Point-BERT

Pre-training 3D Point Cloud Transformers with Masked Point Modeling

Yu, Tang, Rao, Huang, Zhou & Lu

présenté par

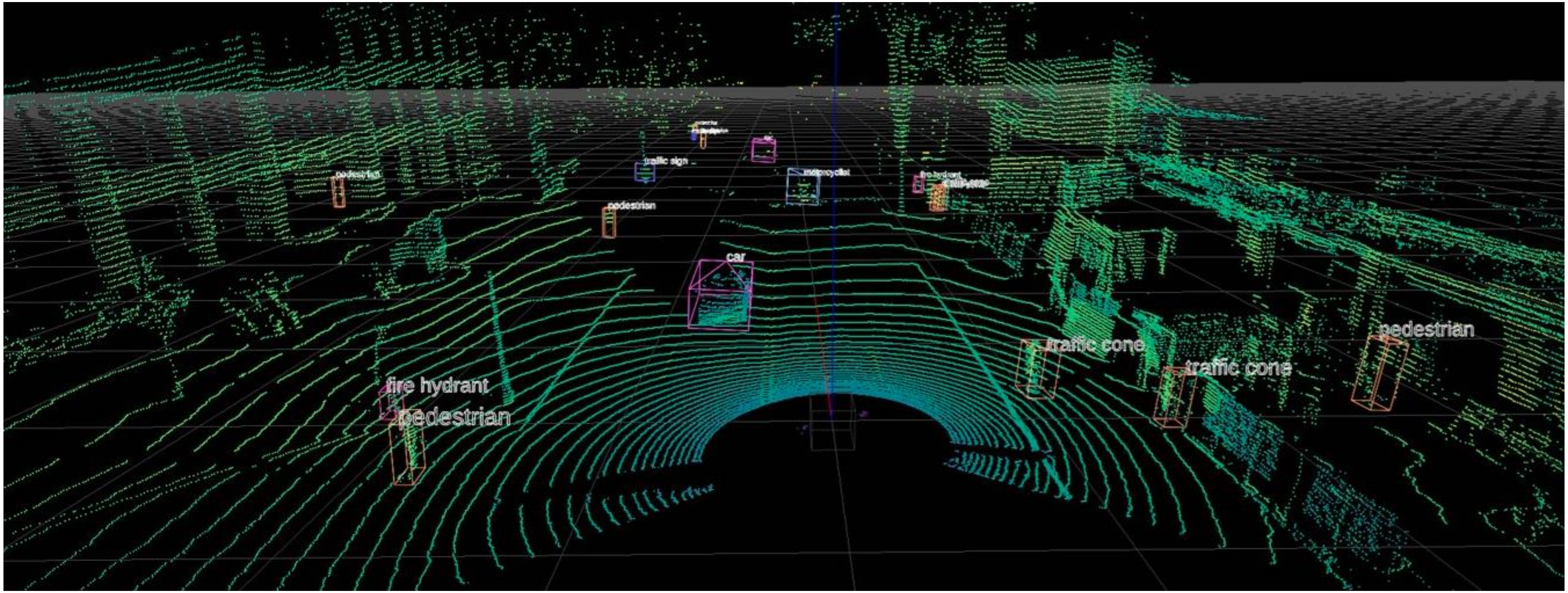
William Guimont-Martin

UNIVERSITÉ LAVAL

NOR Lab

Point-BERT

- Nuages des points



Données extraites de Déziel, Jean–Luc, et al. "PixSet: An Opportunity for 3D Computer Vision to Go Beyond Point Clouds With a Full-Waveform LiDAR Dataset." 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE, 2021.

Point-BERT

- Inspiré du traitement de la langue naturelle
    - Transformer & *Self-supervised learning* de BERT [0]
    - Liens entre les disciplines
- Apprentissage auto-supervisé sur les nuages de points

[0] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Plan de la présentation

- Nuages de points
- Apprentissage auto-supervisé & BERT
- Point-BERT
- Résultats

Nuages de points

- Ensemble de points 3D
- Tâches
  - Classification
  - Segmentation des parties
  - Détection d'objets
  - Segmentation sémantique

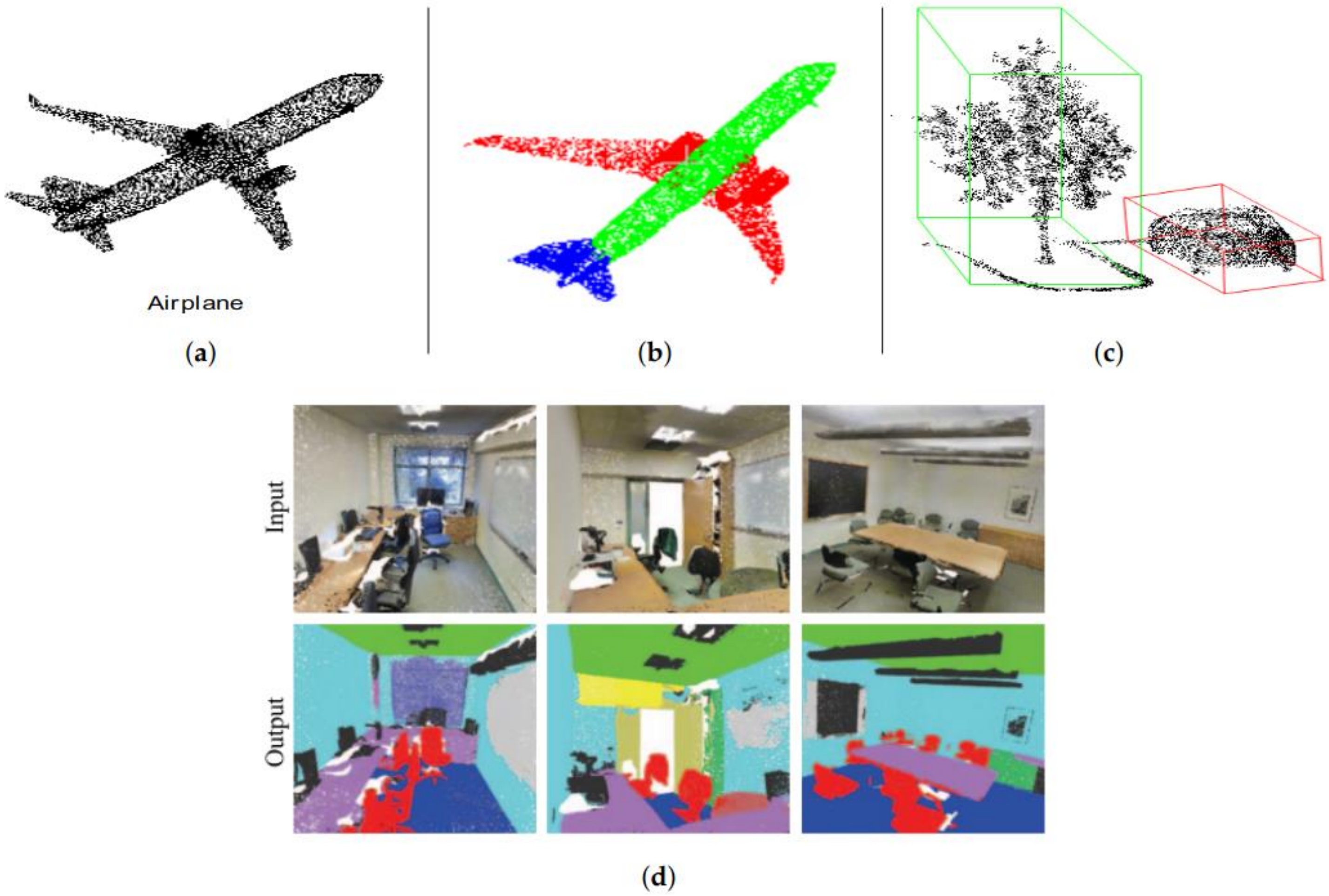
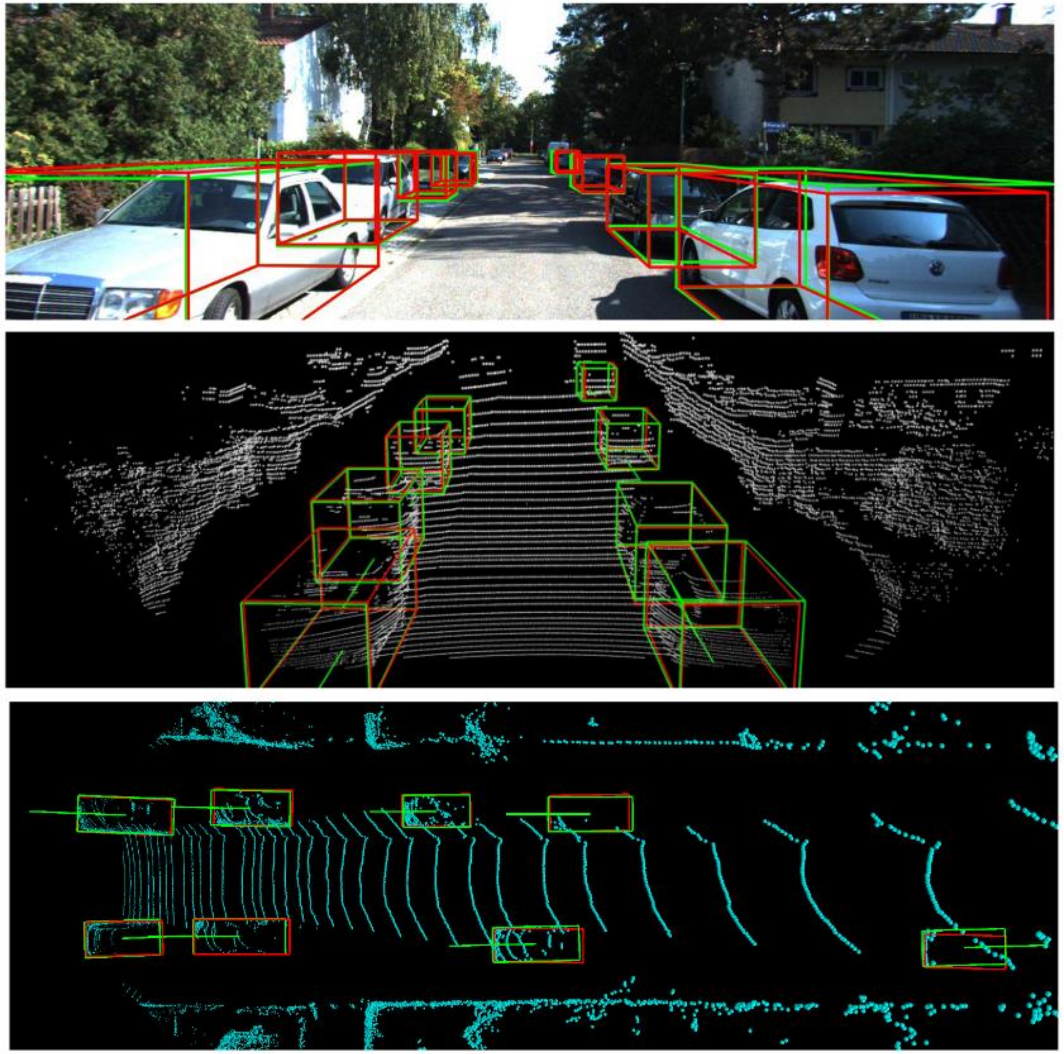


Figure extraite de Bello, Saifullahi Aminu, et al. "Deep learning on 3D point clouds." *Remote Sensing* 12.11 (2020): 1729.

Apprentissage profond et nuages de points

- Plus difficile à travailler que les images [0]
  - Densité irrégulière
  - Non-structuré
  - Non-ordonné
- Et l'annotation...

