# Optimisation de l'apprentissage des représentations pour l'évaluation de la qualité des nuages de points 3D sans référence

Marouane TLIBA[1], Aladine CHETOUANI[1], Giuseppe VALENZISE[2], Frederic DUFAUX[2]

[1]Laboratoire PRISME, Universite d'Orleans, Orleans, France

[2]Universite Paris-Saclay, CNRS, CentraleSupelec, Laboratoire des signaux et systèmes

marouane.tliba@univ-orleans.fr , aladine.chetouani@univ-orleans.fr
giuseppe.valenzise@l2s.centralesupelec.fr, frederic.dufaux@l2s.centralesupelec.fr

**Résumé –** Les récents systèmes d'information et de communication utilisent le nuage de points 3D (PC) comme modalité de représentation géométrique avancée pour les applications immersives. Les PCs sont souvent compressés à des fins de transmission et de visualisation ce qui peut avoir un impact sur la qualité perçue. Le développement de mesures de qualité objectives robustes et efficaces pour les PCs reste ainsi un problème ouvert. Dans cet article, nous proposons une approche pour évaluer les effets perceptuels des solutions de compression des PCs sans référence basée sur l'apprentissage profond. Notre approche se concentre sur l'exploitation des caractéristiques intrinsèques des PCs pour quantifier les déficiences de codage à partir de quelques patchs distants choisis aléatoirement en utilisant des stratégies d'apprentissage supervisée et non supervisée. Pour évaluer les performances de notre méthode, deux bases de données ont été utilisées. Les résultats démontrent l'efficacité et la fiabilité de la méthode proposée par rapport aux méthodes de l'état de l'art.

**Abstract –** Recent information and communication systems have employed 3D Point Cloud (PC) as an advanced geometrical representation modality for immersive applications. Like most multimedia data, PCs are often compressed for transmission and viewing purposes, which can impact the perceived quality. Developing robust and efficient objective quality metrics for PCs is still an open problem. In this paper, we propose an end-to-end deep approach for evaluating the perceptual effects of point cloud compression solutions without reference. Our approach focuses on leveraging the intrinsic point cloud characteristics to quantify the coding impairments from few distant randomly selected patches using supervised and unsupervised training strategies. To evaluate the performance of our method, two well-known datasets have been used. The results demonstrate the effectiveness and reliability of the proposed method compared to to state-of-the-art methods.

## 1 Introduction

Les nuages de point 3D (PCs) ont été adoptés comme l'un des formats 3D les plus préférables. Ils sont des ensembles de points non ordonnés déterminés par leurs coordonnées cartésiennes et des attributs potentiels tels que la couleur, les courbures, la réflectance et les vecteurs normaux. Il est utilisé dans de nombreuses applications récentes telles que la réalité étendue, les communications immersives en temps réel, la robotique, les jeux 3D et le patrimoine culturel. La représentation d'une scène 3D réaliste avec une haute résolution nécessite jusqu'à des millions de points. L'application des schémas de compression émergent devient alors inévitable. Pour évaluer la distorsion de débit ou les performances des schémas de codage de nuages de points existants, des métriques de qualité perceptuelle doivent être adoptées.

Les mesures de qualité existantes pour les PCs peuvent être classées en trois groupes principaux : les métriques basées sur les points, celles basées sur les caractéristiques ou encore les métriques basées sur la projection. Les mesures basées sur les points telles que Point-to-Point (Po2Po) [1], Point-to-Plane (Po2Pl) [2], et Plane-to-Plane prédisent la qualité en fonction de la distance géométrique et/ou des caractéristiques entre le PC de référence et sa version dégradée. Il est intéressant de noter que MPEG a adopté les métriques Po2Po MSE et Po2Pl MSE associées au PSNR comme standard pour la mesure de qualité des PCs. Les métriques de qualité PC basées sur les caractéristiques extraient la géométrie en exploitant des attributs spécifiques d'une manière globale ou locale. Parmi ces métriques, nous pouvons citer PC-MSDM [3] qui est une extension de la métrique SSIM 2D [4] en considérant les statistiques de courbure locale, et PCQM [5] qui combine les caractéristiques de la géométrie et de la couleur. Pour les mesures de qualité des PCs basées sur la projection, les points 3D sont projetés en grilles régulières 2D et des mesures de qualité 2D sont appliquées à ces vues. Le manque d'un grand ensemble de données pour estimer la qualité des PCs empêche le développement de métriques efficaces basées sur l'apprentissage profond. Il est donc impératif de trouver un moyen de pousser l'apprentissage de la représentation à partir d'une quantité limitée de données. De plus, toutes les métriques PCQA citées ci-dessus nécessitent un long pré-traitement ce qui est coûteux en termes de calcul, et la

