Compression sélective d'image basé sur la sémantique

Tom BORDIN Thomas MAUGEY

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

Résumé – Cet article propose un modèle pour une compression d'image sélective. L'image est séparée en deux régions à l'aide d'un masque, traitant l'une sur un critère sémantique et l'autre sur un critère distorsion. L'encodage de la majorité de l'image devient ainsi une simple extraction du contenu sémantique ce qui permet d'atteindre des taux de compression extrêmement bas. Le décodage repose sur un modèle génératif s'appuyant sur le processus de diffusion. La description sémantique est alors faite à partir d'une carte de segmentation et de couleur. Cette méthode permet ainsi d'obtenir des images de haute qualité à bas débit en régénérant seulement une partie de faible intérêt.

Abstract – We propose a framework for selective image compression. The model uses a semantic description rather than distortion metrics on a chosen portion of the image. Encoding the image thus becomes a simple extraction of semantic enabling to reach drastic compression ratio. The decoding side is handled by a generative model relying on the diffusion process for the reconstruction of images. We propose to describe the semantic using low resolution segmentation and color maps as guide. We show that this method is able to produce high quality images while regenerating the negligible content.

1 Introduction

Une image est classiquement compressée en cherchant à minimiser l'erreur commise lors de la reconstruction. La *Mean Square Error* (MSE) s'impose alors naturellement comme un critère simple et efficace d'évaluation [10]. Cependant, la MSE ne permet pas toujours d'évaluer la qualité visuelle et cela s'observe en particulier à très bas débit avec l'apparition de nombreux artefacts de compression. Blau *et al* [2] mettent en évidence ce phénomène au travers l'existence d'un compromis entre la fidélité définie par la MSE et la qualité visuelle de l'image. Ils l'appellent le compromis *perception-distorsion*. Ils ajoutent dans [3] que ce compromis dépend aussi du débit ciblé. En effet, à haut débit, l'optimisation de la MSE suffit à obtenir une bonne qualité d'image. Les méthodes neuronales récentes [1] et [8] proposent d'atténuer ce problème en intégrant des métriques perceptuelles dans leur apprentissage. Leur intérêt reste néanmoins porté sur les hauts débits où l'impact de la MSE est encore importante.

Dans ce travail, nous nous concentrons sur les débits extrêmement faibles, où le critère MSE n'est plus pertinent. La représentation compressée de l'image prend alors la forme d'une description sémantique et le décodage devient une tâche générative conditionnelle, *i.e* générer une image à partir des informations sémantiques [1]. Naturellement, il n'est pas toujours pertinent de générer l'image dans son intégralité, une partie du contenu doit alors rester fidèle à l'entrée. La compression sélective permet alors d'utiliser différentes méthodes et débits sur les zones de l'image. On nomme notre paradigme *compression générative sélective basée sur la sémantique (SSGC)*.

Le schéma de compression SSGC utilise une description sémantique contenant l'information sur la position des couleurs ainsi que des labels. D'un côté, un label donne de l'information sur le contenu, et de l'autre la couleur assure une ressemblance à l'image d'origine. Le modèle génératif utilisé repose sur les travaux d'une publication soumise à EUSIPCO 2023 en cours d'évaluation. La génération est faite à l'aide du processus de diffusion sur une architecture de Latent Diffusion Model (LDM) [9]. Nous présentons en détail le schéma génératif dans un contexte de compression sélective.

Nous présentons la formulation de la compression générative sélective utilisant une représentation sémantique ainsi que l'architecture du modèle sur lequel nous nous appuyons dans la section 2. Dans la section 3, nous présentons les résultats de notre méthode ainsi qu'une comparaison par rapport aux codecs récents.

2 Schéma de compression

2.1 Formulation du problème

La compression sélective permet de distinguer différentes parties de l'image à coder avec différents débits. Dans notre schéma, nous distinguons deux parties dans l'image, une partie sélectionnée S comme importante pour l'observateur, et l'autre qui constitue le reste de l'image. Cette séparation est représentée à l'aide d'un masque m tel que $m_i = \delta_{i,S}$. Une image x est ainsi codée d'une part avec un codec classique pour les pixels de S et d'autre part avec une méthode générative pour les pixels restants.

