Compression quasi sans perte d'images satellites par filtrage de résidus

Pascal BACCHUS[1] Aline ROUMY[1] Renaud FRAISSE[2] Christine GUILLEMOT[1]

[1]INRIA, Campus de Beaulieu, 263 Av. Général Leclerc, 35042 Rennes, France

[2]Airbus Defence and Space, Toulouse, France

Résumé – Nous décrivons un réseau de neurones entraînable de bout en bout pour la compression d'images satellites. Nous introduisons d'abord une fonction de coût combinant une perte perceptuelle et l'erreur quadratique moyenne classique qui améliore de manière significative les performances de la solution de compression proposée. Nous présentons ensuite une stratégie d'équilibrage multi-objectifs pour optimiser l'apprentissage. Enfin, nous ajoutons une compression du résidu de l'image avec un réseau spécialisé pour préserver au mieux les détails à haute fréquence présents dans les images satellites.

Abstract – We describe an end-to-end trainable neural network for satellite image compression. We first introduce a cost function combining a perceptual loss and the classical mean square error that significantly improves the performance of the proposed compression solution. We then present a multi-loss balancing strategy to optimize the learning. Finally, we add a compression of the image residual with a specialised network to best preserve the high frequency details present in the satellite images.

1 Introduction

La nouvelle génération de caméras embarquées sur les satellites permet d'acquérir des images avec une résolution spatiale et spectrale accrue, ce qui entraîne une énorme quantité de données à transmettre au sol. Cette quantité est d'autant plus exacerbée que l'utilisation de données d'observation terrestre pour tous les types d'applications augmente. Il devient nécessaire de développer de nouvelles solutions de compression efficaces pour la transmission des images satellitaires.

Malgré cette augmentation de résolution, les images satellites ont la particularité d'avoir des détails de la taille d'un pixel et donc une fréquence élevée. Le principal défi de la compression de ces images est de pouvoir distinguer, dans les informations à haute fréquence, l'information du signal de ce qui est dû au bruit pour une interprétation précise. La conception d'algorithmes de compression efficaces pour les images satellites doit donc tenir compte de plusieurs contraintes pour obtenir le meilleur compromis débit-distorsion. Tout d'abord, (i) il doit être *adapté aux statistiques de l'image* et à leurs détails pixelliques. Ensuite, (ii) la compression doit être quasiment sans perte pour permettre une interprétation précise sur le terrain. Dans cet article, nous proposons un algorithme de compression efficace qui satisfait ces deux contraintes.

Les caractéristiques statistiques des images satellitaires diffèrent des images naturelles. Cette adaptation (i) peut être réalisée grâce à des auto-encodeurs variationnels. Ils ont d'abord été introduits pour apprendre des algorithmes de compression de bout en bout pour les images naturelles [4, 13] et finalement surpasser les codecs traditionnels [14, 7]. Ici, nous présentons notre architecture qui améliore encore les résultats de compression [1, 2] en concevant une nouvelle fonction de coût basée sur des métriques perceptuelles [2] ainsi qu'une stratégie d'équilibrage multi-objectifs des fonctions de coût [2]. L'utilisation d'auto-encodeurs pour la compression d'images satellites a également été explorée dans [9] afin de réduire la complexité des architectures basées sur les auto-encodeurs. Cependant, l'une des limites de ces approches est qu'elles saturent à haut débits, et ne peuvent pas reconstruire correctement les détails pixelliques. En effet, le flou est un artefact ajouté lors de la compression et entraîne une perte d'information dans les images à grain fin telles que les images satellites, avec une forte distorsion pour ces détails à haute fréquence. Pour atténuer ce comportement et satisfaire (ii) tout en bénéficiant de l'amélioration apportée par les réseaux d'auto-encodeurs, nous proposons de compresser séparément les résidus à haute fréquence du reste de l'information.

2 Schéma de compression

Dans cette section, nous présentons notre architecture de compression générale [2]. Nous passons d'abord en revue l'algorithme de compression de l'état de l'art également appelé architecture « hyperprior » [4]. Nous décrivons enfin les fonctions de coût que nous proposons.

