Pontifícia Universidade Católica
do Rio de Janeiro

Marco Antônio Barbosa Teixeira

Gaussian Splating em ambientes não controlados através do uso de Appearance Embedding e Máscaras de Oclusão

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo Programa de Pós–graduação em Informática, do Departamento de Informática da PUC-Rio.

Orientador : Prof. Alberto Barbosa Raposo
Co-orientador: Dr. Vinicius da Silva

Rio de Janeiro
Abril de 2025

Pontifícia Universidade Católica
do Rio de Janeiro

Marco Antônio Barbosa Teixeira

Gaussian Splating em ambientes não controlados através do uso de Appearance Embedding e Máscaras de Oclusão

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo Programa de Pós–graduação em Informática da PUC-Rio. Aprovada pela Comissão Examinadora abaixo:

Prof. Alberto Barbosa Raposo
Orientador
Departamento de Informática – PUC-Rio

Dr. Vinicius da Silva
Co-orientador
Departamento de Informática – PUC-Rio

Prof. Tiago Novello de Brito
IMPA

Dr. Guilherme Gonçalves Schardong
UC

Rio de Janeiro, 30 de Abril de 2025

Marco Antônio Barbosa Teixeira

Graduou-se no curso de Engenharia de Computação pela Pontifícia Universidade Católica do Rio de Janeiro em 2013. Atualmente trabalha no Instituto Tecgraf, e já trabalhou nas áreas de Modelagem de Sistemas Offshore, desenvolvimento, e Visualização nas áreas de Petróleo e Gás.

Ficha Catalográfica

Teixeira, Marco Antônio Barbosa

Gaussian Splating em ambientes não controlados através do uso de Appearance Embedding e Máscaras de Oclusão / Marco Antônio Barbosa Teixeira; orientador: Alberto Barbosa Raposo; co-orientador: Vinicius da Silva. – 2025.

62 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2025.

Inclui bibliografia

1. Informática – Teses. 2. Renderização neural. 3. Gaussian Splatting. 4. Embeddings de Aparência. 5. Máscaras de Oclusão. 6. Computação Gráfica. 7. Síntese de Novas Visualizações. I. Raposo, Alberto Barbosa. II. da Silva, Vinicius. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 004

To those who taught me how to begin.
To the one who walked beside me with love and quiet strength.
And to myself,
for continuing even when the path seemed to fade.
This work is made of many steps.
Each one, a gesture of gratitude.

Agradecimentos

Concluir esta dissertação foi muito mais do que um exercício acadêmico. Foi uma jornada de crescimento, persistência e apoio coletivo, e nada disso teria sido possível sozinho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Agradeço ao Prof. Alberto Raposo e ao Prof. Vinicius Silva pela orientação generosa, pelos conselhos precisos e pela atenção cuidadosa que me acompanharam em cada etapa deste trabalho.

Ao Departamento de Informática da PUC-Rio e ao Tecgraf, agradeço pelo ambiente estimulante e pelas oportunidades ao longo do caminho.

Aos colegas e amigos, obrigado pelas conversas, trocas de ideias e apoio. Em especial ao Vitor Pinheiro, pelos papos loucos e sempre inspiradores.

Aos colegas de trabalho — Felipe Carvalho, Renato Cherullo, Aldo, Alessandra e toda a equipe do Plan360 e VR — meu reconhecimento pela paciência e apoio constantes.

Aos meus pais e irmãos, obrigado pelo exemplo, pelo carinho e pela força que sempre me inspiraram.

À Alice, minha companheira, obrigado pela presença, pelas palavras certas e por me lembrar de seguir mesmo nos momentos difíceis.

E a mim mesmo, pela coragem de continuar.

A todos, minha sincera gratidão.

Resumo

Teixeira, Marco Antônio Barbosa; Raposo, Alberto Barbosa; da Silva, Vinicius. **Gaussian Splating em ambientes não controlados através do uso de Appearance Embedding e Máscaras de Oclusão**. Rio de Janeiro, 2025. 62p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

O campo da renderização neural tem experimentado avanços significativos nos últimos anos, especialmente com o desenvolvimento de técnicas como Neural Radiance Fields (NeRF) e Gaussian Splatting. Essas abordagens permitem a síntese de novas visualizações de cenas com alta fidelidade, sendo fundamentais para aplicações em realidade virtual, realidade aumentada e computação gráfica. Neste trabalho, propomos uma metodologia que combina embeddings de aparência com Gaussian Splatting para aprimorar a renderização de cenas capturadas em ambientes não controlados (in-the-wild). Nossa abordagem transforma índices de imagens em vetores de alta dimensão dentro de um espaço latente, permitindo a modulação dinâmica das características visuais. Além disso, utilizamos máscaras de oclusão para separar elementos estáticos e transitórios, melhorando a consistência visual e a fidelidade da reconstrução. Os resultados obtidos evidenciam o potencial da técnica na resolução de desafios encontrados em ambientes in-the-wild, permitindo uma melhor adaptação visual e ajustes dinâmicos para diferentes condições de iluminação e oclusão. Isso amplia o alcance de aplicações em computação gráfica e realidade aumentada, além de fortalecer os benefícios da técnica de Gaussian Splatting, elevando ainda mais sua qualidade e expandindo seus potenciais para aplicações futuras.

Palavras-chave

Renderização neural; Gaussian Splatting; Embeddings de Aparência; Máscaras de Oclusão; Computação Gráfica; Síntese de Novas Visualizações.

Abstract

Teixeira, Marco Antônio Barbosa; Raposo, Alberto Barbosa (Advisor); da Silva, Vinicius (Co-Advisor). **Gaussian Splatting In-the-Wild through the use of Appearance Embedding and Occlusion Masks**. Rio de Janeiro, 2025. 62p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The field of neural rendering has seen significant advancements in recent years, particularly with the development of techniques such as Neural Radiance Fields (NeRF) and Gaussian Splatting. These approaches enable the synthesis of novel views with high fidelity, making them essential for applications in virtual reality, augmented reality, and computer graphics. In this work, we propose a methodology that combines appearance embeddings with Gaussian Splatting to enhance the rendering of scenes captured in uncontrolled environments (in-the- wild). Our approach transforms image indices into high-dimensional vectors within a latent space, allowing for dynamic modulation of visual characteristics. Additionally, we use occlusion masks to separate static and transient elements, improving visual consistency and reconstruction fidelity. The results obtained demonstrate the potential of this technique in addressing challenges found in in- the-wild environments, allowing for better visual adaptation and dynamic adjustments under different lighting and occlusion conditions. This broadens the scope of applications in computer graphics and augmented reality while strengthening the advantages of Gaussian Splatting, further improving its quality and expanding its potential for future applications.

Keywords

Neural Rendering; Gaussian Splatting; Appearance Embeddings; Occlusion Masks; Computer Graphics; Novel View Synthesis.

Sumário

Lista de figuras

Figura 5.17 Comparação entre diferentes configurações do método proposto aplicadas a uma cena com movimento. À esquerda, modelos sem uso de appearance embeddings; à direita, com uso de aparência. Na linha superior, sem aplicação de máscara de oclusão; na inferior, com oclusão. Nota-se que a combinação de aparência e oclusão (canto inferior direito) produz uma renderização mais coerente, com menor distorção visual e melhor definição estrutural. 58

Lista de tabelas

Lista de algoritmos

Lista de Códigos

Lista de Abreviaturas

ADI – Análise Digital de Imagens

BIF – *Banded Iron Formation*

A tecnologia é maravilhosa. Mas precisa de consciência para ser usada com sabedoria.

Thich Nhat Hanh , *Mestre Budista.*

1
Introdução

1.1
Motivação

A reconstrução tridimensional precisa de ambientes industriais é uma demanda crescente em setores como petróleo, gás e manufatura, onde a documentação visual e a geração de gêmeos digitais (digital twins) têm impacto direto na eficiência operacional, segurança e tomada de decisão. Projetos como o **Plan360** da **Petrobras** ilustram essa necessidade ao adotar captura fotográfica de ambientes para fins de inspeção, análise e navegação virtual.

No entanto, a aplicação de técnicas avançadas de renderização neural em contextos industriais enfrenta desafios significativos: grande variação de iluminação ao longo do dia, presença de elementos transitórios (pessoas, ferramentas, barreiras móveis), além da alta complexidade estrutural. Ambientes industriais são caracterizados por constante movimentação, mudanças rápidas na cena (como adição ou remoção de equipamentos) e necessidade de agilidade nos registros visuais, por questões de segurança e produtividade.

Figura 1.1: Exemplos de ambientes industriais com variação de iluminação, presença de pessoas e equipamentos, comuns no contexto da captura para inspeção e documentação visual.

A captura de imagens nesses cenários é frequentemente feita com câmeras acopladas a capacetes ou suportes fixos como tripés, o que implica na presença do operador ou dos dispositivos nas imagens capturadas. Isso representa uma

dificuldade adicional para técnicas tradicionais de reconstrução, que assumem cenas estáticas e livres de oclusões ou artefatos inesperados.

Nesse contexto, torna-se fundamental o desenvolvimento de abordagens capazes de lidar com:

- Perturbações fotométricas e de aparência entre capturas;
- Oclusões causadas por elementos móveis;
- Presença de pessoas ou tripés visíveis nas imagens;
- Necessidade de reconstrução com poucos registros e em diferentes horários;
- Adaptação a diferentes tipos de cenas sem ajuste manual.

