Architecture efficiente de Transformer pour l'acquisition d'image à grande dynamique sur système léger.

Steven TEL[1,2] Barthélémy HEYRMAN[1] Dominique GINHAC[2]

[1]ImViA EA7535, Université de Bourgogne, Dijon, France

[2]ICB UMR 6303 CNRS, Université de Bourgogne, Dijon, France

Résumé – L'imagerie à haute gamme dynamique (HDR) reste un défi pour la photographie numérique moderne. Les récentes recherches proposent des solutions d'acquisition HDR de haute qualité, mais au prix d'un grand nombre d'opérations et d'un temps d'inférence très long empêchant la mise en œuvre de ces solutions sur des systèmes légers. Nous proposons une nouvelle architecture efficiente de Transformer pour l'acquisition d'image HDR basée sur un module d'attention additive. À notre connaissance, notre solution est la première architecture Transformer pour l'imagerie HDR pouvant être exécutée sur système léger. En réalisant des comparaisons qualitatives et quantitatives de notre réseau avec l'état de l'art, nous démontrons que notre réseau produit des résultats compétitifs en terme de qualité tout en étant plus rapide que l'état de l'art. Les résultats expérimentaux montrent que notre méthode obtient un score μ-PSNR de 44,13 sur le jeu de données de référence proposé par Kalantari *et al.* [2] et peut être exécuté à 11 images par seconde en utilisant un processeur neural Apple M1.

Abstract – High dynamic range (HDR) imaging remains a challenge for modern digital photography. Recent research proposed high-quality HDR acquisition solutions, but at the cost of a large number of operations and a long inference time, making it difficult to implement these solutions on lightweight systems. We propose a new efficient Transformer architecture for HDR imaging based on an additive attention module. To our knowledge, our solution is the first Transformer architecture for HDR imaging that can be executed on lightweight systems. By performing qualitative and quantitative comparisons of our network with the state of the art, we demonstrate that our network produces competitive results in terms of quality while being faster than the state of the art. Experimental results show our method obtains a μ-PSNR score of 44.13 on the reference dataset proposed by Kalantari *et al.* [2] and can be executed at 11 frames per second using a Apple M1 neural processor.

1 Introduction

La plupart des caméras standards ne sont pas en mesure de reproduire fidèlement la gamme d'illuminations d'une scène naturelle, les limitations de leurs capteurs entraînent une perte d'informations structurelles ou texturales dans les régions sous-exposées et sur-exposées de la scène acquise. Pour relever ce défi, des capteurs avec une plage dynamique plus élevée ont été proposés pour capturer davantage de niveaux d'intensité d'éclairage de la scène, mais ces solutions sont coûteuses, ce qui empêche l'acquisition à grande dynamique (HDR) d'être facilement accessible. Pour rendre l'imagerie HDR plus accessible, des solutions logicielles ont été proposées, basées sur l'émergence de l'apprentissage profond dans les applications de vision par ordinateur. Elles acquièrent une seule image à faible plage dynamique (LDR) et tentent d'étendre sa plage dynamique grâce à un réseau génératif. Bien que ces méthodes produisent des images avec une plage d'illumination plus élevée, elles présentent des limites dans l'extension de la plage dynamique de l'image d'entrée. Une approche plus efficace consiste à acquérir plusieurs images LDR avec des temps d'exposition différents et à les fusionner en une seule image composite présentant une gamme plus étendue de valeurs de luminosité. Cependant, dans de nombreux cas d'utilisation, la scène acquise contient du mouvement et les images sont capturées en séquence rapide à partir d'un appareil tenu à la main, ce qui entraîne des désalignements inévitables entre les acquisitions à faible dynamique. Par conséquent, les scènes avec des mouvements introduisent de nouveaux défis tels que des artefacts de type fantôme pour les régions à mouvement important ou une perte de détails dans les régions occultées.

