Localised Frequency Latent Domain Watermarking of DDIM Generated Images

Qiran Lai and Adrian G. Bors

Department of Computer Science, University of York, York YO10 5GH, UK

ABSTRACT

Stable Diffusion models, relying on iterative generative latent diffusion processes, have recently achieved remarkable results in producing realistic and diverse images. Meanwhile, the widespread application of generative models raised significant concerns about the origins of image content or the infringement of intellectual property rights. Consequently, a method for identifying AI generated images and/or other information about their origins is imperatively necessary. To address these requirements we propose to embed watermarks during one of the diffusion iterative steps of the DDIM. Such watermarks are required to be recoverable while also robust to possible changes to the generated watermarked images. The watermarks are embedded in the localized regions of the latent space frequencies. The binary watermarks are detected from the generated watermarked images by means of a CNN watermark detector. The robustness of the CNN watermark detector is improved through training by considering various distortions to the watermarked images.

***Index Terms*—** Image Generation, Digital Watermarking, Denoising Diffusion Implicit Model, Copyright protection.

1. INTRODUCTION

Image generative models, such as the Variational Autoencoder (VAE) [1], Generative Adversarial Networks (GAN) [2] or the Denoising Diffusion Models (DDIMs) [3], developed for synthesising artificial images, have lately found many applications. Among the generative models, the Stable Diffusion [4] has lately became the most popular generative model, due to the quality as well as the diversity of the generated images. Diffusion models [3, 5, 6], implement gradually adding noise and denoising processes. The image information is learnt during the denoising process, ensuring a high quality of the resulting image. However, the widespread use of AI-generated images raised concerns [7] about their misuse, including for human deception, as well as about raising copyright infringement claims. In order to address such concerns, digital watermarks [8, 9, 10] can be used to identify the AI-generated images. The problem with watermarking Stable Diffusion generated images is represented by trade-off between achieving watermark robustness and their invisibility in the generated images, given that both are dependent on the strength of the embedded watermark changes. Traditional methods add watermarks as post-generation, but such watermarks can easily be removed. In this paper we propose a new method, namely the Localised Frequency Latent Domain Watermarking (LFLDW), which embeds watermarks in one of the internal diffusion steps, ensuring the resulting image quality while enhancing watermark robustness as well.

Previous studies can be divided into three main categories: 1) embedding watermarks directly into images [9, 11]; 2) embedding changes during the VAE stage of the generative stable diffusion process [12]; 3) embedding watermarks during the diffusion process itself [13, 14, 15]. Compared with other approaches, embedding watermarks into the diffusion process has major advantages given that it represents the image generation component during which most DDIM's parameters are updated (ten times more parameters are changed than for the VAE enabling the diffusion), resulting in watermarks that are difficult to remove which are also robust and secure.

A promising direction of research is by embedding tree-rings watermark changes in the Fourier domain [14] of the DDIM [5] initial diffusion latent variables. In addition, the improved frequency domain watermarking method called ZoDiac [15] uses backpropagation to get better trainable diffusion noise when embedding the watermark, resulting in the watermarked image becoming more similar to the one initially intended to be generated. However, this approach has at least two limitations: 1) it requires the original image for embedding the watermark; 2) it employs backpropagation to optimize the initial diffusion noise each time when producing the watermarked images.

The proposed Localised Frequency Latent Domain Watermarking (LFLDW) provides the following advantages over the existing generative image watermarking approaches : 1) Watermarks can be embedded in any iteration of the diffusion denoising mechanism, ensuring control over the watermarked image properties while enhancing the watermark security; 2) Watermarks are embedded in the middle frequency range of the latent domain, ensuring a better trade-off between the watermark visibility and its robustness to image compression and other distortions; 3) A neural network is trained for extracting the watermark, ensuring a high watermark prediction accuracy while achieving significant watermark robustness advantages over other approaches.