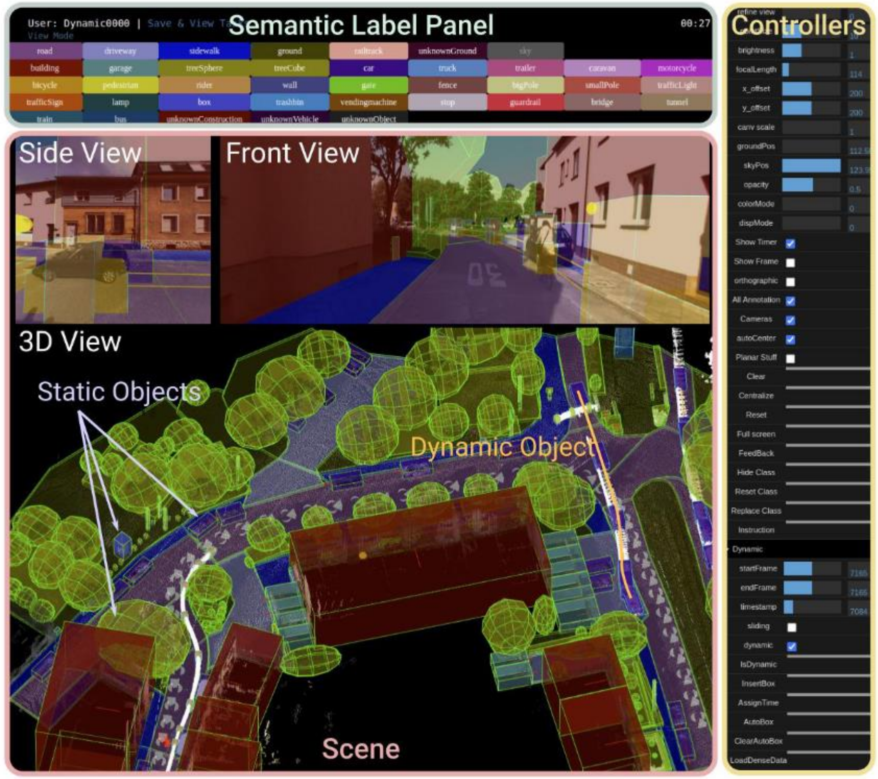
[1]



[0] Bello, Saifullahi Aminu, et al. "Deep learning on 3D point clouds." *Remote Sensing* 12.11 (2020): 1729.

[1] Figure montrant des données de KITTI, extraite de Zheng, Wu, et al. "SE-SSD: Self-ensembling single-stage object detector from point cloud." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2021.

Apprentissage auto-supervisé



[0]

- Annoter des données 3D est dur et coûteux
  - 3h pour 200m! [0]
- *Self-Supervised Learning* (SSL)
- Tâche de pré-entraînement
  - Pré-entraînement sur beaucoup de données non-annotées
  - Fine-tuning sur peu de données annotées
- Génère sa propre supervision à partir des données
  - Pas besoin d'annotation
  - Corruption / reconstruction

[0] Liao, Yiyi, Jun Xie, and Andreas Geiger. "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d." arXiv preprint arXiv:2109.13410 (2021).

BERT: Masked Language Modeling

- Corruption

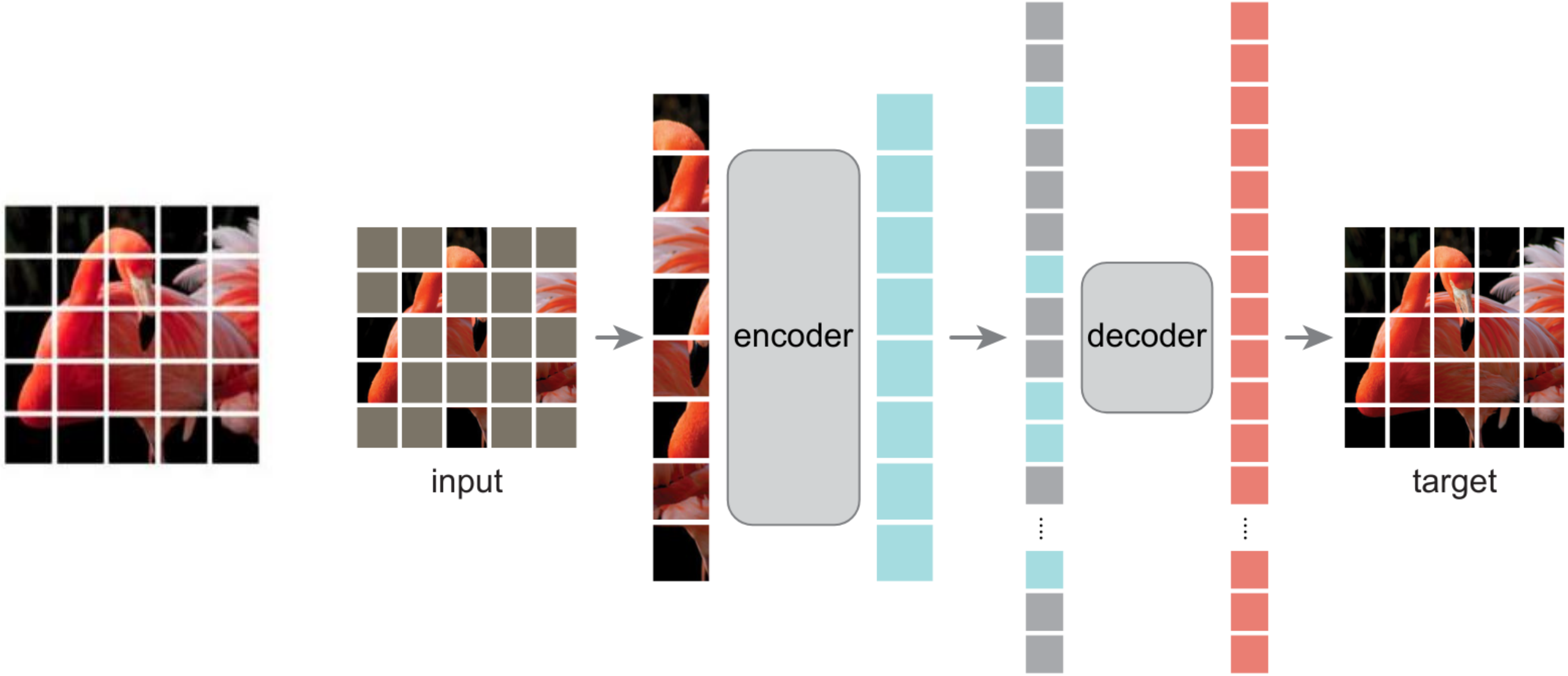
The quick brown fox ~~jumps~~ over the lazy dog

- Reconstruction

The quick brown fox jumps over the lazy dog

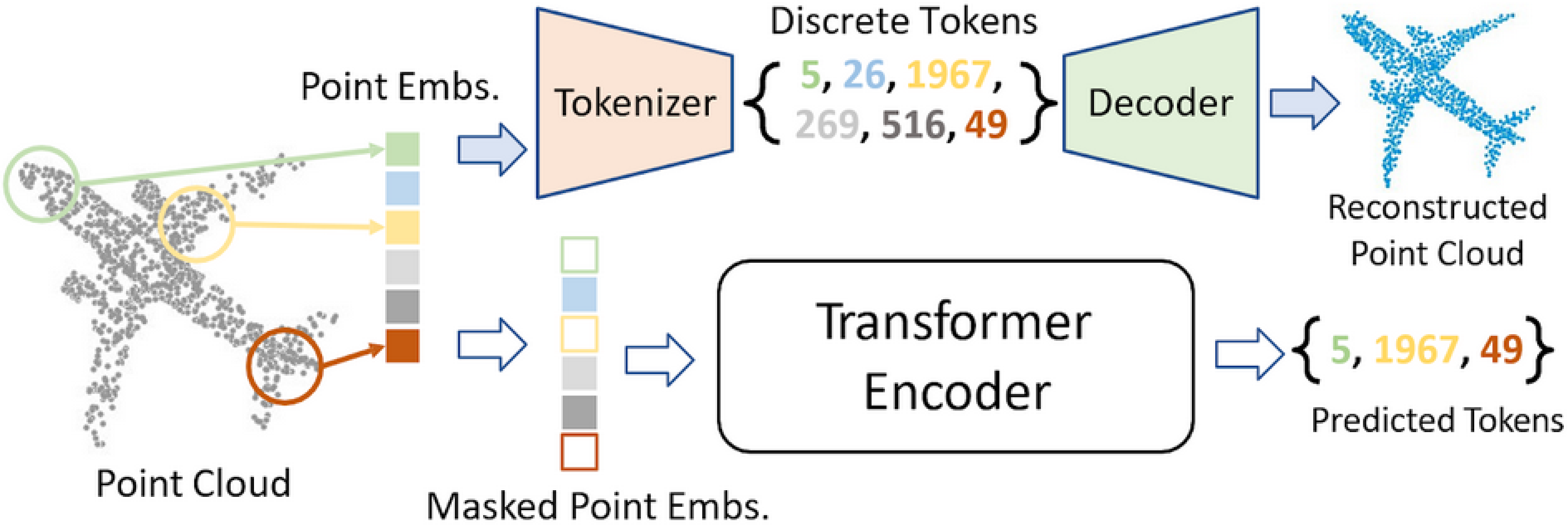
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Masked Autoencoder



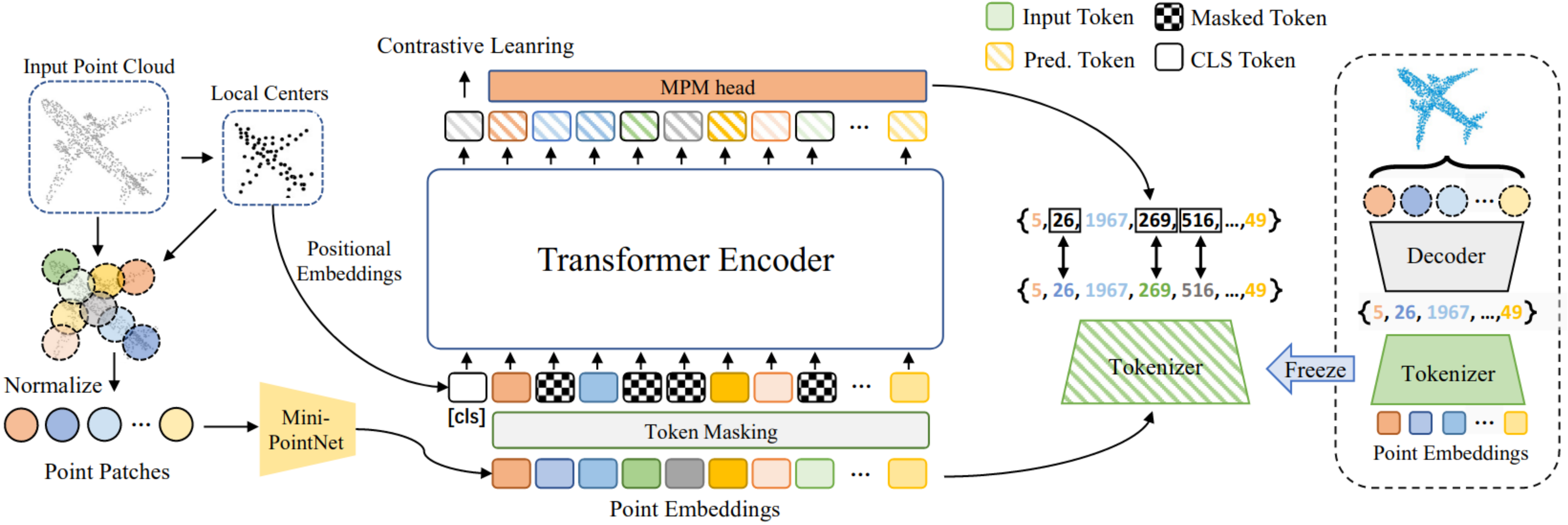
He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *arXiv preprint arXiv:2111.06377* (2021).

