Modèle de diffusion avec guide linéaire pour la compression d'images

Tom BORDIN Thomas MAUGEY

INRIA Rennes, 263 Av. Général Leclerc, 35042 Rennes, France

Résumé – Nous présentons un schéma de compression d'images à très bas débit reposant sur l'utilisation de modèles de diffusion au décodeur, guidés par un filtre linéaire de l'image, sans entraînement. Nous montrons que cette formulation de guidage est versatile, pouvant préserver différentes informations sur l'aspect global de l'image.

Abstract – We present a framework for image compression at extremely low bitrates using diffusion models guided with a linear filter of the source without additional training. We show that this formulation of guide is versatile as it can represent contain several types of information on the general aspect of the image.

1 Introduction

À extrêmement bas débit, et notamment en compression d'images, le compromis entre débit et distorsion (erreur pixel-à-pixel) ne reflète pas de manière idéale la qualité de l'image. En effet, comme mis en évidence par [2], un troisième critère de perception (qualité visuelle de l'image) intervient dans ce compromis. L'introduction des modèles génératifs au décodage permet, à très bas débit, de conserver la perception au détriment de la distorsion, mais souvent mise au profit de la sémantique ou d'informations sur l'aspect général de l'image. Ces modèles génératifs sont alors souvent entraînés conjointement avec l'information, on parle alors de conditionnement de la génération. Ajouter une nouvelle condition au modèle peut alors être coûteux en temps et en calcul.

Contrairement aux GANs [4], les modèles de diffusion ne requièrent pas nécessairement d'être entraînés pour conditionner la génération. Il est en effet possible de guider la génération de l'image. Ce principe a tout d'abord été introduit pour conditionner la génération sur le label de l'image avec la *classifier guidance* [10], réutilisant les gradients d'un classificateur. Mais plus récemment, l'*universal guidance* [1] permet de modifier la génération afin de générer des images qui s'approchent de n'importe quelle condition donnée. Néanmoins, dans le cas où les images générées doivent suivre la condition de manière très précise, cette méthode n'est pas idéale. Les auteurs de [3] ont alors montré qu'il était possible de guider avec peu d'erreur la génération dans le cas où la condition est une réduction d'échelle de l'image d'origine. Plus généralement, il est aussi possible d'estimer précisément la correction à appliquer à un modèle pour le guider dans le cas où la condition s'exprime comme une fonction linéaire du signal à générer. Nous détaillons les calculs du terme de correction pour guider dans ce cas. Nous présentons ensuite un schéma de compression sémantique à très bas débit et explorons plusieurs utilisations de ce guide linéaire sur un modèle de diffusion déjà entraîné à la génération d'images.

2 Guides linéaires pour les modèles de diffusion

Nous rappelons rapidement le fonctionnement des modèles de diffusions et les équations nécessaires pour montrer comment les guider sur une condition linéaire.

2.1 Modèles de diffusion

La génération d'image par un modèle de diffusion passe par une succession de débruitages. Un bruit aléatoire gaussien est progressivement transformé jusqu'à convergence vers une image naturelle. Le modèle est ainsi entraîné à inverser la corruption d'une image à l'ajout de bruit.

Pour une image $\boldsymbol{x}_0$, l'image bruitée suit la loi :

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon \tag{1}$$

où les α_t sont des paramètres du modèle faisant varier la quantité de bruit avec le pas de temps t. Le modèle de diffusion, en lui-même, est un réseau de neurones noté ϵ_θ qui estime le bruit ϵ présent à partir de l'image bruitée $\boldsymbol{x}_t$ et du niveau de bruit évoluant avec t. Le processus de génération est initialisé avec $\boldsymbol{x}_T$, un bruit aléatoire, et la génération se fait en calculant successivement les $\boldsymbol{x}_{t-1}$ jusqu'à atteindre $t = 0$. L'estimation de $\boldsymbol{x}_{t-1}$ peut être fait comme proposé dans les DDIM[9] avec :

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\boldsymbol{x}_t, t)}{\sqrt{\alpha_t}} \right)}_{\text{"prédiction de } \boldsymbol{x}_0 : \hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t)\text{"}} + \underbrace{\sqrt{1-\alpha_{t-1}}\epsilon_\theta(\boldsymbol{x}_t, t)}_{\text{"direction vers } \boldsymbol{x}_t\text{"}} \tag{2}$$

Notons qu'avec cette formulation, il est possible de séparer une estimation de l'image *prédite* : $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t)$ du reste du bruit. Cette prédiction s'améliore progressivement jusqu'à la fin de la génération.