2. GENERATIVE DIFFUSION DOMAIN WATERMARKING

The Denoising Diffusion Model (DDIM) [3, 6, 16] image generation process consists of a succession of dual-step iterative Markov Chain processes of successive noise additions followed by denoising steps. During these steps, the image is synthesized from latent spaces through optimization. The Stable Diffusion [4] uses a pre-trained CLIP model [17] as a text encoder receiving user prompts as inputs and a UNet network [18] as the diffusion network backbone, while a VAE, consisting of an encoder and a decoder, compresses the feature space for accelerating the image generation.

The DDIM [5] improved the Denoising Diffusion Probabilistic Models (DDPM) [3] by reducing the randomness while accelerating the dual sampling process, through :

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \underbrace{\left(\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(\mathbf{x}_t\right)}{\sqrt{\bar{\alpha}_t}} \right)}_{\text{predicted } \mathbf{x}_0} + \underbrace{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} \cdot \epsilon_\theta\left(\mathbf{x}_t\right)}_{\text{direction pointing to } \mathbf{x}_t} + \underbrace{\sigma_t \epsilon_t}_{\text{random noise}} \tag{1}$$

where $\sigma_t^2 = \eta \cdot \sqrt{(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)}\sqrt{(1-\bar{\alpha}_t/\bar{\alpha}_{t-1})}$, and the standard deviation σ is a weighting parameter while $\mathbf{x}_1, \ldots, \mathbf{x}_T$ are latent variables. θ are the network parameters and $\bar{\alpha}_t$ is the cumulative multiplication of α's controlling the noise strength in the forward process.

For $\sigma_t = 0$, the reverse mechanism of the diffusion model becomes a deterministic process without noise, which means that DPM can skip some steps and eliminate the dependency on the Markov chain, according to the relationship :

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(\mathbf{x}_t\right)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t-1}} \cdot \epsilon_\theta \mathbf{x}_t \tag{2}$$

The proposed Localised Frequency Latent Domain Watermarking (LFLDW) method, which is illustrated in Fig. 1, uses the latent space created by the Denoising Diffusion Implicit Model (DDIM) [3] in order to embed watermarks, considered as sequences of bits. The watermarks are embedded in the frequency coefficients, provided by the Fast Fourier Transform (FFT) of the latent code, obtained at a specific denoising step of the DDIM image generation pipeline. The frequency domain watermarking provides an ideal watermarking control environment enabling the trade-off between the watermark visibility and its robustness to JPEG compression, noise addition or blurring attacks [9]. The reconstructed latent code using the Inverse FFT (IFFT) is then fed into the DDIM iterative process, which proceeds accordingly, resulting in an image carrying the given watermark without causing visible changes. Then a specialised neural network is trained to be used to detect the watermark from the generated image, even when the watermarked image is distorted.

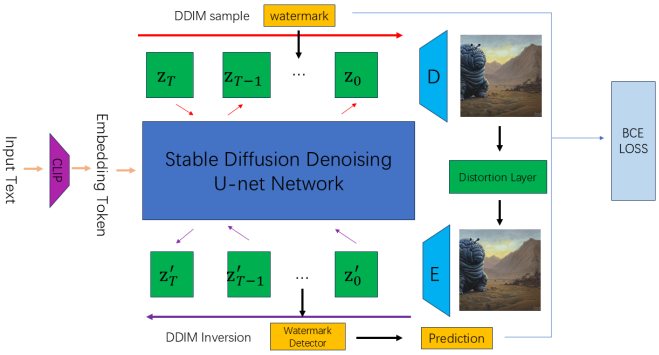

Fig. 1: The Localised Frequency Latent Domain Watermarking (LFLDW) architecture using the Fourier frequency coefficients of the latent space for embedding the watermark.

