

SIMuLDiTex : Un modèle de diffusion léger, mono-image et multi-échelle pour la synthèse de texture.

Pierrick CHATILLON Julien RABIN David TSCHUMPERLÉ

Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC, UMR 6072, F-14000 Caen, France

Résumé – Nous présentons **SIMuLDiTex**, un modèle frugal de diffusion multi-échelle pour la synthèse de textures à partir d’une seule image d’exemple. Cette approche est basée sur une stratégie pyramidale d’échantillonnage et de génération des textures permettant de préserver la fidélité des détails générés, même à haute résolution. Grâce à un mécanisme original de conditionnement multi-échelle d’un U-Net compact (environ un million de paramètres), SIMuLDiTex surpasse les méthodes de l’état de l’art à nombres de paramètres équivalents, offrant ainsi une solution rapide et de qualité pour la génération de textures de très haute résolution.

Abstract – We present SIMuLDiTex, a frugal multi-scale diffusion model for texture synthesis from a single example image. This approach is based on a pyramidal strategy for texture sampling and generation, ensuring the fidelity of generated details even at high resolution. Thanks to an original multi-scale conditioning mechanism within a compact U-Net (approximately one million parameters), SIMuLDiTex outperforms state-of-the-art methods that have an equivalent number of parameters, providing a fast and high-quality solution for ultra-high-resolution texture generation.

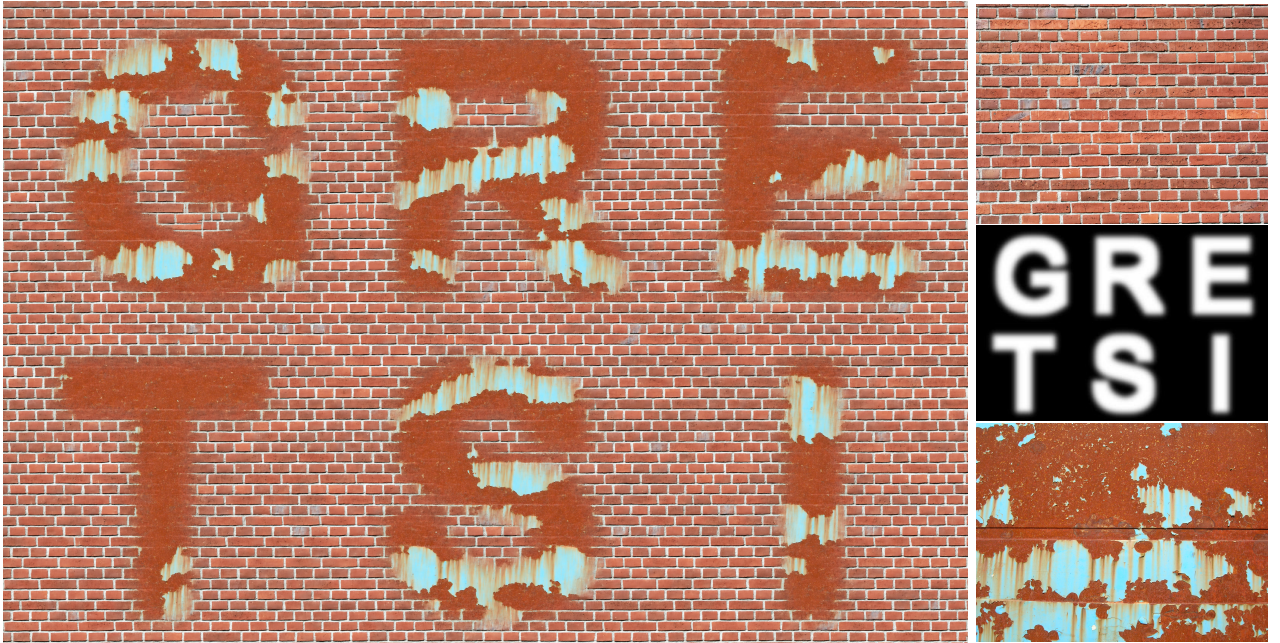


FIGURE 1 : Texture composite synthétisée à la résolution 8K (9216×6144 pixels), en 72 s.

Ce travail a été financé par *IArtist*, un projet “Label Excellence” de la région Normandie et utilise les ressources du CRIANN.