Essas características reforçam a importância de técnicas como o **Gaussian Splatting**, capazes de realizar renderização em tempo real, e motivam o desenvolvimento de um pipeline que incorpore *appearance embeddings* e *máscaras de oclusão*, como proposto neste trabalho.

Figura 1.2: Exemplos de ambientes industriais com variação de iluminação, presença de pessoas e equipamentos, comuns no contexto da captura para inspeção e documentação visual.

1.2 Objetivo

O principal objetivo deste trabalho é tornar o método de Gaussian Splatting mais robusto e adequado para aplicações em ambientes não controlados, nos quais são comuns variações de aparência e a presença de elementos transitórios. A proposta é desenvolver um pipeline que combine mecanismos de embeddings de aparência e máscaras de oclusão, a fim de melhorar a fidelidade visual e a consistência estrutural da reconstrução tridimensional.

Especificamente, buscamos desenvolver um mecanismo de codificação latente capaz de representar variações fotométricas presentes nas imagens de entrada, permitindo que a aparência da cena seja modulada de forma explícita durante o processo de renderização. Além disso, integramos máscaras de oclusão ao pipeline de reconstrução, permitindo que elementos inconsistentes — como pessoas em movimento, tripés ou objetos transitórios — sejam corretamente identificados e descartados.

Para validar a eficácia do modelo proposto, realizamos experimentos com cenas sintéticas e reais, variando as condições de iluminação e presença de oclusões. O desempenho da técnica foi comparado com abordagens convencionais, avaliando-se sua capacidade de preservar estrutura geométrica, aparência visual e estabilidade perceptual.

1.3 Estrutura

Esta dissertação está organizada da seguinte forma vista a seguir. O Capítulo 2 apresenta uma revisão de pesquisas relacionadas com o corrente trabalho. O Capítulo 3 apresenta a fundamentação teórica necessária para o desenvolvimento do trabalho, os conceitos técnicos de Gaussian Splatting, embeddings de aparência, o uso de máscaras de oclusão e seus respectivos contextos de aplicação. O Capítulo 4 descreve em detalhes a metodologia proposta, com destaque para a arquitetura do modelo, o processo de codificação de aparência, os mecanismos de tratamento de oclusões, bem como as adaptações realizadas na implementação original e a função de perda utilizada. No Capítulo 5, são descritos os experimentos realizados, abrangendo os dados utilizados, o processo de treinamento, os resultados obtidos e uma análise comparativa detalhada entre diferentes variações do método. Por fim, o Capítulo 6 reúne as conclusões do trabalho, apresentando um resumo das contribuições alcançadas, as limitações encontradas e possíveis direções para pesquisas futuras.

2
Trabalhos Relacionados

2.1
Reconstrução de Cenas In the Wild

A reconstrução de cenas em ambientes não controlados (in-the-wild) é um desafio considerável devido à presença de iluminação variável, texturas complexas e oclusões dinâmicas. Trabalhos como NeRF in the Wild (MARTIN-BRUALLA et al., 2021) introduzem mecanismos para modelagem de variações fotométricas, permitindo adaptação a mudanças ambientais. No entanto, esses métodos são computacionalmente intensivos, requerendo múltiplas estimativas de mapas latentes para diferentes condições de captura.

Outras abordagens, como Mega-NeRF (TURKI; RAMANAN; SATYANARAYANAN, 2022), exploram a aplicação de modelos de grande escala para síntese de novas vistas em cenas complexas, melhorando a representação de detalhes em larga escala. No entanto, a necessidade de armazenamento e inferência distribuída torna essas técnicas desafiadoras para aplicações interativas.

2.2
Fototurismo e Modelagem Urbana

O Phototourism (SNAVELY; SEITZ; SZELISKI, 2006), um dos primeiros estudos amplamente utilizados na reconstrução 3D, foi um dos primeiros trabalhos a demonstrar reconstrução de larga escala a partir de fotografias capturadas de forma não estruturada. Esse método mostrou eficiência na modelagem de marcos urbanos, mas sua dependência de técnicas tradicionais como SfM limitou a qualidade da reconstrução em superfícies detalhadas e regiões sujeitas a variações atmosféricas.

Avanços recentes, como Neural Points (XU et al., 2022), introduziram representações híbridas que combinam nuvens de pontos com aprendizado profundo para melhorar a eficiência computacional e a qualidade da reconstrução de cenas urbanas.

2.3
Gaussian Splatting

Os trabalhos analisados buscam superar limitações impostas por métodos tradicionais, proporcionando maior flexibilidade na modelagem de cenas tridimensionais e tornando a técnica aplicável a contextos mais diversos. En-

tre as abordagens recentes, destacam-se aquelas voltadas à generalização do método para ambientes dinâmicos, otimização de desempenho para grandes volumes de dados e incorporação de informações temporais para captura de transformações visuais ao longo do tempo.

2.3.1 WildGaussians: 3D Gaussian Splatting in the Wild

Entre os trabalhos mais recentes que tratam da aplicação do *Gaussian Splatting* em ambientes não controlados, destaca-se o WildGaussians (KULHANEK et al., 2024). Essa abordagem busca lidar com a variabilidade fotométrica e estrutural presente em conjuntos de imagens capturados em condições reais, por meio de modificações no processo de otimização das gaussianas, incluindo regularizações e filtragens para atenuar ruídos e inconsistências. Apesar de compartilhar objetivos semelhantes, o presente trabalho se diferencia ao incorporar mecanismos explícitos de controle de aparência e oclusão, baseados em embeddings e segmentações, respectivamente. Enquanto o WildGaussians foca predominantemente em ajustes geométricos e no refinamento direto da representação 3D, nossa proposta explora o uso de informações extras, como vetores de aparência e máscaras semânticas, para adaptar a renderização de forma mais sensível às variações da cena.

2.3.2 WE-GS: An In-the-wild Efficient 3D Gaussian Representation for Unconstrained Photo Collections (2024)

O WE-GS (2024) foca na otimização do Gaussian Splatting para conjuntos de dados não estruturados, permitindo reconstruções a partir de coleções de fotos capturadas de forma não supervisionada. Diferente de métodos convencionais, essa abordagem incorpora aprendizado adaptativo para mitigar inconsistências causadas por iluminação heterogênea e variação de perspectiva. Além disso, o modelo emprega técnicas de compactação que aprimoram a eficiência computacional, tornando-o mais viável para aplicações em larga escala, como fotogrametria e reconstrução de cenas a partir de imagens coletadas em ambientes não controlados. Essa capacidade de adaptação sem necessidade de calibração manual rigorosa consolida o WE-GS como um avanço relevante no uso do Gaussian Splatting para síntese de vistas a partir de dados dispersos (WANG; WANG; QI, 2024).

2.3.3 Outros Trabalhos

Podemos também citar pesquisas complementares que expandem as capacidades do Gaussian Splatting em diferentes domínios. O Gaussian Time Machine (2024) introduz uma abordagem para modelagem de variações temporais, permitindo capturar transformações de iluminação e aparência ao longo do tempo. A técnica propõe um espaço latente probabilístico, no qual transições visuais são modeladas dinamicamente, viabilizando aplicações como reconstrução histórica de ambientes e síntese temporalmente coerente para efeitos visuais (SHEN et al., 2024).

Outra contribuição importante é o SpotlessSplats (2024), que foca na melhoria da qualidade da reconstrução tridimensional por meio da remoção automatizada de artefatos e ruídos provenientes dos dados de entrada. Com a incorporação de aprendizado de máquina para filtragem de elementos distratores, a técnica melhora a fidelidade visual da cena reconstruída, reduzindo interferências sem comprometer a integridade estrutural do modelo gerado (SABOUR et al., 2024).

Outras otimizações significativas incluem o Lighting Every Darkness with 3DGS (2024), que aprimora a renderização em High Dynamic Range (HDR), garantindo melhor captura e reprodução de iluminação em cenas de alta complexidade. Por sua vez, o Superpoint Gaussian Splatting (2024) introduz um refinamento baseado em superpoints, garantindo maior preservação de detalhes e texturas finas em reconstruções dinâmicas, um avanço crucial para aplicações em animação, realidade virtual e reconstrução interativa (JIN et al., 2024).

3
Fundamentação Teórica

3.1
Contextualização

3.1.1
Novel View Synthesis

A síntese de novas vistas tem sido um tópico central na visão computacional, impulsionado pela necessidade de reconstrução tridimensional realista a partir de um número limitado de imagens. Esse campo se expande à medida que novas aplicações exigem modelagem tridimensional precisa para realidade aumentada, reconstrução 3D e síntese de imagens fotorrealistas. Além disso, aplicações emergentes, como digitalização de patrimônio histórico, criação de ambientes imersivos para treinamento e entretenimento, e geração de representações urbanas detalhadas, demandam soluções cada vez mais avançadas e adaptáveis a diferentes cenários. Trabalhos como o Google Earth 3D Reconstruction têm explorado abordagens híbridas que combinam modelagem geométrica e aprendizado profundo para criar representações urbanas detalhadas a partir de grandes volumes de imagens aéreas. Na preservação do patrimônio histórico, projetos como o CyArk utilizam síntese de novas vistas para capturar e reconstruir digitalmente monumentos ameaçados, permitindo sua conservação virtual e reconstituição fidedigna. No domínio de realidade imersiva, pesquisas recentes em VR Scene Reconstruction exploram o uso de redes neurais para reconstrução interativa de ambientes, facilitando a criação de espaços virtuais realistas para aplicações em simulação e entretenimento (LI et al., 2023).