3. WATERMARK EMBEDDING

During the noise generation stage of the DDIM, we have a 4-channel latent code $\mathbf{z}_t$, when generating the image of size $N \times M$, at iteration t, $t \leq T$, where T is the total number of iterations for the DDIM process. The watermark can be embedded in various t iterations of the DDIM and this was studied in an ablation study. Then, the Fast Fourier Transform (FFT) is used to convert a chosen channel of the latent code $\mathbf{z}_t$ to the frequency domain. The real components of a set of coefficients located within a circular ring of frequencies $\mathcal{D} = \{\{\mathbf{c}_{ij}^t\} | r_1 < \|\mathbf{f}_{ij}^t\| = \sqrt{(\mathbf{c}_i^t - \frac{N}{2})^2 + (\mathbf{c}_j^t - \frac{M}{2})^2} < r_2, i = 1, \ldots, N, j = 1, \ldots, M\}$, where $\mathbf{c}_{ij}^t$ are the coefficients corresponding to the real part of the Fourier domain, and r_1 and r_2 represent the inner and outer radius bounds of the ring of $|\mathcal{D}|$ frequencies, chosen to be watermarked. We then consider the average of all real parts of the Fourier coefficients $\mathbf{c}_m^t = \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\mathbf{c}_{ij}}{NM}$. We embed the watermark $\mathbf{w} = \{w_l\}_{l=1}^{L}$, consisting of a sequence of N bits, into the selected coefficients using :

$$\widehat{\mathbf{c}}_{ij}^t = \mathbf{c}_m^t + \gamma(2w_l - 1) \tag{3}$$

where $\widehat{\mathbf{c}}_{ij}^t$, for $\{i, j\} \in \mathcal{D}$, represents the frequencies carrying the watermark signal and for the watermark digits $l = 1, \ldots, L$, where L is the watermark size ($L = 8$ in the experiments) and γ represents the watermark strength.

After embedding all L watermark bits, the latent code is reconstructed by the Inverse Fast Fourier Transform (IFFT) and then the iterative DDIM process continues from this latent code with the image reconstruction and for the further $\{t-1, \ldots, 0\}$ iterations, without any changes from the usual DDIM procedure. Eventually, the watermarked generated image is produced.

4. WATERMARK DETECTOR TRAINING

In order to enhance the watermark robustness to attacks, we employ a Convolution Neural Network (CNN) which is trained for detecting the watermarks after considering a

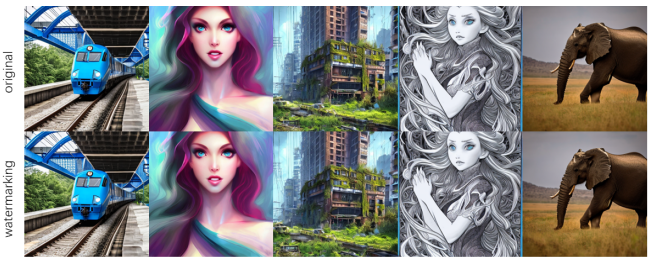

Fig. 2: Watermarked DDIM generated images (bottom) compared to the non-watermarked generated images (top).

variety of distortions. Such distortions are applied to the generated watermarked image and are consistent with possible attacks intended to remove the watermark, or produced by algorithms involved in the usual image processing, such as image compression. We consider additive noise, colour jitter changes, image blurring, cropping, JPEG lossy compression, as well as geometric transformations such as image rotation, as image attacks. During the training, in the distortion layer, one of these attacks is randomly chosen each time in order to be used to distort the watermarked image. The distorted watermarked images are then used for training the watermark detector to enable watermark detection after such distortions and eventually increase robustness to the corresponding attacks. The resulting distorted watermarked images pass through the same VAE encoder associated with the DDIM, and then DDIM inversion is employed for estimating the original latent codes, as in the following :