1 Introduction

Contexte Nous nous intéressons au problème de la synthèse rapide et de haute qualité de textures à partir d’une seule image d’exemple, un problème difficile classique de vision par ordinateur, qui a des applications en édition d’images et de vidéos. Notre objectif est de générer une texture haute résolution (à partir d’une texture potentiellement de résolution moyenne), tout en préservant la cohérence perceptuelle et structurelle de l’exemple d’origine, et en assurant une bonne diversité des textures

générées. Les algorithmes traditionnels, comme les approches basées sur les patches, permettent une synthèse rapide mais manquent souvent de fidélité ou de diversité visuelle, notamment quand les textures d’entrées sont complexes. Les méthodes récentes d’apprentissage automatique, comme les GANs, les VAEs ou encore les modèles de diffusion, offrent une synthèse réaliste, mais nécessitent quant à elles des données et des ressources computationnelles importantes. Le défi reste donc de trouver un bon équilibre entre *légereté*, *fidélité* et *diversité*, surtout si l’on veut appliquer des modèles entraînés sur un seul exemple, qui ont souvent du mal à éviter le phénomène de sur-apprentissage et à extrapoler la géométrie des textures correctement. Les méthodes

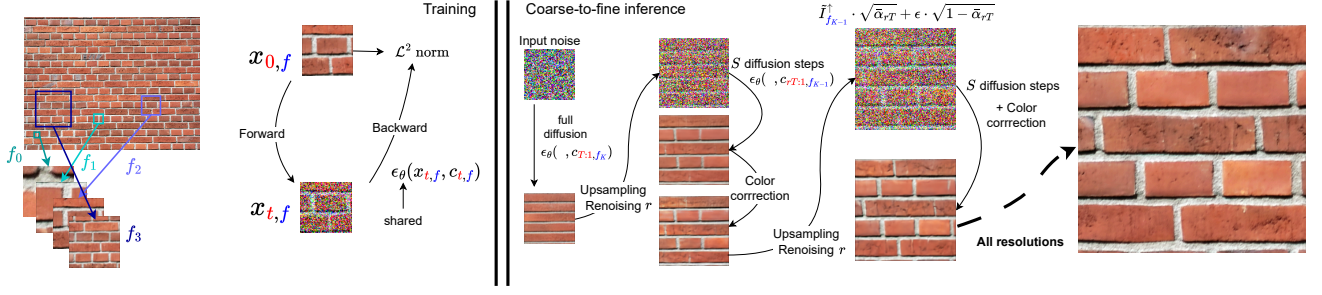


Figure 2: **Vue d'ensemble de la méthode SIMuLDiTex pour l'entraînement et l'échantillonnage de textures.** *Gauche* : Le réseau est entraîné comme un débruiteur, conditionné sur un niveau de bruit t et un facteur d'échelle f . *Droite* : Pendant l'inférence, l'image est synthétisée de manière *coarse-to-fine*, avec sur-échantillonnage, ajout de bruit, étapes de débruitage et correction des couleurs. $\tilde{I}_f$ désigne l'image synthétisée à la résolution f .

s'appuyant sur des modèles de fondations (en utilisation directe ou par *fine-tuning* ou distillation), rencontrent le même défi. Dans tous les cas de figure, générer des images à haute résolution (cf fig. 1) reste prohibitif en raison de l'empreinte mémoire de ces modèles génératifs reposant sur des réseaux de neurones très profonds.

Contributions et organisation Dans ce papier, nous décrivons **SIMuLDiTex**, un modèle de diffusion léger et multi-échelle pour la synthèse rapide de textures à partir d'une seule image. Après une courte revue des modèles génératifs pour la génération de textures (§ 2), notre modèle est décrit en section 3. Ce modèle capture des motifs à longue distance tout en maintenant une qualité de synthèse à haute résolution, en évitant les artefacts classiques des auto-encodeurs. Un mécanisme original de conditionnement à l'échelle nous permet de déployer une approche multi-échelle rapide sans augmenter le nombre de paramètres de la méthode. Nous proposons une architecture U-Net légère, avec seulement un million de paramètres, qui est entraînable efficacement avec peu de données. Nous remplaçons les mécanismes d'attention classiquement utilisés dans ce type de réseaux, par des modules de Fourier [3] afin d'obtenir une synthèse avec une complexité linéaire. En section 4, nous validons notre méthode sur diverses textures, en montrant une performance supérieure en termes de vitesse, de qualité et de scalabilité par rapport aux méthodes équivalentes de l'état de l'art. La figure 1 illustre par exemple le potentiel de la génération de textures haute résolution en temps réel, avec un coût de calcul raisonnable.