2.2 Guider les modèles

Guider un modèle vers une condition donnée se fait en ajoutant un terme de correction à l'estimation du bruit. Ainsi, on peut formuler un modèle de diffusion conditionnel comme la somme d'un modèle non conditionnel et de sa correction :

$$\tilde{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}) = \epsilon_\theta(\boldsymbol{x}_t, t) - \sqrt{1-\alpha_t}\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{c}|\boldsymbol{x}_t) \tag{3}$$

Cette formulation a été proposée pour la *classifier guidance* [10], où le terme d'erreur $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{c}|\boldsymbol{x}_t)$, le gradient de la densité de probabilité d'appartenir à la classe $\boldsymbol{c}$ sachant l'image bruité $\boldsymbol{x}_t$, est estimé par les gradients d'un modèle de classification. La condition $\boldsymbol{c}$ est alors un label d'image. Ceci nécessite néanmoins d'entraîner un classificateur d'image, et même préférablement, un classificateur d'images bruitées aux différents niveaux de bruits correspondants à t.

De manière plus générale, il est possible de guider un modèle de diffusion vers une nouvelle condition en estimant la valeur du terme de correction. Nous noterons par la suite $G(\cdot,\cdot,\cdot)$, l'opérateur de guidage :

$$G(\boldsymbol{x}_t, t, \boldsymbol{c}) = -\sqrt{1-\alpha_t}\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{c}|\boldsymbol{x}_t) \tag{4}$$

2.3 Guide linéaire

Inspirés du guide de carte de couleurs chez [3], nous prolongeons le cas où la condition que l'on souhaite ajouter à la génération de l'image peut s'écrire comme une fonction linéaire de l'image :

$$\boldsymbol{c} = \boldsymbol{A}\boldsymbol{x}. \tag{5}$$

Il est alors possible d'estimer précisément la valeur du guide G sans entraîner ou affiner le modèle.

L'image prédite au temps t se réécrit comme :

$$\begin{aligned}\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t) &= \frac{\sqrt{\alpha_t}\boldsymbol{x}_0 - \sqrt{1-\alpha_t}(\epsilon_\theta(\boldsymbol{x}_t, t) - \epsilon)}{\sqrt{\alpha_t}} \\ &= \boldsymbol{x}_0 - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}}\Delta\epsilon_t,\end{aligned} \tag{6}$$

où $\Delta\epsilon_t$ est l'erreur commise sur l'estimation du bruit par le modèle de diffusion au temps t. Comme montré par [3] on peut mesurer, en échantillonnant sur différentes images et différents bruits, que cette erreur s'approxime par une loi Gaussienne centrée en 0 avec une variance mesurée $\overline{\lambda}_t$:

$$\Delta\epsilon_t \sim \mathcal{N}(0, \overline{\lambda}_t^2 \boldsymbol{I}). \tag{7}$$

On en conclut ainsi que l'image prédite pour chaque niveau de bruit suit aussi une loi Gaussienne centrée sur l'image d'origine :

$$\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t) \sim \mathcal{N}\left(\boldsymbol{x}_0, \overline{\lambda}_t^2 \frac{1-\alpha_t}{\alpha_t}\mathbf{I}\right). \tag{8}$$

On peut alors calculer une condition estimée à partir de l'image prédite, où l'on sait qu'il existe ϵ tel que :

$$\begin{aligned}\hat{\boldsymbol{c}}_t(\boldsymbol{x}_t, t) &= \boldsymbol{A}\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t) \\ &= \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{A}\overline{\lambda}_t\frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}}\epsilon \\ &= \boldsymbol{c} + \overline{\lambda}_t\frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}}\boldsymbol{A}\epsilon.\end{aligned} \tag{9}$$