Point-BERT


Point Embs.
Tokenizer
Discrete Tokens
{ 5, 26, 1967, 269, 516, 49 }
Decoder
Reconstructed Point Cloud
Point Cloud
Masked Point Embs.
Transformer Encoder
{ 5, 1967, 49 }
Predicted Tokens

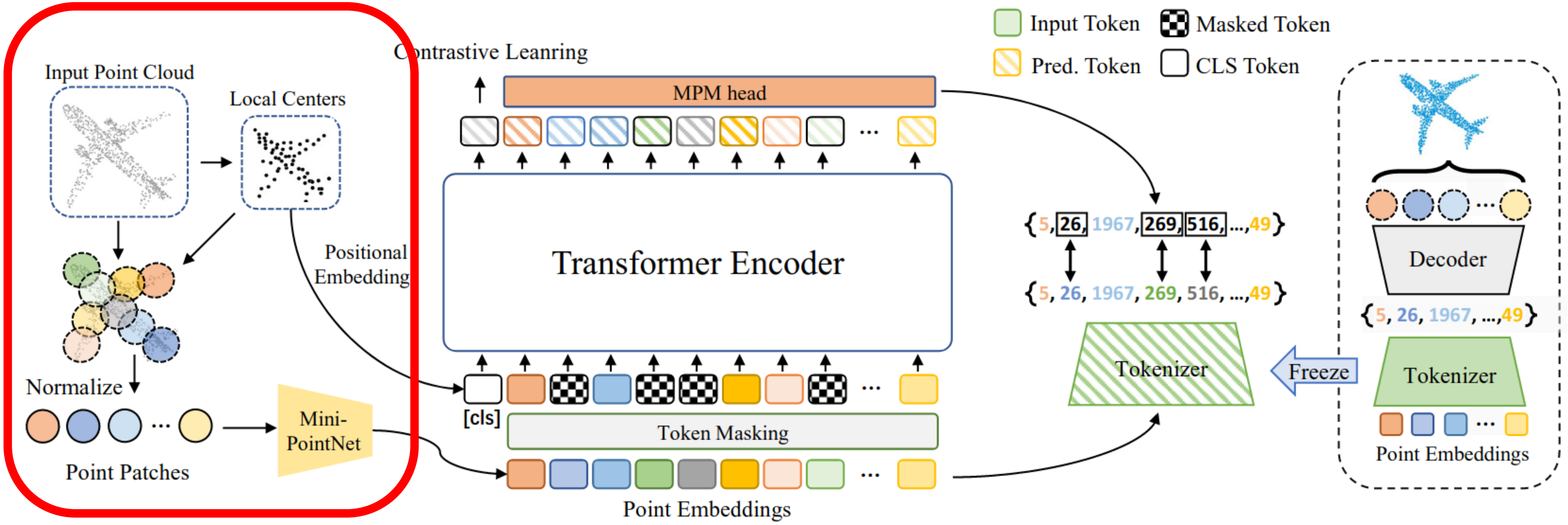
Point-BERT

Pre-training 3D Point Cloud Transformers with Masked Point Modeling



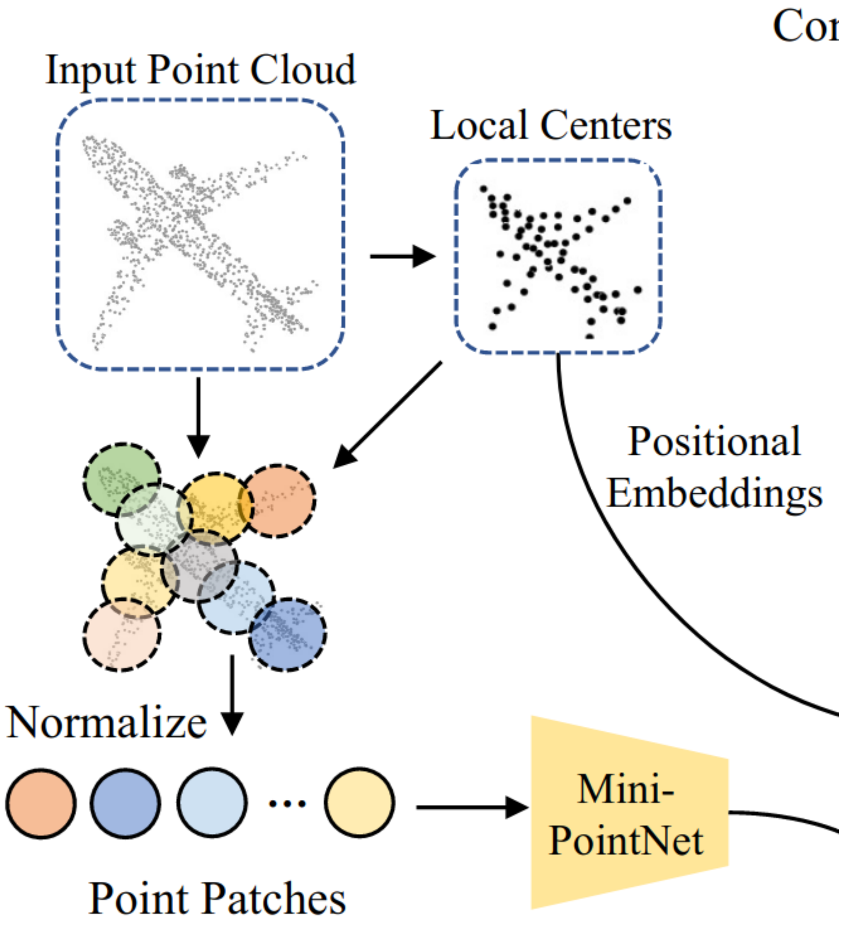
Point-BERT

Pre-training 3D Point Cloud Transformers with Masked Point Modeling



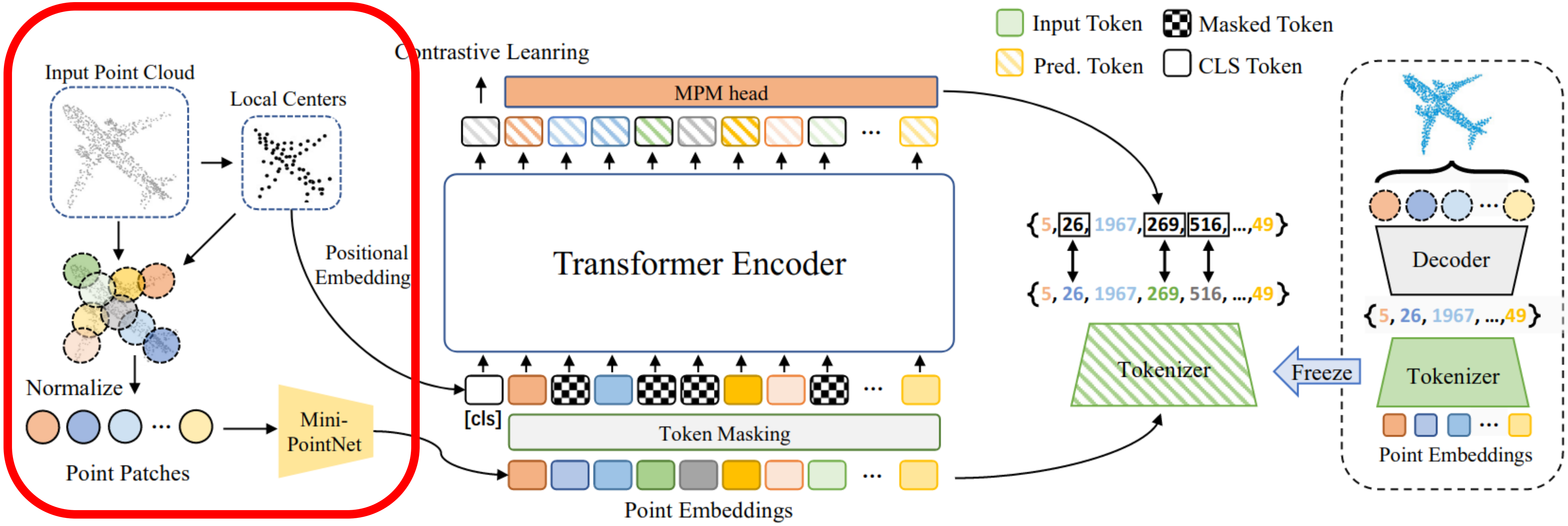
Encodage

- Échantillonnage des points centraux
- Voisinnage
- Normalisation
  - Garder information locale seulement
- Mini-PointNet
  - Invariant aux permuations des points



