Et en 3D, vous y voyez quelque chose ? Mais oui grâce à mes réseaux neuronaux.

Rédigé par Eric Debeir – Directeur scientifique - eric@datalchemy.net

NeRFs, Signed Distance Fields… Derrière ces termes barbares, se cachent une petite révolution scientifique née de l'intelligence artificielle aux nombreuses applications. Nous vous proposons aujourd'hui un focus sur ce nouveau domaine qui est apparu aux alentours de 2021 et qui a récemment connu de nouveaux résultats impressionnants. Nous repartirons ainsi d'un travail incontournable de *NVIDIA*, les *Instant Neural Graphic Primitives* pour nous intéresser ensuite à deux travaux récents : *LeRFs*, et le *Gaussian Splatting*. Mais avant cela, une introduction plus complète s'impose.

De quoi parlons-nous ?

Un enjeu fondamental est celui de la *Computer Vision* : exploiter une information visuelle issue de capteurs, afin de modéliser une représentation du monde. Ces modélisations sont souvent en deux dimensions (à partir d'une photographie classique), mais le passage à une représentation en 3 dimensions est incontournable si l'on veut ensuite pouvoir agir correctement : détection précise d'un élément dans l'espace, manipulation robotique, etc.

Deux grandes approches existent fondamentalement pour représenter une scène en 3 dimensions :

- Une modélisation par des *voxels*, soit un équivalent en 3 dimensions d'un *pixel*. Là où un *pixel* est un point atomique en deux dimensions, un *voxel* est une brique atomique de représentation de l'espace. La modélisation par des *voxels* est souvent inutilement lourde, en ceci que les objets que l'on désire identifier occupent une part minimale de l'espace total. De nombreux travaux ont tenté d'optimiser cette approche, les curieux pourront ainsi s'intéresser au *Minkowski Engine*.
- Une modélisation par de simples points en 3 dimensions, sous l'appellation de « nuage de points ». Cette modélisation est souvent la plus efficace, et de nombreuses architectures *Deep Learning* sont optimisées pour gérer cette forme de la donnée. Le lecteur intéressé pourra se reporter au *PointNet* ainsi qu'à ses nombreux successeurs.

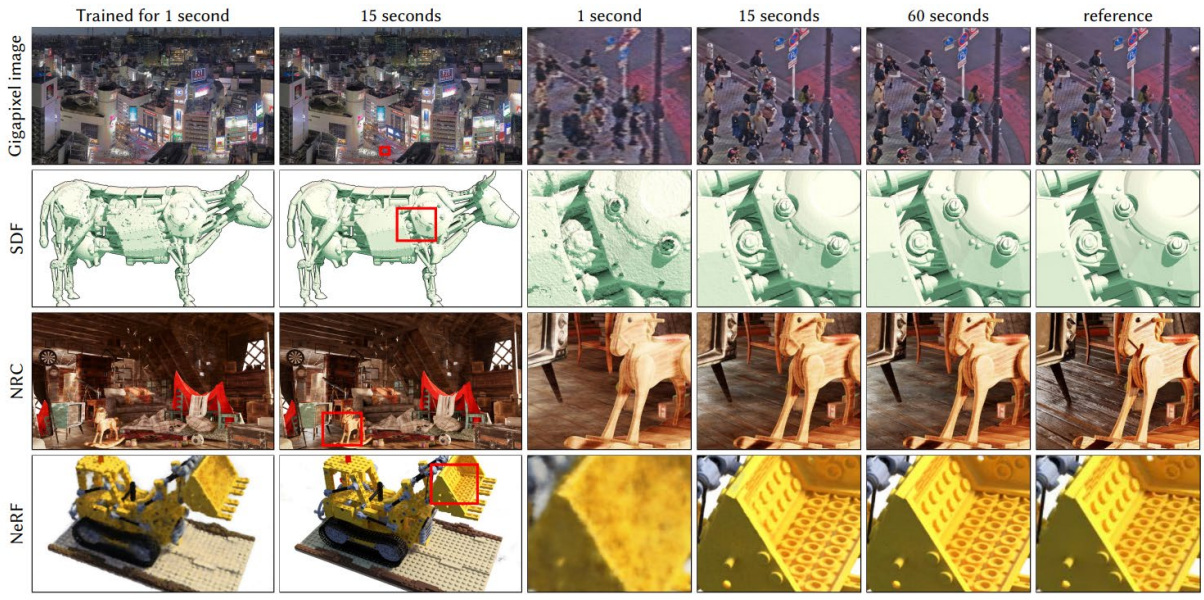
NeRFs (Neural Radience Fields) et *SDFs (Signed Distance Fields)* sont deux nouvelles méthodes pour modéliser une scène en trois dimensions. Chacune a ses avantages et ses défauts, et nous ne rentrerons pas trop dans les détails techniques dans cet article. Retenons surtout que ces approches visent à modéliser une scène via l'entraînement d'un réseau de neurones spécifique, réseau de neurones qui, une fois entraîné, pourra ensuite être utilisé pour questionner l'apparence de chaque « point » à partir d'une caméra virtuelle.

