Geração de Imagens Aéreas com Modelos de Difusão Controláveis: Um Estudo de Caso de *Data Augmentation* com Fine-Tuning de ControlNet para Florestas Contendo Pinus

Thiago Innani Justus
Universidade Tecnológica
Federal do Paraná
Ponta Grossa, Paraná
Email: thiagoinnani@gmail.com

Amanda Gonsalves
Universidade Tecnológica
Federal do Paraná
Ponta Grossa, Paraná
Email: amandagonsalves@alunos.utfpr.edu.br

Gilson Giraldi
Laboratório Nacional
de Computação Científica
Ponta Grossa, Paraná
Email: gilson.giraldi@gmail.com

Rodrigo Minetto
Universidade Tecnológica
Federal do Paraná
Curitiba, Paraná
Email: rodrigo.minetto@gmail.com

Mauren Louise Sguario
Coelho de Andrade
Universidade Tecnológica
Federal do Paraná
Ponta Grossa, Paraná
Email: mlsguario@utfpr.edu.br

***Abstract*—This work presents a method for aerial imagery dataset expansion through the fine-tuning of controllable diffusion models, focusing on the monitoring of pinus forests in Brazil. Starting from the pre-trained Seg2Sat model, which combines Stable Diffusion with ControlNet, this study performs a fine-tuning process on a custom dataset of 29 high-resolution images. It introduces a new semantic class, "pinus", using multi-class segmentation masks and corresponding text prompts. The results, evaluated by Fréchet Inception Distance (FID) and CLIP Score metrics, demonstrate that the model successfully learns to generate the new class according to the spatial control provided by the mask. The study concludes that, even with an extremely limited dataset, fine-tuning ControlNet is a viable and promising approach for data augmentation in remote sensing and forestry monitoring applications.**

***Resumo*—Este trabalho apresenta um método para a expansão de *datasets* de imagens aéreas através do fine-tuning de modelos de difusão controláveis, com foco no monitoramento de florestas de pinus no Brasil. Partindo do modelo pré-treinado Seg2Sat, que combina Stable Diffusion com ControlNet, este estudo realiza um processo de fine-tuning com um dataset customizado de 29 imagens de alta resolução. O trabalho introduz uma nova classe semântica, "pinus", através de máscaras de segmentação multi-classe e *prompts* de texto correspondentes. Os resultados, avaliados pelas métricas FID (Fréchet Inception Distance) e CLIP Score, demonstram que o modelo aprende com sucesso a gerar a nova classe de acordo com o controlo espacial da máscara. O estudo conclui que, mesmo com um dataset extremamente limitado, o fine-tuning do ControlNet é uma abordagem viável e promissora para *data augmentation* em aplicações de sensoriamento remoto e monitoramento florestal.**

I. Introdução

As árvores do gênero Pinus spp. desempenham um papel estratégico na indústria florestal global, sendo cultivadas para a produção de madeira, papel, celulose e resinas [1]. No Brasil, estas plantações possuem importância econômica e ambiental, contribuindo significativamente para a captura de carbono e auxiliando na mitigação dos efeitos das mudanças climáticas. Por outro lado, as árvores de Pinus spp. (*taeda e elliottii*) não são originárias do Brasil e são consideradas árvores exóticas invasoras [2]. Neste caso, as árvores proliferam-se rapidamente sem a necessidade de ação humana. Com isso, interferindo na vegetação nativa, consequentemente, alterando característica da fauna e flora das regiões em que a profileração ocorre de forma descontrolada. O monitoramento preciso e o manejo sustentável destas florestas são, portanto, essenciais para garantir tanto a sua produtividade contínua quanto os seus benefícios ecológicos [3].

A aquisição de imagens para sensoriamento remoto divide-se em duas abordagens principais: imagens de satélite e imagens obtidas por Sistemas de Aeronaves Remotamente Pilotadas (RPAs), ou drones. Imagens de satélite, como as utilizadas por Dias et al. [4] para mapear a invasão de pinus nos Campos Gerais, oferecem a vantagem de uma vasta cobertura e disponibilidade de dados, sendo ideais para monitoramento em larga escala. Contudo, para a tarefa de *data augmentation* focada nas texturas e nos detalhes finos das copas das árvores, a resolução espacial torna-se um fator crítico. Por esta razão, este trabalho optou pela utilização de imagens aéreas de

altíssima resolução, capturando um nível de detalhe que é mais associado a levantamentos com RPAs. Esta escolha garante que o modelo generativo aprenda as características texturais distintivas da espécie-alvo com a maior fidelidade possível, um pré-requisito para a criação de um dataset sintético de alta qualidade.