Métodos clássicos, como Structure from Motion (SfM) (SNAVELY; SEITZ; SZELISKI, 2006) e Multi-View Stereo (MVS) (GOESELE et al., 2007), forneceram as bases para a recuperação da geometria de cenas, utilizando correspondências entre imagens para inferir profundidade e estrutura. No entanto, essas abordagens são limitadas ao lidar com superfícies complexas, condições de iluminação variáveis e oclusões, o que compromete sua eficácia em cenários não controlados. Para contornar essas dificuldades, métodos híbridos que combinam elementos de aprendizado profundo com técnicas geométricas vêm sendo desenvolvidos, explorando o melhor dos dois mundos para aprimorar a acurácia da reconstrução.

3.1.2 Aprendizado de Máquina em Visão Computacional

A adoção de redes neurais profundas revolucionou a visão computacional, permitindo representações contínuas e mais expressivas de cenas tridimensionais. Os Neural Radiance Fields (NeRF) (MILDENHALL et al., 2020) introduziram um novo paradigma ao modelar a cena como uma função contínua de densidade e cor, aprendida diretamente a partir de imagens 2D. Essa abordagem permitiu reconstruções com alta fidelidade de detalhes, mas a um custo computacional elevado, exigindo grande poder de processamento para treinamento e inferência. Além disso, pesquisas recentes propõem variantes do NeRF, como Mip-NeRF (BARRON et al., 2021), NeRF-W (MARTIN-BRUALLA et al., 2021), e Instant NGP (MÜLLER et al., 2022), para lidar com questões como variações fotométricas e eficiência na inferência em tempo real.

Modelos supervisionados e auto-supervisionados vêm sendo empregados para otimizar representações tridimensionais, minimizando a necessidade de grandes conjuntos de dados anotados. Além disso, a integração de arquiteturas como transformers e redes convolucionais tem aprimorado a eficiência da reconstrução de superfícies e da extração de características semânticas. Essas abordagens vêm sendo exploradas principalmente para permitir que modelos generalizem para novos domínios e se adaptem a diferentes fontes de dados.

3.1.3 Gaussian Splatting

O avanço das técnicas de reconstrução tridimensional tem impulsionado novas formas de representar ambientes reais com alta fidelidade visual e geométrica. Técnicas anteriores, como o EWA Splatting(ZWICKER et al., 2002) e a representação por surfels(PFISTER et al., 2000), estabeleceram as bases para métodos mais recentes ao introduzirem o uso de primitivas pontuais rasterizáveis diretamente sobre a imagem. Entre esses métodos contemporâneos, o Gaussian Splatting (KERBL et al., 2023) destaca-se por sua eficiência e capacidade de renderização interativa. Diferente de abordagens volumétricas como o NeRF, o método utiliza uma representação discreta, baseada em primitivas gaussianas 3D, que são diretamente rasterizadas sobre o espaço da imagem. Essa estrutura permite gerar novas imagens em tempo real, com qualidade competitiva e baixo custo computacional.

Graças a essas características, o Gaussian Splatting tem sido explorado em uma gama crescente de aplicações. Em realidade virtual e aumentada, viabiliza a reconstrução imersiva de espaços reais para experiências interativas. Em contextos industriais, é empregado na criação de gêmeos digitais para ins-

peção, manutenção e visualização de ambientes complexos. No campo da documentação e preservação de patrimônio histórico, permite capturar e explorar digitalmente monumentos e sítios culturais com riqueza de detalhes. (CHEN; WANG, 2024)

Outros cenários relevantes incluem a análise urbana com base em imagens aéreas ou registros fotográficos irregulares, produção de conteúdo tridimensional a partir de vídeos, visualização científica de dados espaciais, navegação autônoma em robótica e geração de ambientes realistas para jogos e simulações. A versatilidade do método, aliada à sua eficiência, faz com que ele seja cada vez mais considerado como base para pipelines de reconstrução 3D em tempo real. (CHEN; WANG, 2024; ZHU et al., 2024; ALI et al., 2025)

Nos capítulos seguintes, serão apresentados os fundamentos do Gaussian Splatting, sua formulação matemática e funcionamento, bem como exemplos e discussões sobre sua aplicação prática.

3.1.4 Modelos de Segmentação e Classificação

Modelos de segmentação desempenham um papel crítico na melhoria da reconstrução tridimensional. Algoritmos como U-Net (RONNEBERGER; FISCHER; BROX, 2015) e DeepLab (CHEN et al., 2017) proporcionam segmentação precisa de contornos e superfícies, melhorando a definição estrutural das cenas sintetizadas. Essas abordagens são frequentemente utilizadas para eliminar ruídos e refinar detalhes geométricos em representações reconstruídas. Adicionalmente, arquiteturas baseadas em aprendizado contrastivo têm sido exploradas para melhorar a segmentação de objetos e superfícies em condições adversas.

Além disso, Vision Transformers (ViTs) têm demonstrado desempenho superior na classificação e análise de representações visuais de larga escala. A combinação de técnicas de segmentação e classificação tem sido investigada para aprimorar a reconstrução de elementos parcialmente ocluídos, aumentando a coerência semântica das reconstruções. Em aplicações industriais e biomédicas, a segmentação refinada tem permitido um avanço significativo na precisão da modelagem 3D de estruturas anatômicas e superfícies de engenharia.

3.1.5 Foundation Models e Transfer Learning

Os Foundation Models têm ganhado destaque na visão computacional, permitindo generalização eficiente por meio de treinamento prévio em grandes

volumes de dados. Modelos como CLIP (RADFORD et al., 2021) e DINO (CARON et al., 2021) demonstram alta adaptabilidade a diversas tarefas visuais, incluindo reconstrução e síntese de novas vistas. A integração dessas abordagens ao Gaussian Splatting pode otimizar a inferência e minimizar a necessidade de treinamento extensivo, viabilizando reconstruções mais eficientes. Além disso, o uso de transfer learning em modelos de reconstrução 3D permite a adaptação a novos domínios sem a necessidade de grandes conjuntos de dados de treinamento.

Um exemplo notável no domínio de segmentação é o Segment Anything Model (SAM), proposto por Kirillov et al. (KIRILLOV et al., 2023). O SAM é um modelo fundacional treinado com bilhões de máscaras sobre uma grande variedade de imagens e contextos, capaz de generalizar a novas cenas e objetos com alta robustez. Sua arquitetura combina um encoder visual com um decodificador flexível condicionado por prompts (pontos, caixas ou máscaras), permitindo segmentações precisas mesmo em contextos não supervisionados. Devido à sua versatilidade, o SAM tem sido amplamente utilizado como ponto de partida para tarefas de segmentação em pipelines específicos, incluindo cenários com oclusões parciais ou objetos transitórios. No contexto deste trabalho, o Segment Anything é empregado como etapa inicial no tratamento de oclusões, auxiliando na separação de elementos persistentes e dinâmicos da cena.

3.2 Conceitos Fundamentais

3.2.1 Gaussian Splatting

O Gaussian Splatting (KERBL et al., 2023) é uma técnica recente de renderização neural que representa a geometria da cena tridimensional por meio de primitivas gaussianas 3D. Cada ponto da cena é modelado como uma gaussiana anisotrópica, com atributos como posição, orientação espacial (covariância), cor, opacidade e escala. Essa representação explícita e diferenciável permite renderizações altamente eficientes, com fidelidade visual comparável a métodos volumétricos como o NeRF (MILDENHALL et al., 2020), porém com desempenho substancialmente superior.

A renderização é feita diretamente por um processo de rasterização, no qual as gaussianas são projetadas no espaço da imagem e compostas por meio de uma etapa de *splatting*. Conforme ilustrado na Figura 3.1, enquanto o NeRF utiliza amostragens ao longo de raios e redes neurais para estimar cor

(c) e densidade (α), o Gaussian Splatting projeta as primitivas diretamente, considerando suas formas elípticas e parâmetros radiométricos.

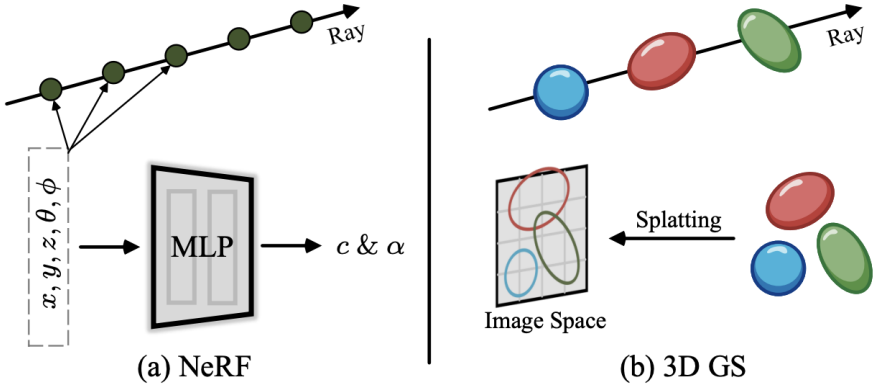

Figura 3.1: Comparação entre renderização por NeRF (esquerda) e Gaussian Splatting (direita). Adaptado de "A Survey on 3D Gaussian Splatting"(2024).