Le codage génératif $(\mathcal{E}, \mathcal{D})$ comme décrit en figure 1, repose sur une description sémantique de l'image σ. Ainsi en notant les reconstructions respectivement $\tilde{x}_C$ pour le codage classique et $\tilde{x}_G = \mathcal{D}(\sigma)$ pour le codage génératif, l'image reconstruite $\tilde{x}$ se formule :

$$\tilde{x} = (1 - m).\tilde{x}_G + m.\tilde{x}_C \tag{1}$$

On cherche alors à optimiser les compromis suivant :

$$\underbrace{R_C + \gamma||m.(x - \tilde{x}_C)||_2^2}_{\text{débit-distorsion}} + \underbrace{R_G + \lambda\Psi_m(x, \tilde{x}_G)}_{\text{débit-perception}} \tag{2}$$

R_G et R_C sont respectivement les débits des deux méthodes de codage et λ et γ sont des coefficients régulateurs. Ψ_m est une mesure de similarité de la sémantique et de la qualité

visuelle en dehors du masque. L'objectif hors du masque ne cherche pas à reconstruire x en tenant compte de la norme. Deux termes se distinguent alors, une partie débit-distorsion correspondant à S et une optimisation d'un débit-perception en dehors de S.

En utilisant la définition de la perception introduite dans [2], nous utilisons de manière similaire une distance entre des distributions statistiques pour évaluer la sémantique entre x et $\tilde{x}$. Ψ est donc estimé en utilisant la distance entre $p(\tilde{x}|\sigma)$, $q(x|\sigma)$, respectivement les distributions des images générées et des images du jeu de données. L'image est générée de manière à être réaliste compte tenu de la contrainte sémantique, conformément à la définition de la perception.

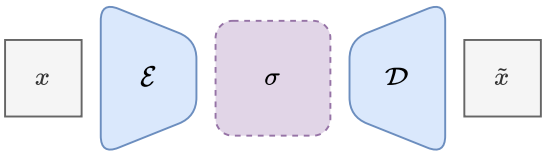

FIGURE 1 : Schéma de compression sémantique sur une structure d'encodeur-décodeur$(\mathcal{E}, \mathcal{D})$ où σ est la représentation sémantique. L'image reconstruite $\tilde{x}$ doit contenir la même sémantique que x.

L'encodage de S requiert l'optimisation d'un compromis classique de débit-distorsion. Pour cela, nous utilisons un codec classique tel que VVC. Nous nous intéressons aux cas où S représente une petite partie connexe de l'image. L'attention est principalement portée sur la partie générative décrite dans ce qui suit. Nous présentons tout d'abord l'encodeur de la sémantique $\mathcal{E}$, puis le décodeur sur un modèle génératif $\mathcal{D}$.

2.2 Encoder la sémantique : $\mathcal{E}$

Nous proposons de décrire σ comme une combinaison d'une carte de segmentation σ_s et d'une carte de couleur σ_c.

2.2.1 La carte de segmentation σ_s

Elle donne des informations sur les objets et leur position dans l'image. Les cartes de segmentation de nos image sont estimées à l'aide du modèle entraîné DeepLabV3 [4]. Elles sont ensuite compressées par sous-échantillonnage. Au niveau du décodage, la carte de segmentation est sur-échantillonnée pour s'adapter à l'entrée du modèle.

Les images compressées en utilisant des cartes de segmentation peuvent présenter d'importantes différences. En effet, une classe ne capture qu'un type de sémantique et à moins que le nombre de classes n'explose, il est difficile de décrire précisément l'image. Ainsi, une photo prise de jour aura les mêmes étiquettes si elle est prise de nuit, cette variance est illustrée en Figure 2 a). Pour résoudre ce problème nous complétons σ par une carte de couleur.