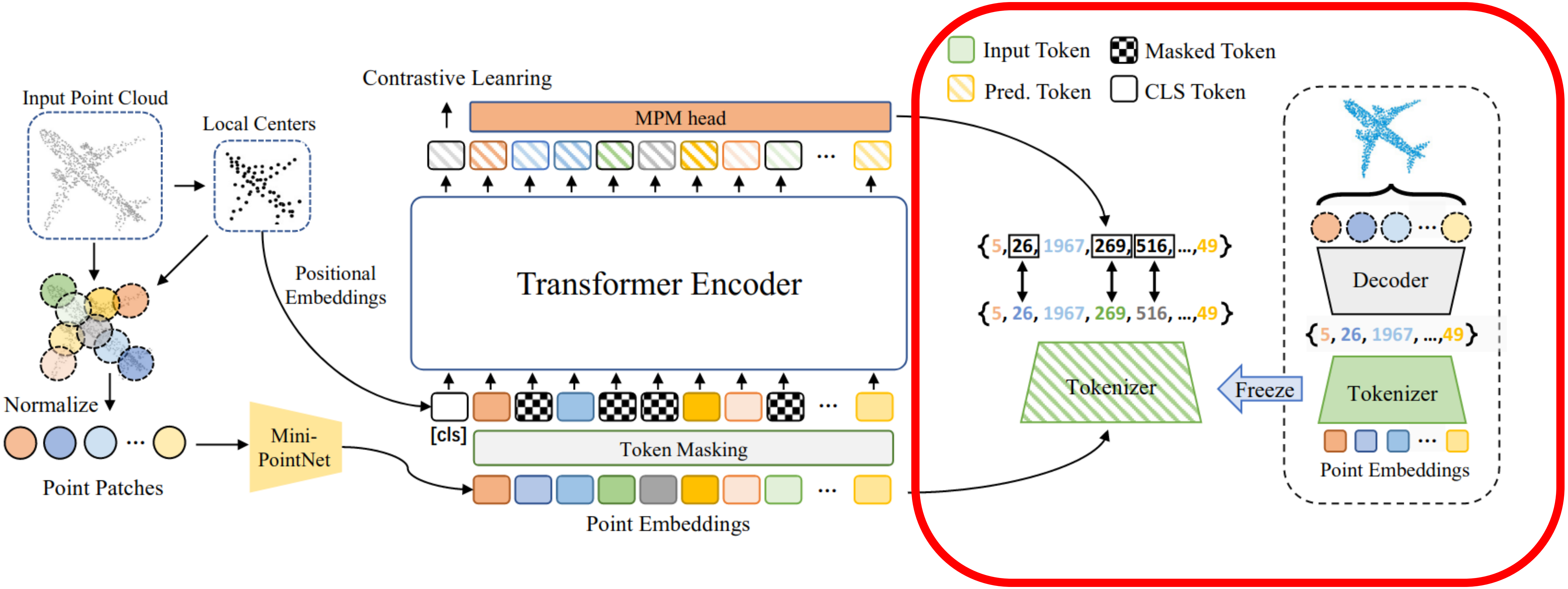
Point-BERT

Pre-training 3D Point Cloud Transformers with Masked Point Modeling



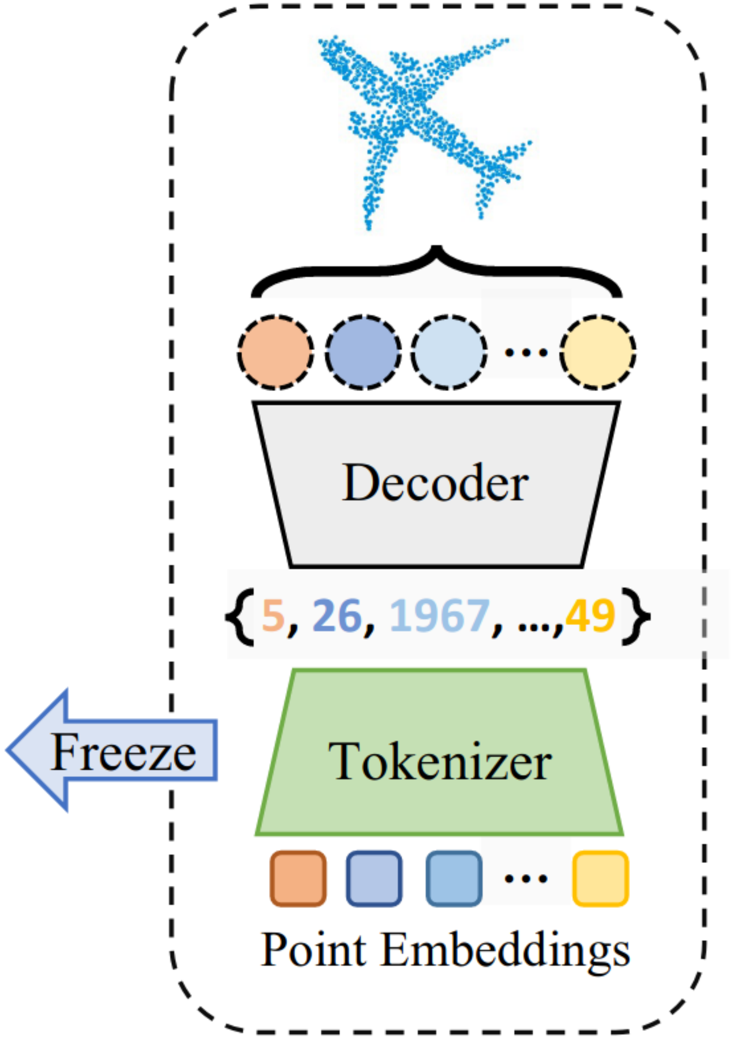
Point-BERT

Pre-training 3D Point Cloud Transformers with Masked Point Modeling

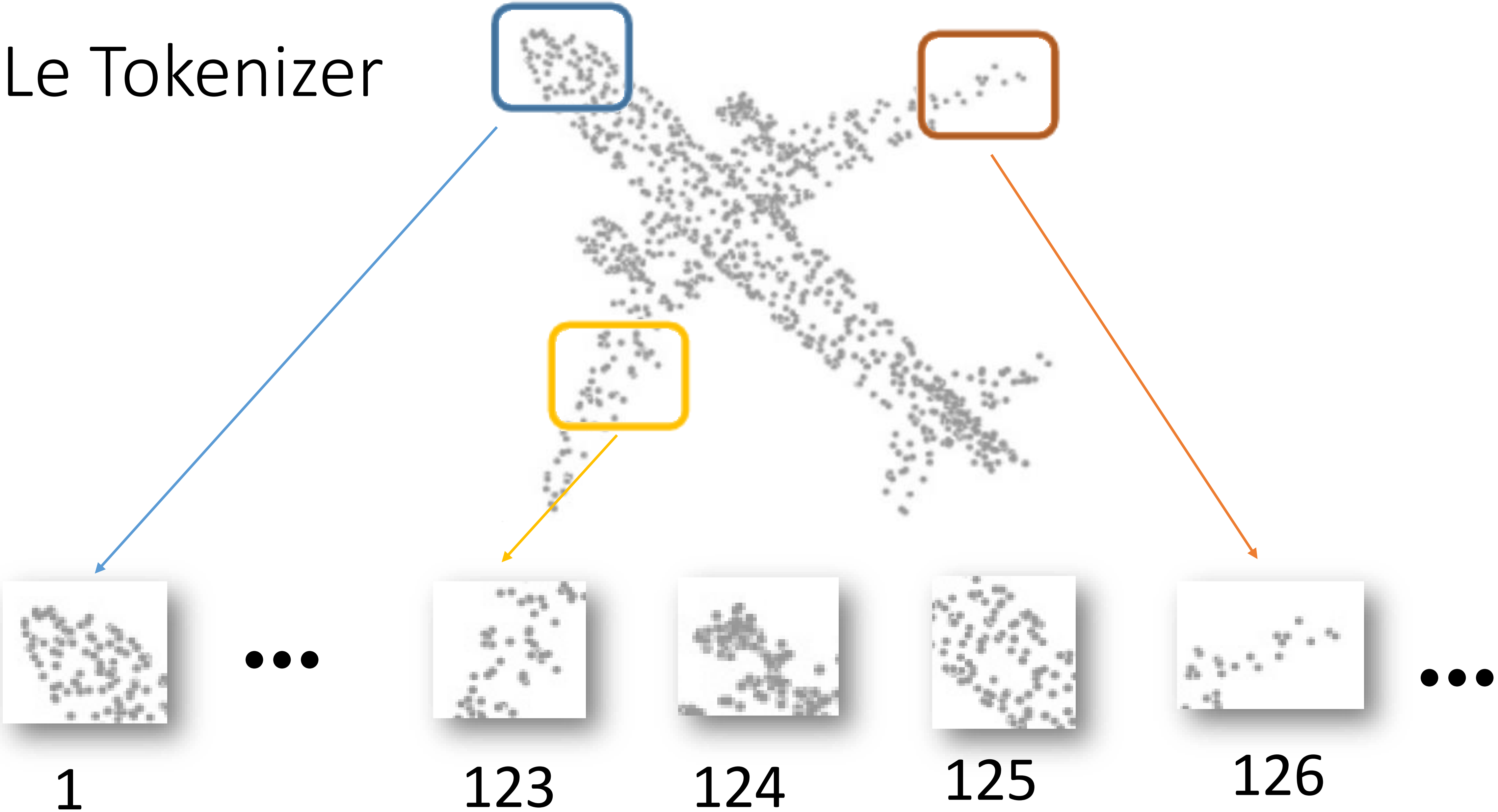


Discrete Variational Autoencoder

- Variational autoencoder
  - Compression
  - Décompression
  - Espace latent discret (tokens)
- Espace latent discret
  - Pont nuages de points -> NLP
  - Vocabulaire fixe
  - Similaire à des "mots"
  - Forme des "phrases" représentant le nuage de points

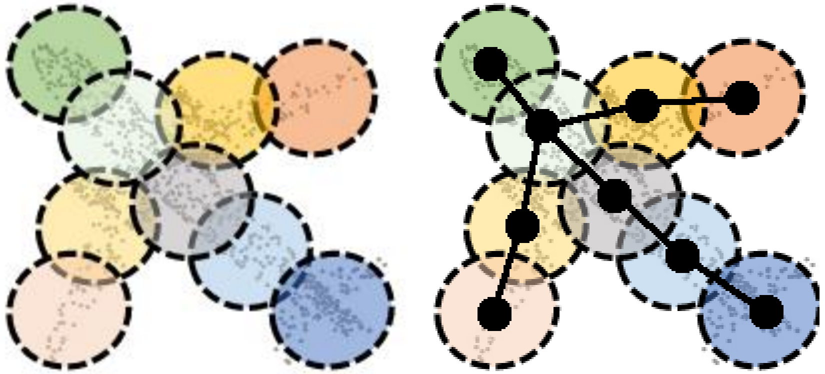


Le Tokenizer

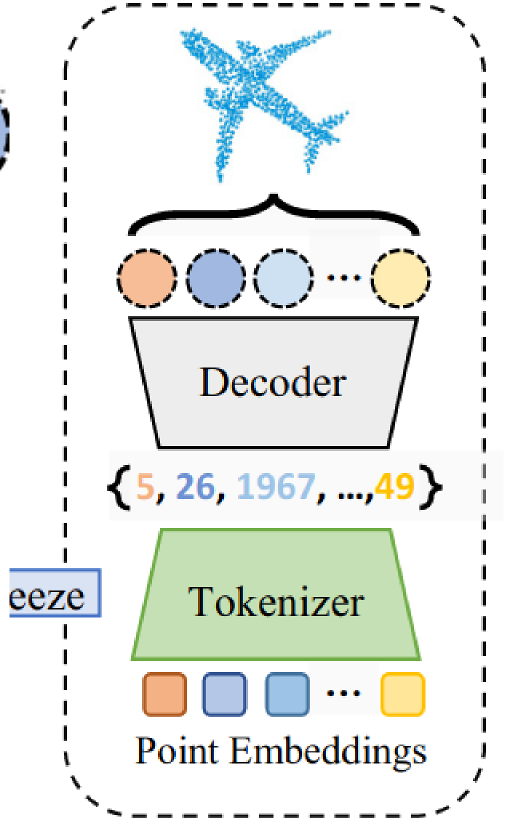


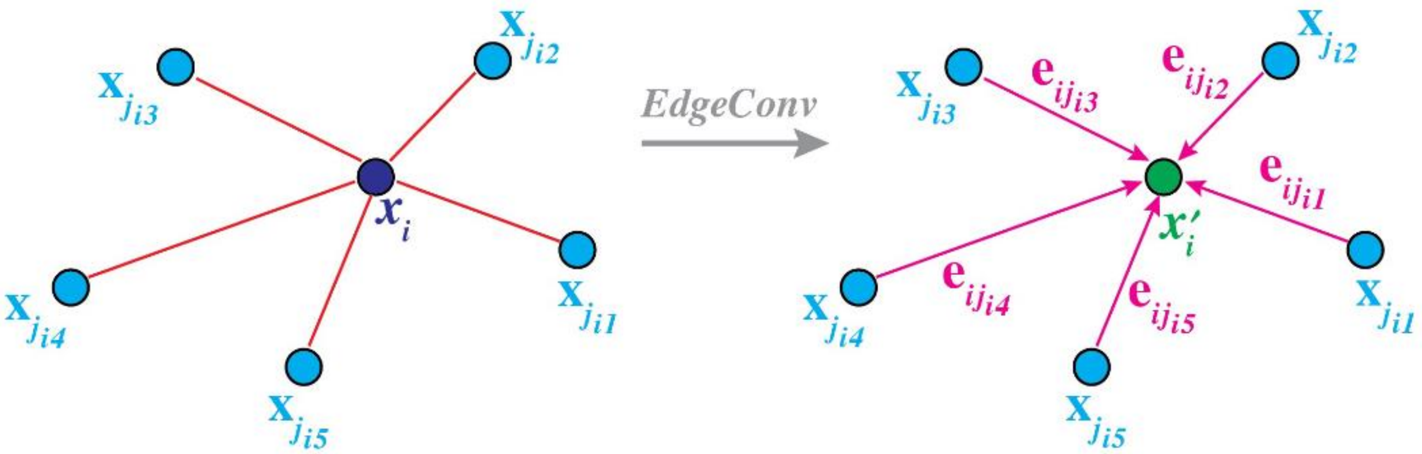
Adapté de Yu, Xumin, et al. "Point-BERT: Pre-training 3D Point Cloud Transformers with Masked Point Modeling." arXiv preprint arXiv:2111.14819 (2021).

The Tokenizer



- Traduit le nuage de points en "mots"
- DGCNN
  - Construction d'un graphe kNN
  - Convolution sur les arêtes

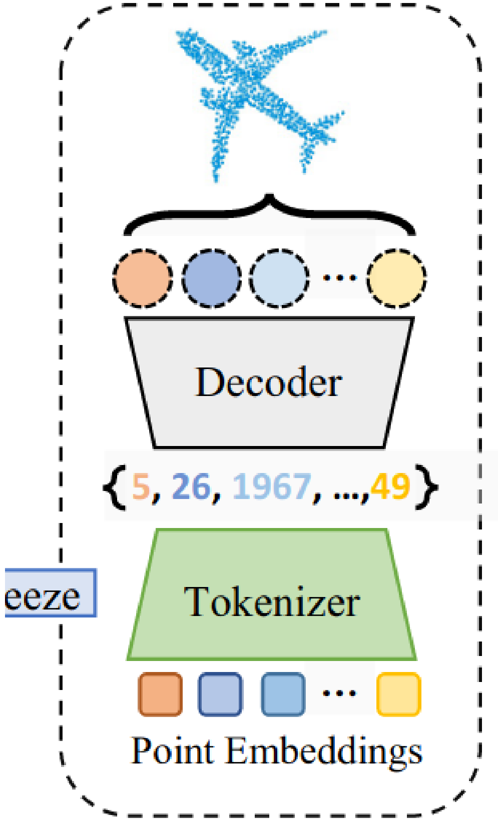




Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. TOG, 2019.