Suivant les récents progrès dans le domaine de l'apprentissage profonds, plusieurs méthodes basées sur des réseaux de neurones convolutifs (CNN) ont d'abord été proposées pour aligner spatialement les images d'entrée sur une image de référence lors de leur fusion en une image HDR finale. L'émergence des architectures de type Transformer dans la vision par ordinateur [1] a mené à la réalisation de nouvelles architectures pour l'imagerie HDR, permettant de corriger au mieux les effets fantômes dus au large mouvement dans la scène lors de l'acquisition. Cependant, celles-ci reposant sur des mécanismes d'attention de complexité quadratique par rapport à la taille d'entrée des images, le coût de calcul et le temps d'exécution de ces solutions logicielles d'imagerie HDR sont conséquents, ce qui empêche leur utilisation sur des systèmes légers et/ou dans des applications en temps réel. Par conséquent, l'objectif principal des solutions logicielles d'imagerie HDR, qui était de rendre l'imagerie HDR plus largement disponible par rapport aux solutions matérielles n'est plus respecté.

Dans ce travail, nous proposons une architecture efficace de Transformer basée sur un mécanisme d'attention additive de complexité linéaire par rapport à la taille des images d'entrée pour la génération d'images HDR, qui peut être exécutée sur des systèmes légers tout en offrant des performances compétitives avec l'état de l'art. En utilisant une architecture efficace et

Modèle disponible : https://steven-tel.github.io/ehdrt

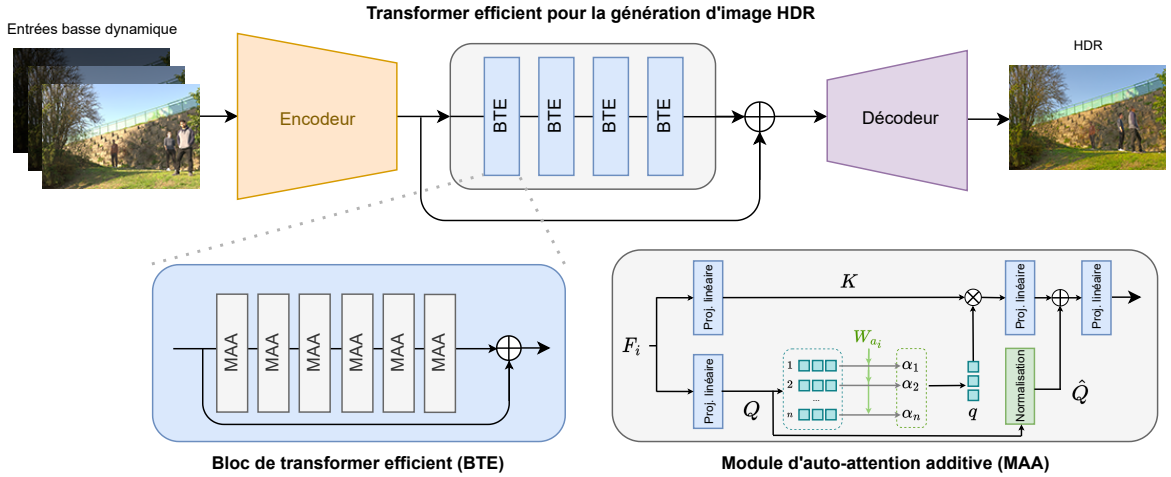

FIGURE 1 : Architecture de la solution proposée.

des techniques d'optimisation avancées, notre méthode permet de capturer des scènes avec une plage dynamique élevée tout en corrigeant les effets fantômes dus aux larges mouvements dans ces scènes.

2 Présentation de la solution proposée

2.1 Pré-traitement

Nous considérons trois images à basse dynamique $I_i \in \mathbb{R}^{3\times H\times W}$ avec leurs temps d'exposition respectifs t_i en tant qu'entrées. L'image HDR générée est alignée spatialement avec l'image d'entrée centrale I_2 sélectionnée comme image de référence. Pour rendre notre solution plus robuste aux différences d'exposition entre les entrées, la projection respective de chaque image d'entrée dans le domaine HDR est calculée à l'aide de la fonction d'encodage gamma décrite dans l'équation 1

$$H_i = \frac{I_i^{\gamma}}{t_i}, \quad \gamma = 2,2, \tag{1}$$

Où $H_i \in \mathbb{R}^{3\times H\times W}$ est l'entrée projetée dans le domaine de l'irradiance. Ensuite, chaque image d'entrée est concaténée avec sa projection respective $L_i \in \mathbb{R}^{6\times H\times W}$:

$$L_i = I_i \oplus H_i, \tag{2}$$

Où $\oplus$ représente l'operation de concatenation.