Un intérêt particulier de ces méthodes et que l'on peut générer la représentation en 3 dimensions à partir d'une collection de photographies en 2 dimensions prises à des angles différents. Au-delà, les *NeRFs* et *SDFs* permettent une reconstruction d'une qualité photo-réaliste inégalée, dépassant tout ce qui existait auparavant.

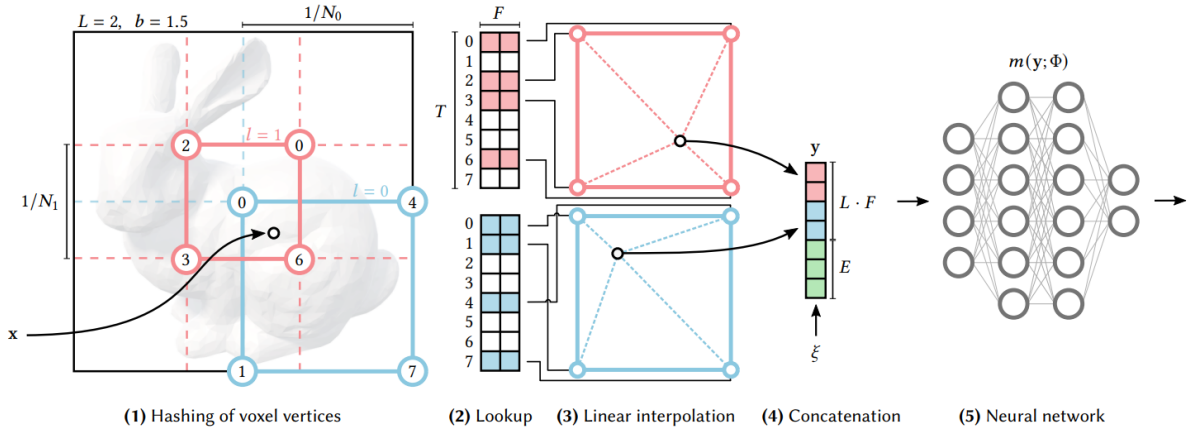
Jusqu'au travail suivant, le principal problème était l'extrême lenteur (due à la lourdeur de calcul) pour générer une nouvelle scène. Mais *NVIDIA* lança une première révolution avec les *Instant NGPs…*

Instant NGP – Instant Neural Graphic Primitives with a Multiresolution Hash Encoding

Ce travail de *Müller et al* [1] a eu l'effet d'un coup de tonnerre dans le monde de la recherche. Les auteurs ont ici travaillé globalement l'idée d'entraîner un réseau de neurones pour représenter une scène, en travaillant sur quatre applications différentes : *Gigapixel image*, *SDF*, *NRC* et *NeRF*. La principale évolution ici est la rapidité d'entraînement et de génération, avec une approche considérée comme « instantanée » (ce qui est peut être légèrement exagéré). À partir d'une collection de photographies de la scène cible, un réseau de neurones est entraîné avec une précision croissante en fonction du temps alloué pour l'entraînement.



La principale révolution apportée par les auteurs porte sur la méthode pour modéliser les détails de la scène en fonction du niveau de résolution visé. Ici, une architecture hiérarchique encode des *hashs* permettant d'identifier la manière dont nous voulons appeler le réseau de neurones entraîné. Cette méthode est extrêmement efficace, et visiblement suffisamment volatile pour être appliquée aux quatre domaines cibles de la publication.