Le Decoder

- Traduire les "mots" en nuages de points
- DGCNN
  - Prendre toute la "phrase" en compte
- FoldingNet pour la reconstruction



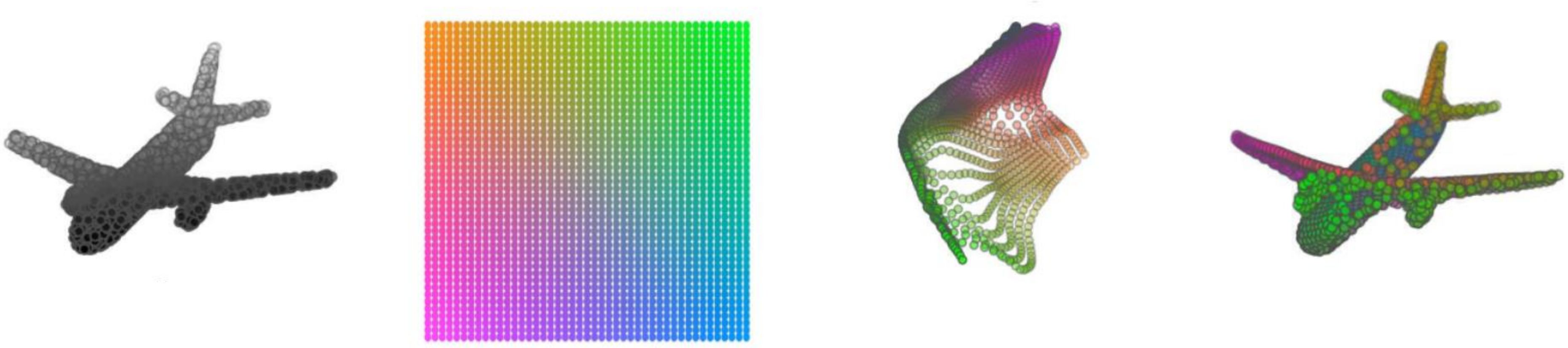
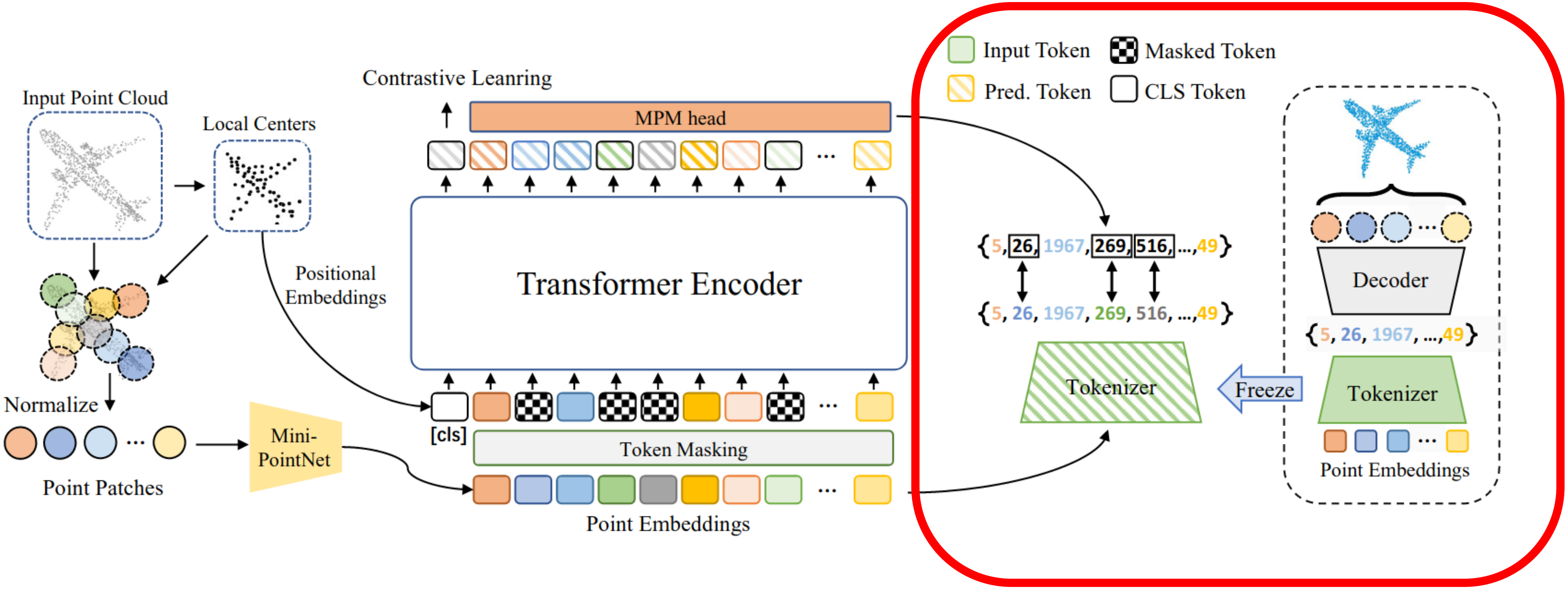


Figure adaptée de Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In CVPR, 2018.