A inteligência artificial (IA), por meio da análise de imagens aéreas e de satélite, oferece uma solução para este monitoramento, permitindo a identificação automática das copas das árvores, a deteção precoce de pragas, doenças ou áreas de *stress* hídrico [5]. Contudo, o desenvolvimento de modelos de IA precisos para esta tarefa enfrenta o obstáculo de escassez de grandes volumes de dados de imagem devidamente rotulados, um processo que é tradicionalmente caro, demorado e que exige conhecimento especializado [6].

A base para soluções avançadas de IA reside no campo de *deep learning* (aprendizagem profunda), a qual utiliza redes neurais artificiais [7], modelos computacionais inspirados na estrutura do cérebro humano. Estas redes são compostas por camadas de ”neurônios” interligados que aprendem a reconhecer padrões complexos diretamente dos dados. O seu grande poder de generalização, no entanto, depende intrinsecamente do acesso a *datasets* massivos e diversificados, o que reforça a criticidade do problema da escassez de dados.

As técnicas de *data augmentation* tornaram-se uma etapa presente no treinamento de modelos de *deep learning*, principalmente em visão computacional, onde a performance das redes neurais depende diretamente da quantidade e da diversidade dos dados de treino [8]. Estas técnicas visam enriquecer artificialmente um *dataset*, gerando novas amostras sintéticas a partir de dados existentes para mitigar o *overfitting* e melhorar a capacidade de generalização do modelo.

As abordagens de *data augmentation* dividem-se em duas categorias principais. A primeira, baseada em manipulação de imagem, inclui tanto transformações geométricas simples (*data warping*) como rotação e espelhamento, quanto técnicas mais complexas de mistura ou remoção de pixels, como *Cutout*, *CutMix* [9], e métodos específicos para sensoriamento remoto como o ChessMix [10]. Embora eficazes para aumentar a variabilidade dos dados, estes métodos operam sobre os pixels existentes e podem não gerar a diversidade necessária para representar plenamente a complexidade do mundo real [8].

Para superar tais limitações, modelos generativos baseados em *deep learning* são utilizados para criar dados sintéticos completamente novos. As Redes Generativas Adversariais (GANs) são um marco nesta categoria, utilizando um processo competitivo entre uma rede geradora e uma discriminadora para produzir amostras de alta fidelidade [11], [12]. Outras arquiteturas, como os *autoencoders* variacionais (VAEs), também aprendem a gerar novas imagens a partir de uma representação latente dos dados. Apesar do seu poder, estes modelos generativos são notoriamente difíceis de treinar, podendo sofrer de instabilidade, colapso de modos e exigir grandes volumes de dados para convergir [13], [14].

Desta forma, a fim de superar a limitação do volume de dados anotados, especificamente relacionados a imagens aéreas da árvore Pinus, este trabalho explora o uso de modelos generativos de IA. A idéia central é geração de imagens sintéticas para utilização como técnica de *data augmentation*. Para tanto, o estudo utiliza o Stable Diffusion [15], um modelo VAE de geração de imagem em alta resolução, em conjunto com a tecnologia ControlNet [16], para criar imagens aéreas fotorrealistas da espécie Pinus spp. a partir de mapas de segmentação semântica. O objetivo é desenvolver uma metodologia para expandir a base de dados de Pinus [4], possibilitando o treino futuro de modelos de deteção eficientes e escaláveis.

II. Metodologia

A base da abordagem metodológica é o modelo Seg2Sat [17], um projeto de código aberto que integra o Stable Diffusion v2.1 com uma rede neural ControlNet [16]. Esta arquitetura permite uma geração de imagens condicionada tanto por texto quanto por uma entrada espacial.

A. Stable Diffusion e o Processo de Difusão

O Stable Diffusion é um modelo de difusão latente (*Latent Diffusion Model* - LDM) que se destaca por sua eficiência e qualidade na geração de imagens [15]. O seu processo de treino e geração opera em duas etapas principais conforme ilustrado na Figura 1, estas etapas incluem o processo de difusão e o processo de denoising.