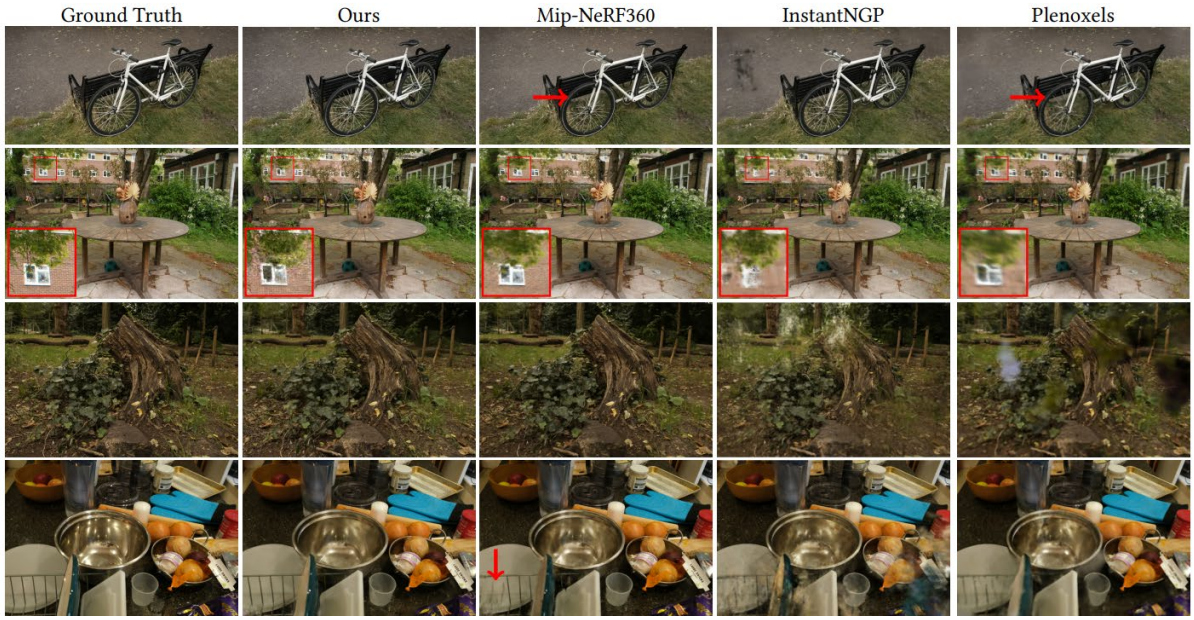


Autre argument très fort de ce travail, la rapidité d'entraînement du modèle qui était bien plus intéressante que ce qui existait jusqu'ici, rendant l'approche pratiquement exploitable. À noter que cette rapidité n'est pas étrangère à l'affiliation des auteurs (*NVIDIA*). L'algorithme de *multi-hash encoding* a ainsi été implémenté directement en *CUDA* pour être le plus efficace possible sur un *GPU*. Ce travail est disponible sur *github* aujourd'hui, quand bien même la licence du code source ne permet pas une utilisation commerciale…

On notera que *NVIDIA* a notamment accompagné un grand nombre de créations artistiques exploitant ce travail, en encourageant l'utilisation dans différents cadres de création… Néanmoins, à l'époque, ce travail s'est avéré insuffisant pour imaginer une application à l'industrie réaliste et efficace. De nombreux chercheurs ont travaillé le sujet, et nous vous proposons de vous présenter deux travaux récents passionnants…

3D Gaussian Splatting for Real-Time Radiance Field Rendering

Ce travail de *Kerbl et al* [2] de l'*Inria* (Cocorico de rigueur) a récemment chamboulé le domaine en proposant des résultats de bien meilleure qualité avec un rendu temps réel beaucoup plus exploitable. On a ainsi vu récemment de nombreux acteurs se jeter sur cette approche pour créer des visualiseurs *web* de *NeRFs*, ou pour faire du rendu temps réel animé de *meshs* tout en conservant une qualité proche du photo-réalisme. Ici, les auteurs arrivent à générer de nouvelles vues à partir d'une scène enregistrée à 30 images par secondes, à une résolution de 1080p. Les résultats sont impressionnants :