2.1 Architecture du réseau

L'architecture « hyperprior » [4] est composée de deux réseaux auto-encodeurs comme le montre la figure 1. Le premier produit une représentation latente y des données d'entrée x. Des opérations de compression standard telles que la quantification et le codage entropique sont effectuées sur cette représentation latente pour produire un flux binaire, qui est ensuite décodé par le décodeur entropique sous la forme $\hat{y}$. La dérivée de la fonction de quantification est nulle ou indéfinie, elle est remplacée par un bruit uniforme pour l'apprentissage [3]. Le décodeur reconstruit le signal $\hat{x}$ à l'aide de transformées inverses. L'autre auto-encodeur (l'« hyperprior ») modélise les paramètres de la distribution de la représentation latente pour améliorer le modèle entropique. Ce modèle partagé est adapté aux caractéristiques des données d'entrée, les paramètres étant ré-estimés à chaque entrée.

Nous appliquons un paramètre d'échelle avant la quantification qui agit comme un paramètre de qualité au moment de l'exécution afin que le modèle puisse donner de bons résultats dans une petite plage de débit autour du débit cible. Cela

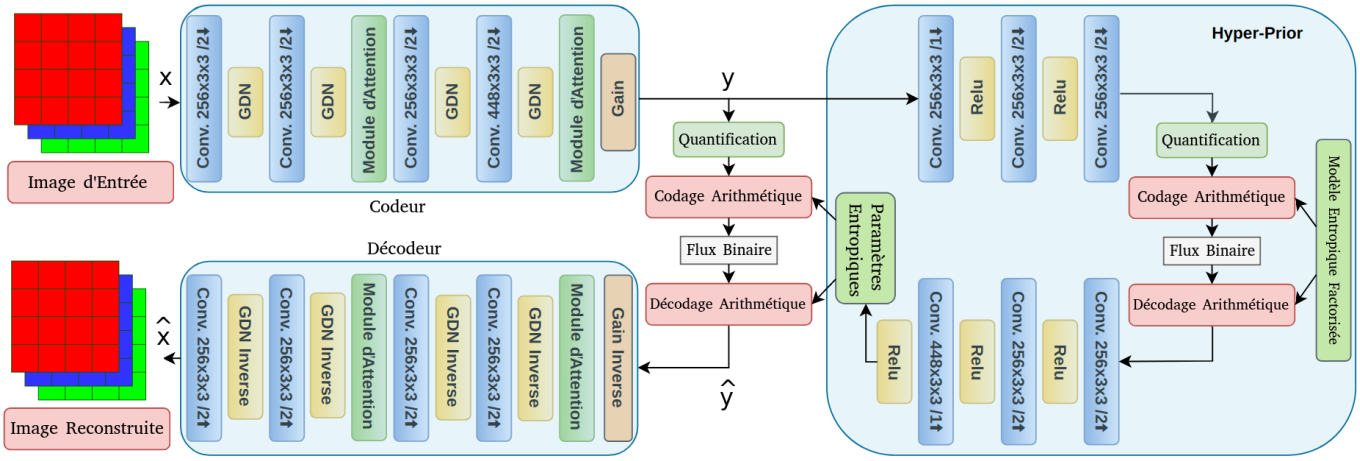


FIGURE 1 : Architecture de compression général [2]. L'architecture spécialisée à la même structure avec une taille des filtres réduite.

permet une plus grande flexibilité car les modèles formés ne sont plus bloqués à un point de distorsion de débit fixe comme c'est le cas avec la plupart [4, 13, 14].

2.2 Fonctions de coût

La principale source d'erreur dans nos images reconstruites provient des motifs rayés à haute fréquence [1]. Ils ont une fréquence spatiale de la taille du pixel et disparaissent en raison du flou généré par la métrique de distorsion, la norme $L2$.

Pour mieux s'adapter aux caractéristiques des données et mieux reconstruire ces zones très typées lors de la compression, nous avons incorporé un a priori qui localise ces zones d'intérêt en ajoutant une métrique perceptuelle dans la fonction de coût. Cette métrique diffère des métriques basées sur les pixels car elle vise à calculer une distance entre des caractéristiques extraites sur les images reconstruites $\hat{x}$ et sur les images vérité terrain x.