L_i est ensuite fournie à notre réseau de fusion HDR dont l'architecture globale est représentée dans la Figure 1.

2.2 Fusion

Notre réseau de fusion a pour objectif de produire l'image HDR finale tout en corrigeant les effets fantômes dus aux mouvements entre les images d'entrée. Pour l'encodage des données, 60 cartes de caractéristiques sont produites depuis L_i à l'aide d'une opération de convolution :

$$F_i = conv_E(L_i), \tag{3}$$

Où $F_i \in \mathbb{R}^{60\times H\times W}$ représente les cartes de caractéristiques extraites.

Ensuite, contrairement aux méthodes existantes qui calculent des matrices de similarité sur les caractéristiques peu profondes afin de les aligner spatialement sur celles correspondantes à l'image centrale à l'aide de convolutions, nous introduisons directement les caractéristiques fusionnées dans 4 blocs de Transformer contenant chacun 6 têtes d'auto-attention efficace.

Pour modéliser la corrélation spatiale entre les différentes images, nous adoptons l'attention multi-têtes basée sur une fenêtre. Les caractéristiques fusionnées sont d'abord divisées en patchs. Chaque couche d'attention de notre modèle prend en entrée F_i qui est projetée en une matrice Q (query), une matrice K (key) et une matrice V (value) à l'aide de trois matrices de pondération W_Q, W_K et W_V. Chacune de ces couches d'attention est composée de 6 têtes d'attention, permettant au modèle de considérer différentes représentations d'une même entrée. Le mécanisme d'attention permet alors au modèle de pondérer l'importance de chaque élément de l'entrée en fonction de sa relation avec les autres éléments. Généralement, le mécanisme d'attention utilisé est décrit comme suit :

$$F = softmax\left(\frac{Q\cdot K^T}{\sqrt{d}}\right)\cdot V, \tag{4}$$

Les scores d'attention entre chaque paire de jetons Q et K sont d'habitude calculés à l'aide de l'opération de produit scalaire. Ensuite, ces scores sont normalisés par Softmax pour pondérer les interactions entre les tokens. Enfin, les interactions pondérées sont multipliées par V pour produire la sortie pondérée finale. Dans l'ensemble, la complexité de l'auto-attention est $O(n^2\cdot d)$, où n est le nombre de tokens et d est la dimension des tokens choisis.

Afin de réduire la complexité de notre réseau par rapport à la taille des images d'entrée, nous proposons de réaliser notre fusion HDR à l'aide du module d'attention additive efficace[5]. Cette méthode se distingue par sa capacité à produire des représentations contextuelles fiables tout en étant plus rapide lors de l'inférence.

L'entrée F_i est transformée en tenseur de requête $Q \in \mathbf{R}^{n\times d}$ et une matrice de clé $K \in \mathbf{R}^{n\times d}$ en utilisant deux matrices $W_{\{q,k\}} \in \mathbf{R}^{d\times d}$. Ensuite, la requête Q est multipliée