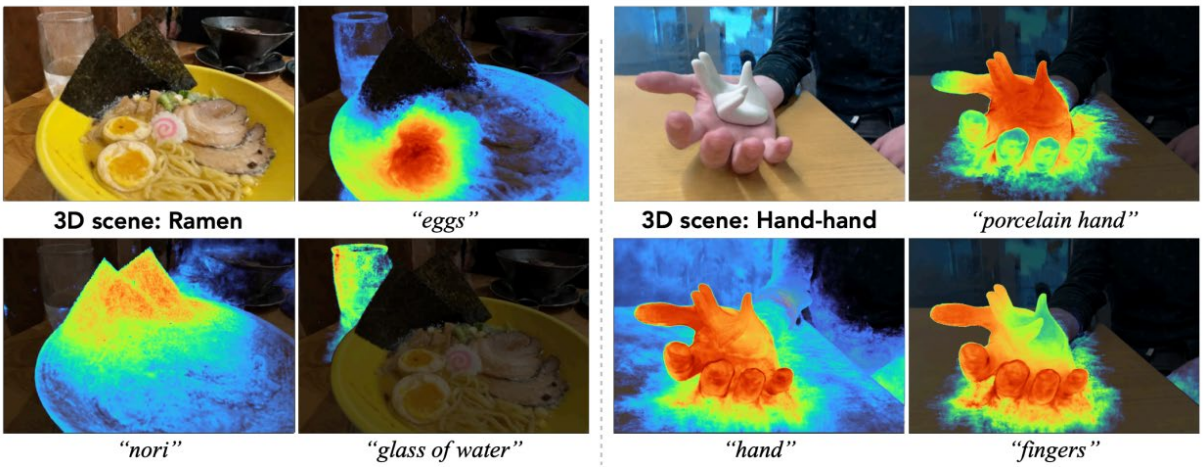


À ce jour, cette approche est la plus pertinente pour entraîner et utiliser un *NeRF*, en attendant les travaux ultérieurs. Il faut noter que cette notion de rapidité de rendu est fondamentale pour utiliser correctement ces outils, par exemple en rendu d'animation (pour explorer une scène acquise, voire pour animer une scène). Nous suivons avec beaucoup d'attention ces travaux d'optimisation, et attendons les prochains pour continuer d'identifier les meilleures opportunités pour nos clients.

LERF : Language embedded Radiance Field

Le travail suivant de *Kerr et* al [3] fera sourire ceux qui suivent le domaine du *Deep Learning*, car après tout, existe-t-il ne serait-ce qu'un domaine où les récents progrès en traitement du langage n'ont pas été appliqué ?

Ici, l'idée est de coupler l'apprentissage d'un *NeRF* avec le modèle *CLIP* d'*OpenAI*. Pour rappel, ce dernier a été entraîné à rapprocher, dans un espace latent, une description textuelle d'une image. *CLIP* a déjà été extensivement utilisé pour questionner le contenu d'une image à partir d'un *prompt* textuel. Ici, les auteurs entraînent le modèle de manière à pouvoir interroger un *NeRF* à partir d'un texte de description. L'image ci-dessous devrait éclaircir le sujet :



Attention : c'est bien une scène en 3 dimensions qui est travaillé, et non une simple image. Ce travail nous permet de mettre en évidence une limite aujourd'hui des *NeRFs*, à savoir, leur utilisation. C'est une chose de modéliser une scène en 3d d'une manière photo-réaliste, c'en est une autre de pouvoir ensuite extraire de l'information ou de pouvoir localiser une partie à partir d'une requête

sémantique. Ici, l'approche permet directement de pouvoir requêter un modèle entraîné pour isoler un contenu intéressant à partir d'une description précise ou non : description d'un objet, d'une couleur, d'une notion, etc. Evidemment, les alertes et limites que nous avons donné sur le contrôle par le langage restent ici valables.

Les auteurs ont d'ailleurs l'élégance de préciser certaines de ces limites, notamment l'aspect « *bag of word* » : une requête textuelle ne respectera pas nécessairement le sens logique de la phrase. *« not red »* est similaire à « *red »*

Ce travail n'en reste pas moins une étape importante dans la recherche académique sur le sujet des *NeRFs*

Faut-il aujourd'hui « faire du NeRF » ?

Suivre la recherche académique est une chose, l'appliquer à un problème concret en est une autre. Des premières applications sont déjà concrètes pour acquérir et restituer une scène en 3 dimensions. Des acteurs visent ainsi à créer de nouveaux outils basés sur ces techniques, pour par exemple diffuser l'apparence en trois dimensions d'un produit sur Internet, ou pour faciliter une capture. Néanmoins, ces outils sont encore très jeunes et souffrent encore d'une réelle lourdeur dans leur utilisation. On remarque aussi qu'il y a, à date, encore peu de travaux pertinents (d'un point de vue industriel/applicatif) pour résoudre un problème technique à base de *NeRF/SDF*. C'est une chose d'acquérir une représentation en 3 dimensions d'une scène, c'en est une autre d'extraire des informations concrètes avec un contrôle réel des limites et des résultats de l'outil.

Nous avons l'impression que nous approchons du moment où les choses vont basculer. De nombreux défis techniques et scientifiques ont été relevés ces deux dernières années, et un bon pari est que 2024 sera l'année où le traitement *Computer Vision* d'une scène par ces approches deviendra un réflexe.

Si vous désirez suivre ces travaux, nous pouvons recommander l'excellent site https://neuralradiancefields.io/ qui suit et commente l'actualité de ce nouveau domaine scientifique.

[1] : https://nvlabs.github.io/instant-ngp/assets/mueller2022instant.pdf

[2] : https://arxiv.org/abs/2308.04079

[3] : https://www.lerf.io/