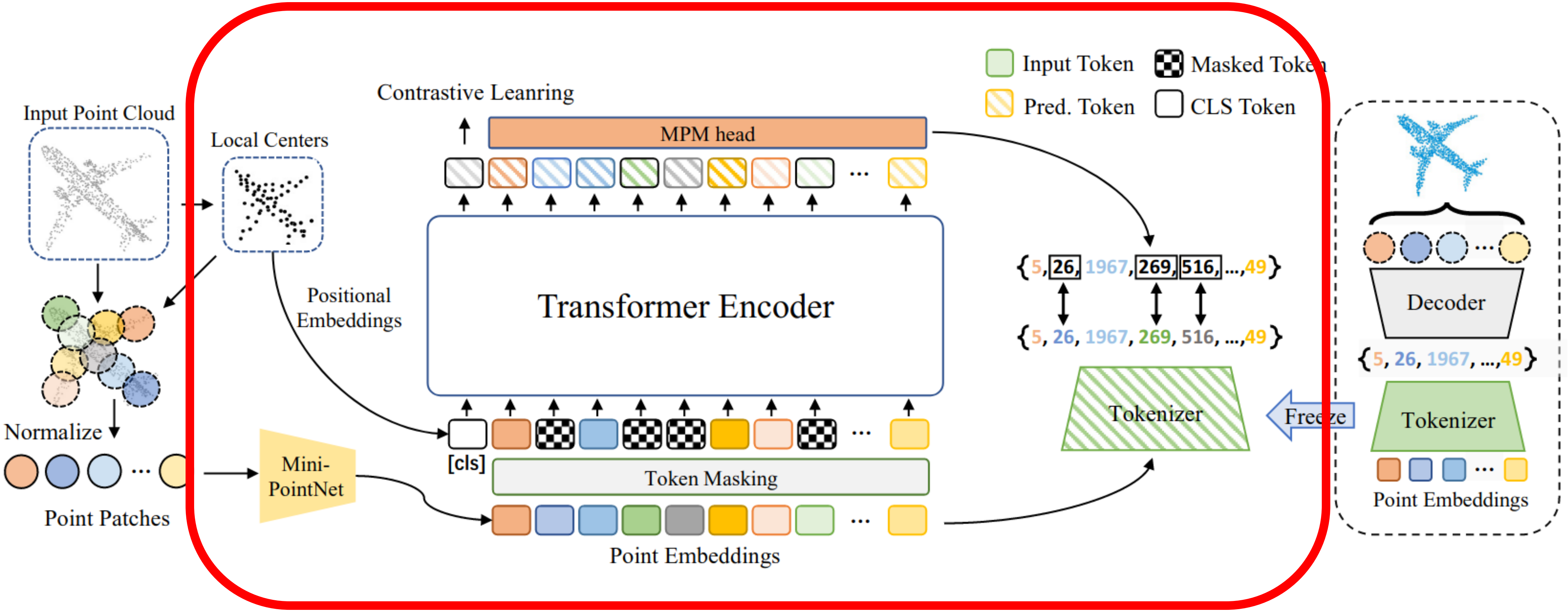
Point-BERT

Pre-training 3D Point Cloud Transformers with Masked Point Modeling



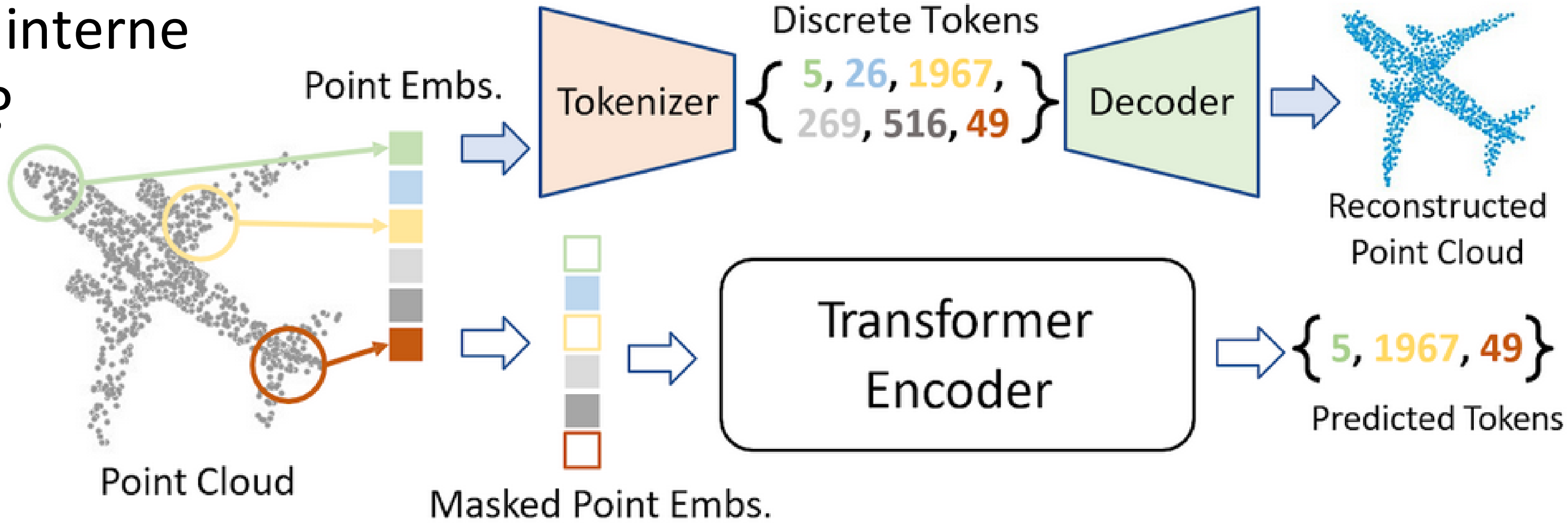
Point-BERT

Pre-training 3D Point Cloud Transformers with Masked Point Modeling



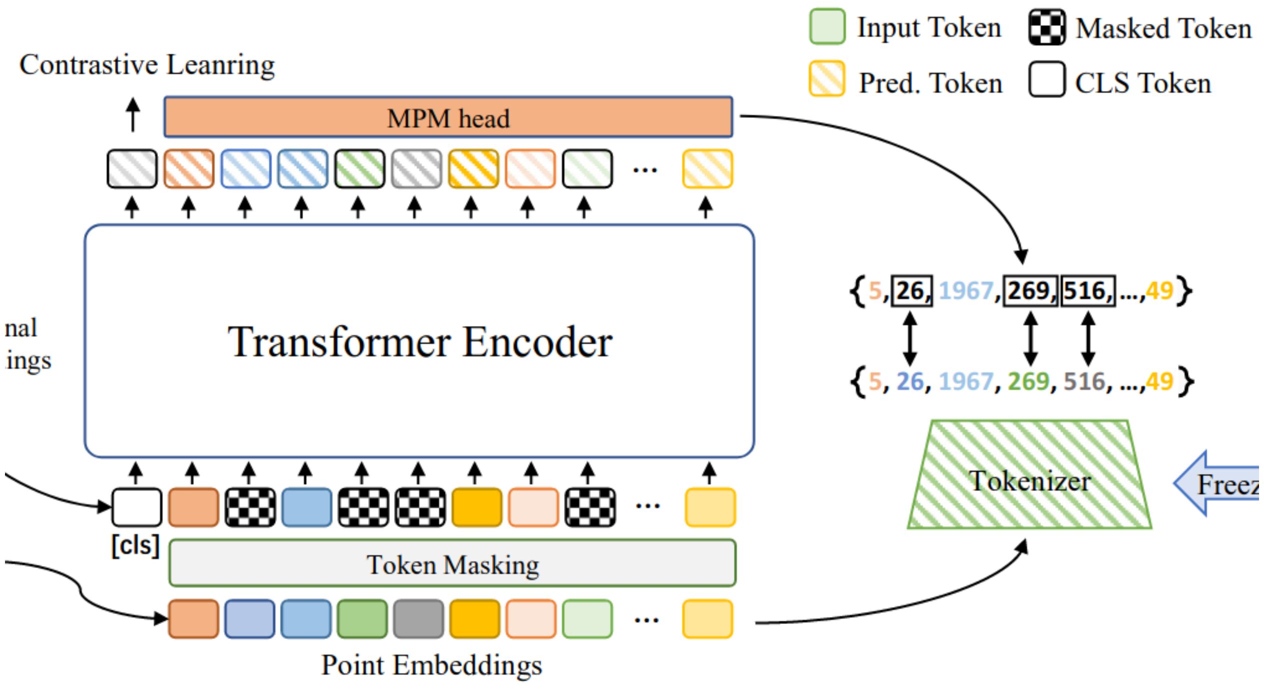
Point-BERT

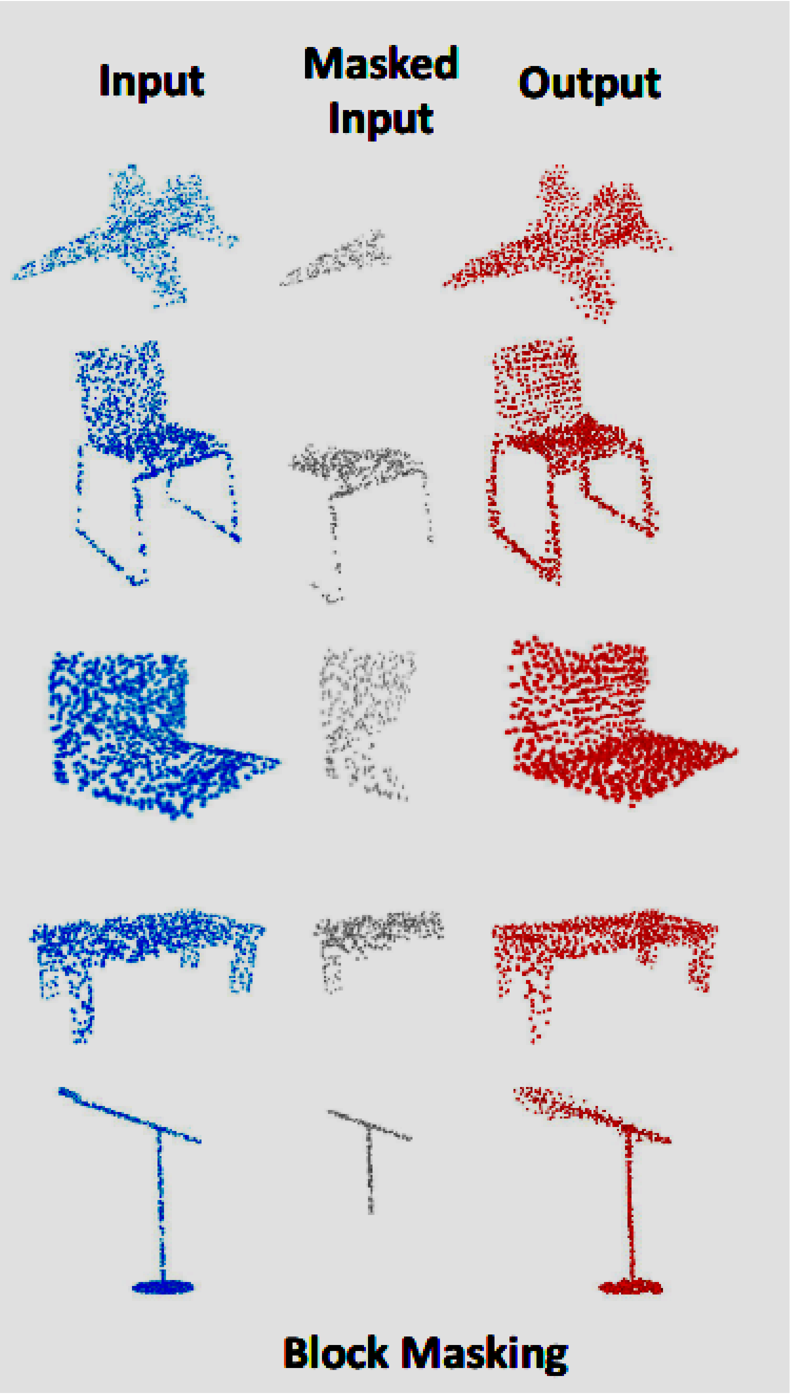
- Encodage des nuages de points
- Tokenizer pour utiliser des concepts de NLP
  - Pont entre les nuages de points et le NLP
- Masked Point Modeling
  - "Compréhension" des nuages
  - Bonne représentation interne
  - Comment on masque?



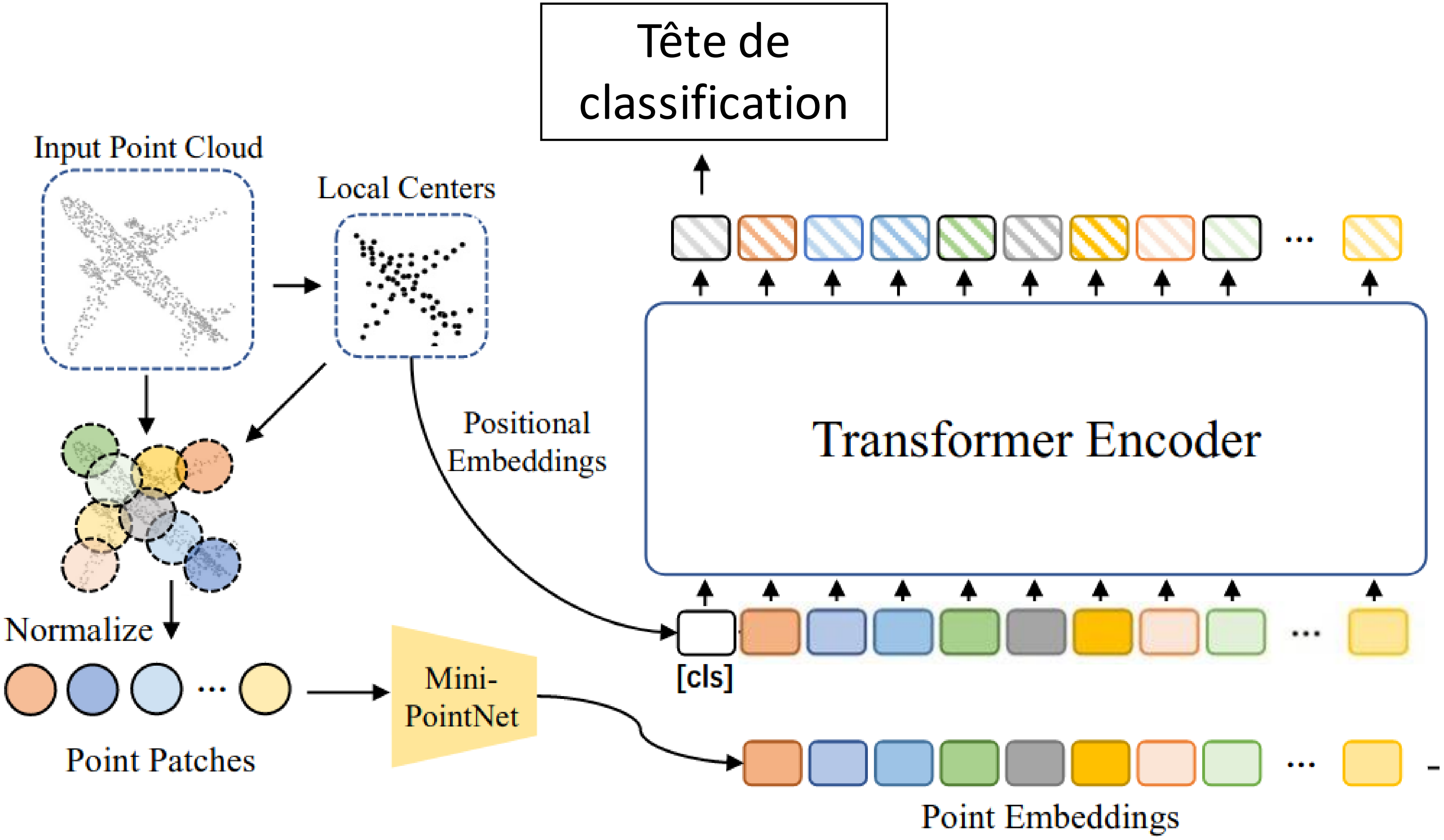
Masked Point Modeling

- Masque en bloc
- Masque 25%-45% des tokens





Fine-tuning: Classification



Classification

- Pré-entraînement sur ShapeNet
- Test sur ModelNet40
- [T] = Transformer + biais inductifs
- Patches = Plus de points

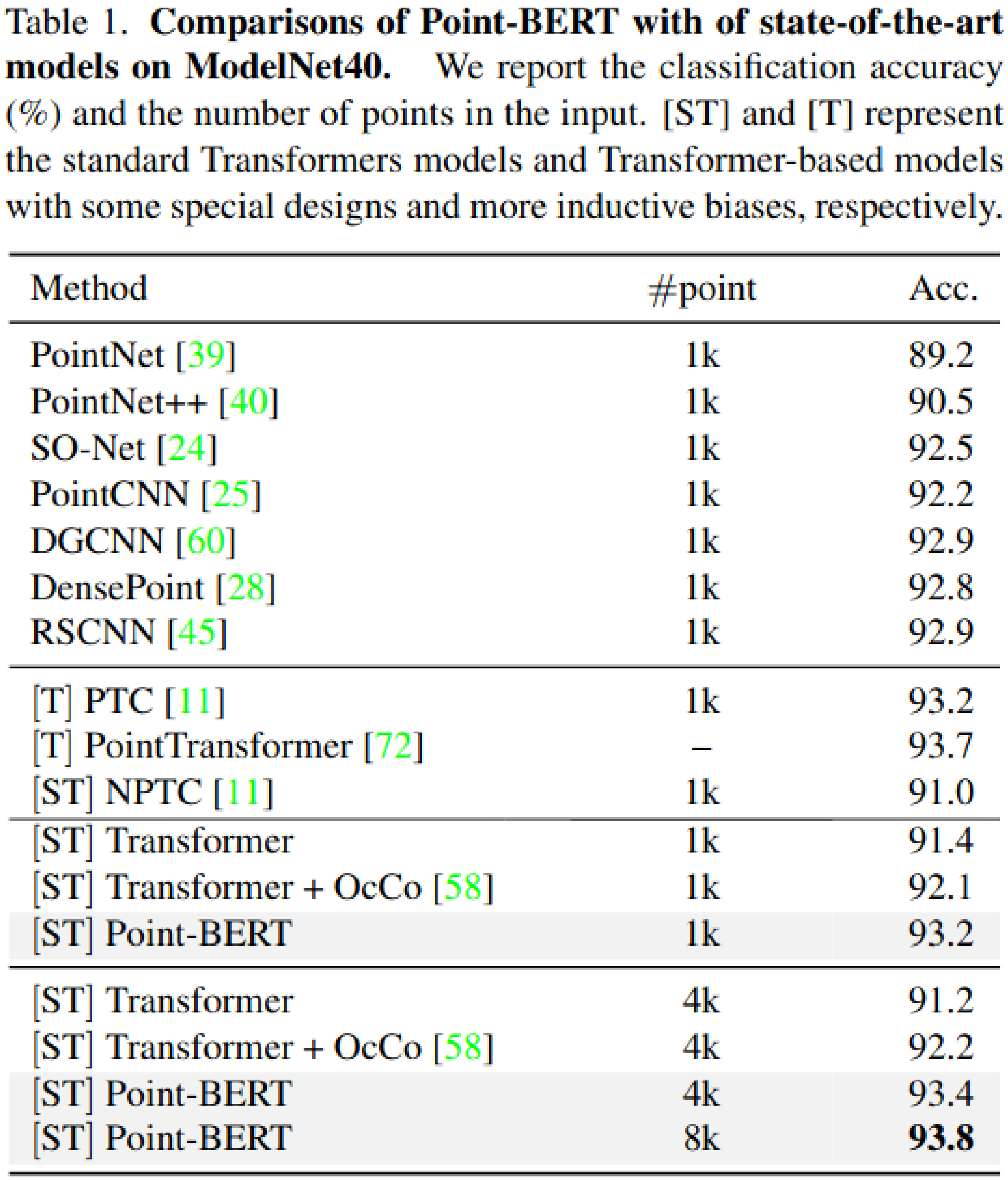


Table 1. **Comparisons of Point-BERT with of state-of-the-art models on ModelNet40.** We report the classification accuracy (%) and the number of points in the input. [ST] and [T] represent the standard Transformers models and Transformer-based models with some special designs and more inductive biases, respectively.