O processo de difusão (forward process) ocorre durante o treinamento, o modelo aprende a transformar imagens limpas em ruído puro. Ele pega uma imagem do *dataset* de treino (representado na Figura 1) pela imagem de entrada, a comprime para um espaço latente de menor dimensão através de um codificador, e depois adiciona gradualmente ruído gaussiano a esta representação latente ao longo de vários passos de tempo (*timesteps*) até virar ruído aleatório [15].

O processo de denoising (reverse process) faz a geração de uma nova imagem, para isso é executado o processo inverso da difusão, conforme ilustrado na Figura 1. O modelo inicia com uma matriz de ruído aleatório no espaço latente, denominada z_t. Em paralelo, as informações de condicionamento — que neste trabalho incluem tanto os *prompts* de texto quanto os mapas de segmentação semântica — são processadas por um codificador de condição dedicado. Este codificador transforma as condições de entrada em uma representação vetorial que a rede neural principal pode entender.

O núcleo do processo de geração reside em uma rede neural U-Net, que opera de forma iterativa para remover o ruído da representação latente. A cada passo de tempo (*timestep*), de t para $t-1$, a U-Net recebe como entrada a representação latente ruidosa e a representação de condicionamento. A tarefa da U-Net é prever o ruído presente na imagem latente, que é então subtraído para produzir uma versão mais “limpa” (z_{t-1}). Este condicionamento é injetado nos blocos da U-Net através de um mecanismo de atenção cruzada (*cross-attention*), o que força o modelo a gerar características visuais que são fiéis tanto à descrição textual quanto à estrutura espacial da máscara de

segmentação. Após um número pré-definido de passos, este processo resulta numa representação latente final "limpa" (z). Finalmente, um decodificador converte esta representação do espaço latente de volta para o espaço de pixel, resultando na imagem final em alta resolução [15].

B. ControlNet

Enquanto o *prompt* de texto controla o que gerar, o *ControlNet* controla onde gerar. O *ControlNet* é uma estrutura de rede neural que adiciona um condicionamento espacial extra ao processo de difusão [16]. A sua arquitetura, mostrada na Figura 2, é desenhada para preservar o conhecimento do modelo original:

- Ele cria duas cópias dos blocos de codificação da U-Net do Stable Diffusion: uma cópia "trancada" (*locked*), que mantém os pesos originais intactos, e uma cópia "treinável" (*trainable*).
- Apenas a cópia treinável aprende a interpretar a condição de controle (neste caso, a máscara de segmentação)
- As saídas da cópia treinável são então injetadas na cópia trancada através de camadas especiais de "convolução zero", que no início do treino têm peso zero, garantindo que o modelo original não seja corrompido.

Este método permite fazer um fine-tuning, mesmo com o *dataset* pequeno que está sendo utilizado.

C. Base de dados

A área de estudo para esta pesquisa está situada no sul do Brasil, dentro do Parque Estadual de Vila Velha, localizado no segundo planalto do Paraná, na região conhecida como Campos Gerais, no município de Ponta Grossa, às margens da rodovia BR-376 (Figura 3). A área a ser considerada para análise inclui o entorno imediato do parque e sua zona de amortecimento. A zona de amortecimento cobre uma área de 38.112 ha, variando de 2 km (distância mínima ao perímetro do parque) a 16 km (distância máxima). Com aproximadamente 3.222 hectares, o parque representa uma importante zona de transição entre o bioma Cerrado e a Mata Atlântica, composta principalmente por savana herbácea e floresta tropical mista.

As imagens foram adquiridas utilizando-se um drone DJI Mini 3 Pro equipado com um sensor de câmera CMOS 1/2.3 capaz de capturar vídeo 4K (3840 × 2160 pixels, Ultra HD) e fotos de 48 MP, com arquivos de vídeo salvos no formato MP4. Duas campanhas de voo foram conduzidas sobre o Parque Estadual de Vila Velha em duas altutides diferentes em relação ao solo, 50 e 100 metros. As imagens utilizadas neste trabalho, foram adquiridas no dia 17 de fevereiro de 2024 aproximadamente 10h00 (meados do verão) a 50 metros do nível do solo. Os voos do drone, operados manualmente para pesquisar três ambientes distintos — a plantação de Pinus, a área aberta de Cerrado e a floresta nativa — seguiram trajetórias de voo predeterminadas. Informações detalhadas sobre o conjunto de dados e anotações são fornecidas na Tabela I.