O splatting consiste na projeção de cada gaussiana sobre o plano da câmera, gerando elipses que se sobrepõem nos pixels da imagem. Cada splat contribui com uma cor ponderada pela opacidade, e a combinação dessas contribuições resulta na imagem final. Esse processo é implementado com suporte a ordenação por profundidade (*depth peeling*), garantindo visibilidade correta entre os splats.

A contribuição espacial de cada gaussiana no espaço da imagem é determinada pela sua projeção e forma elíptica, controladas pela matriz de covariância Σ. Essa influência é modelada por uma função gaussiana 2D projetada, centrada na posição μ_i com covariância Σ_i, cuja densidade define o peso w_i de cada splat em um pixel p. Essa formulação segue a abordagem proposta por Kerbl et al. (KERBL et al., 2023), sendo essencial para calcular a contribuição de cada ponto na composição da imagem final, como descrito na Equação 3-1:

$$w_i(p) = \exp\left(-\frac{1}{2}(p-\mu_i)^T\Sigma_i^{-1}(p-\mu_i)\right) \tag{3-1}$$

Esse peso define a influência de cada splat no pixel e é usado na composição da imagem. Como a função gaussiana decai exponencialmente à medida que a distância em relação ao centro μ_i aumenta, a maior contribuição ocorre nas regiões próximas ao centro da projeção, enquanto pontos mais afastados têm peso reduzido. Esse comportamento favorece uma interpolação suave entre splats próximos e reduz interferências de elementos distantes no plano da imagem.

Durante a projeção da gaussiana 3D para a imagem 2D, a matriz de covariância sofre uma transformação para refletir corretamente a distorção

pela perspectiva da câmera. Essa transformação é dada por:

$$\Sigma' = JW\Sigma W^T J^T \tag{3-2}$$

onde:

- Σ' é a nova matriz de covariância no espaço da imagem,
- W representa a transformação de rotação e escala da gaussiana no espaço 3D,
- J é o Jacobiano da transformação de projeção (do 3D para o 2D).

Essa operação garante que a forma projetada da gaussiana preserve corretamente sua orientação e escala, permitindo a rasterização precisa dos splats como elipses no plano da imagem.

O cálculo da cor final observada em um pixel pode ser descrito por uma fórmula de alpha blending, que leva em conta a contribuição acumulada das gaussianas ordenadas por profundidade:

$$C = \sum_{i=1}^{N} T_i \left(1 - \exp(-\sigma_i \delta_i)\right) c_i, \quad \text{com} \quad T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) \tag{3-3}$$

Essa expressão pode ser reescrita de forma equivalente como:

$$C = \sum_{i=1}^{N} T_i \alpha_i c_i, \tag{3-4}$$

$$\begin{aligned} \text{com} \quad \alpha_i &= 1 - \exp(-\sigma_i \delta_i), \\ T_i &= \prod_{j=1}^{i-1} (1 - \alpha_j) \end{aligned}$$

ou ainda:

$$C = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{3-5}$$

Essas equações refletem o princípio de composição das cores por atenuação ao longo do raio, considerando a transparência relativa de cada splat. O termo α_i representa a opacidade efetiva da gaussiana i, enquanto o fator T_i quantifica a transmissão acumulada dos splats anteriores.

Durante o treinamento, os parâmetros das gaussianas (posição, covariância, cor, opacidade) são ajustados para minimizar a diferença entre a imagem renderizada e a imagem de referência. A função de perda proposta por Kerbl et al. (KERBL et al., 2023) combina o erro absoluto por pixel com a diferença estrutural perceptual (DSSIM), resultando na seguinte formulação:

$$\mathcal{L}_G = (1 - \lambda) \cdot \mathcal{L}_1 + \lambda \cdot \mathcal{L}_{DSSIM} \tag{3-6}$$

O termo $\mathcal{L}_1$ mede o erro fotométrico entre pixels, enquanto o $\mathcal{L}_{DSSIM}$ avalia diferenças estruturais baseadas na percepção visual. O hiperparâmetro λ controla o peso entre os dois termos, e sua escolha influencia diretamente o equilíbrio entre fidelidade de cor e preservação de estrutura.

A principal inovação do Gaussian Splatting está na representação discreta e eficiente da geometria e na rasterização direta dos splats, o que permite renderizações interativas com qualidade comparável a métodos muito mais custosos computacionalmente. Este modelo serve como base para o presente trabalho, que propõe extensões voltadas à robustez em ambientes não controlados.

3.2.2 Embeddings de Aparência

Embeddings de aparência são representações vetoriais compactas que capturam características visuais de uma imagem, como cor, iluminação, saturação e estilo. Esses embeddings são aprendidos por redes neurais durante o treinamento e podem ser usados para condicionar a saída de modelos de síntese ou reconstrução. Em nosso caso, o embedding permite adaptar o Gaussian Splatting à aparência de cada imagem individual, reduzindo variações cromáticas e melhorando a consistência perceptual entre diferentes ângulos de visão.

3.2.3 Espaço Latente de Aparência

O espaço latente é um conceito fundamental em modelos neurais modernos, representando um domínio de menor dimensionalidade onde informações complexas são codificadas de forma densa e estruturada. No contexto da aparência visual, esse espaço organiza atributos como iluminação, saturação, tonalidade e estilo de forma contínua e manipulável.

Durante o treinamento do modelo, cada imagem de entrada é associada a um vetor latente z_i, que é otimizado para reduzir a diferença entre a imagem sintetizada e a imagem real correspondente. O conjunto desses vetores forma um espaço latente que pode ser explorado para diversas finalidades: interpolação entre estilos, clustering de aparências semelhantes ou análise de variações intra-cena.

No trabalho proposto, esse espaço atua como um mecanismo de adaptação da renderização à aparência de cada imagem individual. Ele permite que o modelo aprenda a ajustar cores e iluminação de forma diferenciada, man-

tendo a consistência geométrica entre diferentes perspectivas. Além disso, o espaço latente favorece a generalização para pontos de vista não observados anteriormente, funcionando como uma forma de "memória visual" do sistema.

A semântica do espaço pode ser explorada para estudar a influência de diferentes condições visuais sobre a reconstrução, bem como para gerar visualizações intermediárias entre diferentes embeddings, com potencial para aplicações em realidade aumentada e simulações.

3.2.4 Perturbações Visuais e Separação Geometria-Aparência

A fidelidade visual de uma reconstrução depende diretamente da capacidade do modelo de lidar com variações não controladas no conjunto de imagens de entrada. Nesse contexto, identificamos duas categorias principais de perturbações que impactam a renderização:

- **Perturbações Estáticas**: Variações inerentes à cena que precisam ser preservadas para garantir consistência estrutural. Exemplos incluem variações na iluminação ou mudanças sutis na aparência das superfícies ao longo do tempo.
- **Perturbações Dinâmicas**: Elementos transitórios na cena que não fazem parte da estrutura permanente e podem ser tratados como ruído na reconstrução. Exemplos incluem pedestres transitando em um ambiente urbano ou veículos momentaneamente obstruindo a visão da câmera.

Figura 3.2: Conjunto de imagens do mesmo monumento (Portão de Brandemburgo), parte do dataset PhotoTourism (SNAVELY; SEITZ; SZELISKI, 2006), registradas em diferentes horários, condições climáticas, pontos de vista e configurações fotográficas. A variedade de iluminação, coloração e contraste evidencia os desafios enfrentados na reconstrução 3D a partir de coleções fotográficas não controladas.

Variações de aparência são frequentemente causadas por mudanças nas condições ambientais entre diferentes capturas da mesma cena. Em ambientes externos, isso inclui variações fotométricas ao longo do dia, alterações atmosféricas como névoa ou chuva, e mudanças sazonais. Mesmo em ambientes controlados, sombras móveis ou reflexos podem introduzir inconsistências visuais.

Essas variações resultam em instabilidades na reconstrução, como artefatos cromáticos, borrões ou perdas de definição, sendo especialmente desafiadoras para métodos que dependem de consistência entre diferentes pontos de observação. Portanto, é essencial tratá-las explicitamente, seja por meio de embeddings de aparência ou pelo uso de máscaras de oclusão que separem os elementos persistentes dos transitórios.

Essas perturbações podem ser classificadas como:

- **Perturbação de Aparência**: Alterações fotométricas que afetam a cor, iluminação e reflexos da cena.
- **Perturbação de Oclusão**: Presença de elementos transitórios que bloqueiam parcial ou totalmente partes da cena.

Figura 3.3: Exemplos de obstruções parciais em imagens de cenas reais, extraídas do dataset PhotoTourism (SNAVELY; SEITZ; SZELISKI, 2006). Em ambos os casos, pessoas ou objetos transitórios interferem na visibilidade da estrutura principal (Portão de Brandemburgo), ocultando parcialmente partes da geometria. Esse tipo de variação é comum em ambientes não controlados e representa um desafio para métodos de reconstrução 3D que dependem de visibilidade consistente para gerar representações fotorrealistas e completas da cena.

A distinção entre esses tipos de perturbação é essencial para desenvolver mecanismos que permitam a filtragem seletiva de elementos indesejados e a preservação de características estruturais da cena.