Method	#point	Acc.
PointNet [39]	1k	89.2
PointNet++ [40]	1k	90.5
SO-Net [24]	1k	92.5
PointCNN [25]	1k	92.2
DGCNN [60]	1k	92.9
DensePoint [28]	1k	92.8
RSCNN [45]	1k	92.9
[T] PTC [11]	1k	93.2
[T] PointTransformer [72]	–	93.7
[ST] NPTC [11]	1k	91.0
[ST] Transformer	1k	91.4
[ST] Transformer + OcCo [58]	1k	92.1
[ST] Point-BERT	1k	93.2
[ST] Transformer	4k	91.2
[ST] Transformer + OcCo [58]	4k	92.2
[ST] Point-BERT	4k	93.4
[ST] Point-BERT	8k	**93.8**

Few-Shot Classification

- K-way N-shot
  - K classes
  - N exemples par classe
- Meilleure performance

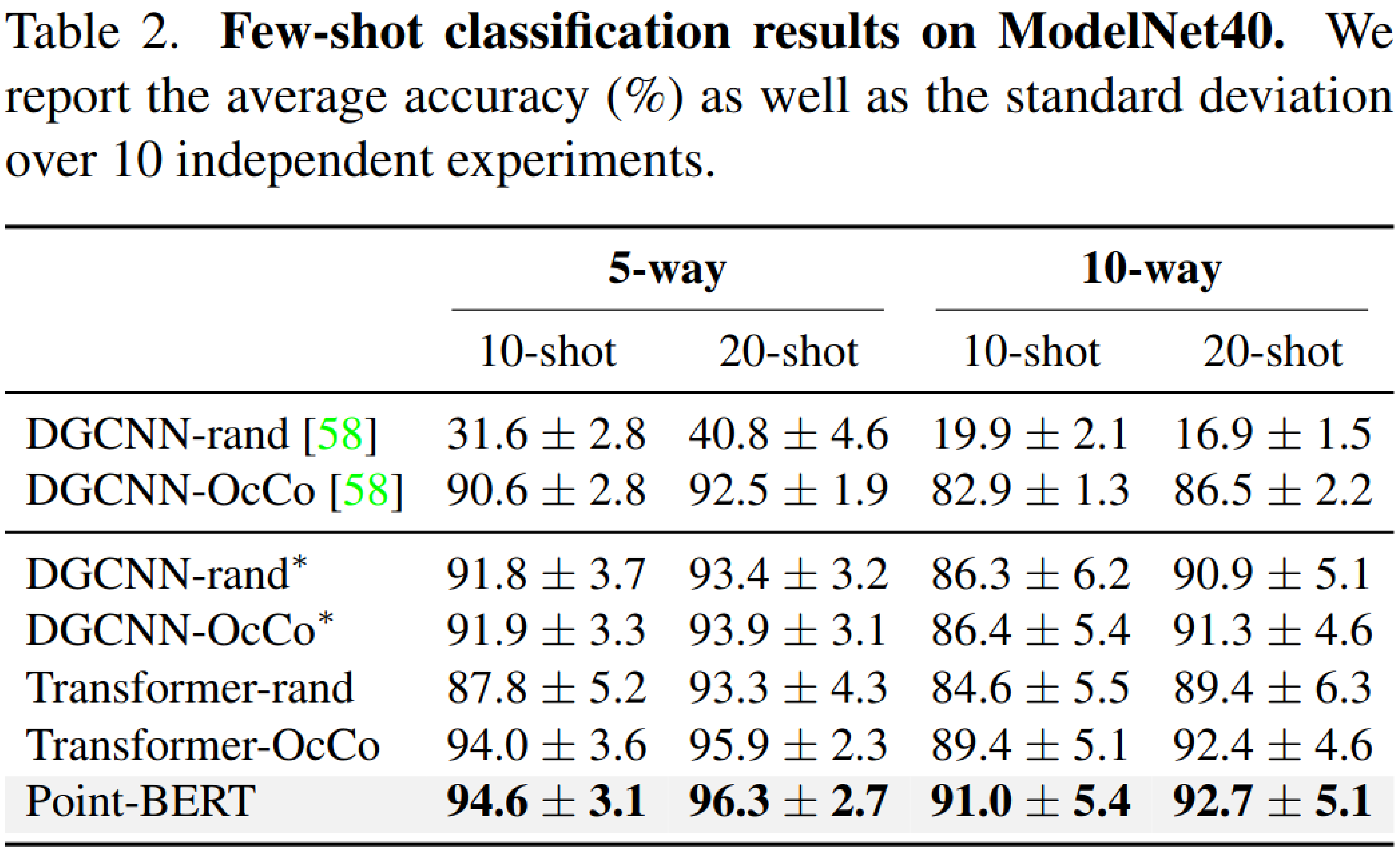
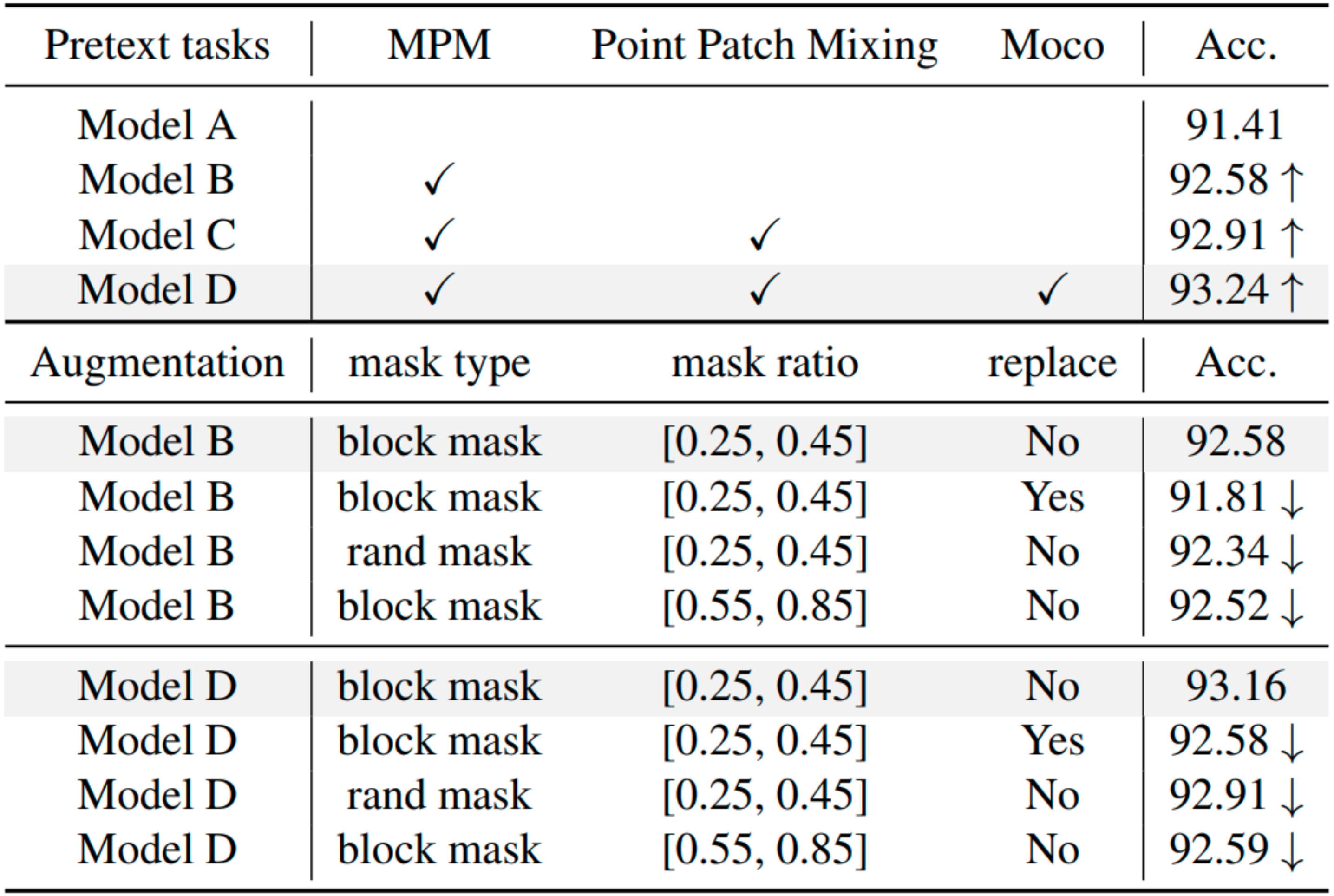


Table 2. **Few-shot classification results on ModelNet40.** We report the average accuracy (%) as well as the standard deviation over 10 independent experiments.

	5-way		**10-way**	
	10-shot	20-shot	10-shot	20-shot
DGCNN-rand [58]	31.6 ± 2.8	40.8 ± 4.6	19.9 ± 2.1	16.9 ± 1.5
DGCNN-OcCo [58]	90.6 ± 2.8	92.5 ± 1.9	82.9 ± 1.3	86.5 ± 2.2
DGCNN-rand*	91.8 ± 3.7	93.4 ± 3.2	86.3 ± 6.2	90.9 ± 5.1
DGCNN-OcCo*	91.9 ± 3.3	93.9 ± 3.1	86.4 ± 5.4	91.3 ± 4.6
Transformer-rand	87.8 ± 5.2	93.3 ± 4.3	84.6 ± 5.5	89.4 ± 6.3
Transformer-OcCo	94.0 ± 3.6	95.9 ± 2.3	89.4 ± 5.1	92.4 ± 4.6
Point-BERT	**94.6** ± **3.1**	**96.3** ± **2.7**	**91.0** ± **5.4**	**92.7** ± **5.1**

Étude d'ablation



Pretext tasks	MPM	Point Patch Mixing	Moco	Acc.
Model A				91.41
Model B	✓			92.58 ↑
Model C	✓	✓		92.91 ↑
Model D	✓	✓	✓	93.24 ↑

Augmentation	mask type	mask ratio	replace	Acc.
Model B	block mask	[0.25, 0.45]	No	92.58
Model B	block mask	[0.25, 0.45]	Yes	91.81 ↓
Model B	rand mask	[0.25, 0.45]	No	92.34 ↓
Model B	block mask	[0.55, 0.85]	No	92.52 ↓
Model D	block mask	[0.25, 0.45]	No	93.16
Model D	block mask	[0.25, 0.45]	Yes	92.58 ↓
Model D	rand mask	[0.25, 0.45]	No	92.91 ↓
Model D	block mask	[0.55, 0.85]	No	92.59 ↓

Yu, Xumin, et al. "Point-BERT: Pre-training 3D Point Cloud Transformers with Masked Point Modeling." arXiv preprint arXiv:2111.14819 (2021).