TABLE I
RESUMO DAS CARACTERÍSTICAS DO CONJUNTO DE DADOS, DETALHANDO OS PARÂMETROS DE VOO, O NÚMERO DE VÍDEOS GRAVADOS AO LONGO DE ROTAS DISTINTAS, A CONTAGEM DE IMAGENS TOTALMENTE ANOTADAS E O NÚMERO DE ÁRVORES PINUS SEGMENTADAS MANUALMENTE PARA A ÁREA PESQUISADA DENTRO DO PARQUE ESTADUAL DE VILA VELHA.

Altitude	# videos rotas	# rotulagem imagens	# pinus
50 (metros, aqui utilizadas)	2	29	1,040
100 (metros)	7	33	2,667

D. Fine-Tuning para o Ecossistema Brasileiro

O modelo Seg2Sat original foi treinado sobre o dataset francês FLAIR [19], que contém milhares de imagens aéreas e as suas respetivas máscaras de segmentação com 19 classes (edifícios, água, entre outros), todas com resolução de 512x512 pixels.

Para especializar este modelo para o bioma brasileiro e adicionar a classe de interesse (neste caso *Pinus spp.*), realizou-se um processo de *fine-tuning*. O dataset utilizado consiste em 29 imagens aéreas de alta resolução (3840x2160) adquiridas a 50 metros de altura (descrito na Seção II-C). O processo de preparação de dados se divide em duas etapas, a rotulagem das máscaras e a criação de seus metadados.

Para a rotulagem das máscaras, realiza-se para cada uma das 29 imagens a criação de uma máscara de segmentação multi-classe. A nova classe "pinus" foi definida com uma cor RGB única (roxo, #5f50e7). As outras classes presentes (florestas nativas, estradas, campos) foram rotuladas reutilizando as cores de classes pré-existentes do dataset FLAIR, tais como *coniferous*, *impervious surface* e *herbaceous vegetation*. Para a criação de metadados criou-se um ficheiro chamado `metadata.csv` a fim de associar cada par de imagem/máscara a um *prompt* de texto. Este *prompt* descreve o conteúdo da imagem e as classes que ela contém.

O treino foi realizado com as imagens e máscaras redimensionadas para a resolução nativa do modelo, 512x512. O êxito da abordagem foi avaliado através das métricas FID (Fréchet Inception Distance) e CLIP Score [20].

E. Parâmetros da rede

Para a interação com o modelo e a geração de imagens, desenvolveu-se uma interface web local (*web UI*) utilizando a biblioteca Gradio [21]. A interface permite ao utilizador criar um mapa de segmentação semântica manualmente, pintando sobre um *canvas* com uma paleta de cores pré-definida, onde cada cor corresponde a uma classe. Este processo pode ser entendido como uma forma de *inpainting semântico*, onde o modelo generativo recebe a máscara como uma condição espacial e tem a tarefa de "preencher" as áreas designadas com as texturas e os objetos correspondentes, como a classe "pinus".

Para garantir a consistência e a alta qualidade dos resultados apresentados neste trabalho, os parâmetros de inferência foram padronizados. O *prompt* de texto segue a estrutura de uma descrição base (ex: "Aerial view of a forest with pinus trees")