4
Método

Neste capítulo, apresentamos os detalhes do método proposto para lidar com os desafios inerentes à renderização neural em ambientes não controlados. Como mencionado anteriormente, o objetivo é desenvolver um método robusto para tratar inconsistências visuais e distorções que ocorrem em cenas capturadas sob condições variáveis.

Este capítulo detalha os componentes da metodologia desenvolvida, incluindo as adaptações feitas sobre o *Gaussian Splatting* original e os mecanismos introduzidos para aumentar sua robustez em cenários variáveis.

A abordagem proposta busca aprimorar a consistência visual e reduzir artefatos por meio de ajustes na modelagem da aparência e da manipulação de variações fotométricas.

4.1
Arquitetura

Nossa metodologia combina técnicas de appearance embedding com Gaussian Splatting para modular características de aparência durante a renderização de cenas.

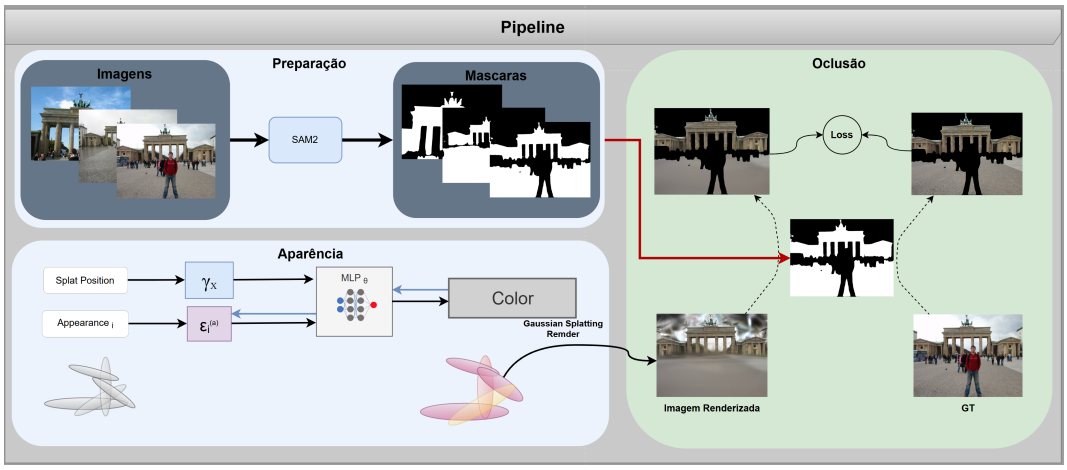

Figura 4.1: Arquitetura geral do método proposto, composta por dois módulos: um responsável pela estimação de cores condicionada à aparência (à esquerda) e outro dedicado ao tratamento de oclusões (à direita). A renderização final é supervisionada tanto pela imagem de referência quanto por uma máscara de oclusão extraída automaticamente a partir do ground truth.

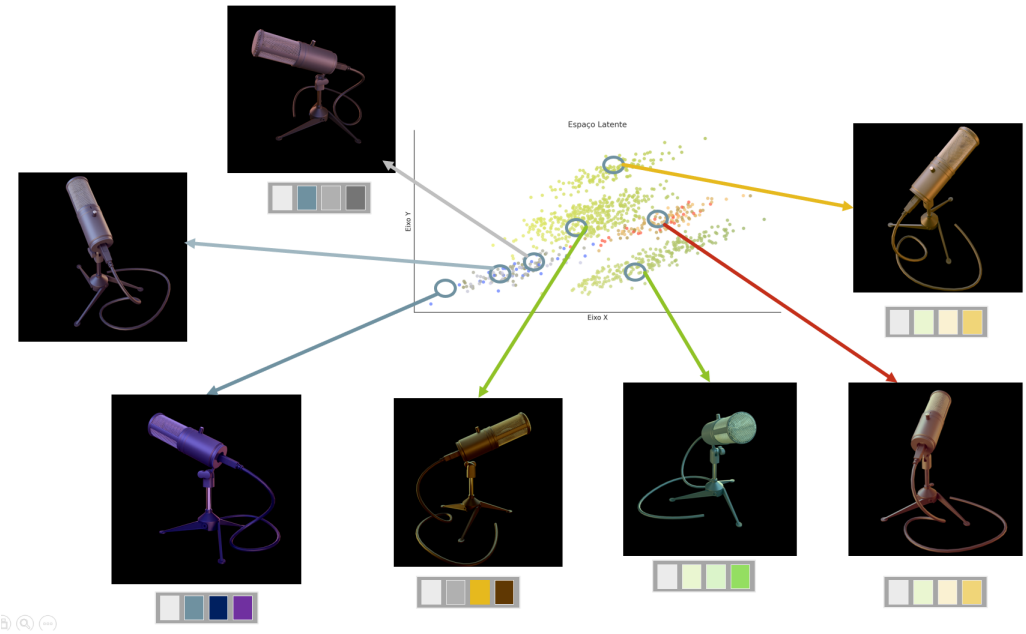

Figura 4.2: Representação bidimensional ilustrativa do espaço latente onde cada ponto corresponde a uma condição visual distinta associada a uma imagem renderizada. As setas indicam a projeção de diferentes aparências no espaço, destacando como a distribuição dos embeddings permite organizar e agrupar as imagens de acordo com suas características visuais, como tonalidade e iluminação.

4.2 Modelagem de Variações de Aparência

As variações de aparência são um dos desafios mais críticos para a reconstrução de cenas fotorrealistas. No *NeRF-W* (MARTIN-BRUALLA et al., 2021), essas inconsistências são tratadas por meio de um modelo latente que ajusta as diferenças fotométricas entre múltiplas imagens de uma mesma cena. No entanto, no contexto do *Gaussian Splatting*, a representação baseada em distribuições Gaussianas projetadas diretamente no espaço da imagem torna o método altamente sensível a essas variações.

Para mitigar esses efeitos, utilizamos *appearance embeddings* que parametrizam explicitamente a aparência de cada vista da cena. Essa abordagem permite uma reconstrução mais robusta e consistente ao longo de diferentes condições de iluminação e variação atmosférica.

4.2.1 Representação Latente da Aparência

Para capturar a variabilidade de aparência das imagens de entrada, adotamos um fator latente i que parametriza a distribuição das características visuais específicas da imagem I. Esse fator permite modelar e incorporar as

características específicas da imagem de entrada i durante o treinamento, garantindo que a rede seja capaz de aprender e reproduzir as diferenças de iluminação, tonalidade e outras variações visuais que possam existir entre as capturas da cena.

O pipeline inicia-se com a entrada de uma imagem I e um identificador único id associado a essa imagem. Esse identificador opera como um índice que possibilita a projeção da imagem em um espaço latente específico. A partir de id, um vetor latente $z_i \in \mathbb{R}^d$ é inferido, onde d representa a dimensionalidade do espaço latente. Esse vetor codifica informações semânticas e estilísticas da imagem, servindo como um controle explícito sobre a modulação da aparência na renderização das gaussianas.

4.2.2 Modulação da Aparência e Ajuste da Cor

O vetor latente z_i é incorporado aos parâmetros das gaussianas que compõem a cena. Esses parâmetros incluem coordenadas espaciais (μ_x, μ_y, μ_z), tensores de covariância Σ, além de propriedades radiométricas como opacidade e escala. O conjunto dessas características compõe a entrada para uma rede neural f, responsável por realizar um mapeamento não linear sobre os dados de entrada:

$$c_i = f(G, z_i) \tag{4-1}$$

onde G representa o conjunto completo de parâmetros das gaussianas, e c_i é o vetor de cores ajustado, utilizado diretamente na fase de renderização do *Gaussian Splatting*.

O propósito dessa rede neural f é aprender uma transformação no espaço latente que permita modificar as propriedades das gaussianas de maneira a otimizar a fidelidade visual da cena reconstruída. Essa modificação inclui ajustes contextuais na distribuição de cores, influências adaptativas na iluminação e a preservação das características estilísticas intrínsecas à imagem original.

No contexto do método proposto, a função f é desempenhada por um módulo denominado *Color Processor*, integrado diretamente ao pipeline de renderização do Gaussian Splatting. Esse módulo estima as cores das gaussianas com base em suas posições espaciais e no embedding de aparência associado à imagem de entrada.

A operação do Color Processor pode ser descrita pela seguinte equação:

$$c_{\theta_{is}} = \mathrm{MLP}_\theta(\mu_s, \ell_i) \tag{4-2}$$

onde μ_s representa a posição da gaussiana s, ℓ_i é o vetor latente de aparência referente à imagem i, e MLP_θ é uma rede neural parametrizada por

θ. Essa rede aprende a mapear as características espaciais e visuais para uma cor refinada $c_{\theta_{is}}$, que será utilizada diretamente no processo de renderização.

Esse modelo permite a modulação seletiva da aparência da cena, de modo que diferentes condições visuais possam ser representadas com fidelidade e estabilidade perceptual, mesmo em ambientes não controlados.

Essa estratégia de modulação é ilustrada na Figura 4.2, que apresenta uma projeção bidimensional do espaço latente aprendido. Cada ponto no gráfico representa uma imagem do conjunto de treinamento, organizada de acordo com suas características visuais predominantes. As setas conectam os vetores latentes às respectivas renderizações resultantes, demonstrando como variações sutis de tonalidade, iluminação e estilo são codificadas e reproduzidas pelo modelo. Essa visualização reforça a capacidade do sistema em agrupar condições visuais semelhantes, facilitando a interpolação coerente entre diferentes aparências e contribuindo para uma renderização mais estável e realista em cenários desafiadores.

