

Periodično vzorčenje za semantično vodeno generiranje velikih tekstur z latentnim difuzijskim modelom

Jakob Kreft¹, Matija Marolt², Janez Pers¹, Žiga Lesar³

¹Univerza v Ljubljani, Fakulteta za elektrotehniko

²Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

³Univerza v Novi Gorici

E-pošta: jk6684@student.uni-lj.si, matija.marolt@fri.uni-lj.si, janez.pers@fe.uni-lj.si, ziga.lesar@ung.si

Periodic Sampling for Semantically Guided Large-Texture Generation with a Latent Diffusion Model

Diffusion models are trained on fixed-size crops, whereas seamless textures for cylindrical and toroidal surfaces span thousands of pixels. We present a two-stage pipeline that decouples global semantic layout from local appearance. First, a Markov random field synthesises a coarse class map with controllable class proportions. Second, a mask-conditioned latent diffusion model generates a high-resolution texture by asynchronously denoising randomly chosen overlapping latent patches. We extend the sampling procedure with modular patch indexing and periodic Hann-window blending during VAE decoding, which imposes cylindrical or toroidal boundary conditions purely at sampling time, with no architectural changes and no additional retraining. The U-Net uses the DiffInfinite architecture and is trained from random initialisation on our data. The only pretrained component is the frozen Stable Diffusion VAE. A dataset of 270 spruce-bark image and mask pairs was prepared from 140 field photographs of 15 logs. Over 200,000 training steps the loss plateaued after about 20,000 steps, while patch-based FID and KID kept improving until about 150,000 and proved the more reliable progress indicators. The final model reaches FID 139 and KID 0.072 on 512×512 patches at 250 DDIM steps. A further 1000 steps gain little at four times the cost. Stress tests with synthetic masks confirm that semantic regions are respected, and the texture wraps seamlessly around a cylindrical log model. Periodicity is exact by construction, because modular indexing removes the notion of a canvas edge. Both components are needed.

1 Uvod

Študentje gozdarstva se ocenjevanja kakovosti hlodov učijo na terenu, kjer je število dostopnih primerov omejeno z lokacijo in letnim časom. To delo je korak k učnemu orodju, ki bi generiralo poljubno število navideznih smrekovih hlodov z realistično porazdeljenimi značilnostmi lubja. Tekstura takega hloda mora biti lokalno fotorealistična, brezšivna po obodu valja in mora ohranjati naravno globalno razporeditev semantičnih razredov: lubja, slepic in mehanskih poškodb. Slepica je zaraščena sled odpadle veje in se na lubju kaže kot bradavičasta izboklina

z značilnim vzorcem kitajskih brkov. V tem delu enako obravnavamo tudi odrezane veje.

Difuzijski modeli praviloma generirajo izseke enake velikosti, kot so jih videli pri učenju [1, 2], obod hloda pa v visoki ločljivosti meri več tisoč pikslov. Metode za združevanje več difuzijskih poti [3, 4] zmanjšajo meje med zaplatami. Sorodni pristopi obravnavajo poljubno ločljivost [5], usklajeno ploščičenje [6] ali vodoravno periodičnost panoram [7]. Prostorsko pogojevanje je mogoče dodati tudi predučenim modelom [8]. Nobena od naštetih metod ne združuje splošnih valjnih oziroma toroidnih robnih pogojev z eksplicitnim nadzorom globalnih deležev semantičnih razredov.

Predlagani postopek zato loči globalno kompozicijo od lokalnega videza. Grob semantični zemljevid določa razporeditev razredov, maskirno pogojeni latentni difuzijski model pa iz njega generira fotorealistično teksturo. Osrednji prispevek je razširitev vzorčenja DiffInfinite [9] z modularnim indeksiranjem zaplat in periodičnim mešanjem pri dekodiranju, ki deluje na poljubno velikem valjnem ali toroidnem platnu. Razširitev spremeni zgolj vzorčenje, zato ne zahteva dodatnega učenja ali posega v arhitekturo mreže U-Net. Dodatni prispevki so generator semantičnih zemljevidov z nadzorom deležev razredov, javno objavljen podatkovni nabor slik lubja s semantičnimi maskami in empirična analiza učenja na našem naboru.