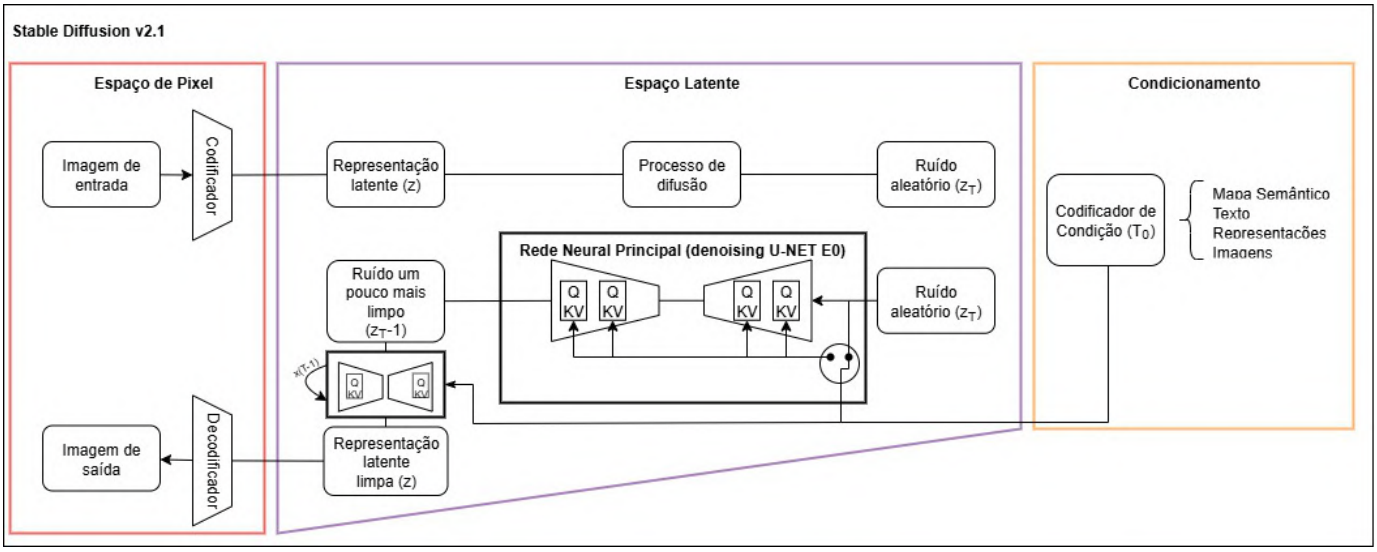


Fig. 1. Diagrama da arquitetura geral do Stable Diffusion, ilustrando o processo de difusão (treino) e o processo de denoising (geração), condicionado por *prompts* e, neste trabalho, por máscaras de segmentação via ControlNet.

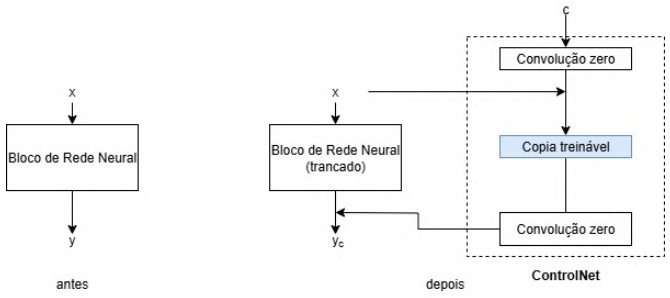


Fig. 2. Arquitetura do ControlNet, mostrando a cópia "trancada" que preserva o modelo original e a cópia "treinável" que aprende a condição espacial [16].

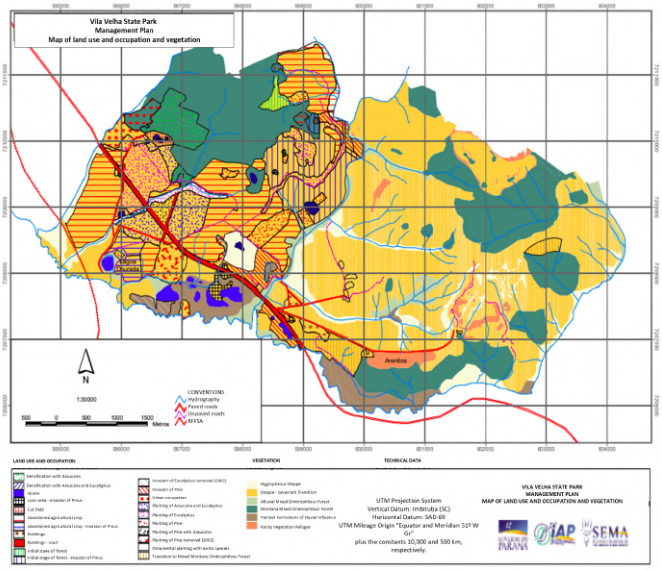


Fig. 3. Plano de Manejo da Vegetação do Parque Estadual de Vila Velha [18]

.

complementada por modificadores de qualidade (ex: "photo-realistic, 4k, high detail"). O número de passos de inferência (*steps*) foi fixado em 30, e a semente (*seed*) para o gerador de números aleatórios foi padronizada em 1024 para garantir a reprodutibilidade dos resultados. Adicionalmente, o fine-tuning do modelo ControlNet foi realizado por 150 épocas (*epochs*) sobre o *dataset* customizado.