$$\begin{aligned} \widehat{\mathbf{x}}_0^t &= \left(\widehat{\mathbf{x}}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t)\right)/\sqrt{\bar{\alpha}_t} \\ \widehat{\mathbf{x}}_{t+1} &= \sqrt{\bar{\alpha}_{t+1}}\widehat{\mathbf{x}}_0^t + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_\theta(\widehat{\mathbf{x}}_{\tilde{t}}), \end{aligned} \tag{4}$$

where $\widehat{\mathbf{x}}_{\tilde{t}}$ is the estimation of $\widehat{\mathbf{x}}_{t+1}$, which is replaced by $\widehat{\mathbf{x}}_t$ in the practical DDIM inversion because their values are very close to each other for consecutive timesteps. The relevant latent code channel is chosen from the four channels of the generated watermarked image and converted to the frequency domain by FFT. Finally, the CNN watermark detector can predict the watermark's real FFT component, denoted as $\widehat{\mathbf{w}}$, from the resulting images, and the loss function used for training the CNN network considers the difference between the watermark prediction $\widehat{\mathbf{w}}$ and the original watermark code $\mathbf{w}$, as :

$$\mathcal{L} = -\frac{1}{L}\sum_{l=1}^{L}\left[\widehat{\mathbf{w}}_l \log\left(\mathbf{w}_l\right) + \left(1-\widehat{\mathbf{w}}_l\right)\log\left(1-\mathbf{w}_l\right)\right], \tag{5}$$

where L is the number of bits for the watermark $\mathbf{w}$ and predicted watermark $\widehat{\mathbf{w}}$, after considering the distortions. Following the training by considering Eq (5) the watermark robustness to various attacks is significantly increased.

5. EXPERIMENTS

In the following we apply the proposed Localised Frequency Latent Domain Watermarking (LFLDW) for watermarking Denoising Diffusion Implicit Model (DDIM) generated images and test the watermark visibility and robustness to a wide variety of processing algorithms or attacks. Randomly generated watermarks of $L = 8$ bits each, with the watermark strength given by $\gamma = 30$ in Eq. (3), are embedded into 500 different generated images. For generating the images we had used prompt texts randomly selected from the COCO dataset, [19] as well as the dataset Stable-Diffusion-Prompts (SDP) [20], which generates both graphics and photo-realistic images. We train the CNN network, according to the proposed approach from Section 4, using one A40 GPU for about 10 days. The following attacks are considered in the CNN's distortion layer to increase watermark's robustness : a random crop of 70% or 30%; Gaussian blur, with a kernel of size of 3×3 and intensity of 2; for JPEG compression, we use lossy compression with quality factors of 50 or 80; rotation with 90 degrees, -90 degrees; contrast factor 1.5, saturation factor 1.5, and hue factor 0.25; for additive Gaussian noise, we consider a standard deviation of 0.1 or 0.3.

Following the methodology from Sections 2 and 3 we embed watermarks into a variety of generated images using different prompts and some watermarked images are shown in the bottom row of Fig. 2, while on the top row, we show the same generated images without an embedded watermark. In the following, we perform experiments for choosing the diffusion iteration and the range of latent space frequencies, for optimally embedding the watermark. Two watermarked generated images, when considering watermark embedding in different DDIM denoising steps t, are shown on two rows from Fig. 3 indicating the step t on top. When embedding the watermark in a denoising step t near the initial iteration T (denoising iterations are considered in the decreasing order), the changes caused by the watermark are propagated through the following iterations while causing some visible changes when compared to the image generation without watermark, named the "Original Image" in Fig. 3. However, when the watermarking is performed in an iteration close to the final step 0, then some visible changes caused by the watermark appear as some unusual shadows, as observed in the images from the bottom row. Given that no distortions are visible in the third image from each row of Fig. 3, we conclude that the best iteration for the DDIM watermark embedding is a mid-range diffusion iteration, such as $t = 15$.

In Fig. 4, we show two generated images, and their corresponding watermarked counterparts, when considering the frequency ranges of $[r_1, r_2] = \{[4, 13], [19, 28], [31, 40]\}$ for the latent spaces. In the case of the bottom image, we observe the presence of ripples in the images' corners produced, for certain frequency ranges, by watermarks.