Le résultat ci-dessus peut alors se reformuler en : la condition $\boldsymbol{c}$ sachant l'image bruitée $\boldsymbol{x}_t$ suit une loi normale centrée sur la condition prédite $\hat{\boldsymbol{c}}_t(\boldsymbol{x}_t, t)$:

$$\boldsymbol{c}|\boldsymbol{x}_t \sim \mathcal{N}\left(\hat{\boldsymbol{c}}_t, \overline{\lambda}_t^2 \frac{1-\alpha_t}{\alpha_t}\boldsymbol{A}\boldsymbol{A}^T\right). \tag{10}$$

Nous avons ainsi la valeur de $p_t(\boldsymbol{c}|\boldsymbol{x}_t)$, la probabilité de densité d'une loi normale multivariée :

$$p_t(\boldsymbol{c}|\boldsymbol{x}_t) = \frac{1}{\sqrt{(2\pi)^m}} \exp\left(-\frac{\alpha_t}{2\overline{\lambda}_t^2(1-\alpha_t)}(\boldsymbol{c}-\hat{\boldsymbol{c}}_t)^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}(\boldsymbol{c}-\hat{\boldsymbol{c}}_t)\right). \tag{11}$$

D'où la formulation de l'opérateur de guidage explicite :

$$\boxed{G(\boldsymbol{x}_t, t, \boldsymbol{c}) = \frac{\sqrt{\alpha_t}}{2\overline{\lambda}_t}\nabla_{\boldsymbol{x}_t}(\boldsymbol{c}-\hat{\boldsymbol{c}}_t)^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}(\boldsymbol{c}-\hat{\boldsymbol{c}}_t)} \tag{12}$$

Dans le cas où $\boldsymbol{A}\boldsymbol{A}^T$ ne serait pas inversible, nous pouvons prendre la pseudo-inverse pour le calcul du guide.

La formulation de l'opérateur de guidage que l'on obtient est similaire à celle proposée par *universal guidance* [1] où il est proposé d'estimer G par :

$$G(x_t, t, c) = s\sqrt{1-\alpha_t}\nabla_{\boldsymbol{x}_t}\|\boldsymbol{c} - \hat{\boldsymbol{c}}_t\|_2^2 \tag{13}$$

avec s un coefficient d'ajustement de guidage. Dans le cas où $\boldsymbol{A}\boldsymbol{A}^T$ est l'identité, la valeur du gradient diffère seulement de par un coefficient fonction du temps.

3 Exemples de guides

En plus de permettre de contrôler la couleur de en utilisant une réduction de résolution de l'image d'origine en guidant avec une vignette comme chez [3], il est aussi possible de conditionner sur des objets plus complexes. Dans le cas où l'on a accès à une segmentation de l'image, on peut alors calculer des filtres linéaires qui prennent en compte la structure de l'image d'origine.

3.1 Retouche d'image

La formulation de l'opérateur de guidage permet de ne choisir de générer qu'une partie de l'image. Des versions des modèles de diffusions ont été *fine-tune* pour la retouche d'image, mais un résultat équivalent peut être obtenu en les guidant, avec notre approche décrite précédemment. En effet, en définissant $\boldsymbol{A}$ comme un masque sur les pixels de l'image, on peut guider la génération sur une zone restreinte. On ne communique ainsi qu'une partie des pixels de l'image au décodeur. Ne guidant alors pas sur le reste du contenu, il est généré uniquement à partir de l'information de la sémantique. Nous illustrons en Fig. 1 la génération guidée par un contenu masqué, en pratique, la condition est en fait la vectorisation des pixels du masque. La génération du reste de l'image est ainsi faite simultanément à l'intégration du contenu du masque permettant de conserver la cohérence de l'image. Nous comparons notre formulation du guide avec celle de *universal guidance* [1].

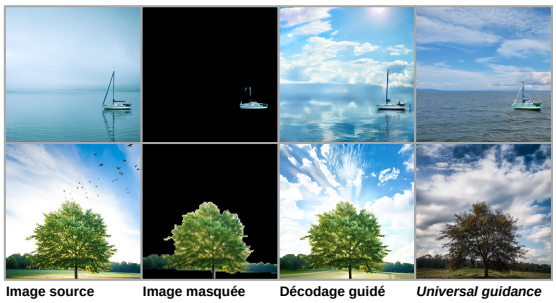

FIGURE 1 : Comparaison entre la formulation du guide linéaire et *universal guidance* sur une condition masquée.