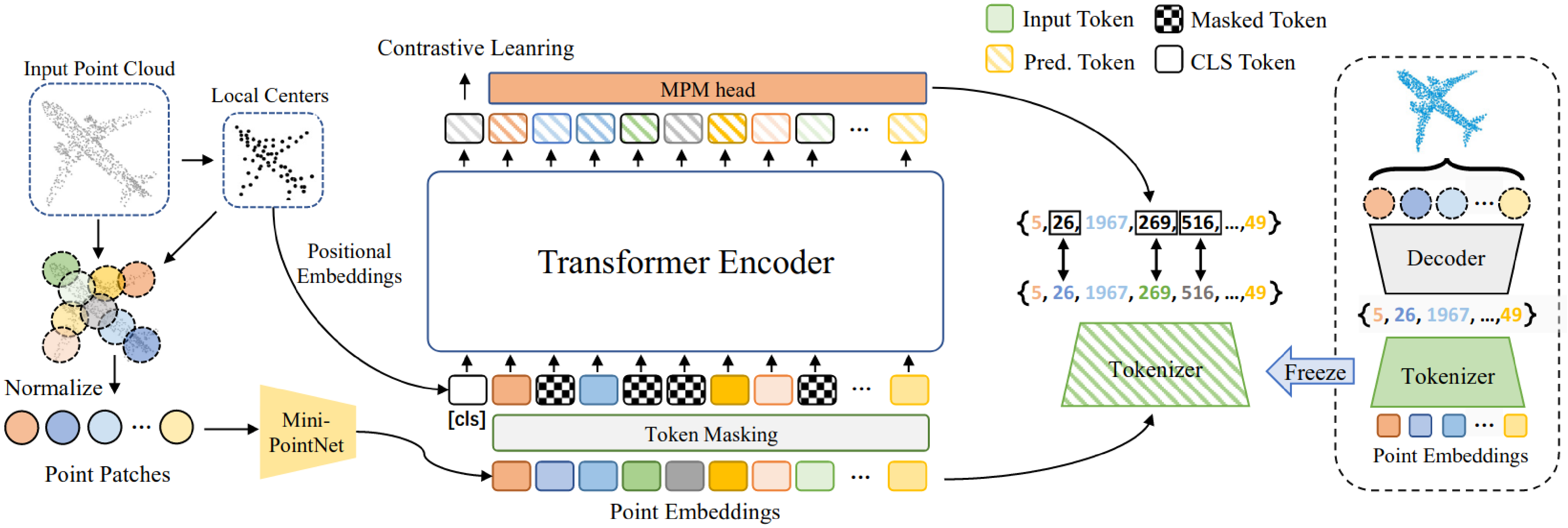
Conclusion

- Touche à beaucoup de techniques
  - PointNet
  - Transformers
  - GNN
  - FoldingNet
- SSL semble être le futur
  - Acquisition peu coûteuse
  - Coûts d'annotation très grands
  - Multi-modalité
- Liens entre les disciplines de l'apprentissage profond
  - Avenue de recherche intéressante
  - Plus d'interdisciplinarité

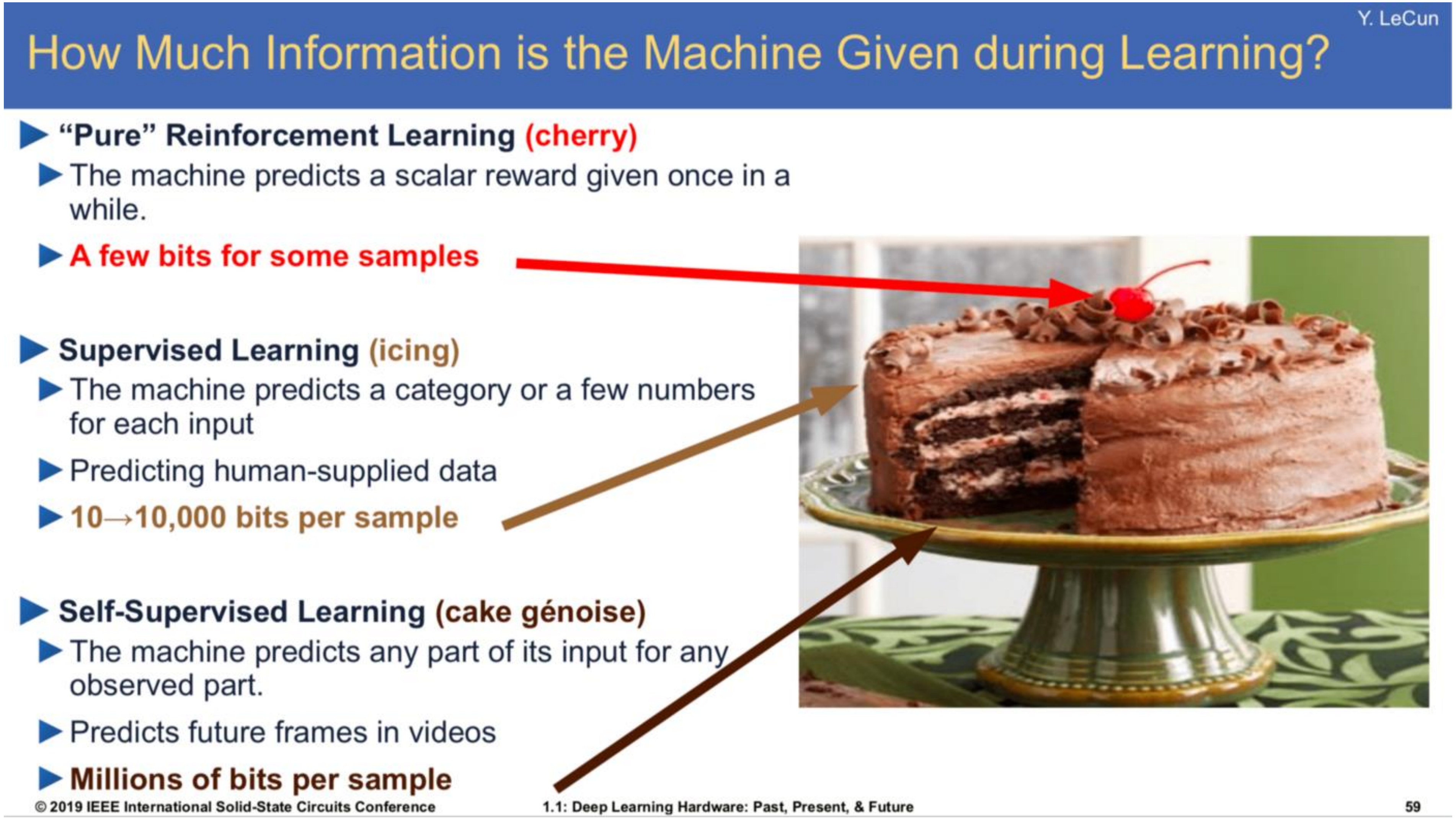
Références

- Yu, Xumin, et al. "Point-BERT: Pre-training 3D Point Cloud Transformers with Masked Point Modeling." arXiv preprint arXiv:2111.14819 (2021).
- Déziel, Jean–Luc, et al. "PixSet: An Opportunity for 3D Computer Vision to Go Beyond Point Clouds With a Full-Waveform LiDAR Dataset." 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE, 2021.
- Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
- Bello, Saifullahi Aminu, et al. "Deep learning on 3D point clouds." *Remote Sensing* 12.11 (2020): 1729.
- Zheng, Wu, et al. "SE-SSD: Self-ensembling single-stage object detector from point cloud." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
- Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. TOG, 2019.
- Wang, Yue, et al. "Dynamic graph cnn for learning on point clouds." Acm Transactions On Graphics (tog) 38.5 (2019): 1-12.
- He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
- Guimont-Martin, William. "Présentation de Point-BERT" William Guimont-Martin, 2022, https://willguimont.github.io/cs/2022/01/28/point-bert.html

Questions?


Input Token
Masked Token
Pred. Token
CLS Token
Input Point Cloud
Local Centers
Contrastive Leanring
MPM head
Positional Embeddings
Transformer Encoder
Normalize
Point Patches
Mini-PointNet
[cls]
Token Masking
Point Embeddings
{5, 26, 1967, 269, 516, ...,49}
{5, 26, 1967, 269, 516, ...,49}
Tokenizer
Freeze
Decoder
{5, 26, 1967, ...,49}
Tokenizer
Point Embeddings

LeCake



Y. LeCun

How Much Information is the Machine Given during Learning?

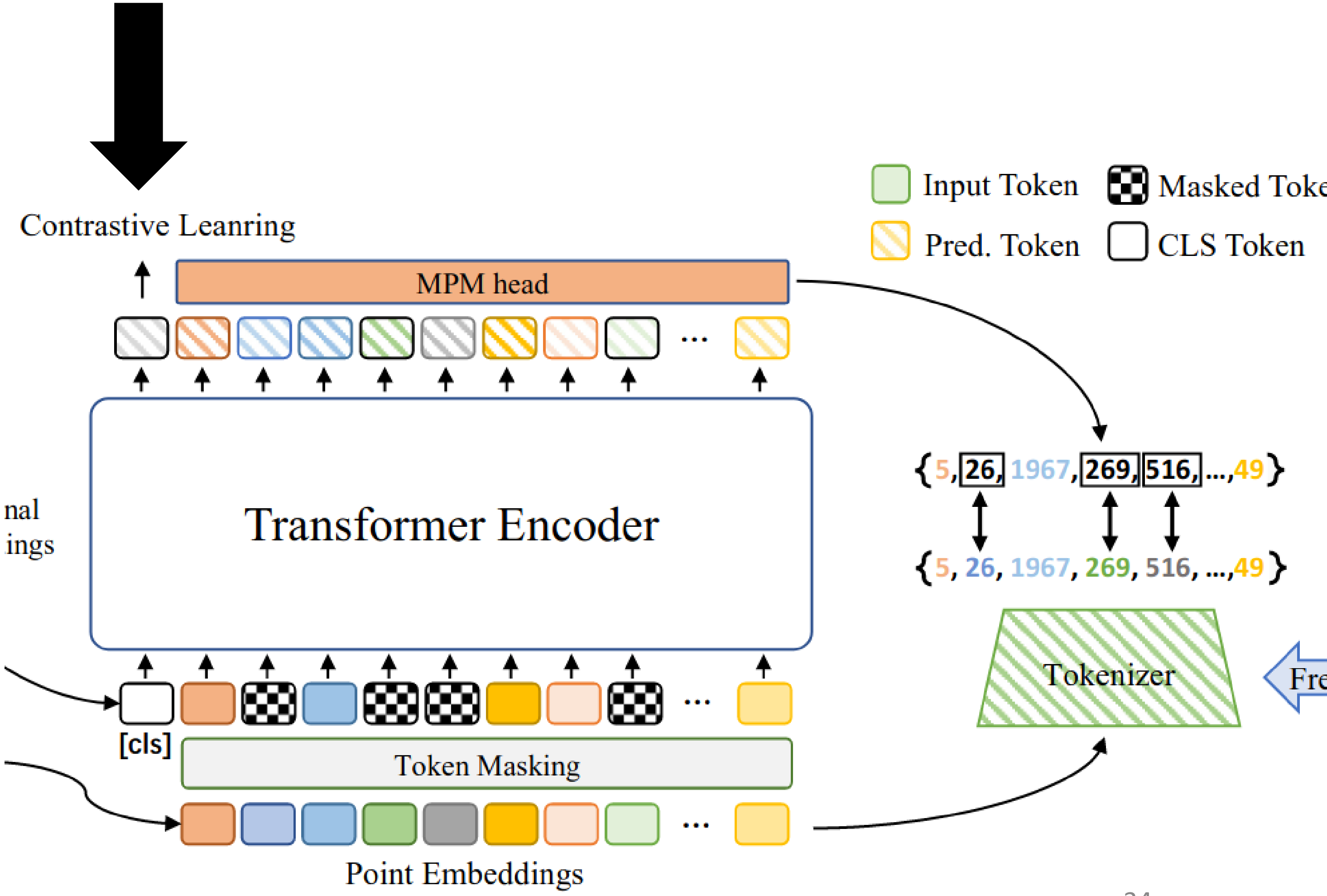
- **“Pure” Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

© 2019 IEEE International Solid-State Circuits Conference 1.1: Deep Learning Hardware: Past, Present, & Future 59

LeCun, Yann. "Predictive Learning" NIPS 2016.

Contrastive Learning

- Autre type de SSL
- Sémantique de haut niveau
    - CLS token
- MoCo [0]

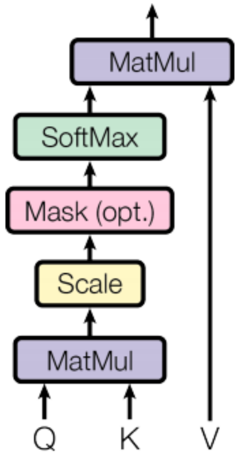


[0] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In CVPR, 2020.

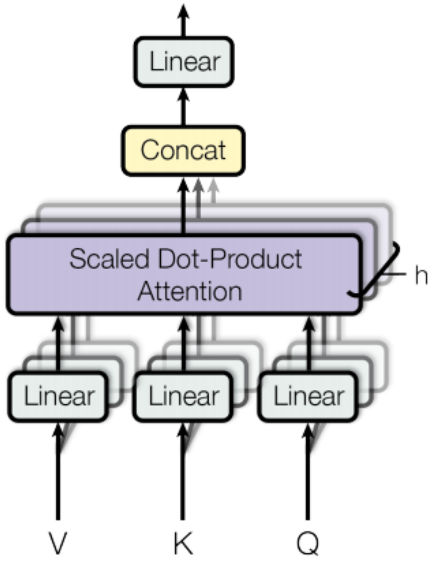
QKV Attention

- Query
- Key
- Value

Scaled Dot-Product Attention



Multi-Head Attention



Want more transformers?

- Transformers in Computer Vision (French)