4.3 Tratamento de Oclusão

O tratamento de oclusões na reconstrução de cenas tridimensionais é um desafio fundamental, especialmente em cenários não controlados. Oclusões podem ocorrer devido a elementos transitórios na cena, como veículos ou pedestres, ou podem ser causadas por estruturas fixas que bloqueiam parcialmente a visão da câmera. Para garantir a robustez da reconstrução, utilizamos uma abordagem baseada em máscaras de segmentação, permitindo a seleção de informações relevantes e a desconsideração de elementos transientes.

A estratégia adotada fundamenta-se em um processo de pré-processamento das imagens de entrada. Inicialmente, um conjunto de pontos é selecionado para definir regiões de interesse na cena. A partir dessa seleção, são extraídas máscaras de segmentação que identificam os elementos estruturais predominantes. Essas máscaras são posteriormente analisadas em múltiplas imagens para determinar a recorrência dos elementos extraídos e estabelecer um critério de confiabilidade na reconstrução.

O pipeline de tratamento de oclusão adotado neste trabalho foi desenvolvido com o objetivo de isolar e remover elementos transitórios da cena, preservando apenas as regiões estruturalmente consistentes. Esse processo é composto pelas seguintes etapas:

1. **Seleção inicial de pontos de interesse**: Um ou mais pontos são definidos manualmente sobre a imagem ou quadro de vídeo, indicando regiões relevantes a serem segmentadas.

2. **Execução do Segment Anything**: A partir dos pontos definidos, o modelo Segment Anything (KIRILLOV et al., 2023) é aplicado sobre todas as imagens ou quadros do vídeo, produzindo máscaras de segmentação correspondentes.

3. **Verificação das segmentações**: Cada resultado é avaliado visualmente. Caso uma segmentação seja inadequada ou imprecisa, o ponto correspondente é removido da entrada.

4. **Ajustes incrementais**: Novos pontos são adicionados conforme necessário, de forma a complementar regiões que não tenham sido corretamente segmentadas na iteração anterior.

5. **Iteração até convergência**: O processo é repetido até que todas as imagens tenham sido verificadas e estejam adequadamente segmentadas, garantindo consistência na separação dos elementos persistentes e transitórios da cena.

Essa abordagem favorece uma reconstrução mais robusta e visualmente coerente, ao reduzir o impacto de elementos não estruturais ou inconsistentes na composição final. A utilização do Segment Anything como ferramenta central de segmentação garante alta cobertura e adaptabilidade a diferentes tipos de conteúdo visual, contribuindo para a precisão e estabilidade da reconstrução tridimensional.

4.4 Adaptações da Implementação Base

Em nosso trabalho, adotamos a implementação original do Gaussian Splatting para treinamento e renderização. Em vez de utilizar a computação interna tradicional e harmônicos esféricos para representação de cores, optamos por usar cores pré-computadas.

Além disso, adaptamos o processo de treinamento para considerar todas as imagens de entrada em cada iteração. Essa modificação permite o cálculo e a otimização da incorporação da aparência, garantindo que o modelo capture melhor as variações na aparência em diferentes visualizações.

Mantivemos a perda original proposta no artigo (KERBL et al., 2023) para calcular a qualidade por imagem, conforme definido na Equação 4-3 a seguir:

$$\mathcal{L}_G = (1 - \lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{\text{DSSIM}} \tag{4-3}$$

onde λ é um fator de regularização, $\mathcal{L}_1$ é a perda L1 e $\mathcal{L}_{\text{DSSIM}}$ é um termo de similaridade estrutural (DSSIM).

Os experimentos realizados com a técnica proposta, bem como os datasets utilizados e o processo de treinamento, são descritos com mais detalhes no Capítulo 5.

4.5 Função de Perda

Adaptamos a função de perda para considerar todas as imagens do conjunto de treinamento, garantindo que a função de cor possa ser otimizada de maneira robusta.

Precisávamos adaptar a função de perda para considerar o conjunto de todas as imagens. Isso garante que a função de cor possa ser otimizada considerando todos os índices, direcionando assim a variância para ser distribuída com o menor erro possível. A equação 4-4 mostra a função de perda em detalhes:

$$\mathcal{L} = \frac{1}{N} \min_{\theta \in \Theta} \sum_{i}^{N} \mathcal{L}_G(\mathbf{x}_i, \mathcal{G}(S, \mathbb{C}_{\theta i S})) \tag{4-4}$$

onde $\mathcal{L}_G$ é a perda de Splatting Gaussiano Original na Equação 4-3, $\mathbf{x}_i$ é a imagem de referência(ground truth), $\mathcal{G}$ é a função de renderização de Splatting Gaussiano que considera todo o conjunto S e o conjunto de cores gaussianas $\mathbb{C}_{\theta i S}$.

5
Experimentos

Este capítulo apresenta os experimentos realizados para avaliar o desempenho do modelo proposto em diferentes contextos. A proposta foi testada tanto em cenas sintéticas quanto reais, considerando cenários com variabilidade visual, presença de oclusões transitórias e condições fotométricas não controladas. Os experimentos têm como objetivo verificar a robustez da técnica frente a esses desafios, bem como analisar a contribuição individual de cada componente introduzido — embeddings de aparência e máscaras de oclusão — por meio de estudos de ablação.

A avaliação inclui tanto análises qualitativas quanto quantitativas, utilizando métricas perceptuais amplamente adotadas na literatura. As seções a seguir detalham os conjuntos de dados utilizados, a configuração de treinamento, os resultados obtidos e uma discussão comparativa entre diferentes variantes do modelo.

5.1
Visão Geral dos Experimentos

Os experimentos foram organizados de forma a permitir uma análise progressiva dos efeitos de cada componente do pipeline proposto. Inicialmente, são apresentados resultados em cenas sintéticas, onde há maior controle sobre variabilidade de aparência e geometria. Em seguida, são explorados casos reais, oriundos de conjuntos fotográficos não estruturados, que refletem situações mais próximas de aplicações práticas.

Os testes buscam responder a três principais questões:

- O modelo proposto é capaz de manter consistência visual e estrutural em ambientes com variação de aparência?
- A introdução de máscaras de oclusão contribui para a remoção de elementos transitórios e melhora a qualidade da reconstrução?
- Como cada componente do modelo influencia o desempenho global, tanto em termos qualitativos quanto quantitativos?

Com base nessas diretrizes, foram conduzidos experimentos controlados, seguidos por análises de ablação e comparações com variações do modelo sem os módulos propostos.

5.2 Treinamento

Nesta seção descreveremos o processo de treinamento do modelo, abordando os dados utilizados, a configuração dos experimentos, os hiperparâmetros e os critérios de convergência. O objetivo é detalhar como o modelo foi ajustado e treinado para alcançar os melhores resultados possíveis.

5.2.1 Configuração do Treinamento

Os experimentos foram conduzidos em ambiente computacional controlado, com o objetivo de garantir reprodutibilidade e estabilidade ao longo do processo de treinamento. As redes foram implementadas utilizando o framework PyTorch, mantendo compatibilidade com a implementação base do *Gaussian Splatting*. Para a execução, foram utilizadas duas GPUs distintas: uma NVIDIA RTX 3080 e uma Quadro RTX 6000, garantindo ampla capacidade de memória e paralelismo.

O tempo médio de treinamento para cada experimento foi de aproximadamente 30 minutos, considerando 50 mil iterações por cena. Essa configuração foi mantida constante em todas as avaliações para possibilitar comparações justas entre diferentes variantes do modelo.

5.2.2 Hiperparâmetros e Estratégias de Regularização

Os hiperparâmetros foram definidos de forma a equilibrar a velocidade de convergência com a estabilidade do treinamento, respeitando as configurações padrão do método base e estendendo-as para os novos módulos introduzidos. A taxa de aprendizado utilizada para o embedding de aparência (`embed_lr`) e para o módulo MLP (`mlp_lr`) foi de 0,0025 em ambos os casos. O vocabulário latente foi configurado com 1500 vetores distintos, cada um com dimensão de 48. Para o mapeamento posicional, foi adotado um encoding de Fourier com 10 frequências.

A função de perda utilizada foi composta por dois termos: um erro absoluto (L1) e a métrica perceptual DSSIM, ponderados conforme descrito no Capítulo 4. A otimização foi realizada utilizando o otimizador Adam, mantendo-se os parâmetros originais da implementação do *Gaussian Splatting*, com extensões aplicadas aos módulos adicionais de embedding e rede MLP.

Durante o treinamento, foi utilizada uma política de regularização com densificação progressiva. O processo de densificação foi aplicado até a iteração 25.000, com checkpoints intermediários salvos a cada 5.000 iterações e passos de

salvamento a cada 10.000 iterações. Além disso, foi atualizada a estratégia de escalonamento da taxa de aprendizado das posições, limitada a 50.000 iterações para garantir a estabilização espacial das gaussianas ao final do treinamento.

5.2.3 Dataset

Os experimentos desta dissertação foram conduzidos com dois cenários principais de dados, organizados para permitir uma análise progressiva da robustez da abordagem proposta. Foram utilizados tanto dados sintéticos com perturbações controladas quanto dados reais com características complexas e variação natural.