2.2.2 La carte de couleur σ_c

Elle décrit de manière succincte l'information sur les couleurs de l'image. La carte de couleur est une version très basse résolution de l'image σ_c. Mais à elle seule, ce n'est pas suffisant pour guider la génération, du jaune pouvant exemple être interprété comme un taxi aussi bien qu'un poussin. Les deux sémantiques sont donc complémentaires et ne se suffisent pas à elles seules.

Pour décoder une image à partir de la représentation sémantique, nous proposons l'utilisation d'un décodeur génératif reposant sur le processus de diffusion.

2.3 Décodage génératif avec les LDM : $\mathcal{D}$

Les modèles de diffusion [6] sont entraînés à maximiser la fonction de vraisemblance, i.e la probabilité que l'image générée appartiennent à la distribution du jeu de données. Ils débruitent de manière itérative un vecteur initialisé aléatoirement jusqu'à la convergence vers une image. Au lieu de faire la diffusion dans l'espace des pixels, le LDM fait la diffusion dans l'espace latent d'un *variational auto-encoder* (VAE) entraîné séparément $(\mathcal{E}', \mathcal{D}')$ [5]. Le rôle du modèle de diffusion ϵ_θ est alors de générer des projections d'images dans cet espace latent. Ils présentent et fournissent plusieurs modèles entraînés conditionnellement sur diverses entrées telles que du texte ou des cartes de segmentation.

Nous illustrons le processus de diffusion conditionnel en Figure 2 a). Un bruit z_T est d'abord tiré aléatoirement, puis débruité de manière itérative à l'aide d'un modèle de diffusion : $z_{t-1} = \frac{1}{\sqrt{\alpha_t}}(z_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(z_t, t, \sigma_s))$, où α_t et β_t sont des paramètres du modèle de diffusion. $\mathcal{D}'$ le décodeur du VAE est ensuite utilisé pour générer $\tilde{x}$.

Les LDMs n'ont pas été entraînés sur des cartes de couleur. Nous proposons une alternative pour imposer une contraintes sur les couleurs sans ré-entraînement. Cette méthode est illustrée dans la figure 2 b). Inspirés par [7], à la place d'initialiser la diffusion avec un bruit gaussien aléatoire z_T, nous utilisons la carte des couleurs encodée dans l'espace latent à l'aide de l'encodeur VAE $\mathcal{E}'(\sigma_c)$, ce qui correspond à z_0 dans la figure 2 b). En ajoutant ensuite du bruit, le processus de diffusion est intercepté à un pas de temps $t < T$ correspondant à la quantité de bruit ajoutée. Le reste du processus de diffusion se déroule normalement pendant t pas de temps.

Notons que le choix du pas de temps t auxquels nous interceptons la diffusion a un impact sur la génération. En effet, plus t est petit, plus z_t est proche de la distribution finale, et si t est proche de T alors z_t est principalement composé de bruit. Nous constatons empiriquement que l'utilisation de 70% de T est idéale pour la génération en ce qui concerne les couleurs.

3 Expériences et résultats

Dans cette section, nous comparons notre méthode aux codecs standard à des débits similaires. Puisque la diffusion n'est pas déterministe dépendant de la graine de l'aléatoire, nous choisissons le premier échantillon généré pour chaque image présentée afin d'éviter le *cherry picking*. Notez que toutes les images présentées dans les figures sont en 512×512 et devraient être vues en zoomant pour une comparaison plus précise.