Nous savons qu'il existe un compromis entre distorsion et distance perceptuelle [6]. Il existe néanmoins une métrique pour laquelle ce compromis est moins sévère : les réseaux de classification VGG [6]. Nous définissons donc une fonction de coût basée sur ce réseau pour extraire les structures à l'intérieur de nos caractéristiques et guider l'apprentissage vers une meilleure reconstruction à haute fréquence. Nous prenons une distance $L2$ sur les représentations latentes de x et $\hat{x}$ aux profondeurs 2 et 4 de VGG.

$$P(x,\hat{x}) = \frac{1}{nm}(VGG_{0:2}(x,\hat{x})^2 + VGG_{0:4}(x,\hat{x})^2) \quad (1)$$

Nous avons décidé d'utiliser les premières couches de VGG car elles supervisent l'apprentissage des caractéristiques spatiales de bas niveau [15] alors que les couches plus profondes se concentrent sur des caractéristiques plus abstraites. Nous obtenons alors un compromis débit-distorsion-perception pour optimiser les paramètres du réseau :

$$\mathcal{L} = \lambda_a D(x,\hat{x}) + \lambda_b P(x,\hat{x}) + \alpha R(\hat{y}) \quad (2)$$

α est défini pour cibler un débit lors de ce compromis. D est notre métrique de distorsion, une distance $L2$ entre les images vérité terrains et reconstruite. R est la somme des deux flux binaires du réseau, celui de la représentation latente et celui des paramètres entropiques issus de l'« hyperprior ».

2.3 Équilibrage multi-objectifs

Un apprentissage conjoint de plusieurs tâches, avec leurs fonctions de coûts respectives, peut conduire à un meilleur résultat pour toutes les tâches que l'apprentissage pour chaque tâche individuellement, comme le montre [11] où la classification sémantique et l'estimation de la profondeur induisent de meilleures performances ensemble que séparément. Cependant, cela implique d'ajouter plus de termes d'équilibrage dans la fonction de coût et il devient alors plus difficile d'optimiser le réseau pour ce compromis débit-distorsion-perception.

Une solution pour définir les différents paramètres consiste à régler conjointement tous les paramètres de la fonction de coût [8] à l'aide d'un schéma de contrôle automatique. Il n'est plus nécessaire de procéder à un réglage manuel et cela garantit un compromis optimal entre tous les termes de la fonction de coût. La fonction de coût devient :

$$\mathcal{L} = \lambda_1 L_1(x,\hat{x}) + \lambda_2 L_2(x,\hat{x}) + \alpha R(\hat{y}) \quad (3)$$

avec $L_1 = \lambda_a D; L_2 = \lambda_b P$.

Pour évaluer automatiquement les λ_k, nous suivons l'approche de la moyenne de poids dynamique [12] pour calculer à chaque pas d'entraînement un nouveau λ_k basé sur les mesures de coût précédentes pour la distorsion et la métrique perceptuelle :

$$\lambda_k = 2.\frac{exp(\frac{w_k(t-1)}{T})}{\sum_i exp(\frac{w_i(t-1)}{T})}, w_k = \frac{L_k(t-1)}{L_k(t-2)} \quad (4)$$

T représente la température de recuit et est fixée à $T = 2$. Cette valeur contrôle la pondération des tâches entre elles. Une valeur élevée de T se traduit par une répartition plus homogène entre les différentes tâches.

3 Filtrage de résidu

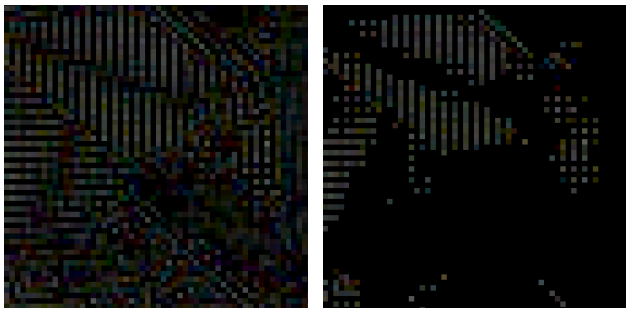
Au lieu de s'appuyer uniquement sur un réseau de compression général, nous préférons des réseaux spécialisés, chacun responsable d'une partie déterminée du spectre de fréquences pour obtenir une reconstruction plus fidèle. Cela conduit à un réseau lourd générique pour une compression générale et à un réseau secondaire plus léger pour capturer les erreurs de haute fréquence dans l'image résiduelle. Ces erreurs résultent