- **Cenário 1 — Dados Sintéticos com Perturbações Controladas**: Baseado nas cenas do conjunto Blender do NeRF (MILDENHALL et al., 2020), este cenário foi utilizado para simular diferentes tipos de distorções visuais em ambiente controlado. As perturbações incluíram variações cromáticas, alterações na iluminação e inserções artificiais de objetos oclusores, inspiradas nas estratégias adotadas pelo NeRF-W (MARTIN-BRUALLA et al., 2021). Essa configuração permitiu avaliar de forma isolada o impacto de cada fator sobre a qualidade da reconstrução.
- **Cenário 2 — Dados Reais em Ambientes Complexos**: O foco principal neste cenário foi a cena do caminhão (*Truck*) do conjunto Tanks and Temples (KNAPITSCH et al., 2017), que apresenta desafios típicos de reconstrução em ambientes não controlados, como variações de iluminação natural, presença de elementos transitórios (pessoas e veículos), diferentes ângulos de captura e texturas com níveis variados de detalhe. Essa cena foi selecionada por representar um caso realista, com geometria densa e alto nível de ruído visual.

A Figura 5.1 apresenta exemplos de imagens do primeiro cenário, onde a cena do microfone foi distorcida com diferentes alterações visuais para testar a robustez do modelo proposto.

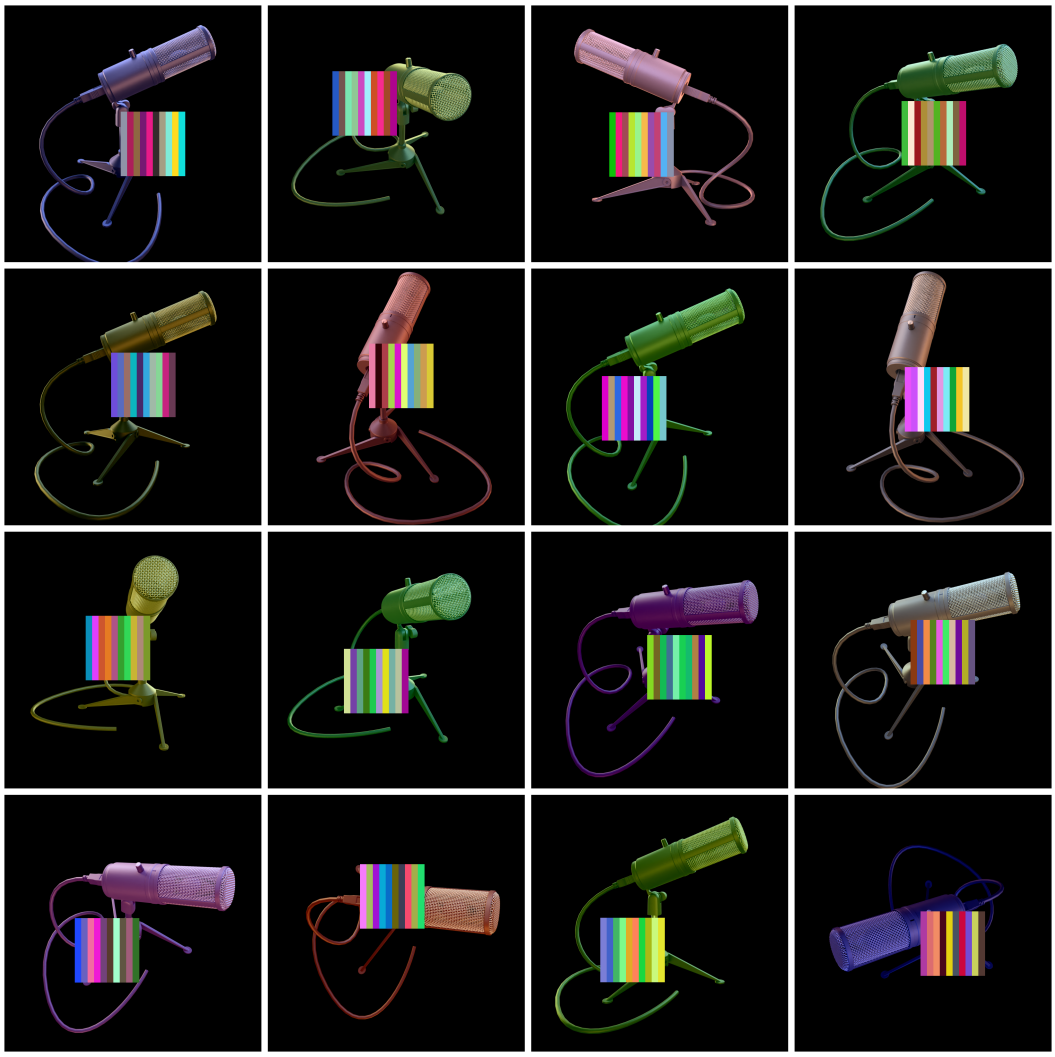

Figura 5.1: Exemplo de amostras do dataset sintético baseado no NeRF (MILDENHALL et al., 2020), contendo múltiplas imagens da cena do microfone com variações controladas de aparência e oclusão. Cada imagem apresenta distorções cromáticas e inserções artificiais que simulam alterações de iluminação, materiais e obstruções parciais, permitindo a análise do impacto de inconsistências visuais sobre a reconstrução.

Em ambos os cenários, as imagens foram acompanhadas de máscaras de segmentação geradas com o modelo Segment Anything (KIRILLOV et al., 2023), aplicadas para filtrar regiões não estruturais e objetos transitórios durante o treinamento. A utilização dessas máscaras está detalhada no Capítulo 4.

Para simular variações de aparência entre diferentes pontos de vista, utilizamos uma abordagem inspirada no NeRF-W (MARTIN-BRUALLA et al., 2021), aplicando distorções fotométricas nas imagens originais. Cada imagem $I_i \in [0,1]^{800\times800\times3}$ foi modificada para $\tilde{I}_i$ por meio da expressão:

$$\tilde{I}_i = \min(1, \max(0, s_i I_i + b_i)) \tag{5-1}$$

onde $s_{ij} \sim \mathcal{U}(0.8, 1.2)$ e $b_{ij} \sim \mathcal{U}(-0.2, 0.2)$ são amostrados aleatoriamente

para cada canal RGB da imagem i. Essa modificação gera variações controladas de brilho, saturação e coloração, permitindo avaliar a robustez do modelo proposto frente a perturbações visuais.

Por fim, as imagens de referência (ground truth) foram utilizadas para comparação direta e cálculo de métricas perceptuais (SSIM, PSNR, LPIPS), permitindo uma avaliação quantitativa objetiva da qualidade das reconstruções obtidas.

5.2.4 Produção das Máscaras

A produção de máscaras constitui uma etapa essencial no tratamento de oclusões, pois permite isolar elementos transitórios que poderiam comprometer a consistência da reconstrução. Para essa tarefa, utilizamos o modelo *Segment Anything*, que permite segmentar automaticamente objetos relevantes nas imagens de entrada, gerando máscaras binárias que destacam regiões potencialmente ocluidoras, como pessoas, veículos ou objetos móveis.

O processo consiste em aplicar a segmentação diretamente sobre os frames do conjunto de dados, com base em pontos de interesse previamente definidos ou em estratégias automáticas. O *Segment Anything* mostrou-se eficaz na identificação de estruturas não estáticas mesmo em condições desafiadoras de iluminação, movimento e complexidade de cena.

A Figura 5.2 apresenta exemplos desse processo: cada par mostra, à esquerda, a imagem original da cena, e à direita, a máscara correspondente gerada automaticamente. Tais máscaras são posteriormente utilizadas para filtrar as contribuições dos elementos transitórios durante o processo de reconstrução, aumentando a fidelidade e a estabilidade da renderização final.