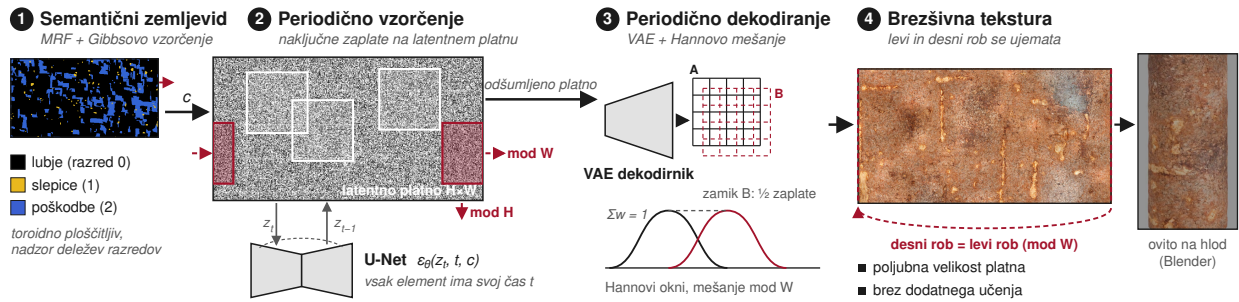
2 Predlagani postopek

Postopek ima dve fazi, ki ju povzema slika 1. Prva generira grob semantični zemljevid z nadzorovanimi deleži razredov, druga pa iz njega sintetizira teksturo na velikem latentnem platnu. Zadnji podrazdelek platno opremi s periodičnimi robnimi pogoji, in sicer zgolj s spremembo vzorčenja in dekodiranja.

2.1 Semantični zemljevid

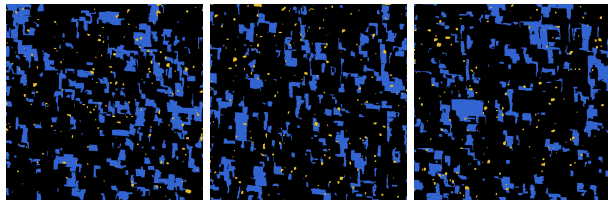
Vsaka celica grobega zemljevida pripada enemu od treh razredov. Semantični zemljevid je namenoma nizkoločljiv. Modelira globalno kompozicijo razredov in ne drobne teksture. Zemljevid generiramo z markovskim naključnim poljem (MRF) [10, 11] in Gibbsovim vzorčenjem. Energija konfiguracije $\mathbf{x}$ je

$$E(\mathbf{x}) = \sum_i \phi_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) + \lambda R_{\pi}(\mathbf{x}), \quad (1)$$



Slika 1: Pregled postopka. MRF generira toroidno ploščitljiv semantični zemljevid (1), ki pogojuje difuzijsko vzorčenje naključnih prekrivajočih se zaplat na latentnem platnu z modularnim indeksiranjem (2). Sledi dekodiranje VAE v dveh zamaknjenih mrežah s Hannovim mešanjem (3); rezultat je brezšivna toroidna tekstura, ovita na valjasti hlood (4). Periodičnost je dosežena zgolj s spremembo vzorčenja, brez dodatnega učenja.

kjer unarni potenciali ϕ_i povzamejo empirične frekvence razredov, parni potenciali ψ_{ij} skladnost parov sosednjih celic iz množice $\mathcal{E}$, R_{π} pa kaznuje odstopanje od ciljnih deležev π . Utež $\lambda > 0$ uravnava vpliv tega globalnega člena. Pogoje porazdelitve ne modeliramo parametrično. Razred celice vzorčimo po statistiki enakih sosedstev (jedro 7×7) v ročno označenih maskah. Če nastane sosedstvo, ki ga v maskah ni, razred vzorčimo po najbolj podobnih sosedstvih. Vsi indeksi so modularni, zato je že pogojni signal periodičen. V primerjavi z algoritmom kolapsa valovne funkcije (WFC) [12] je MRF stabilnejše ohranjal globalne deleže razredov in večje povezane regije. Naravne vzorce je dalo vzorčenje pri temperaturi $\tau = 1$ (slika 2).



Slika 2: Trije semantični zemljevidi (1024×1024 celic), generirani z MRF. Zaradi modularnih indeksov so toroidno ploščitljivi. Črna je lubje, rumena slepice in modra poškodbe.

2.2 Veliko latentno platno

Zamrznjeni variacijski samokodirnik (VAE) preslika slike velikosti 512×512 v latentne predstavitve $64 \times 64 \times 4$. U-Net na vsakem koraku obratne difuzije napove šum ob pogojevanju na semantično masko [2, 9]. Za veliko ciljno sliko celotno latentno platno inicializiramo z Gaussovimi šumom. V vsaki iteraciji izberemo zaplato 64×64 , centrirano na naključnem mestu med najmanj odšumljenimi, na njej izvedemo korak obratne difuzije in posodobitev zapišemo nazaj. Vsak latentni element šteje svoje opravljene časovne korake. Ker isto mesto nastopa v mnogo različnih zaplatah, se lokalni prehodi usklajujejo po celotnem platnu [9].