plupart d'entre elles requièrent le PC de référence.

Dans cet article, nous cherchons à apprendre une représentation efficace à partir des caractéristiques intrinsèques locales des patchs du PC. Plus précisément, nous extrayons les patchs PC en sélectionnant d'abord $M$ points centroïdes à l'aide de l'algorithme d'échantillonnage du point le plus éloigné [6]. Ensuite, nous appliquons la méthode de clustering des $K$ plus proche voisins pour former un patch autour de chaque centroïde. Enfin, nous apprenons une fonction de coût de descripteur de caractéristiques invariant par permutation, suivie d'un regresseur peu profond qui estime le score de qualité final. Pour optimiser davantage le modèle, en plus d'entraîner notre réseau peu profond à l'aide de la perte supervisée, nous utilisons également une perte de classement non-supervisée. L'objectif est de forcer le modèle à apprendre des représentations descriptives plus riches capturant les caractéristiques intrinsèques locales des nuages de points.

Les principales contributions de ce document sont résumées dans ce qui suit :

— Nous proposons un nouveau modèle efficace de bout en bout et peu profond pour évaluer les effets perceptuels des méthodes de compression en nous appuyant sur les caractéristiques intrinsèques locales de sous-ensembles de points.

— Nous étendons la stratégie avec un apprentissage non-supervisé en se basant sur un critère de rang comme pseudo-étiquette afin d'optimiser l'apprentissage sur les caractéristiques intrinsèques d'un sous-ensemble de points vers une représentation descriptive potentielle liée à notre tâche.

# 2 Méthode Proposée

L'objectif principal de ce travail est de fournir une métrique sans référence efficace pour estimer la qualité des PCs compressés sans ajouter d'étape de projection ou appliquer des changements sur la spécificité des PCs. Motivés par la nature des schémas de compression qui produisent un effet distribué uniformément sur les caractéristiques locales, nous supposons que la distribution des PCs compressés pourrait être représentée comme un ensemble de caractéristiques cachées ou de descripteurs liés aux effets de la compression tels que la cohérence des points, la rareté et la densité. Ajouté à cela un ensemble de descripteurs liés à d'autres caractéristiques telles que la position, la forme globale et la structure. Afin d'apprendre efficacement la représentation de ces caractéristiques cachées ou intrinsèques et en s'inspirant de PointNet [6], nous utilisons ici un encodeur invariant par permutation peu profond par le biais d'une série de convolutions 1D partagées sur tous les points. Nous employons ensuite un regresseur peu profond pour prédire la qualité du PC. Inspiré des travaux antérieurs qui apprennent la représentation à partir du classement [7], notre réseau est optimisé pour apprendre simultanément la représentation interne à partir des données étiquetées, ainsi que le rang implicite connu entre les échantillons de données comme pseudo-labels.