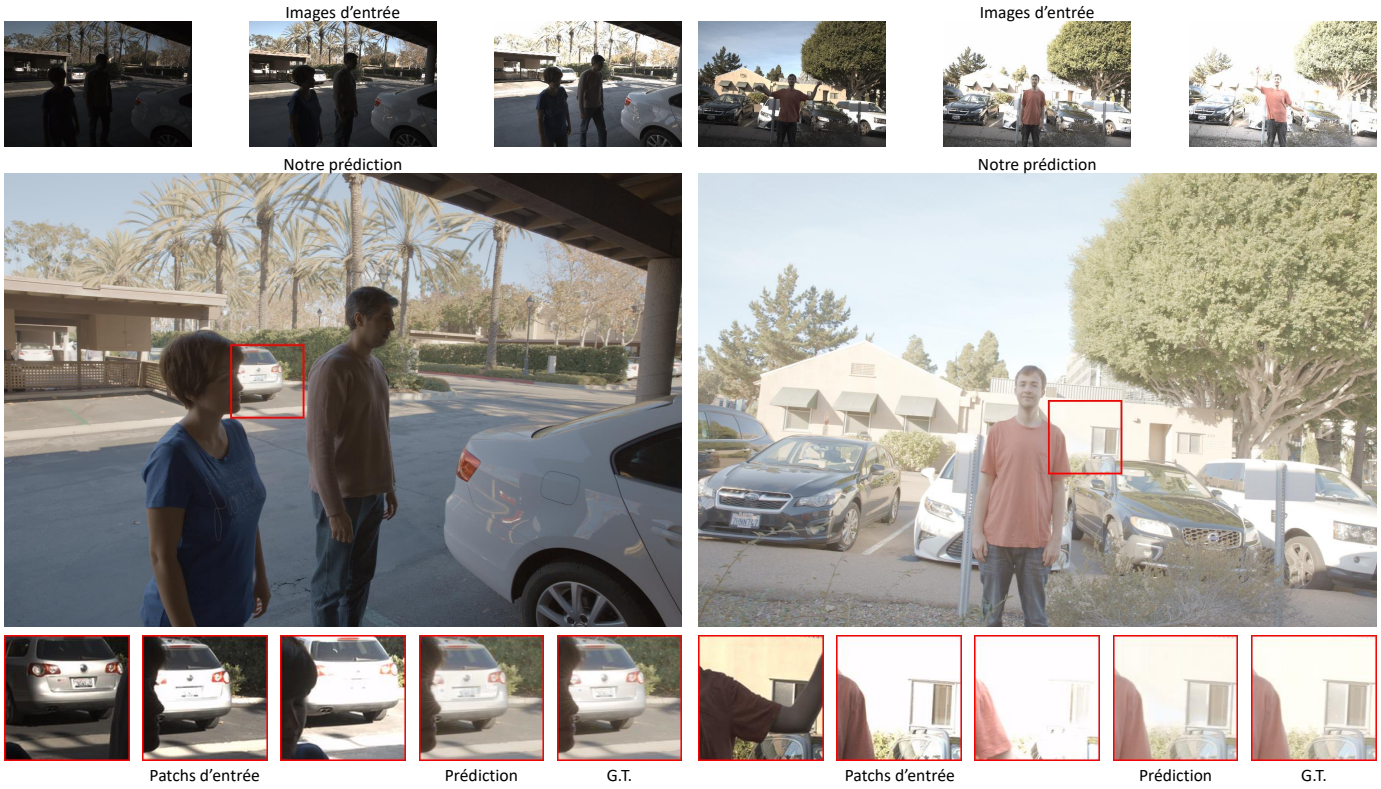

FIGURE 2 : Prédiction de la solution proposée pour des exemples de tests provenant du jeu de données Kalantari2017[2].

par le vecteur de paramètres $W_a \in \mathbf{R}^d$ pour produire les poids d'attention de la requête, suivie d'une opération de Softmax pour produire le vecteur d'attention global $\alpha \in \mathbf{R}^n$:

$$\alpha_i = \frac{exp(Q \cdot W_{a_i}/\sqrt{d})}{\sum_{j=1}^{n} exp(Q \cdot W_{a_j}/\sqrt{d})}, \tag{5}$$

Le tenseur de requête global $q \in \mathbf{R}^d$ est ensuite obtenu grâce à des poids d'attention appris, comme suit :

$$q = \sum_{i=1}^{n} \alpha_i * Q_i, \tag{6}$$

Ensuite, les interactions entre le vecteur de requête global q et la matrice de clé K sont encodées en utilisant un produit élément par élément. Finalement, on peut représenter le mécanisme d'attention additive efficient par :

$$F = \hat{Q} + l(K * q), \tag{7}$$

Où $\hat{Q}$ est la valeur normalisée du tenseur de requête et l l'opération de projection linéaire.