In Table 1, we provide a comparative study for the proposed watermarking methodology when considering various attacks. We consider JPEG lossy compression with a quality setting of 50, while a central crop of 90% and rotation with 90 degrees is considered for geometric attacks. We consider

METHODS	Bit accuracy results when considering various image processing algorithms							PSNR ↑	SSIM ↑
	CLEAN	BRIGHTNESS	CROP&RESIZE	BLUR	GAUSSIAN NOISE	JPEG	ROTATION		
STABLESIG (COCO) [12]	100.00	96.28	97.39	90.55	71.78	85.94	50.0	30.0	0.89
AquaLoRA (COCO) [13]	95.79	93.38	91.44	95.85	93.00	94.92	/	29.85	0.92
LFLDW (COCO)	100.0	**98.9**	70.4	**100.0**	**99.7**	**99.8**	**85.0**	**34.3**	0.849
LFLDW (SDP)	100.00	**98.7**	70.1	**99.9**	**99.4**	**99.3**	**91.4**	**34.5**	0.844
	Watermark true positive detection rate								
STABLESIG (COCO) [12]	0.993	0.984	0.988	0.903	0.347	0.833	0.580	/	/
TREE-RING (COCO) [14]	1.00	1.00	0.140	0.968	0.619	0.946	0.810	11.0	0.52
ZoDiac (COCO) [15]	0.992	0.990	/	0.988	0.984	0.978	0.106	29.4	0.92
AquaLoRA (COCO) [13]	0.990	0.941	0.919	0.994	0.958	0.998	/	/	/
LFLDW (COCO)	1.00	0.994	0.280	**1.00**	**0.992**	**0.998**	0.670	/	/
LFLDW (SDP)	1.00	0.984	0.260	**1.00**	**0.998**	**0.998**	0.830	/	/

Table 1: Robustness assessment to various attacks. Bit accuracy is defined as the percentage of matching bits between embedded watermark $\mathbf{w}$ and its prediction $\widehat{\mathbf{w}}$ as in [12, 13].

Fig. 3: Watermarked DDIM generated images when considering embedding at different denoising iterations, with the non-watermarked images shown as the first on each row. The PSNRs for the resulting watermarked generated images for the timesteps $t = \{35, 15, 2\}$, are of 28.6, 34.0, and 34.3.

Gaussian blur with a kernel size of 3×3 and a strength of 4. We also consider additive Gaussian noise with the mean of 0 and variance 0.1, while pixels are normalised to [0, 1]. For changing the brightness of the watermarked image, we consider a brightness strength factor of 2.0.

We also test the watermark robustness when varying the strength of various image attacks. The robustness tests for COCO and SDP datasets for Brightness, Additive Noise, JPEG compression and blurring are provided in the Figures 5a, 5b, 5c and 5d, respectively. The results from Table 1 and Fig. 5 indicate a high level of robustness to various attacks for the proposed LFLDW.

6. CONCLUSION AND DISCUSSION

In this paper, we propose a novel method for inserting watermarks into DDIM generated images, by considering a specific range of frequencies of the latent space, obtained during one of the DDIM denoising steps. The watermarks are detected by means of a convolution neural network (CNN), trained for ensuring watermark robustness to various attacks. The proposed approach is shown to be efficient and robust, while it does not produces significant distortions in the generated watermarked

Fig. 4: Changing the latent space frequency ranges for embedding the watermarks.