## 2.1 Extraction des Patchs et Pré-traitement

Le PC est d'abord divisé en petits patchs en considérant uniquement les coordonnées $(x, y, z)$ et les informations de couleur RGB. Pour cela, l'échantillonnage du point le plus éloigné [6] ainsi que l'algorithme des $K$ plus proches voisins (K-NN) sont utilisés comme suit :

— A partir d'un PC donné $X$, nous appliquons d'abord l'algorithme d'échantillonnage du point le plus éloigné pour sélectionner $M$ centroïdes (i.e. C = $\{ P_1, P_2, ., P_m \}$)

— Afin de former un patch autour de chaque centroïde, nous employons ensuite la méthode du $K - NN$. Le sous-ensemble de points de la $i_{th}$ centroïde, dénoté ici $\{ P_i^1, P_i^2, P_i^3, ., P_i^K \}$, représente le $i_{th}$ patch. $K$ a été fixé ici à $512$.

— Enfin, nous calculons les différences de position et de couleur entre chaque sous-ensemble de points et son centroïde correspondant. Nous obtenons le sous-ensemble final de points, noté ici $S_i$ qui sert d'entrée au modèle $\{ P_i^1 - P_i, P_i^2 - P_i, P_i^3 - P_i, ., P_i^K - P_i \}$.

## 2.2 Architecture de réseau

La conception de notre réseau s'inspire du moodèle PointNet [6]. Comme nous travaillons sur des patchs, nous nous concentrons sur l'apprentissage d'une représentation utile uniquement à partir de caractéristiques intrinsèques locales plutôt que de la forme globale. En résumé, l'architecture de notre modèle est composée d'un extracteur de caractéristiques invariantes symétriques (c'est-à-dire une série de convolution 1D suivie d'une opération de poling max sur la représentation des points) et d'un régresseur peu profond suivi d'une activation sigmoïde afin de produire une probabilité qui représente le score de qualité estimé.

## 2.3 Apprentissage Auto-Supervisé et Supervisé

Comme mentionné précédemment, deux stratégies ont été considérées pour optimiser notre modèle : un apprentissage supervisé à partir des scores subjectifs et un apprentissage non-supervisé à partir du rang. L'étape d'apprentissage supervisé vise à apprendre une fonction de correspondance avec le score d'opinion moyen (MOS), tandis que l'objectif de l'étape d'apprentissage non-supervisé est d'apprendre une meilleure

représentation qui maximise l'apprentissage à partir des caractéristiques intrinsèques. Nous cherchons ici à forcer le modèle à maximiser l'apprentissage de caractéristiques intrinsèques locales à partir d'un sous-ensemble de points de manière non-supervisée par étiquette comme suit :

$$S_d = Q_d + N_d \tag{1}$$

où $S_d$ représente la distribution d'un sous-ensemble de points. $Q_d$ est la distribution de caractéristiques intrinsèques liées à l'estimation de la qualité perceptuelle et $N_d$ désigne la distribution d'autres caractéristiques qui ne sont pas liées à notre tâche.

#### 2.3.1 Apprentissage Supervisé

Le but de l'étape supervisée est de trouver une fonction $f(S_d : \theta)$ avec des paramètres $\theta$ qui capture la relation entre $S_d$ et les scores de qualité $Y$ où $(X, Y) = \{(S_{d0}, Y), (S_{d1}, Y), ..(S_{dm}, Y)\}$.

Étant donné que la sortie de notre modèle est une valeur de probabilité qui représente la qualité perceptuelle, nous considérons l'entropie croisée binaire (ECB) comme fonction de perte pour minimiser le risque empirique de la régression sur l'ensemble d'échantillons. Elle est définie comme suit :

$$Rg\_loss(P, Y) = -Y \log(P) - (1 - Y) \log(1 - P) \tag{2}$$

où $P = f(S_d : \theta)$ et $Y$ est le score subjectif des échantillons du PC.

#### 2.3.2 Apprentissage Auto-Supervisé

L'objectif de l'étape non-supervisée est d'obtenir une tâche supervisée par procuration avec des pseudo-étiquettes riches. Cela permet d'apprendre la séparation à partir de différentes dégradations et en même temps de maximiser l'apprentissage sur les $Q_d$ distributions (i.e. caractéristiques intrinsèques).