Enfin, les caractéristiques produites sont concaténées avec celles extraites par l'encodeur à l'aide d'un raccourci global afin de stabiliser l'entraînement. L'image HDR finale est produite à l'aide d'un décodeur composé d'une unique couche de convolution.

2.3 Entrainement

Notre réseau est supervisé à l'aide d'une fonction de perte $\mathcal{L}_1$. On calcule cette erreur après mappage des tons en utilisant la fonction loi-μ :

$$T(H) = \frac{log(1+\mu H)}{log(1+\mu)}, \quad \mu = 5000, \tag{8}$$

Où H est l'image dans le domaine linéaire et μ le taux de compression. Notre fonction de perte $\mathcal{L}_1$ devient alors :

$$\mathcal{L}_1 = \|T(H) - T(\hat{H})\|_1, \tag{9}$$

Où T est la fonction de tone-mapping loi-μ, H et $\hat{H}$ sont respectivement l'image prédite par notre réseau et la vérité terrain dans le domaine HDR.

De plus, nous supervisons notre réseau avec une fonction de perte perceptuelle supplémentaire. Celle-ci mesure la différence entre les caractéristiques de notre prédiction et de la vérité terrain prédite à différents étages d'un CNN pré-entrainé. Cette mesure est réalisée en calculant l'erreur quadratique comme suit :

$$\mathcal{L}_p = \sum_j \|\Phi(T(H)) - \Phi(T(\hat{H}))\|_1, \tag{10}$$

Où Φ sont les cartes de caractéristiques prédites par un modèle VGG-16 pré-entrainé et j est l'indice de la couche du réseau. Enfin, la fonction de perte utilisée peut être définie par :

$$\mathcal{L} = \mathcal{L}_1(T, T_{GT}) + \alpha \times \mathcal{L}_p(T, T_{GT}), \tag{11}$$

Où α est un hyper-parametre fixé à 0.01.

Nous utilisons l'optimiseur ADAM avec un taux d'apprentissage initial de $2\mathrm{e}{-4}$ et β_1 et β_2 fixés à $0,9$ et $0,999$ respectivement, ainsi qu'une valeur ϵ de $1\mathrm{e}{-8}$. Nous entraînons notre réseau à partir de zéro avec une taille de batch de 16 pendant 100 epochs en utilisant le jeu de données proposé par Kalantari *et al.* [2].

TABLE 1 : Comparaison quantitative de notre solution avec l'état de l'art. Chaque solutions est évaluée sur les 15 examples de test fournis dans le jeu de données Kalantari*et al.* [2]

Méthode	μ-PSNR	PU-PSNR	PSNR	μ-SSIM	PU-SSIM	SSIM	HDR-VDP2
Sen *et al.*	40.80	40.47	38.11	0.9808	0.9775	0.9721	59.38
Hu *et al.*	35.79	34.96	30.76	0.9717	0.9615	0.9503	57.05
DHDRNet	42.67	41.83	41.23	0.9888	0.9832	0.9846	65.05
DeepHDR	41.65	41.35	40.88	0.9860	0.9815	0.9858	64.90
AHDRNet	43.63	42.93	41.14	0.9900	0.9849	0.9702	64.61
NHDRRNet	42.41	42.97	41.43	0.9877	0.9855	0.9857	61.21
CEN-HDR	43.05	43.24	40.53	0.9908	0.9821	0.9856	64.34
HDRGAN	43.92	44.03	41.57	0.9905	0.9851	0.9865	65.45
HDR-Transformer	**44.32**	**44.23**	**42.18**	0.9916	**0.9924**	**0.9884**	66.03
Notre solution	44.13	44.21	41.86	**0.9921**	0.9923	0.9881	**66.12**