2 Méthodes de l'état de l'art

Synthèse de textures avec statistiques de CNN

Les méthodes basées sur les filtres d'image classiques ont récemment évolué vers l'utilisation de réseaux neuronaux pré-entraînés, par exemple avec le travail pionnier de [5], mais l'optimisation de ces méthodes reste coûteuse, comme dans [7, 11]. Des modèles adversariaux, comme SinGAN [16], ont permis de générer des textures rapidement, mais peinent à capturer les structures à grande échelle [16]. D'autres méthodes, telles que [22], ont combiné le critère d'optimisation de texture et la supervision par GAN pour la synthèse de textures non-stationnaires

à haute résolution.

Génération d'images à partir d'une seule image SinGAN [16] et ses variantes utilisent des réseaux génératifs multi-échelles [7, 11, 17] pour synthétiser des images à partir d'un seul exemple.

Modèles de diffusion pour la synthèse de textures Plus récemment, les modèles de diffusion [8] ont montré des capacités de génération d'images impressionnantes, mais ces méthodes nécessitent en pratique de nombreux pas d'échantillonnage. Des travaux récents ont permis de réduire cette charge computationnelle, en particulier par l'utilisation de la diffusion latente [12]. En outre, des modèles comme [4] et [15] créent de grandes images en assemblant des patches dans l'espace latent.

Pour la synthèse de textures, l'approche *Infinite Texture* [21] ajuste un modèle pré-entraîné de diffusion latente pour générer des variations d'une texture donnée. Cependant, ces modèles restent coûteux en calcul et limités en résolution. Des alternatives, comme *SinDiffusion* [20], combinent diffusion et approches de type *SinGAN*, mais sans décomposition multi-échelle. D'autres, comme [9] et [14] utilisent des réseaux dédiés pour la montée en résolution après une synthèse par diffusion à faible résolution.

3 Méthode proposée

Dans cette section, nous présentons notre méthode pour l'entraînement et l'inférence illustrée en fig. 2. Nous décrivons dans un premier temps la représentation multi-échelle des images, puis la conception du réseau de débruitage conditionnel frugal entraîné pour le processus de diffusion inverse à toutes les résolutions de l'image. L'inférence suit une approche *coarse-to-fine*, avec des images à des résolutions grossières qui sont progressivement agrandies et rebruitées, avant d'effectuer quelques étapes du processus inverse de diffusion. Par ailleurs, la distribution des couleurs est ajustée durant le processus de diffusion afin de mieux contrôler la synthèse.

Extraction multi-échelle des patches Nous utilisons les puissances k d'un facteur de zoom f pour créer une pyramide d'images $\{I_{f_k}\}$ à différentes résolutions. Des patches carrés de taille fixe sont extraits de cette pyramide pour être traités par le modèle.



Figure 3: **Synthèse de texture avec SIMuLDiTex.** *Gauche* : image d'exemple pour l'entraînement. *Droite* : Synthèse avec $S = 10$ et $r = 0.8$. L'inférence prend environ 20 secondes pour 10 000 pixels de large au total.

Réseau de débruitage Le modèle génératif est fondé sur un processus de diffusion gaussien. Nous utilisons un réseau de débruitage de type U-Net [13], mais sans couches d'attention pour réduire l'empreinte mémoire. Nous appliquons une convolution rapide dans le domaine de Fourier dans le *bottleneck* du réseau [3] pour mieux capturer les structures périodiques des textures. Le réseau ϵ_θ est conditionné par le facteur d'échelle f et le temps t via un vecteur $c_{t,f}$ [10].