3.1 Modèle génératif

L'architecture et les poids du modèle LDM sont ceux partagés par [9] qui peuvent être trouvés sur leur dépôt github. Nous utilisons plus particulièrement le modèle de génération

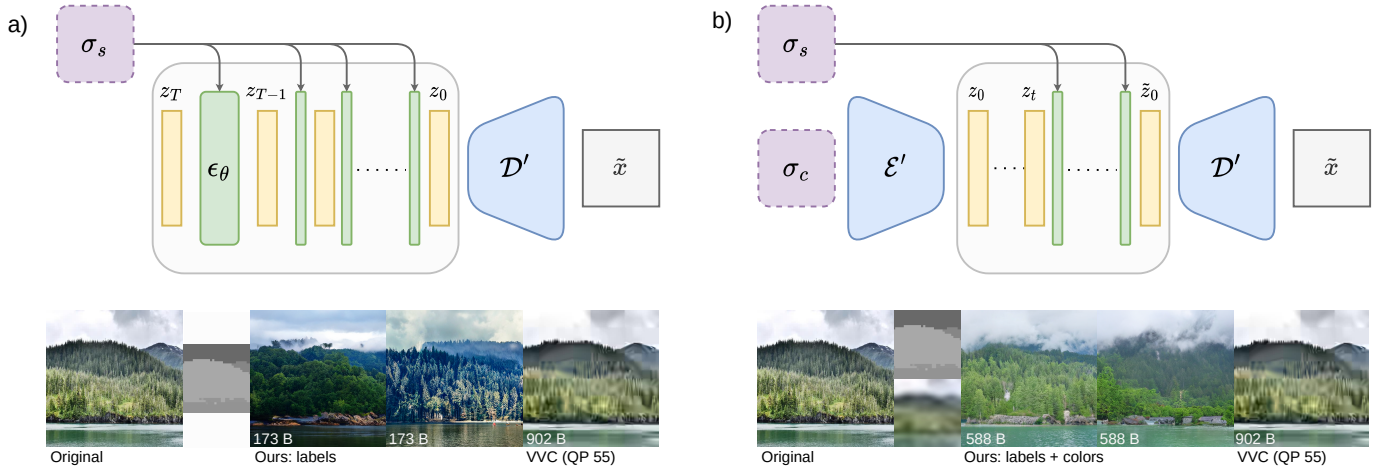

FIGURE 2 : a) Illustration de la génération conditionnelle d'image avec des cartes de segmentation. Entre deux générations la variance entre les images peut être importante. b) Intégration de la carte de couleur dans le conditionnement sans ré-entraînement du modèle. Les images générées sont stabilisées.

sémantique conditionnelle entraîné sur le jeu de données Landscape [5] pour la génération d'images 512×512. Les échantillons sont générés en utilisant 200 pas de temps avec un ordonnanceur DDIM et un paramètre de *free guidance* de 2. Augmenter ce paramètre force l'image a correspondre plus à la carte de segmentation ce qui est ici indésirable vu son aspect pixélisé.

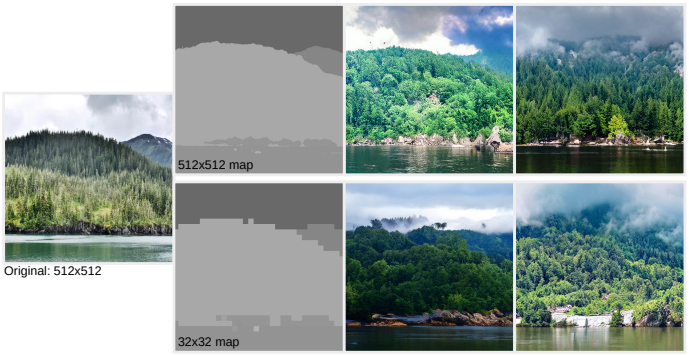

FIGURE 3 : L'utilisation de cartes de segmentation faible résolution n'endommage pas la qualité de la génération d'images et seulement faiblement la fidélité sémantique.

Les cartes de segmentation ont été obtenues à l'aide de DeepLabV3 [4]. Ce modèle de segmentation joue ainsi une partie du rôle de l'encodeur, qui se réfère à $\mathcal{E}$ dans la figure 1. Nous montrons que le sous-échantillonnage de la carte de segmentation par un facteur 16 n'affecte pas la qualité de la génération, illustré en Figure 3. Le réseau est capable d'extrapoler en dehors des limites données afin de générer des échantillons plus réalistes.