Pour un module d'encodage donné (c'est-à-dire MPEG G-PCC [8] ou MPEG V-PCC [9]), nous appliquons plusieurs niveaux de compression $l_j$ et $l_{j+}$ sur le PC $X$ dont les PCs résultants sont $X_j$ et $X_{j+}$, respectivement. Ainsi, nous pouvons facilement connaître le rang sur la base des distorsions introduites.

Considérant que la dégradation de la qualité visuelle de ces modules de codage est uniforme sur la géométrie, nous voulons que notre modèle se concentre sur les caractéristiques intrinsèques et qu'il ignore l'apprentissage d'autres caractéristiques telles que la position des patchs (i.e., $f(S_d : \theta) = f(Q_d : \theta)$). En d'autres termes, quels que soient les patchs sélectionnés de $X_j$ et $X_{j+}$, le modèle doit donner des sorties ordonnées selon l'ordre inverse induit par le niveau de compression :

$$l_j < l_{j+} mean(f({S_d}^j)) > mean(f({S_d}^{j+})) \tag{3}$$

À cette fin, nous utilisons un réseau Siamois [10] pour minimiser la perte de classement entre l'activation des paires d'entrées. Chaque branche de ce réseau est basée sur les paramètres de notre modèle.

$$Rank\_Loss = max(0, \Delta f + margin) \tag{4}$$

Le gradient de cette fonction est représenté comme suit :

$\nabla_\theta Rank\_Loss = \{\, 0 \, , if \Delta f + margin \leq 0$
$\nabla_\theta(\Delta f), otherwise$ où $\Delta f = mean(f({S_d}^{j+})) - mean(f({S_d}^j))$

Le gradient tend vers zéro lorsque l'activation du réseau est d'ordre inverse de celle induite par la compression. En revanche, le gradient de l'activation supérieure diminue tandis que le gradient de l'activation inférieure augmente lorsque les activations ont le même ordre que celui induit par la compression.

### 2.4 Protocole d'Apprentissage

À chaque étape de l'apprentissage, nous mettons à jour les poids du modèle en minimisant à la fois la fonction de perte de classement auto-supervisée et celle supervisée sous forme d'un apprentissage multi-tâches.

$$MultiTask\_loss = Rank\_Loss + 0.01 * Rg\_loss \tag{5}$$

Il convient de noter qu'à chaque étape de l'apprentissage, les centroïdes sélectionnées pour l'extraction des patchs changent (i.e., $(C^j) \neq (C^{j+})$). En d'autres termes, nous forçons implicitement le modèle d'écarter les informations relatives à la position et à la forme globale afin de maximiser l'apprentissage sur les informations relatives à la qualité qui représentent l'information mutuelle entre les différents éléments du modèle. En outre, ce protocole strict (i.e., les données d'apprentissage changent à chaque époque) rend le modèle plus robuste et augmente ainsi sa capacité de généralisation. La valeur de la marge change en fonction de la distance entre les niveaux de compression variant de $0.1$ à $0.6$. Nous fixons les poids initiaux du modèle de manière aléatoire et entraînons l'ensemble des paramètres de bout en bout pendant $500$ époques en utilisant l'optimiseur Adam.

## 3 Résultats

Nous évaluons notre modèle à l'aide d'un ensemble de données bien connus **ICIP20** qui utilisent différents schémas de compression avec plusieurs niveaux d'encodage. Cette base est composée de 6 PCs de référence à partir desquels 90 versions dégradées ont été dérivées par trois types de compression : V-PCC, G-PCC avec codage en soupe de triangle et G-PCC avec codage en octree. Chaque PC de référence a été compressé en utilisant cinq niveaux différents.

| Modèle | PLCC ↑ | SROCC ↑ |
|---|---|---|
| po2pointMSE | 0.945 | 0.950 |
| po2planeMSE | 0.945 | 0.959 |
| PSNRpo2pointMSE | 0.880 | 0.934 |
| PSNRpo2planeMSE | 0.916 | 0.953 |
| Our-S | 0.745 | 0.621 |
| Our-(S+SSL) P2 | **0.908** | **0.955** |