d'une mauvaise reconstruction des hautes fréquences en général (mauvaise prise en compte du bruit) et des motifs rayés en particulier. Le réseau utilisé pour compresser l'image résiduelle est une version allégée du schéma de compression général [1]. Toute l'image résiduelle n'est pas compressée de la même manière, car nous utilisons un masque pour éliminer autant que possible le bruit non structuré. Ce masque est obtenu par filtrage et seuillage de la luminance de l'image résiduelle RGB. Les détails à haute fréquence sont mal reconstruits par rapport aux textures. Nous essayons donc de filtrer le bruit aléatoire induit par la compression et de conserver les blocs de détails significatifs. Le filtre suivant est utilisé pour ne conserver que les motifs rayés.

$$Kernel = \frac{1}{28}\begin{pmatrix} 2 & 0 & 2 & 0 & 2 \\ 0 & 2 & 0 & 2 & 0 \\ 2 & 0 & 4 & 0 & 2 \\ 0 & 2 & 0 & 2 & 0 \\ 2 & 0 & 2 & 0 & 2 \end{pmatrix}$$

Les images sur lesquelles nous travaillons ont une résolution spatiale de 50 cm, ce qui se traduit par des images avec une forte entropie et des détails pixelliques. Nous nous concentrons donc sur les motifs qui ont une période de 2 pixels. Ces motifs sont affichés dans le filtre avec un espacement de chaque composante. Ce filtre met en évidence les zones d'un certain motif au détriment de l'autre. Nous seuillons le résultat pour conserver des zones d'erreurs homogènes. Enfin, ce masque est utilisé sur l'image résiduelle originale pour supprimer le bruit non structuré coûteux à compresser.

Dans la figure 2.a, nous distinguons facilement les erreurs structurées. L'ensemble du processus de filtrage et de seuillage consiste à mettre l'accent sur les zones significatives qui n'ont pas été correctement compressées, de sorte que le réseau de compression spécialisé ne se concentre que sur ces quelques zones d'erreurs.



(a) Résidu pré-traitement (b) Résidu post-traitement

FIGURE 2 : Effet de notre masque de filtrage/seuillage

4 Expérimentations

4.1 Détails de l'entraînement

L'ensemble de données utilisé comprend 300 images satellites de paysages variés, RGB 12 bits (2000x2000) avec une résolution géométrique de 50 cm comme dans [1], 5% sont utilisées pour les tests, le reste pour l'entraînement. Chaque lot d'images est découpé en patchs et augmenté de manière aléatoire avec une rotation pour assurer l'invariance par rotation. Les réseaux ont été conçus à l'aide de la bibliothèque Python CompressAI [5], une sur-couche PyTorch pour les modèles de compression de réseaux de neurones.

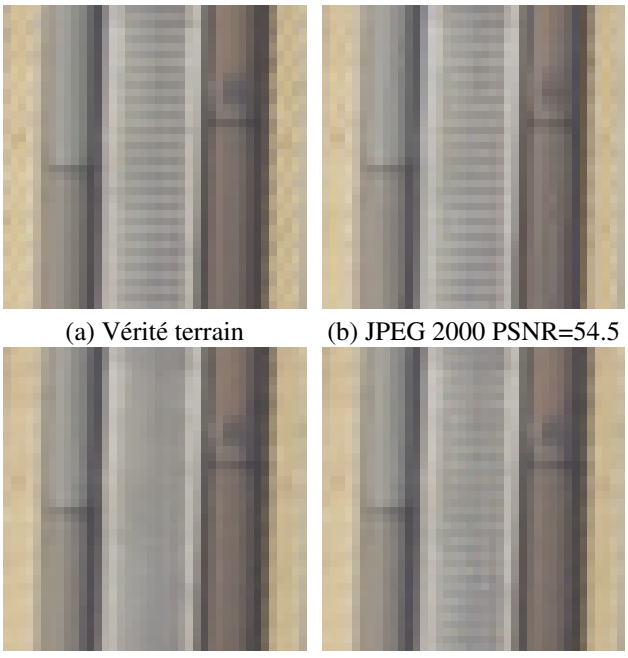
Nous utilisons des méthodes de référence proches des performances des satellites embarqués. Pour la compression, nous considérons JPEG 2000 car ce codec est similaire à la norme utilisée pour les images [10] utilisant des transformées DCT.