Les cartes de couleur sont estimées à l'aide d'une série d'opérations de flou gaussien et de sous-échantillonnage afin d'obtenir une représentation 16×16. Le débit est donc mesuré sur la représentation de ces deux cartes.

Le décodage des images est fait sur la base de leur qualité visuelle directement. Pour le décodeur LDM, des mesures utilisant la *Fréchet Inception Distance* (FID) sont présentes dans [9]. Toutefois, cette mesure ne peut pas être utilisée de manière équitable pour une comparaison avec VVC, l'objectif de VVC étant l'optimisation de la MSE. Nous utilisons la dernière version du codeur intra de VVC (v1.6) avec l'implémentation de VVenc [11]. Lors de l'encodage, nous utilisons un QP de 55 ou bien un débit cible similaire à celui obtenu par notre méthode.

Nous montrons en Figure 4(gauche) que le débit nécessaire pour décrire la représentation sémantique est assez peu coûteux pour la qualité et la fidélité qu'il apporte sur l'image source. De plus, aux mêmes débits l'optimisation de la MSE faite par VVC ou les autres codecs classique avec une abstraction totale de la sémantique, conduit à la création d'artefacts visuels important.

3.2 Compression sélective

La figure 4(droite) illustre des scénarios dans lequel une zone S a été sélectionné comme importante pour l'observateur. En dehors de S l'image peut ainsi être reconstruite afin de consacrer plus de débit à la description du contenu de S. Les masques sont ici construits à la main sur l'image en ajoutant S sur une image existante du jeu de données.

On peut voir sur les exemples fournis que les images encodées avec VVC perdent l'information du contenu défini comme intéressant, ici le visage dans la première image et la rose dans la seconde. Le débit dédié à l'arrière plan ne permet toujours pas de reconstruire une image visiblement acceptable. Notre méthode au contraire propose une perte d'information sur l'arrière plan au profit d'une bonne reconstruction du contenu de S. Évidemment, l'efficacité de notre méthode va varier avec la proportion de S. En effet, le débit R_C croit avec S.

Comme l'illustre la figure, notre méthode permet d'obtenir une qualité visuelle homogène et cohérente sur l'ensemble de l'image. Adopter la même approche en utilisant VVC à différents débits en fonction de la zone en question résulte en une image hétérogène en terme de qualité. De même, un encodage uniforme de l'image avec un codec classique, optimisant la MSE, entraîne un flou et des artefacts sur l'ensemble de l'image. Il y alors une perte de qualité à la fois dans la reconstruction du contenu important mais aussi sur le reste de l'image, ici représenté par l'arrière plan.