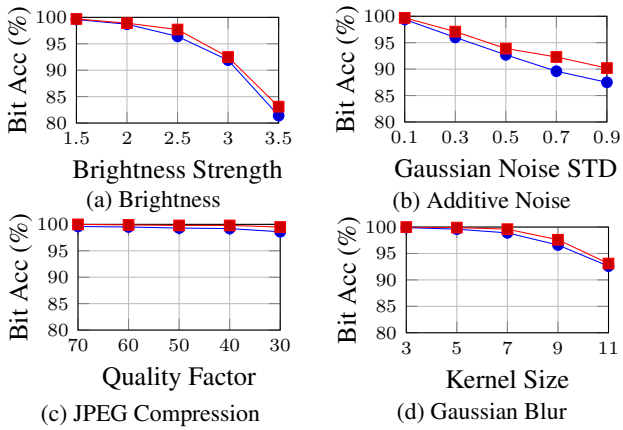

Fig. 5: Robustness tests when increasing the attacks' strength. Red and blue lines are the results for COCO and SDP datasets.

image. The watermarking method is specifically designed to be robust to image compression and smoothing. One limitation is represented by the fact that the embedded watermarks are not robust to image cropping. Only 8 bits were embedded during the experiments and more bits can be embedded in frequencies of multiple latent spaces; however, some distortions may result in this case. Watermarks can be used to identify the origins of the generated images or specifics of its generator training, among others.

7. REFERENCES

[1] Diederik P Kingma and Max Welling, "Auto-encoding variational Bayes," in *International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1312.6114*, 2014.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672–2680.

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 6840–6851.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695.

[5] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Systems (ICLR), arXiv preprint arXiv:2010.02502*, 2021.

[6] Tom Tirer, "Iteratively preconditioned guidance of denoising (diffusion) models for image restoration," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 2465–2469.

[7] Leslie Katz, "AI photo contest winner disqualified because it's real," *Forbes*, 2024, https://www.forbes.com/sites/lesliekatz/2024/06/13/real-photo-wins-ai-photography-contest/.

[8] Anatol Z Tirkel, GA Rankin, RM Van Schyndel, WJ Ho, NRA Mee, and Charles F Osborne, "Electronic watermark," *Digital Image Computing, Technology and Applications (DICTA'93)*, pp. 666–673, 1993.

[9] Adrian G Borş and Ioannis Pitas, "Image watermarking using block site selection and DCT domain constraints," *Optics Express*, vol. 3, no. 12, pp. 512–523, 1998.

[10] Ingemar J. Cox, Matthew L. Miller, and Jeffrey A. Bloom, *Digital Watermarkimg*, Morgan Kaufmann, 2011.

[11] Vojtěch Holub and Jessica Fridrich, "Designing steganographic distortion using directional filters," in *Proc. of the IEEE International Workshop on Information Forensics and Security (WIFS)*, 2012, pp. 234–239.

[12] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon, "The stable signature: Rooting watermarks in latent diffusion models," in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22466–22477.

[13] Weitao Feng, Wenbo Zhou, Jiyan He, Jie Zhang, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu, "AquaLoRA: Toward white-box protection for customized stable diffusion models via watermark lora," in *Proc. of International Conference on Machine Learning (ICML), vol. PMLR 235*, 2024, pp. 13423–13444.

[14] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein, "Tree-rings watermarks: Invisible fingerprints for diffusion images," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023, pp. 58047–58063.

[15] Lijun Zhang, Xiao Liu, Antoni Viros Martin, Cindy Xiong Bearfield, Yuriy Brun, and Hui Guan, "Robust image watermarking using stable diffusion," *arXiv preprint arXiv:2401.04247*, 2024.

[16] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning (ICML), vol. PMLR 37*, 2015, pp. 2256–2265.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. International Conference on Machine Learning (ICML), vol. PMLR 139*, 2021, pp. 8748–8763.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Medical image computing and computer-assisted intervention (MICCAI), vol. LNCS 9351*, 2015, pp. 234–241.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft COCO: Common objects in context," in *Proc. European Conference on Computer Vision (ECCV), vol. LNCS 8693*, 2014, pp. 740–755.

[20] Gustavosta, "Stable diffusion prompts," 2022, https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts.