaux de dégradation de qualité différents.
- **SJTU** [16] contient 378 objets 3DPC dégradés, dérivés de 9 3DPC de référence. Chaque 3DPC est dégradé par 7 types de distorsion avec 6 niveaux de qualité différents : la compression basée sur Octree, le bruit de couleur, le bruit gaussien de la géométrie, le bruit de couleur et le bruit gaussien de la géométrie.

Pour comparer notre méthode à l'état de l'art, nous avons appliqué une approche de validation croisée de type 4-fold pour ICIP2020 et 9-fold pour SJTU. Plus précisément, chaque fold est constitué des versions dégradées du même objet source. Par ailleurs, un fold est utilisé pour la phase de validation, un autre pour la phase de test et le reste des folds est exploité pour entraîner le modèle. Cette approche nous permet de garantir que notre modèle évalue la qualité de l'objet plutôt que celle du contenu. Il convient de souligner que ce protocole a été appliqué de manière cohérente à toutes les méthodes comparées à notre méthode proposée.

TABLE 1 : Comparaison des performances de la méthode proposée par rapport aux méthodes de pointe sur les bases de données ICIP2020 et SJTU.

	Methodes	ICIP2020		SJTU	
		SROCC	PLCC	SROCC	PLCC
FR	P2P MSE [9]	0.954	0.615	0.803	0.606
	P2P Hausdorff [4]	0.682	0.615	0.687	0.606
	P2Pl MSE [12]	**0.971**	0.618	0.715	0.568
	P2Pl Hausdorff [4]	0.735	0.491	0.683	0.562
	Pl2Pl[1]	0.902	0.626	0.772	0.615
	P2D-G MMD [6]	0.960	0.784	0.604	0.628
	JP2D-GC MMD[5]	0.965	0.881	0.755	0.667
	PCQM [10]	**0.977**	**0.942**	**0.807**	**0.805**
	PointSSIM [2]	0.795	0.717	0.685	0.652
RR	PCM [14]	0.882	0.627	0.219	0.263
NR	**Méthode proposée**	0.960	**0.918**	**0.931**	**0.926**

Pour calculer les corrélations entre les scores subjectifs (MOS) et les scores objectifs, nous avons utilisé 3 critères : le coefficient de corrélation de rang de Spearman (SROCC) qui mesure la monotonie de la prédiction du modèle, le coefficient de corrélation linéaire (PLCC) qui mesure la précision de la prédiction du modèle et l'erreur quadratique moyenne (RMSE) qui calcule l'erreur de distance entre le score prédit et la vérité terrain de l'objet 3DPC. Les valeurs élevées des deux critères SROCC et PLCC (proches de 1) indiquent une bonne précision, contrairement à RMSE (proche de 0).

3.2 Comparaison avec les méthodes de l'état de l'art

Dans cette section, nous présentons une étude comparative entre la méthode proposée et les méthodes de l'état de l'art, y compris les méthodes de référence complète (FR) et réduite (RR) sur deux bases de données : ICIP2020 et SJTU. Les résultats de comparaison sont rapportés dans le tableau 1. Les deux performances les plus élevées sont mises en gras. Notre modèle présente des performances prometteuses, comparées aux métriques de référence complète et réduite qui exigent la présence de l'information complète ou réduite de l'objet de référence pour évaluer le 3DPC dégradé. Il convient de noter que la performance de toutes les méthodes de l'état de l'art sur la base de données ICIP2020 est plus élevée par rapport à la base de données SJTU. Cela pourrait être expliqué par le type de distorsion dans les deux bases de données. Sur la base de données SJTU, il y a des types de dégradation plus difficiles tels que l'échantillonnage et le bruit de couleur, tandis que sur ICIP2020, il y a seulement des types de distorsion de compression. Néanmoins, notre modèle a montré de bonnes performances sur la base de données STJU par rapport aux autres métriques de l'état de l'art. Cela montre la stabilité de

TABLE 2 : Étude d'ablation sur ICIP2020 : Comparaison des performances du modèle proposé avec et sans pondération de carte de saillance.

	PLCC	SROCC	RMSE
Notre méthode sans pondération en utilisant la carte de saillance	**0.970**	0.901	0.285
Notre méthode avec pondération en utilisant la carte de saillance	0.960	**0.918**	**0.275**

la performance de notre modèle dans l'évaluation de la qualité des dégradations, indépendamment des types de dégradation dans les bases de données.
Nous avons également réalisé une étude d'ablation pour montrer la pertinence de l'utilisation des cartes de saillance en entraînant le modèle uniquement avec les vues projetées (c'est-à-dire sans l'étape de pondération à l'aide des cartes de saillance). Comme le montre le tableau 2, les résultats montrent que la performance est améliorée sur SROCC et RMSE lorsque nous pondérons les images projetées avec les cartes de saillance et les présentons en tant qu'entrée au modèle d'apprentissage en profondeur.