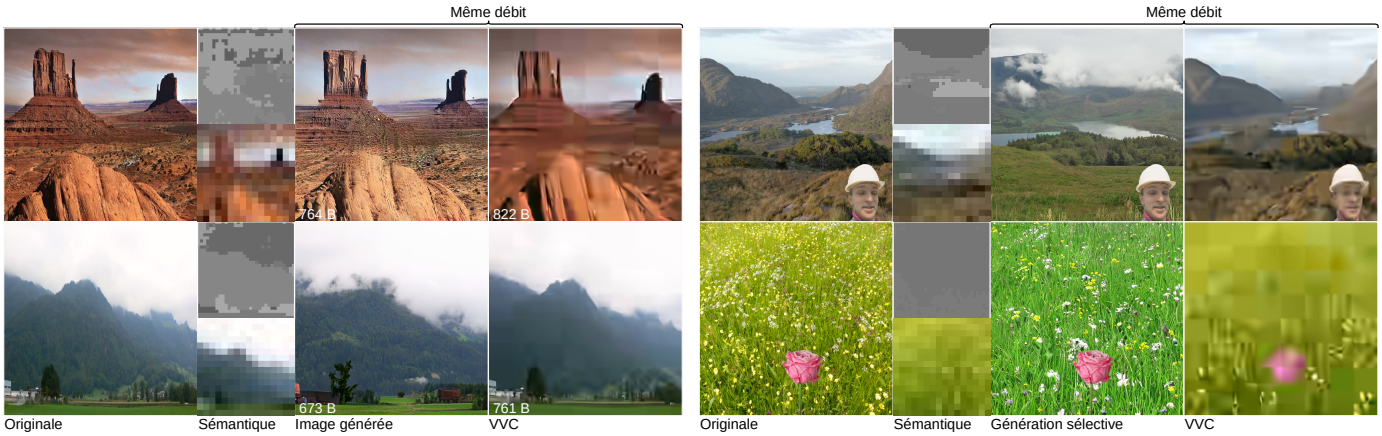

FIGURE 4 : Illustration des résultats de génération. $S = \emptyset$ (gauche) permet de générer une image intégralement. Pour la génération sélective (droite) S contient un visage dans la première image et une rose dans la seconde. La génération permet de conserver la sémantique fournie dans les cartes de sémantique 32×32 et couleur 16×16.

4 Conclusion

Nous avons proposé un nouveau schéma pour la compression d'images basé sur une représentation sémantique. La génération est conditionnée en utilisant des cartes de couleur sur un modèle entraîné sur des cartes de segmentation. Le décodage s'appuie sur un processus de diffusion conditionnelle pour générer des images fidèles à la sémantique. Nous avons montré que l'utilisation d'une description sémantique de l'image est suffisante pour produire des échantillons proches de l'image avec une haute qualité visuelle. Par rapport aux codecs récents, les images décodées sont très détaillées grâce aux informations synthétiques. Cette méthode s'applique à une compression sélective, en générant uniquement les informations de faible importance.

Une amélioration pour des travaux futurs serait de disposer d'un calcul du masque automatique plutôt qu'une étape de détourage pour l'instant effectuée à la main. Ce calcul est néanmoins difficile puisqu'il dépend de l'observateur de l'image.

Références

[1] Eirikur AGUSTSSON, Michael TSCHANNEN, Fabian MENTZER, Radu TIMOFTE et Luc Van GOOL : Generative adversarial networks for extreme learned image compression. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 221–231, 2019.

[2] Yochai BLAU et Tomer MICHAELI : The perception-distortion tradeoff. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018.

[3] Yochai BLAU et Tomer MICHAELI : Rethinking lossy compression : The rate-distortion-perception tradeoff. *In International Conference on Machine Learning*, pages 675–685. PMLR, 2019.

[4] Liang-Chieh CHEN, Yukun ZHU, George PAPANDREOU, Florian SCHROFF et Hartwig ADAM : Encoder-decoder with atrous separable convolution for semantic image segmentation. *In Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[5] Patrick ESSER, Robin ROMBACH et Bjorn OMMER : Taming transformers for high-resolution image synthesis. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[6] Jonathan HO, Ajay JAIN et Pieter ABBEEL : Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[7] Chenlin MENG, Yang SONG, Jiaming SONG, Jiajun WU, Jun-Yan ZHU et Stefano ERMON : Sdedit : Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv :2108.01073*, 2021.

[8] Fabian MENTZER, George D TODERICI, Michael TSCHANNEN et Eirikur AGUSTSSON : High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, 33:11913–11924, 2020.

[9] Robin ROMBACH, Andreas BLATTMANN, Dominik LORENZ, Patrick ESSER et Björn OMMER : High-resolution image synthesis with latent diffusion models. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[10] Zhou WANG, Eero P SIMONCELLI et Alan C BOVIK : Multiscale structural similarity for image quality assessment. *In The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[11] Adam WIECKOWSKI, Jens BRANDENBURG, Tobias HINZ, Christian BARTNIK, Valeri GEORGE, Gabriel HEGE, Christian HELMRICH, Anastasia HENKEL, Christian LEHMANN, Christian STOFFERS, Ivan ZUPANCIC, Benjamin BROSS et Detlev MARPE : Vvenc : An open and optimized vvc encoder implementation. *In Proc. IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–2.