F. Avaliação e Métricas

Para avaliar quantitativamente a performance do modelo, este trabalho utiliza duas métricas estabelecidas na avaliação de modelos generativos, a Fréchet Inception Distance (FID) e o CLIP Score. A seleção destas métricas permite uma análise tanto da fidelidade visual das imagens quanto da sua correspondência com a intenção textual [20].

A métrica FID serve para avaliar a qualidade e a diversidade de imagens geradas por modelos como as GANs (Generative Adversarial Networks) e os modelos de difusão. Ela quantifica a "distância" entre a distribuição estatística de um conjunto de imagens reais e um de imagens geradas. Para isso, ambas as coleções de imagens são processadas por uma rede neural pré-treinada (Inception-v3), e a distância de Fréchet é calculada entre as duas distribuições de características extraídas [14]. Um score FID mais baixo indica que as imagens geradas são estatisticamente mais semelhantes às reais, sugerindo maior fotorrealismo e diversidade.

O CLIP Score é uma métrica que avalia a correspondência semântica entre uma imagem e uma descrição de texto (*prompt*). Utilizando o modelo CLIP (Contrastive Language-Image Pre-Training), que foi treinado com milhões de pares de imagem-texto, o *score* mede a similaridade de cosseno entre as representações vetoriais da imagem gerada e do *prompt* de entrada [22]. Um CLIP Score mais alto significa uma maior

compatibilidade entre o conteúdo visual e a intenção textual, indicando que o modelo seguiu corretamente as instruções do *prompt*.

III. Resultados Preliminares

A. Geração a Partir de Máscaras Semânticas

O resultado do fine-tuning é um novo modelo ControlNet capaz de gerar imagens aéreas de Pinus a partir de mapas de segmentação semântica (a técnica de *inpainting* sobre uma máscara). A nova classe "pinus" foi integrada na paleta de classes disponíveis para pintura em uma interface de *debug* customizada, criada com a biblioteca Gradio [21]. A Figura 4 representa a *web UI* criada.

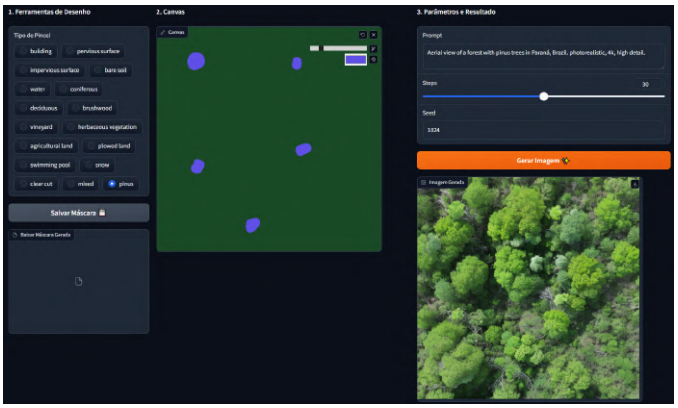
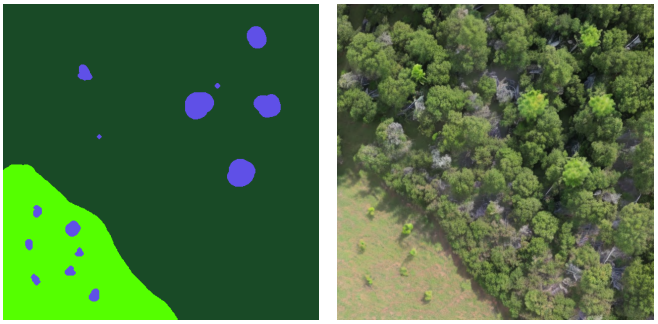


Fig. 4. Interface de debug customizada desenvolvida com a biblioteca Gradio. À esquerda, a paleta de cores interativa com as classes de segmentação, incluindo a nova classe "pinus". À direita, a tela de desenho (canvas) e os parâmetros de inferência como *prompts*, passos e *seeds* para a geração da imagem.

O modelo gera os imagens na resolução de 512x512. A Figura 5 descreve como o modelo aprendeu a renderizar a textura e a forma dos Pinus nas áreas designadas pela cor roxa. O texto do *prompt* utilizado para a geração das imagens foi "Aerial view of a forest with Pinus trees in Paraná, Brazil. photorealistic, 4k, high detail".