Entraînement Nous désignerons par $x_{t,f}$ une image indexée par un facteur d'échelle f et un pas de temps t du processus de diffusion caractérisé par la variance cumulée correspondante $\bar{\alpha}_t$ [8]. Le but du modèle ϵ_θ est de prédire le bruit ε ajouté à un patch aléatoire p provenant de I_f , à l'instant t du processus de diffusion, conditionnellement à t et f . La fonction de perte correspondante, que nous cherchons à minimiser, est la suivante :

$$\mathcal{L}(\theta) = \mathbb{E}_{t,f,p,\varepsilon} \left\| \epsilon_\theta(\sqrt{\bar{\alpha}_t}p + \sqrt{1 - \bar{\alpha}_t}\varepsilon \mid c_{t,f}) - \varepsilon \right\|^2 \quad (1)$$

Dans cette expression, l'espérance est calculée sur les variables aléatoires suivantes : le pas de temps $t \in [1..T]$ et le facteur d'échelle $f \in \mathcal{F}$, les patches uniformément choisis $p \in \mathcal{P}_f$ définis dans section 3, et le bruit gaussien unitaire $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Échantillonnage *coarse-to-fine* À la résolution la plus grossière f_K , une inférence de diffusion complète est effectuée en utilisant notre réseau conditionné sur l'échelle grossière $\epsilon_\theta(x_{t-1,f_K} \mid c_{t-1,f_K})$ de $t = T$ à $t = 1$. Cette étape est rapide grâce à la résolution limitée et à l'utilisation de l'échantillonnage de bruit déterministe DDIM [18], ce qui permet de limiter le nombre d'étapes à $T = 200$. Ensuite, chaque résolution, de f_{K-1} à $f_0 = 1$, est traitée de manière itérative. Pour chaque résolution f_k , la résolution précédente $x_{0,f_{k-1}}$ est mise à l'échelle par interpolation bilinéaire pour définir $\tilde{x}_{0,f_k}$. Plutôt que de commencer à partir du bruit pur x_{T,f_k} , on démarre à un temps intermédiaire $t = rT$ en utilisant le modèle de diffusion inverse pour définir x_{rT,f_k} en combinant $\tilde{x}_{0,f_k}$ avec x_{T,f_k} , avec le niveau de bruit approprié. Le paramètre $0 \leq r \leq 1$ contrôle le niveau de bruit (fig. 2).

Nous effectuons ensuite S étapes d'échantillonnage DDIM, avec des étapes temporelles échantillonnées linéairement entre $s = rT$ et $s = 1$, pour mettre à jour la résolution actuelle f_k . Nous choisissons typiquement $S = 3$ pour une vitesse d'échantillonnage rapide. Enfin, l'histogramme des couleurs de l'image est ajusté pour

correspondre à celui de l'exemple à l'aide de la distance de Wasserstein projetée. Pour les grandes images (par exemple, 8192 px de large), l'image est générée par grands morceaux (limite de mémoire GPU) et assemblée.

Interpolation Nous générons des textures intermédiaires en combinant les réseaux de diffusion de deux images sources, ce qui permet d'obtenir des transitions douces entre deux textures [6].

Stylisation SIMuLDiTex peut être adapté à la transfert de style, où une image de contenu impose les grandes structures et une image de style les motifs locaux.

4 Résultats

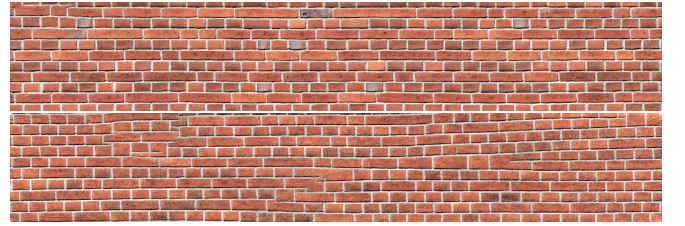


Figure 4: Ablation sur l'utilisation du module Fourier : utilisé pour l'image du haut et pas en bas.