Table 1 – Résultats obtenus sur le jeu de données ICIP20

Afin d'étudier équitablement les performances, nous sélectionnons de manière aléatoire 64 centroïdes pour la validation ainsi que pendant l'apprentissage. Nous adoptons un protocole de validation croisée de 6 où 6 fait référence au nombre d'échantillons de PCs de référence. Plus précisément, à chaque itération, 5 échantillons de PCs de référence et leurs versions compressées sont utilisées pour l' apprentissage, et 1 échantillon de PC de référence et ses versions compressées sont utilisées pour le test. Les coefficients de corrélation de Pearson (PCC) et de rang de Spearman (SROCC) sont calculées pour évaluer notre méthode. Ces corrélations sont calculées à chaque itération indépendamment pendant la validation et finalement la moyenne de ces corrélations est reportée.

Le tableau 1 affiche les performances obtenues par notre méthode pour l'approche supervisée, désignée ici par Our-(S) et supervisée avec l'apprentissage multitâche désignée ici par Our-(S+SSL). Les résultats sont également comparés à un ensemble de méthodes de l'état de l'art. Comme on peut le constater, l'utilisation des fonctions de perte auto-supervisées et supervisées permettent d'améliorer considérablement les performances. De plus, notre méthode (S+SSL) est assez compétitive par rapport aux autres modèles, notamment en termes de corrélation de rang (i.e. SROCC). Il convient également de noter que contrairement à notre méthode, toutes les méthodes énumérées nécessitent le PC de référence ainsi qu'un temps de calcul important.

# 4 Conclusion

Dans cet article, nous avons proposé une approche efficace de bout en bout basée sur un modèle peu profond pour interpoler les scores de qualité induits par les effets de compression des PCs. Contrairement aux méthodes précédentes, l'approche proposée fonctionne directement sur les patchs des nuages de point sans nécessiter d'étape de traitement préalable coûteuse en termes de calcul. Nous avons forcé notre modèle à apprendre la représentation potentielle d'un sous ensemble de points en se concentrant sur les caractéristiques intrinsèques locales liées à notre tâche en aval. Pour optimiser le modèle, nous avons également utilisé une stratégie d'apprentissage auto-supervisée par rangs afin d'apprendre une correspondance directe avec les scores de qualité prédits, tout en maximisant la distance entre la représentation de différentes dégradations. Les résultats obtenus à l'aide d'un protocole strict de validation croisée (i.e. k-fold) ont démontré l'efficacité de notre modèle peu profond. Ils ont également montré que notre modèle est compétitif par rapport aux méthodes avec référence de l'état de l'art.

Nous envisageons dans nos travaux futurs d'étendre notre modèle par un mécanisme d'agrégation robuste par apprentissage afin de mieux prendre en compte les effets locaux et globaux sur la distribution des points.

# References


[1] C. Tulvan R. Mekuria, Z. Li and P. Chou, "Evaluation criteria for pcc (point cloud compression)," in *in ISO/IEC MPEG Doc*. N16332, 2016.

[2] C. Feng R. Cohen D. Tian, H. Ochimizu and A. Vetro, "Geometric distortion metrics for point cloud compression," in *IEEE International Conference on Image Processing (ICIP)*, China, 2017.

[3] C. Rochinni P. Cignoni and R. Scopigno, "Metro: measuring errors on simplified surfaces,," in *Computer Graphics Forum*, 1998, vol. 17, pp. 167–174.

[4] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[5] J. Digne G. Meynet, Y. Nehmé and G. Lavoué, "Pcqm: A full-reference quality metric for colored 3d point clouds," 2020.

[6] C. Qi, L. Yi, Hao Su, and Leonidas J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," 2017.

[7] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," pp. 1040–1049, 2017.

[8] et al S. Schwarz, "Emerging mpeg standards for point cloud compression," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 133–148, 2019.

[9] "V-pcc codec description," *ISO/IEC JTC1/SC29/WG11 Doc. N18892*, Geneva, Switzerland, Nov 2019.

[10] R. Hadsell S. Chopra and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 539–546 vol. 1.