Les paramètres de la fonction de coût λ_a et λ_b sont réglés de manière à ce que la distorsion liée à la norme $L2$ et la perte perceptuelle VGG soient du même ordre de grandeur.

La relation entre α et le débit cible est empirique. α est fixé à 0,6 pour toutes les expériences afin de viser 2 bpp (nombre de bits par pixel nécessaires pour compresser le signal) pour que la reconstruction soit d'une qualité suffisante pour les applications satellitaires. Les expériences ont été menées sur des GPU NVIDIA A40. Le temps de calcul est d'environ 1 seconde pour l'encodeur et de 1,5 seconde pour le décodeur.

4.2 Résultats qualitatifs

La figure 3 montre les résultats obtenus avec différentes méthodes en comparaison de la vérité terrain, pour une image satellite d'une résolution géométrique de 50 cm d'un toit d'abris de gare. L'image de référence est compressée à 2 bpp avec le traitement de référence JPEG 2000, le réseau de compression et le réseau de compression avec ajout de la compression du résidu. Toutes les images sont bien reconstruites puisque nous visons un débit élevé. Néanmoins, les détails à haute fréquence, tels que les motifs rayés sur le toit de l'abri, ont disparu pour le modèle (c) malgré un meilleur PSNR que JPEG 2000. Le modèle (d) avec l'ajout de la compression du résidu est proche de récupérer tous ces détails. Ce modèle avec résidu gagne en reconstruction sur une partie minimale de l'image puisqu'on s'aperçoit d'un PSNR pratiquement inchangé.



(a) Vérité terrain (b) JPEG 2000 PSNR=54.5

(c) Réseau PSNR=56.1 (d) Réseau + résidu PSNR=56.2

FIGURE 3 : Comparaison visuelle d'images compressées à même débit (2 bpp) avec la vérité terrain.

4.3 Résultats quantitatifs

Nous évaluons le gain de performance que la perte perceptuelle et l'équilibrage multi-objectifs apportent au modèle commun dans la figure 4. Nous comparons notre modèle pour différentes fonctions de coût basées sur la norme $L2$, VGG ou les deux, ainsi qu'avec l'équilibrage multi-objectifs lorsqu'il est effectué pendant l'apprentissage. Le traitement séquentiel utilisé comme référence est proche des normes d'imagerie satellite avec JPEG 2000 comme codec [10]. Nous nous comparons aussi au modèle [9] ré-entrainé sur nos données.

La fonction VGG seule donne toujours des résultats satisfaisants avec la métrique PSNR même si elle n'est pas adaptée à la distorsion de la norme $L2$. Lorsqu'il est entraîné uniquement avec une norme $L2$, le réseau obtient sans surprise de meilleurs résultats quand il est évalué à l'aide de la métrique PSNR. Les deux métriques combinées conduisent à des performances encore meilleures, en particulier à des débits binaires élevés. La combinaison d'une métrique optimisée sur la norme $L2$ et d'une métrique axée sur l'extraction de la structure réduit les effets de flou induits par le schéma de compression. Le système d'équilibrage permet au modèle précédent d'obtenir un meilleur compromis entre débit et distorsion sur l'ensemble de la plage de débits. Au cours de la formation, les paramètres λ_k s'adaptent à l'importance relative accordée à leur tâche respective au cours des époques précédentes. Cela permet au réseau d'échapper à certains minima locaux, car le paramètre principal λ_k qui conduit le gradient change au fil du temps.