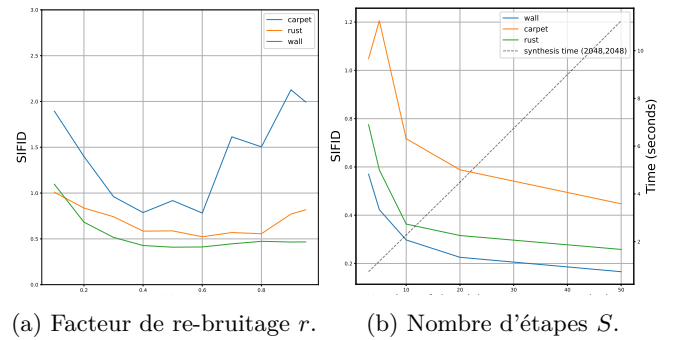


Figure 5: **Ablation sur r et S (en abscisses.** Sélection des hyper-paramètres par le critère SiFID [16] (en ordonnées) pour la synthèse *coarse-to-fine*, et pour 3 exemples différents.

SIMuLDiTex est capable de générer des textures en très haute résolution, comme illustré en fig. 1 et fig. 3.

Interpolation La fig. 1 illustre une interpolation spatiale contrôlable par un masque, obtenue en pondérant les débruitages de deux réseaux entraînés.

Métrique	Temps (s)	VGG-Gram (3 couches) ↓			SIFID ↓			Auto-corrélation ↓		
		1	$\frac{1}{4}$	$\frac{1}{16}$	1	$\frac{1}{4}$	$\frac{1}{16}$	1	$\frac{1}{4}$	$\frac{1}{16}$
Facteur d'échelle										
Snelgrove et al. [17]	120	1	1	1	0.09	0.05	0.17	1842.4	113.9	6.6
Gatys et al. [5]	600	348.2	72.6	69.6	4.14	0.65	1.32	4225.0	262.8	15.8
PSin [2]	780	3.6	5.8	17.3	0.16	0.7	0.97	3568.1	221.5	13.3
SimGAN [16]	1	337.2	292.5	63.7	1.95	1.87	2.02	4315.5	268.4	16.8
Texture Networks [19] ¹	0.3	0.8	5.0	27.0	0.06	0.25	1.48	2333.3	143.5	8.2
PSGAN [1]	0.0003	7.0	16.0	15.8	0.35	1.03	1.63	4874.4	301.5	17.7
SIMuLDiTex S=3	3	66.8	9.7	8.7	0.61	0.3	0.6	2202.3	135.2	7.8
sans correction d'histogramme	3	67.3	18.2	18.5	1.03	0.72	0.85	18958.8	1181.9	73.4
SIMuLDiTex S=10	7	27.2	3.6	4.7	0.27	0.19	0.53	1585.0	97.5	5.7
sans correction d'histogramme	7	29.0	10.5	19.0	0.52	0.56	0.75	19698.9	1227.8	76.3
SIMuLDiTex S=50	28	8.0	2.1	3.7	0.14	0.18	0.6	1807.1	111.2	6.5
SIMuLDiTex 4M S=3	5	47.0	4.5	4.3	0.47	0.22	0.5	1487.9	91.2	5.2
SIMuLDiTex 4M S=10	13	14.1	1.5	3.1	0.19	0.19	0.54	1498.7	91.7	5.2

Table 1: **Comparaison quantitative de fidélité de texture.** Les images sont ici synthétisées en 2048×2048 .

Métriques La table 1 présente une comparaison quantitative à d'autres méthodes, pour trois métriques : Gram-VGG [5], SiFID [16] et distance d'auto-corrélation spatiale. Elles sont calculées à trois résolutions pour évaluer la cohérence multi-échelle.

Architecture U-Net Le module de Fourier améliore l'alignement global des motifs sans le coût des mécanismes d'attention comme illustré en fig. 4.

Nombre d'itérations La fig. 5 étudie l'impact du facteur de rebruitage r et du nombre d'étapes S . Un compromis est nécessaire entre qualité et rapidité d'échantillonnage.

5 Conclusion

SIMuLDiTex est un nouveau modèle de diffusion léger et multi-échelle pour une synthèse de texture rapide et de haute qualité à partir d'une seule image. En tirant parti d'une stratégie de génération *coarse-to-fine* et d'un mécanisme de conditionnement d'échelle original, notre approche accélère l'inférence de manière significative par rapport à d'autres approches d'apprentissage de l'état de l'art. Nos travaux futurs exploreront le cas des textures non-stationnaires et l'intégration avec des cadres de génération contrôlables.