3.2 Carte de couleurs

En utilisant les contours des objets de la carte de segmentation de l'image, on peut définir une carte de couleurs, donnant une couleur moyenne par objet. Pour n objets dans l'image, contenant respectivement $(m_i)_{i\in[1,n]}$ pixels, le filtre $\boldsymbol{A}$ se définit alors par :

$$\boldsymbol{A}_{i,j} = \begin{cases} 1/m_i & \text{si le pixel } j \text{ est dans l'objet } i \\ 0 & \text{sinon.} \end{cases} \tag{14}$$

En définissant ainsi la matrice $\boldsymbol{A}$, la condition est alors un vecteur contenant pour chaque objet la couleur moyenne de la zone. En guidant, la carte de couleurs permet d'avoir la même couleur moyenne pour chaque objet sur l'image décodée que sur celle d'origine. Le débit varie alors avec le nombre d'objets et la complexité de la forme des objets pour l'encodage des contours. Nous illustrons en Fig. 2 le guidage sur cette forme de condition.

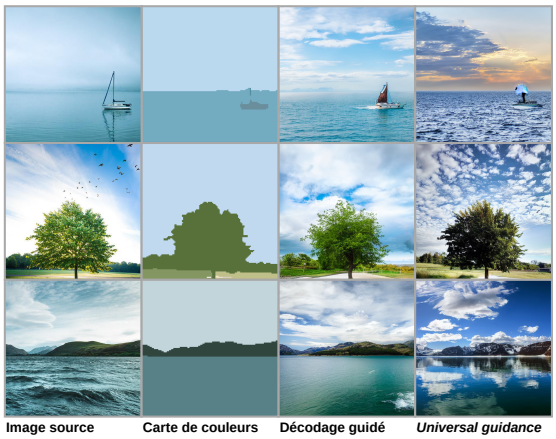

FIGURE 2 : Comparaison avec *universal guidance* en guidant sur une couleur moyenne par objet. La carte de couleur est ici uniquement une visualisation de la condition $\boldsymbol{c}$.

4 Application à la compression

Nous appliquons la formulation du guide de couleur par objet présenté ci-dessus dans un contexte de compression sémantique très bas débit. Nous nous basons pour cela sur un schéma de compression simple utilisant uniquement l'information d'une carte de segmentation de l'image au décodeur. Le guide, ajouté au modèle de diffusion entraîné sur des cartes de segmentation uniquement, permet d'ajouter une information de couleur à faible coût. Nous illustrons en Fig. 3 le schéma de compression utilisé. La carte de segmentation et le guide sont encodés séparément et transmis au modèle de diffusion, d'un côté en conditionnant le modèle et de l'autre en le guidant.

Notons qu'ici, la carte de segmentation qui est nécessaire pour le calcul du filtre $\boldsymbol{A}$ dans la formulation de G est déjà disponible au décodeur. La carte de segmentation peut aussi être remplacée par une autre représentation sémantique de l'image comme en utilisant le modèle de fondation CLIP [8].

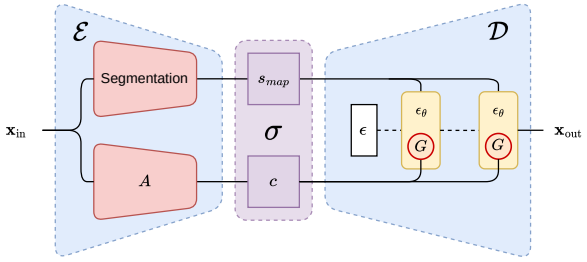

FIGURE 3 : Schéma de compression : la sémantique de l'image, représentée par la carte de segmentation et la moyenne de couleur par objet, est transmise au décodeur.