2.3 Periodični robni pogoji

Naj bo latentno platno velikosti $H \times W$, zaplata pa $h \times w$ z začetkom (i, j) . Za torus element zaplate (a, b) beremo in zapišemo na mestu

$$\mathcal{I}_T(a, b) = ((i + a) \bmod H, (j + b) \bmod W), \quad (2)$$

za valj pa je periodična le vodoravna os,

$$\mathcal{I}_C(a, b) = (i + a, (j + b) \bmod W). \quad (3)$$

Zaplata, ki sega čez rob, se nadaljuje na nasprotni strani platna (slika 1, stopnja 2). Robova platna zato nista posebna primera. Med vzorčenjem lahko nastopita v isti zaplati in se uskladita enako kot katerikoli drugi sosednji mesti.

Po odšumljanju platno dekodiramo po prekrivajočih se zaplatah v dveh mrežah, zamaknjenih za polovico zaplate. Vsako dekodirano zaplato utežimo z dvodimenzionalnim Hannovim oknom ter vsebino in uteži akumuliramo z istim modularnim indeksiranjem. Končni piksel je

$$\hat{x}(p) = \frac{\sum_k w_k(p) x_k(p)}{\sum_k w_k(p) + \varepsilon}, \quad (4)$$

kjer je x_k k -ta dekodirana zaplata, w_k njeno Hannovo okno, ε pa majhna pozitivna konstanta, ki prepreči deljenje z nič na mestih brez prekrivanja. Samo modularno latentno vzorčenje ne odpravi šiva, ki bi ga ponovno vneslo neperiodično dekodiranje. Razširitev ne posega v mrežo U-Net, njene uteži, urnik DDIM ali pogojevanje, zato je vzorčenje v notranjosti platna identično izvirnemu DiffInfinite [9]. Metriki FID in KID na zaplatah 512×512 sta zato za oba postopka enaki in merita realizem, ne periodičnosti.

3 Podatki in izvedba

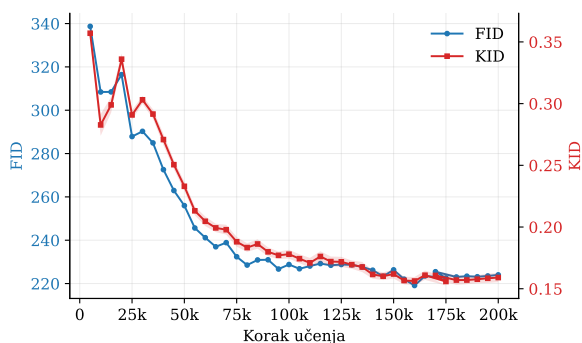
Na dveh terenskih vajah smo v sodelovanju z Oddelkom za gozdarstvo in obnovljive gozdne vire Biotehniške fakultete fotografirali 15 smrekovih hloodov in zajeli 140 slik, od katerih je po odstranitvi neostrih posnetkov ostalo 133 uporabnih. Slike smo poravnali, cilindrično popačenje odpravili z odvijanjem plašča v ravnino, normalizirali osvetlitev in ročno označili tri razrede. Nastalo je 270 parov slik in mask velikosti 1024×1024 . Groba ročna maska ima eno celico na približno 50×50 pikslov, kar omogoči hitro označevanje, vendar omeji ostrino razrednih meja. Segmentacijski model, naučen na tem naboru, pri petkratnem prečnem preverjanju doseže mIoU 0,843 (U-Net [13]) oziroma 0,847 (DeepLabv3+ [14]).

Difuzijski U-Net smo učili 200.000 korakov pri efektivnem svežnju 16 na eni grafični kartici z 20 GB po-