4 Conclusion

Dans cet article, nous avons présenté une approche basée sur l'apprentissage en profondeur qui combine les avantages de l'architecture d'encodeur Transformer et de la saillance visuelle pour prédire la qualité visuelle perçue des nuages de points dégradés. En comparant notre méthode avec les méthodes de référence complètes et réduites existantes de l'état de l'art, notre modèle présente des résultats prometteurs pour l'évaluation de la qualité des nuages de points. Dans le cadre de travaux futurs, nous comptons étudier le degré de liberté du modèle proposé, y compris le modèle utilisé pour la prédiction des cartes de saillance.

Références

[1] Evangelos ALEXIOU et Touradj EBRAHIMI : Point cloud quality assessment metric based on angular similarity. *In 2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2018.

[2] Evangelos ALEXIOU et Touradj EBRAHIMI : Towards a point cloud structural similarity metric. *In 2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–6, 2020.

[3] Alexey DOSOVITSKIY, Lucas BEYER, Alexander KOLESNIKOV, Dirk WEISSENBORN, Xiaohua ZHAI, Thomas UNTERTHINER, Mostafa DEHGHANI, Matthias MINDERER, Georg HEIGOLD, Sylvain GELLY *et al.* : An image is worth 16x16 words : Transformers for image recognition at scale. *arXiv preprint arXiv :2010.11929*, 2020.

[4] Alireza JAVAHERI, Catarina BRITES, Fernando PEREIRA et João ASCENSO : A generalized hausdorff distance based quality metric for point cloud geometry. *In 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2020.

[5] Alireza JAVAHERI, Catarina BRITES, Fernando PEREIRA et João ASCENSO : A point-to-distribution joint geometry and color metric for point cloud quality assessment. *arXiv preprint arXiv :2108.00054*, 2021.

[6] Alireza JAVAHERI, Catarina BRITES, Fernando PEREIRA et João ASCENSO : Mahalanobis based point to distribution metric for point cloud geometry quality evaluation. *IEEE Signal Processing Letters*, 27:1350–1354, 2020.

[7] Alexander KRONER, Mario SENDEN, Kurt DRIESSENS et Rainer GOEBEL : Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020.

[8] Qi LIU, Hui YUAN, Honglei SU, Hao LIU, Yu WANG, Huan YANG et Junhui HOU : Pqa-net : Deep no reference point cloud quality assessment via multi-view projection. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(12):4645–4660, 2021.

[9] RN MEKURIA, Zhu LI, C TULVAN et P CHOU : Evaluation criteria for pcc (point cloud compression). 2016.

[10] Gabriel MEYNET, Yana NEHMÉ, Julie DIGNE et Guillaume LAVOUÉ : Pcqm : A full-reference quality metric for colored 3d point clouds. *In 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2020.

[11] Stuart PERRY, Huy Phi CONG, Luís A. da SILVA CRUZ, João PRAZERES, Manuela PEREIRA, Antonio PINHEIRO, Emil DUMIC, Evangelos ALEXIOU et Touradj EBRAHIMI : Quality evaluation of static point clouds encoded using mpeg codecs. *In 2020 IEEE International Conference on Image Processing (ICIP)*, pages 3428–3432, 2020.

[12] Dong TIAN, Hideaki OCHIMIZU, Chen FENG, Robert COHEN et Anthony VETRO : Geometric distortion metrics for point cloud compression. *In 2017 IEEE International Conference on Image Processing (ICIP)*, pages 3460–3464, 2017.

[13] Yubing TONG, Hubert KONIK, Faouzi CHEIKH et Alain TREMEAU : Full reference image quality assessment based on saliency map analysis. *Journal of Imaging Science and Technology*, 54(3):30503–1, 2010.

[14] Irene VIOLA et Pablo CESAR : A reduced reference metric for visual quality evaluation of point cloud contents. *IEEE Signal Processing Letters*, 27:1660–1664, 2020.

[15] Irene VIOLA, Shishir SUBRAMANYAM et Pablo CESAR : A color-based objective quality metric for point cloud contents. *In 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2020.

[16] Qi YANG, Hao CHEN, Zhan MA, Yiling XU, Rongjun TANG et Jun SUN : Predicting the perceptual quality of point cloud : A 3d-to-2d projection-based exploration. *IEEE Transactions on Multimedia*, pages 1–1, 2020.