4.1 Codage de l'information sémantique

La carte de segmentation est codée en faible résolution 128×128, sans perte, en utilisant un codage en chaîne différentiel. Les contours sont codés indépendamment des labels en répétant la méthode suivante : on fixe les coordonnées d'un point de départ, puis, tant que c'est possible, on code la direction vers le point suivant (nord, est, sud, ouest). En pratique, nous codons la différence avec la direction précédente avant d'effectuer le codage entropique. Le débit des cartes de segmentation varie ainsi avec la complexité des formes et la quantité d'objets détectés dans l'image.

Le guide de couleur par objet est le plus souvent de très faible dimension. Nous l'encodons simplement avec une quantification uniforme avant le codage entropique. Le débit total de l'image est alors la somme des débits de la carte de segmentation et du guide de couleur.

4.2 Résultats

Nous montrons le schéma de compression sur deux approches, une guidée avec une carte de couleur prenant en compte la segmentation, et une autre basée sur CLIP où la carte de couleur est une version très faible résolution (16×16) de l'image d'origine. Nous nous comparons aussi d'une part avec Text+Sketch [6] proposant une sémantique préservant les contours de l'image et une description textuelle, et d'autre part, avec une méthode de compression classique optimisant la distorsion : VVC [11] avec le plus bas paramètre de qualité. Nous illustrons en Fig. 4 la comparaison entre les différentes méthodes. Notre méthode permet d'atteindre des débits très bas, en conservant la position des objets dans l'image décodée.

Nous comparons les différentes méthodes sur différents critères. Sur la conservation de la sémantique, nous utilisons pour cela deux métriques différentes : d'une part la BCE (*binary cross entropy*) entre les cartes de segmentation d'entrée et de sortie, et d'autre part la métrique déduite du modèle de fondation CLIP mesurant la similarité sinus entre les projections des

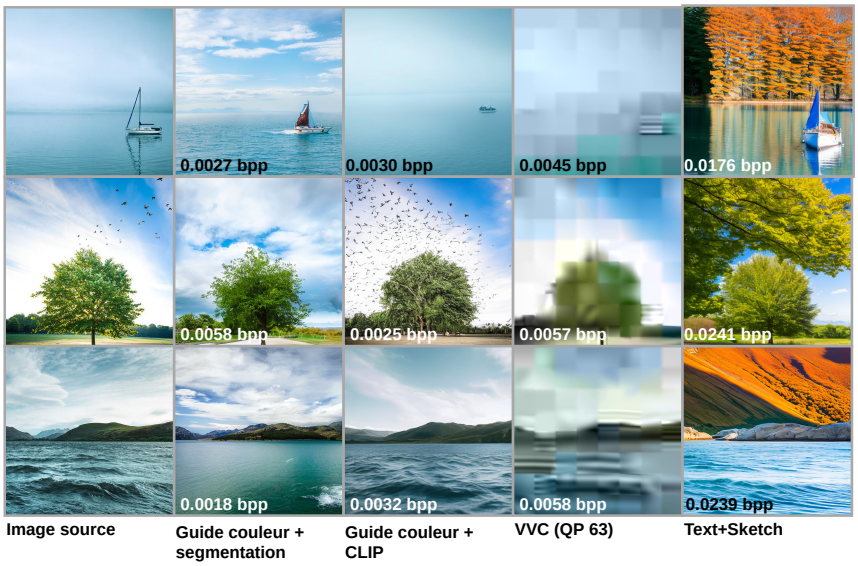

FIGURE 4 : Comparaison du guidage couleur par rapport à d'autres méthodes de compression sémantique très bas débit et à un codeur classique : VVC.

	MSE↓	FID↓	DBCNN↑	BCE↓	CLIP↓	bpp(10^{-3})
VVC(QP 63)	**0.011**	21.8	0.20	1.32	0.69	5.3
Text+Sketch	0.153	12.2	<u>0.61</u>	1.2	0.19	11.2
Guide couleur + CLIP	<u>0.021</u>	<u>4.1</u>	0.62	<u>0.37</u>	**0.12**	**2.2**
Guide couleur + segmentation	0.052	**3.5**	**0.71**	**0.19**	**0.12**	<u>3.8</u>

TABLE 1 : Comparaison des différentes méthodes à très bas débits. Premiers et seconds scores sont respectivement mis en gras et soulignés.

images par le modèle. Un autre critère qui intervient est la qualité visuelle des images décodées, plusieurs métriques existent ainsi comme DBCNN [12] (métrique sans référence) et FID [5]. Et dernièrement, nous évaluons aussi sur la distorsion en mesurant l'erreur quadratique moyenne, métrique classique de la compression, même si cette métrique ne reflète pas toujours la sémantique encodée.