(a) Máscara pintada manualmente através da UI.

(b) Imagem gerada pelo modelo a partir da máscara.

Fig. 5. Exemplo de geração de imagem a partir de uma máscara semântica.

B. Análise Quantitativa de Métricas

A performance do modelo foi avaliada quantitativamente em duas condições, usando sempre uma *seed* de 1024 e 30 passos de inferência. A métrica foi baseada nas 29 imagens de treinamento, utilizando as suas máscaras para a geração de novas imagens e comparando elas as imagens originais do *dataset*. O *prompt* utilizado é relativo ao arquivo `metadata.csv`, o qual foi utilizado previamente para o treinamento do modelo. A Tabela II resume os resultados.

TABLE II
Métricas do Modelo

	FID ↓	CLIP Score ↑	Resolução
Primeira Avaliação	280,34	26,61	3840x2160
Segunda Avaliação	182,35	25,54	512x512

Conforme pode ser observado na Tabela II a primeira avaliação obteve o valor 280,34 no FID e 26,61 para o CLIP Score, esta avaliação foi realizada em máscaras e imagens geradas na resolução 3840x2160 pixels, a segunda avaliação alcançou o valor menor de 182,35 no FID e 25,54 para o CLIP Score, com as suas imagens avaliadas contendo a resolução 512x512 pixels. Os resultados promissores observados pelo valor do FID ao gerar imagens na resolução 512x512 pixels, sugerem que o modelo opera de forma eficiente na resolução nativa, evitando artefatos de escala. Entretanto, o valor de 25,54 no CLIP Score revela uma pequena queda de qualidade de correspondência semântica do *prompt* com a imagem, isso é um sintoma esperado da "confusão semântica" criada ao reutilizar cores de classes existentes para representar a vegetação local, o que enfraqueceu a associação do modelo com as palavras do *prompt* original.

A execução da inferência em resoluções significativamente superiores à resolução de treino nativa do modelo francês FLAIR (512x512 pixels) revelou um artefato em relação a escala da imagem gerada. Ao gerar uma imagem com uma máscara de 3840x2160 pixels, o modelo reproduziu corretamente as classes semânticas, incluindo a nova classe "pinus". Contudo, as características visuais, como as copas das árvores, foram renderizadas numa escala muito menor conforme observado na Figura 6 em proporção ao *canvas* total, resultando numa aparência de imagem aérea capturada de uma altitude muito superior. Este efeito ocorre porque o modelo Stable Diffusion aprendeu a distribuição espacial e o tamanho relativo dos objetos a partir de exemplos de 512x512 pixels. Ao ser aplicado a um *canvas* maior, ele replica estas mesmas características de baixa resolução em vez de gerar detalhes de alta frequência adequados à nova escala.

IV. Conclusão e Trabalhos Futuros

Neste trabalho foram investigadas as potencialidades e os desafios do uso de modelos de difusão controláveis, especificamente o ControlNet sobre o Stable Diffusion, para a tarefa de data augmentation em imagens aéreas de sensoriamento remoto. O estudo focou na metodologia de fine-tuning de um

Fig. 6. Exemplo de imagem gerada na resolução 3840x2160. Observa-se o artefato de escala, onde as árvores são renderizadas em um tamanho desproporcionalmente pequeno, resultado da incompatibilidade entre a resolução de treino (512x512) e a de inferência.

modelo pré-treinado (Seg2Sat) com um dataset customizado e de tamanho reduzido para introduzir uma nova classe semântica ("pinus") relevante para o monitoramento florestal no Brasil. Com isso, pode-se verificar que o fine-tuning de um modelo ControlNet pré-treinado é uma abordagem que pode ser utilizada para geração de imagens aéreas de novas classes específicas, mesmo com um dataset limitado. O modelo aprendeu a gerar a classe "pinus" de forma controlada, validado tanto qualitativamente (imagens geradas) como quantitativamente (CLIP Score promissor). A capacidade de gerar imagens a partir de máscaras aleatórias reforça o seu potencial para *data augmentation*.

A geração de imagens de resolução superior a 512x512 serem na mesma escala também abre uma possibilidade de criação de *datasets* emulando outras alturas (100 metros, por exemplo, cuja resolução prejudica a identificação das copas da espécie Pinus).