References

- [1] Urs BERGMANN, Nikolay JETCHEV et Roland VOLLGRAF : Learning texture manifolds with the periodic spatial GAN. *In ICML*, volume 70, pages 469–477, 2017.
- [2] Nicolas CHEREL, Andrés ALMANSA, Yann GOUSSEAU et Alasdair NEWSON : A patch-based algorithm for diverse and high fidelity single image generation. *In ICIP*, pages 3221–3225, 2022.
- [3] Lu CHI, Borui JIANG et Yadong MU : Fast fourier convolution. *In H. LAROCHELLE, M. RANZATO, R. HADSELL, M.F. BALCAN et H. LIN, éditeurs : Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 4479–4488. Curran Associates, Inc., 2020.
- [4] Zheng DING, Mengqi ZHANG, Jiajun WU et Zhuowen TU : Patched denoising diffusion models for high-resolution image synthesis. *In ICLR*, 2024.
- [5] Leon A. GATYS, Alexander S. ECKER et Matthias BETHGE : Texture synthesis using convolutional neural networks. *In Advances in Neural Information Processing Systems (NeurIPS)*, page 262–270, 2015.

- [6] Leon A. GATYS, Alexander S. ECKER et Matthias BETHGE : Image style transfer using convolutional neural networks. *In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [7] David J. HEEGER et James R. BERGEN : Pyramid-based texture analysis/synthesis. *SIGGRAPH*, page 229–238, 1995.
- [8] Jonathan HO, Ajay JAIN et Pieter ABBEEL : Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [9] Jonathan HO, Chitwan SAHARIA, William CHAN, David J FLEET, Mohammad NOROUZI et Tim SALIMANS : Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.
- [10] Gabriel MESEGUER-BROCAL et Geoffroy PEETERS : Conditioned-u-net: Introducing a control mechanism in the u-net for multiple source separations. *In ISMIR*, November 2019.
- [11] Javier PORTILLA et Eero P SIMONCELLI : A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000.
- [12] Robin ROMBACH, Andreas BLATTMANN, Dominik LORENZ, Patrick ESSER et Björn OMMER : High-resolution image synthesis with latent diffusion models. *In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.
- [13] Olaf RONNEBERGER, Philipp FISCHER et Thomas BROX : U-net: Convolutional networks for biomedical image segmentation. *In MICCAI*, pages 234–241, 2015.
- [14] Dohoon RYU et Jong-Chul YE : Pyramidal denoising diffusion probabilistic models. *ArXiv*, abs/2208.01864, 2022.
- [15] Benjamin SAMUTH, Julien RABIN, David TSCHUMPERLÉ et Frédéric JURIE : A non-parametric latent model for face generation. *Available at SSRN 4772632*, 2024.
- [16] Tamar Rott SHAHAM, Tali DEKEL et Tomer MICHAELI : Singan: Learning a generative model from a single natural image. *In ICCV*, pages 4570–4580, 2019.
- [17] Xavier SNELGROVE : High-resolution multi-scale neural texture synthesis. *In SIGGRAPH Asia 2017 Technical Briefs*, pages 1–4. 2017.
- [18] Jiaming SONG, Chenlin MENG et Stefano ERMON : Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020.
- [19] Dmitry ULYANOV, Vadim LEBEDEV, Andrea VEDALDI et Victor LEMPITSKY : Texture networks: Feed-forward synthesis of textures and stylized images. *ICML*, page 1349–1357, 2016.
- [20] Weilun WANG, Jianmin BAO, Wengang ZHOU, Dongdong CHEN, Dong CHEN, Lu YUAN et Houqiang LI : Sindiffusion: Learning a diffusion model from a single natural image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [21] Yifan WANG, Aleksander HOLYSKI, Brian L CURLESS et Steven M SEITZ : Infinite texture: Text-guided high resolution diffusion texture synthesis. *arXiv preprint arXiv:2405.08210*, 2024.
- [22] Yang ZHOU, Zhen ZHU, Xiang BAI, Dani LISCHINSKI, Daniel COHEN-OR et Hui HUANG : Non-stationary texture synthesis by adversarial expansion. *ACM Trans. Graph.*, 37(4), juillet 2018.