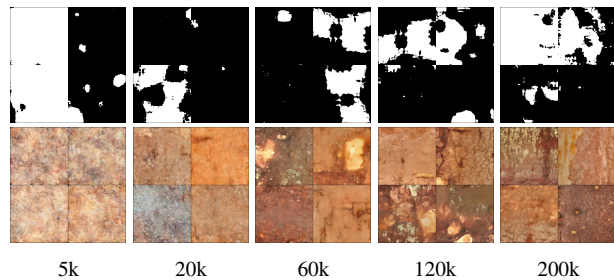
mnilnika. Za ponovljivost navajamo izhodišče uteži. U-Net ima arhitekturo DiffInfinite [9] in ga učimo od naključne inicializacije. Edini predučeni in med učenjem zamrznjeni del je samokodirnik `sd-vae-ft-mse` modela Stable Diffusion [2]. Pri majhnem naboru se je kot ključno izkazalo bogatenje učne množice: zaplate 512×512 med učenjem naključno izrezujemo iz večjih zaplat 1024×1024 ter dodajamo zrcaljenja, rotacije za 90° in zmerne barvne spremembe, kar je občutno znižalo FID in KID. Uporabljamo P2-utežitev kriterijske funkcije [15] in urnik učne stopnje OneCycleLR [16] z največjo vrednostjo $5 \cdot 10^{-5}$. Vzorčenje uporablja DDIM [17]. FID [18] in KID [19] računamo na 1076 generiranih zaplatah 512×512 , pogojenih na maske iz nabora, in jih primerjamo z realnimi sredinskimi izrezi. Koda, naučene uteži in podatkovni nabor so dostopni prek javnega repozitorija <https://github.com/jakobkreft/infinite-bark-thesis>, podrobnosti pa v magistrskem delu [20].

4 Rezultati

Kriterijska funkcija (napaka napovedi šuma) se je ustalila že okoli 20.000 korakov, metriki FID in KID pa sta se izboljševali še do približno 150.000 korakov (slika 3). Pri 50 korakih DDIM je FID med kontrolnima točkama 5k in 200k padel s približno 340 na 225, KID pa z 0,36 na 0,16. Razkorak pojasni povprečenje izgube čez vse šumne nivoje. Izgube posameznih nivojev se razlikujejo za tri velikostne rede, zato skupno povprečje poznih izboljšav ne pokaže. Ustaletev izgube torej ni primeren kriterij za ustavitev učenja. Boljša izbira je periodični izračun FID ali KID na zaplatah. Napredek kvalitativno prikazuje slika 4. Med 5k in 60k koraki model preide iz grobe barvne osnove v prepoznavno teksturo lubja. Kasneje se izostrijo robovi slepic in barvni prehodi poškodb, na lubju pa se pojavijo lišaji.



Slika 3: FID in KID skozi kontrolne točke učenja pri 50 korakih DDIM. Obe metriki se izboljšujeta še dolgo po ustalitvi izgube; osenčeni pas ob KID je standardni odklon prek naključnih podvzorcev.



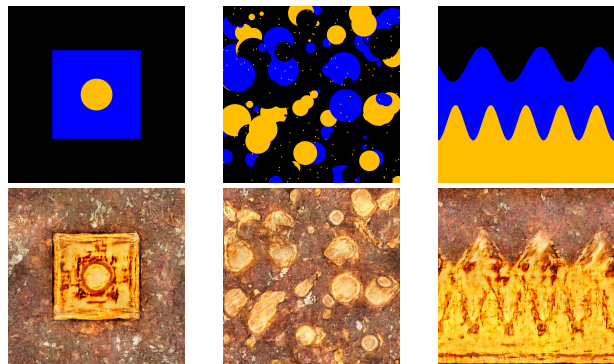
Slika 4: Napredek modela skozi učenje pri petih kontrolnih točkah. Zgoraj vhodna binarizirana maska, spodaj generirani vzorec.

Tabela 1 prikazuje kompromis med kakovostjo in ceno vzorčenja pri končni kontrolni točki. Prehod z 10 na 250 korakov DDIM močno izboljša obe metriki, med 250 in 1000 koraki pa je napredek majhen glede na štirikrat daljši čas: 250 korakov ujame približno 95 % celotne izboljšave FID pri četrtini računske cene.

Tabela 1: Kakovost in čas glede na število korakov DDIM.

Koraki	FID ↓	KID ↓	čas/slika [s]
10	295	0,237	< 1
250	139	0,072	13
1000	131	0,061	52

Slika 5 preverja odziv na sintetične maske z namerno nenaravnimi geometrijami, ki jih model med učenjem ni videl. Razredi so prostorsko upoštevani tudi daleč zunaj porazdelitve učnih mask, kar potrjuje, da rezultat ni le nepogojena tekstura.



Slika 5: Pogojevanje na sintetičnih maskah zunaj učnega nabora. Zgoraj vhodna maska (črna lubje, rumena slepice, modra poškodbe), spodaj generirana tekstura.

Teksturna statistika to dopolni. Po lastnostih GLCM [21] so generirane slike bolj gladke od realnih. Kontrast pri zamiku petih pikslov je 34,8 proti 63,6, višje pa so tudi homogenost, energija in korelacija. Relativna Frobeniusova razdalja med povprečnima Gramovima matrikama [22] v sloju `relu1_1` mreže VGG-19 je 0,64, med polovicama realne množice pa 0,07. Model torej dobro reproducira razredno pogojeni videz, izgubi pa del visokofrekvenčne zrnatosti lubja.

