Plan prévisionnel

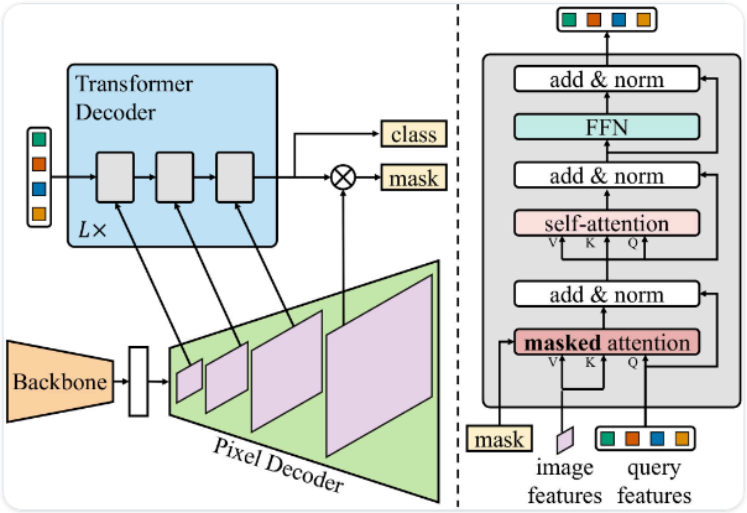
Dataset retenu

Le dataset **Cityscapes** est une référence majeure pour la segmentation sémantique urbaine. Il contient **5 000 images haute résolution** capturées dans plus de **50 villes européennes** (principalement en Allemagne), accompagnées de masques annotés pixel par pixel.
Il est largement utilisé pour entraîner et évaluer des modèles de vision par ordinateur destinés à la **conduite autonome**, à l'analyse de scènes routières et aux systèmes embarqués.

Pour ce POC, une version simplifiée du dataset a été utilisée afin de réduire le coût de calcul tout en conservant la diversité visuelle nécessaire.

Modèle envisagé

Pour ce POC, le modèle envisagé est **Mask2Former**, une architecture récente basée sur les Transformers.
Il se distingue par une approche unifiée de la segmentation (sémantique, instance, panoptique) grâce à un mécanisme d'attention masquée et un pixel decoder multi-échelle.
Ce modèle a été retenu pour sa **capacité à mieux capturer le contexte global**, la finesse des contours, et ses performances reconnues sur les scènes urbaines, ce qui en fait un candidat pertinent pour une future utilisation embarquée.



Mask2Former architecture. Taken from the original paper.

Références bibliographiques

Pour cette preuve de concept, le modèle sélectionné est **Mask2Former**, un modèle de segmentation universelle basé sur les Transformers. Son choix repose sur l'analyse de deux références essentielles permettant de comprendre son intérêt par rapport aux approches traditionnelles.

1. Article Medium — MaskFormer2 : Masked-attention Mask Transformer for Universal Image Segmentation
https://medium.com/@HannaMergui/maskformer2-masked-attention-mask-transformer-for-universal-image-segmentation-c3d14c546d6b

L'article consulté présente de manière pédagogique l'évolution des architectures de segmentation vers des modèles dits **universels**, capables de traiter simultanément la segmentation sémantique, instance et panoptique. Après avoir rappelé les limites des approches spécialisées (FCN pour le sémantique, Mask R-CNN pour l'instance, architectures hybrides pour le panoptique), l'article introduit le concept de **mask classification**, popularisé par MaskFormer, qui remplace la classification pixel-par-pixel par une prédiction de segments associés à des embeddings. L'auteur explique comment Mask2Former améliore cette approche grâce à deux innovations majeures : le **masked attention**, qui restreint l'attention du Transformer aux régions pertinentes du masque pour accélérer l'apprentissage et améliorer la précision, et l'utilisation de **features multi-échelle**, traitées couche par couche dans le décodeur, afin de mieux capter petits objets et contextes globaux. L'article montre que cette combinaison permet à Mask2Former de dépasser les architectures spécialisées sur l'ensemble des tâches et de devenir l'un des standards actuels de la segmentation. Cette synthèse constitue une base claire pour comprendre pourquoi Mask2Former représente aujourd'hui une solution performante et polyvalente pour la segmentation d'images complexes.

2. Documentation Hugging Face — *Mask2Former*
https://huggingface.co/docs/transformers/model_doc/mask2former

La documentation officielle Hugging Face présente **Mask2Former** comme l'évolution directe de MaskFormer.
Elle met en avant plusieurs améliorations majeures :

- **Masked attention**, qui limite la zone d'attention aux régions prédictives pour une segmentation plus précise.
- Un **pixel decoder multirésolution**, permettant de récupérer efficacement les détails fins.
- Une capacité **universelle** : un seul modèle pour la segmentation sémantique, panoptique ou par instance.

- Des performances SOTA sur plusieurs benchmarks (COCO, ADE20K, Cityscapes).

Cette ressource fournit la description technique officielle du modèle, ses hyperparamètres, ses sorties, ainsi que les fonctions de post-traitement, garantissant une implémentation fiable pour le POC.

Démarche de test

Pour évaluer le nouvel algorithme, j'ai défini une démarche centrée sur la comparaison objective avec une **baseline**. La première étape consiste à préparer une version réduite du dataset Cityscapes pour accélérer les expérimentations (100 images). J'entraîne ensuite un modèle de référence — **DeepLabV3+** — afin d'établir un premier niveau de performance sur la segmentation.
Le modèle récent choisi — **Mask2Former** — est entraîné avec le même pipeline, les mêmes métriques (loss, pixel accuracy, mIoU, images/sec) et les mêmes conditions de validation.

Enfin, j'exécute un **Random Search** identique pour les deux modèles afin d'optimiser leurs hyperparamètres et permettre une comparaison équitable. Cette démarche POC permet d'évaluer rapidement la viabilité du modèle Mask2Former et de vérifier s'il surpasse la baseline sur un sous-ensemble représentatif avant d'envisager un entraînement complet.