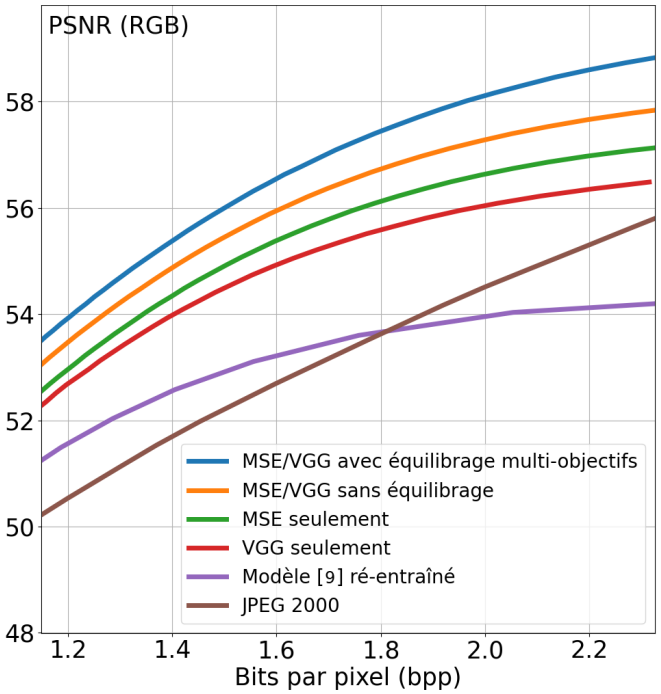


FIGURE 4 : Effet des fonctions de coût et de l'équilibrage multi-objectifs sur les performances du modèle

L'effet du résidu sur les performances quantitatives permet de gagner quelques dixièmes de points de PSNR au prix d'une légère augmentation du débit de 0,3 bpp.

5 Conclusion

Dans cet article, nous avons proposé un modèle de compression conçu pour les images satellites RGB avec des performances de distorsion accrues par rapport au traitement séquentiel traditionnel. La reconstruction est encore améliorée par l'ajout d'une métrique perceptuelle pour extraire les structures à haute fréquence et la stratégie d'équilibrage multi-objectifs pour régler chaque paramètre de la fonction de coût. De plus, l'ajout du résidu compressé permet de récupérer localement beaucoup plus d'informations puisque certains motifs rayés trouvés dans les villes (passages piétons, toits) qui n'avaient pas pu être totalement reconstruits sont maintenant compressés avec une faible distorsion.

Références

[1] P. BACCHUS, R. FRAISSE, A. ROUMY et C. GUILLEMOT : Quasi lossless satellite image compression. *In IGARSS 2022*, pages 1532–1535, 2022.

[2] P. BACCHUS, R. FRAISSE, A. ROUMY et C. GUILLEMOT : Joint Compression and Demosaicking for Satellite Images. *In ICASSP 2023*, juin 2023.

[3] J. BALLÉ, V. LAPARRA et E. P. SIMONCELLI : End-to-end optimization of nonlinear transform codes for perceptual quality. *CoRR*, abs/1607.05006, 2016.

[4] J. BALLÉ, D. MINNEN, S. SINGH, S. J. HWANG et N. JOHNSTON : Variational image compression with a scale hyperprior. *In ICLR*, 2018.

[5] J. BÉGAINT, F. RACAPÉ, S. FELTMAN et A. PUSHPARAJA : Compressai : a pytorch library and evaluation platform for end-to-end compression research. 2020.

[6] Y. BLAU et T. MICHAELI : The perception-distortion tradeoff. *In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018.

[7] Z. CHENG, H. SUN, M. TAKEUCHI et J. KATTO : Learned image compression with discretized gaussian mixture likelihoods and attention modules. *In CVPR*, 2020.

[8] M. CRAWSHAW : Multi-task learning with deep neural networks : A survey. *CoRR*, abs/2009.09796, 2020.

[9] V. Alves de OLIVEIRA, M. CHABERT, T. OBERLIN, C. POULLIAT, M. BRUNO, C. LATRY, M. CARLAVAN, S. HENROT, F. FALZON et R. CAMARERO : Satellite image compression and denoising with neural networks. *IEEE Geoscience and Remote Sensing Letters*, 19, 2022.

[10] Consultative Committee for SPACE DATA SYSTEMS (CCSDS) : *Image data compression CCSDS 122.0-B-1*. CCSDS, 2005.

[11] A. KENDALL, Y. GAL et R. CIPOLLA : Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[12] S. LIU, E. JOHNS et A. J. DAVISON : End-to-end multi-task learning with attention. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] D. MINNEN, J. BALLÉ et G. D. TODERICI : Joint autoregressive and hierarchical priors for learned image compression. *In NeurIPS*, 2018.

[14] D. MINNEN et S. SINGH : Channel-wise autoregressive entropy models for learned image compression. *In ICIP*, 2020.

[15] M. Saeed RAD, B. BOZORGTABAR, U. V. MARTI, M. BASLER, H. Kemal EKENEL et J. P. THIRAN : SROBB : targeted perceptual loss for single image super-resolution. *CoRR*, 2019.