Brezšivnost izhaja iz konstrukcije in ni empirična ugotovitev. Vse lokacije na platnu so enakovredne, zato šiva ni kje meriti. Generirana tekstura se okoli valjastega 3D-modela hloda ovije brez vidnega stika (slika 1, stopnja 4). Nabor 15 hlodov oziroma 270 parov je za difuzijski model majhen in vsi hlodi sodelujejo pri učenju, zato rezultatov

ne razlagamo kot dokaz posploševanja na nova drevesa ali vrste. Čista delitev po hlohah pri 15 drevesih ni izvedljiva. Prednostni dopolnitvi sta zato zajem bistveno večjega nabora z več drevesnimi vrstami in vrednotenje na hlohah, v celoti izločenih iz učenja.

5 Zaključek

Predstavili smo dvofazni postopek za semantično vodeno generiranje brezšivnih tekstur. MRF nadzira globalno kompozicijo razredov, maskirno pogojeni latentni difuzijski model pa lokalni videz. Modularno indeksiranje zaplat in periodično Hannovo mešanje pri dekodiranju razširita DiffInfinite na poljubno velika valjna in toroidna platna zgolj s spremembo vzorčenja, brez dodatnega učenja. Brezšivnost je s tem zagotovljena po konstrukciji. Potrebna sta oba dela razširitve, saj samo modularno indeksiranje šiva ne odpravi. Na lubju smrekovih hlohov postopek pri 250 korakih DDIM dosega FID 139 in KID 0,072. Pri majhnem naboru se je najbolj obrestovalo razširjeno bogatenje podatkov, za spremljanje učenja pa sta se namesto kriterijske funkcije izkazala FID in KID.

Izjava o uporabi umetne inteligence

Pri pripravi članka je bilo orodje generativne umetne inteligence uporabljeno kot pomoč pri zgoščevanju in strukturiranju vsebine magistrskega dela [20], jezikovnem urejanju, pripravi zapisa $\LaTeX$ ter preverjanju dosegljivosti in ustreznosti bibliografskih referenc. Prvi avtor je preveril vse navedbe, rezultate in vire ter v celoti odgovarja za končno vsebino članka.

Zahvala

Delo, objavljeno v tem članku, je bilo sofinancirano iz programov ARIS P2-0095 in P6-0436.

Literatura

- [1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
- [3] O. Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel, "MultiDiffusion: Fusing diffusion paths for controlled image generation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 1737–1752.
- [4] Q. Zhang, J. Song, X. Huang, Y. Chen, and M.-Y. Liu, "DiffCollage: Parallel generation of large content with diffusion models," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 10 188–10 198.
- [5] S. Bond-Taylor and C. G. Willcocks, " ∞ -Diff: Infinite resolution diffusion with subsampled mollified states," in *The Twelfth International Conference on Learning Representations*, 2024.
- [6] O. Madar and O. Fried, "Tiled diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 7795–7804.
- [7] T. Wu, C. Zheng, and T.-J. Cham, "PanoDiffusion: 360-degree panorama outpainting via diffusion," in *International Conference on Learning Representations*, 2024.
- [8] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.
- [9] M. Aversa, G. Nobis, M. Hägele, K. Standvoss, M. Chirica, R. Murray-Smith, A. Alaa, L. Ruff, D. Ivanova, W. Samek, F. Klauschen, B. Sanguinetti, and L. Oala, "DiffInfinite: Large mask-image synthesis via parallel random patch diffusion in histopathology," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. [Online]. Available: <https://openreview.net/forum?id=QXTjde8evS>
- [10] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 721–741, 1984.
- [11] S. Z. Li, *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.
- [12] M. Gumin, "WaveFunctionCollapse algorithm," <https://github.com/mxgmn/WaveFunctionCollapse>, 2016, accessed: 2026-03-11.
- [13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.
- [15] J. Choi, J. Lee, C. Shin, S. Kim, H. Kim, and S. Yoon, "Perception prioritized training of diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 472–11 481.
- [16] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications*, vol. 11006. SPIE, 2019, pp. 369–386.
- [17] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=StlgiaRCHLP>
- [18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.
- [19] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *International Conference on Learning Representations*, 2018.
- [20] J. Kreft, "Semantično vodeni difuzijski modeli za neskončno velike toroidne tekture," Magistrsko delo, Univerza v Ljubljani, Fakulteta za elektrotehniko, 2026. [Online]. Available: <https://hdl.handle.net/20.500.12556/RUL-185875>
- [21] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.
- [22] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," *Advances in neural information processing systems*, vol. 28, 2015.