La comparaison des performances des différentes méthodes est rapportée en Tab. 4.2. Nous pouvons voir que notre méthode permet d'atteindre un très bas débit, ceci est notamment le cas sur les images simples contenant peu d'objets. La qualité visuelle des images décodées ainsi que la sémantique de l'image est mieux préservée par notre méthode. De la même manière que Text+Sketch, nous pouvons transmettre une information sur les contours de l'image au travers de la carte de segmentation, néanmoins l'encodage sous la forme de codage en chaîne différentielle nous permet de grandement diminuer son coût.

5 Conclusion

À travers le guidage, nous montrons qu'il est possible de conditionner la génération d'images sur une information obtenue par filtrage linéaire, sans ré-entraînement et avec peu d'erreur. De plus, par linéarité, les conditions peuvent s'ajouter facilement sans changer la formulation. Cette méthode peut être particulièrement intéressante dans un contexte de codage pour la machine[7], permettant d'adapter l'information transmise à la tâche. Le codage de la condition peut être amélioré de façon à réduire le débit dans le schéma de compression en fonction de la forme du filtre.

6 Remerciements

Ce travail a été financé par l'Agence National de la Recherche. (MADARE, Project-ANR-21-CE48-0002)

Références

[1] A. BANSAL, H. CHU, A. SCHWARZSCHILD, S. SENGUPTA, M. GOLDBLUM, J. GEIPING et T. GOLDSTEIN : Universal guidance for diffusion models. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[2] Yochai BLAU et Tomer MICHAELI : The perception-distortion tradeoff. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[3] Tom BORDIN et Thomas MAUGEY : Linearly transformed color guide for low-bitrate diffusion based image compression. *IEEE Transactions on Image Processing*, 2024.

[4] I. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE et Y. BENGIO : Generative adversarial networks. *Communications of the ACM*, 2020.

[5] M. HEUSEL, H. RAMSAUER, T. UNTERTHINER, B. NESSLER et S. HOCHREITER : Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 2017.

[6] Eric LEI, Yiğit Berkay USLU, Hamed HASSANI et Shirin Saeedi BIDOKHTI : Text+ sketch : Image compression at ultra low rates. *arXiv preprint arXiv :2307.01944*, 2023.

[7] Rémi PIAU, Thomas MAUGEY et Aline ROUMY : Predicting cnn learning accuracy using chaos measurement. *In 2023 IEEE International*

Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), pages 1–4. IEEE, 2023.

[8] A. RADFORD, J. KIM, C. HALLACY, A. RAMESH, G. GOH, S. AGARWAL, G. SASTRY, A. ASKELL, P. MISHKIN, J. CLARK *et al.* : Learning transferable visual models from natural language supervision. *In International conference on machine learning*. PMLR, 2021.

[9] Jiaming SONG, Chenlin MENG et Stefano ERMON : Denoising diffusion implicit models. *arXiv preprint arXiv :2010.02502*, 2020.

[10] Y. SONG, J. SOHL-DICKSTEIN, D. P KINGMA, A. KUMAR, S. ERMON et B. POOLE : Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv :2011.13456*, 2020.

[11] A. WIECKOWSKI, J. BRANDENBURG, T. HINZ, C. BARTNIK, V. GEORGE, G. HEGE, C. HELMRICH, A. HENKEL, C. LEHMANN, C. STOFFERS, I. ZUPANCIC, B. BROSS et D. MARPE : Vvenc : An open and optimized vvc encoder implementation. *In Proc. IEEE International Conference on Multimedia Expo Workshops (ICMEW)*.

[12] Weixia ZHANG, Kede MA, Jia YAN, Dexiang DENG et Zhou WANG : Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.