3 Résultats

Comme nous souhaitons proposer une méthode efficace de génération de HDR, nous comparons le coût de calcul et le temps d'inférence de notre réseau avec l'état de l'art dans le tableau 2. Pour évaluer les durées d'exécution, tous les réseaux comparés sont exécutés sur l'unité de traitement neuronal (NPU) d'un Apple MacBook Pro (2021) équipé d'une puce M1. Le temps indiqué est la moyenne de 500 inférences après un échauffement de 50 exécutions. La taille d'entrée est fixée à 1280 × 720 pixels. La projection gamma des entrées LDR et la fonction de tone-mapping HDR sont incluses dans le processus d'inférence. HDR-Transformer [3] étant trop lourd pour être exécuté sur plateforme embarquée, notre réseau est uniquement comparé avec des réseaux convolutifs. On observe que notre solution est une des moins coûteuses en terme d'opérations et la plus rapide après CEN-HDR [6].

TABLE 2 : Comparaison du coût d'inférence de notre solution avec les solutions supervisées de l'état de l'art. Pour mesurer le temps d'inférence, tous les réseaux comparés sont exécutés sur une unité de traitement neuronal M1. La taille d'entrée est définie à 1280×720 pixels.

Méthode	Nb. Param.	GMAccs	IPS
DeepHDR	14618755	843.16	9.30
AHDRNet	1441283	1334.95	2.18
HDR-GAN	2631011	479.78	4.14
CEN-HDR	282883	78.36	36.38
Notre solution	5774355	256.34	11.24

De plus, afin de valider notre méthode, dans le tableau 1, nous réalisons une comparaison quantitative de notre solution avec les réseaux de l'état de l'art. Pour chaque solution, nous calculons le rapport signal bruit (PSNR) et la similarité structurelle (SSIM) dans le domaine linéaire et après mappage des tons en utilisant la fonction loi-μ et pu [4] ainsi que HDR-VDP2. On observe que notre solution produit un résultat supérieur à toutes les solutions basées sur des réseaux de convolutions. Dans la figure 2, nous réalisons une évaluation qualitative de notre réseau pour des exemples du jeu de test proposé par Kalantari *et al.* [2]. Nous remarquons que notre réseau produit des résultats fidèles à la vérité terrain.

4 Conclusion

Dans cet article, nous avons présenté une nouvelle solution de fusion HDR capable de corriger l'effet fantôme causé par les mouvements d'objets importants dans la scène et les mouvements de la caméra. Nous avons démontré que notre nouvelle architecture de type Transformer basée sur un mécanisme d'attention additive de complexité linéaire est capable de produire des résultats fidèles à la réalité tout en pouvant être exécuté sur un système léger.

Références

[1] Alexey DOSOVITSKIY, Lucas BEYER, Alexander KOLESNIKOV, Dirk WEISSENBORN, Xiaohua ZHAI, Thomas UNTERTHINER, Mostafa DEHGHANI, Matthias MINDERER, Georg HEIGOLD, Sylvain GELLY, Jakob USZKOREIT et Neil HOULSBY : An image is worth 16x16 words : Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[2] Nima Khademi KALANTARI, Ravi RAMAMOORTHI *et al.* : Deep high dynamic range imaging of dynamic scenes. *ACM TOG*, 36(4):144–1, 2017.

[3] Zhen LIU, Yinglong WANG, Bing ZENG et Shuaicheng LIU : Ghost-free high dynamic range imaging with context-aware transformer. *In European Conference on Computer Vision*, pages 344–360. Springer, 2022.

[4] Rafal K. MANTIUK et Maryam AZIMI : Pu21 : A novel perceptually uniform encoding for adapting existing quality metrics for hdr. *In 2021 Picture Coding Symposium (PCS)*, pages 1–5, 2021.

[5] Abdelrahman SHAKER, Muhammad MAAZ, Hanoona RASHEED, Salman KHAN, Ming-Hsuan YANG et Fahad Shahbaz KHAN : Swiftformer : Efficient additive attention for transformer-based real-time mobile vision applications, 2023.

[6] Steven TEL, Barthélémy HEYRMAN et Dominique GINHAC : Cen-hdr : Computationally efficient neural network for real-time high dynamic range imaging. *In ECCVW*, 2022.