Para melhorar os scores de FID e CLIP, este trabalho pretende como próximos passos refazer o processo de treinamento do modelo, introduzindo as imagens do *dataset* com uma legenda de cores totalmente nova para as classes do bioma brasileiro (como "mata_nativa" e "campo"), em vez de reutilizar as do dataset FLAIR. A aparência de outras classes, como campos e estradas, também pode ser melhorada com mais exemplos de treino focados nelas. O passo mais impactante seria aumentar o *dataset* de treino com centenas de exemplos da flora local.

REFERENCES

[1] W.-X. Peng, Y. Liu, Y.-Q. Wu, J.-Z. Qiao, and W.-B. Wei, "Determination of biomedicine resource of benzene/ethanol extractives of masson pine(pinus massoniana l.) wood by py-gc/ms," in *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*, 2008, pp. 1241–1243.

[2] M. P. dos Santos, M. J. de Araujo, and P. H. M. da Silva, "Natural establishment of pinus spp. around seed production areas and orchards," *Forest Ecology and Management*, vol. 494, p. 119333, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378112721004217

[3] J. M. Moreira, E. Oliveira, D. Liebsch, and S. Mikich, "Avaliação econômica do cultivo de pinus spp. para um sistema de produção modal no sul do brasil," 10 2015.

[4] A. N. Dias, M. E. G. P. Gianisella, A. D. S. Gonçalves, R. Minetto, and M. L. S. Coelho de Andrade, "Exploring machine learning and remote sensing techniques for mapping pinus invasion beyond crop areas," in *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 3: VISAPP*, INSTICC. SciTePress, 2025, pp. 873–879.

[5] R. Huang, Y. Shi, Y. He, Y. Zheng, G. Xiao, and Z. Liu, "Semantic circle detection and circle-inner segmentation for tree-wise citrus summer shoot management in aerial images," in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 1090–1094.

[6] H.-T. Chen, C.-H. Liu, and W.-J. Tsai, "Data augmentation for cnn-based people detection in aerial images," in *2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2018, pp. 1–6.

[7] M. Hassaballah and A. I. Awad, *Deep Learning in Computer Vision: Principles and Applications*, 03 2020.

[8] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen, "Image data augmentation for deep learning: A survey," 2023. [Online]. Available: https://arxiv.org/abs/2204.08610

[9] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," 2019. [Online]. Available: https://arxiv.org/abs/1905.04899

[10] M. Pereira and J. Santos, "Chessmix: Spatial context data augmentation for remote sensing semantic segmentation," in *Anais da XXXIV Conference on Graphics, Patterns and Images*. Porto Alegre, RS, Brasil: SBC, 2021. [Online]. Available: https://sol.sbc.org.br/index.php/sibgrapi/article/view/19964

[11] A. Biswas, M. A. A. Nasim, A. Imran, A. T. Sejuty, F. Fairooz, S. Puppala, and S. Talukder, "Generative adversarial networks for data augmentation," 2023. [Online]. Available: https://arxiv.org/abs/2306.02019

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[13] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," 2016. [Online]. Available: https://arxiv.org/abs/1511.05644

[14] 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," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2112.10752

[16] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2302.05543

[17] R. Gres, "Seg2sat - segmentation to aerial view using pretrained diffuser models," 2023. [Online]. Available: https://github.com/RubenGres/Seg2Sat

[18] I. Água e Terra Governo do Estado do Paraná. (2004) Plano de manejo parque estadual de vila velha - 2004. [Online]. Available: https://www.iat.pr.gov.br/sites/agua-terra/arquivos_restritos/files/documento/2020-07/pevv_anexos_final.pdf

[19] IGN, "Flair: French land cover from aerospace imagery." [Online]. Available: https://ignf.github.io/FLAIR/

[20] M. Park, J. Yun, S. Choi, and J. Choo, "Learning to generate semantic layouts for higher text-image correspondence in text-to-image synthesis," 08 2023.

[21] A. Abid, A. Abdalla, A. Abid, D. Khan, A. Alfozan, and J. Zou, "Gradio: Hassle-free sharing and testing of ml models in the wild," 2019. [Online]. Available: https://arxiv.org/abs/1906.02569

[22] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," 2022. [Online]. Available: https://arxiv.org/abs/2104.08718