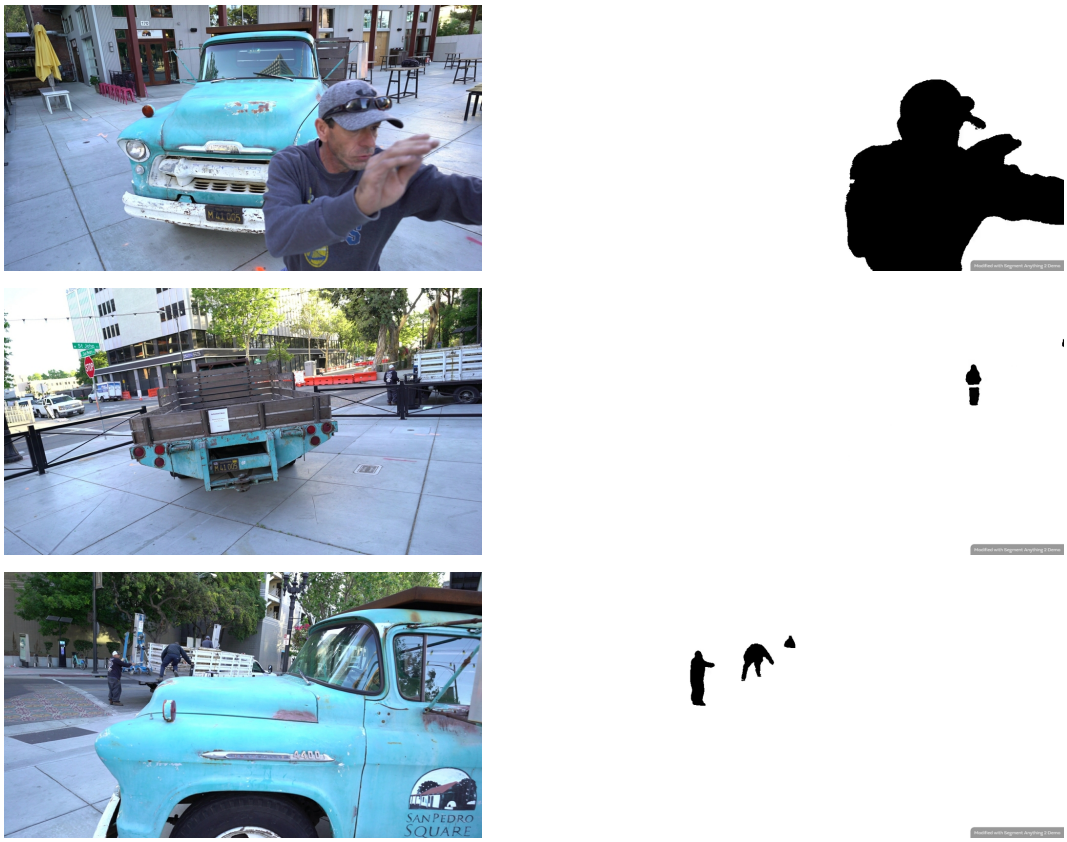

Figura 5.2: Exemplos de máscaras geradas com Segment Anything a partir de imagens da cena do caminhão do conjunto Tanks and Temples (KNAPITSCH et al., 2017). As máscaras destacam regiões com objetos transitórios (como pessoas), que podem interferir na reconstrução fotorrealista.

5.3 Resultados Obtidos

Esta seção apresenta os principais resultados qualitativos obtidos com o modelo proposto, avaliando sua capacidade de reconstruir cenas em diferentes condições visuais. Os exemplos incluem tanto dados sintéticos com distorções controladas quanto cenas reais sujeitas a oclusões transitórias, variações fotométricas e presença de objetos móveis.

A Figura 5.3 exibe o resultado da reconstrução de um objeto sintético do tipo microfone, utilizado em um cenário com distorções artificiais de cor e aparência. Mesmo diante de alterações significativas na coloração das imagens de entrada, o modelo foi capaz de preservar a estrutura geométrica e manter uma aparência estável. Elementos como a grade metálica do microfone, seu suporte e o cabo foram representados com nitidez e continuidade entre diferentes ângulos de câmera.

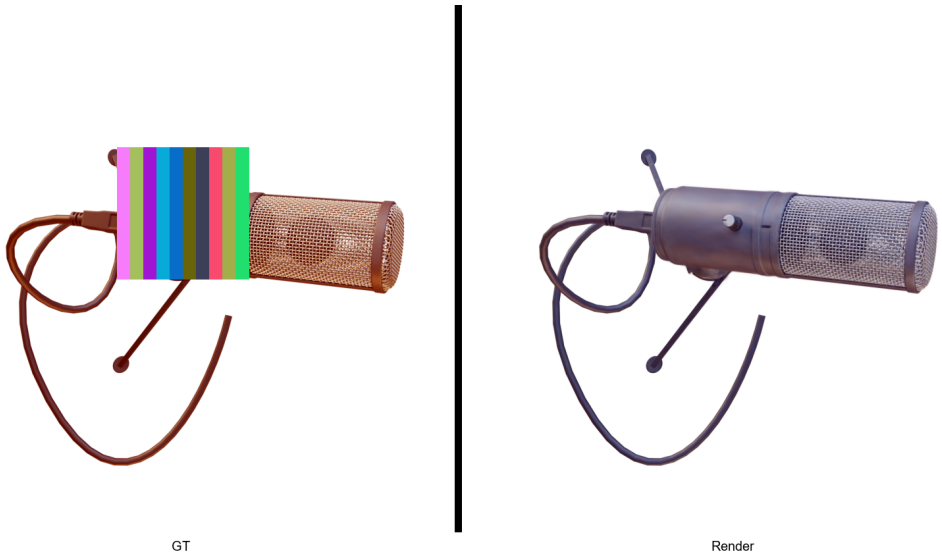

Figura 5.3: Resultado comparando GroundTruth à esquerda com o Resultado obtido após o treinamento

Já nas Figuras 5.4 e 5.5, são apresentados exemplos em ambientes reais, extraídos da cena Truck do dataset Tanks And Temples (KNAPITSCH et al., 2017). A Figura 5.4 mostra uma cena onde um pedestre parcialmente obstrui o objeto de interesse. Ainda assim, a renderização gerada pelo modelo é capaz de eliminar essa oclusão de forma natural, reconstruindo a cena como se o elemento transitório não estivesse presente.

Na Figura 5.5, o desafio é ampliado pela presença de pessoas em movimento. A abordagem proposta novamente demonstra sua robustez, ignorando os elementos móveis e mantendo apenas a geometria estática do local. A combinação entre a filtragem de máscaras de oclusão e o condicionamento por embeddings de aparência permite uma reconstrução mais limpa, coerente e visualmente realista.

Esses exemplos ilustram o potencial da técnica para aplicações em ambientes não controlados, como reconstrução urbana, preservação de patrimônio, realidade aumentada e geração de cenas virtuais consistentes a partir de coleções fotográficas heterogêneas.

Figura 5.4: Exemplo de remoção de oclusão causada por pedestre em cena real. A reconstrução final ignora a presença transitória, preservando a geometria estrutural da imagem.

Figura 5.5: Cenário urbano com objetos em movimento. O método remove com sucesso veículos transitórios, mantendo apenas elementos estáticos da cena.

5.4 Avaliação

Esta seção apresenta uma análise detalhada dos resultados obtidos pelo modelo após o treinamento. São discutidas métricas quantitativas e qualitativas, comparações com abordagens do estado da arte e interpretações sobre o desempenho em diferentes cenários. O objetivo é avaliar a eficácia da abordagem proposta e identificar possíveis direções para otimizações futuras.

5.4.1 Avaliação Qualitativa

Nesta subseção, analisamos qualitativamente os resultados produzidos pelo modelo proposto, comparando as imagens renderizadas com os dados de referência (ground truth) e observando o comportamento em diferentes tipos de desafio visual, especialmente em cenas com oclusões parciais ou elementos transitórios.

Figura 5.6: Na parte superior, a imagem original com destaque em vermelho para a região-alvo da remoção. Na parte inferior, as imagens de referência e os resultados: (A) imagem original contendo o objeto (ground truth); (B) resultado obtido utilizando a técnica padrão de *Gaussian Splatting*; e (C) resultado obtido pelo método proposto. Observa-se no resultado (C) uma reconstrução mais fiel do fundo, com melhor preservação da estrutura e continuidade visual.

A Figura 5.6 apresenta uma situação comum em ambientes reais: a presença de pessoas em movimento próximo à câmera. Na imagem de entrada, um indivíduo atravessa o campo de visão, obstruindo parcialmente a estrutura de fundo. A imagem renderizada demonstra que o modelo foi capaz de eliminar esse elemento transitório, reconstruindo a geometria da cena de forma limpa e contínua. Esse resultado evidencia a eficácia do pipeline de segmentação e filtragem no tratamento de oclusões temporárias.

Figura 5.7: Na parte superior, a imagem original com destaque em vermelho para a região contendo os objetos a serem removidos. Na parte inferior, os resultados: (A) imagem original com os objetos presentes (ground truth); (B) resultado obtido pela técnica de *Gaussian Splatting*; e (C) resultado obtido pelo método proposto.

De maneira semelhante, a Figura 5.7 mostra uma cena urbana com múltiplas interferências visuais, como veículos e equipamentos em movimento. A reconstrução final ignora esses elementos não estruturais e preserva apenas as superfícies fixas e consistentes da cena. Isso confirma a capacidade do método em distinguir entre informações confiáveis e ruído visual, mesmo sem supervisão direta.

Essas observações qualitativas são particularmente relevantes para aplicações em reconstrução urbana, digital twins e geração de ambientes limpos para simulação ou realidade virtual, onde a presença de objetos efêmeros pode comprometer a integridade da cena.

5.4.2 Métricas Quantitativas

A avaliação quantitativa dos métodos foi realizada com base em três métricas amplamente adotadas na literatura de reconstrução de imagens:

- **SSIM (Structural Similarity Index)**: avalia a similaridade estrutural entre imagens, considerando luminância, contraste e estrutura. Valores mais próximos de 1 indicam maior similaridade.
- **PSNR (Peak Signal-to-Noise Ratio)**: mede a razão entre o sinal e o ruído introduzido pela reconstrução, em decibéis (dB). Quanto maior o valor, melhor a fidelidade da imagem.
- **LPIPS (Learned Perceptual Image Patch Similarity)**: métrica perceptual baseada em redes neurais, que correlaciona melhor com julgamentos humanos. Valores menores indicam maior similaridade visual.

As métricas foram aplicadas às renderizações produzidas por dois métodos: **Ours**, baseado em renderização neural com embeddings de aparência, e **3DGS**, baseado em *3D Gaussian Splatting* direto. As comparações foram realizadas sobre as cenas *LEGO* e *MIC*, considerando dois cenários distintos: *Aparência Base sem Distorção* e *Aparência Variável com Distorção de Cor*.

A Tabela 5.1 resume os resultados quantitativos. Os melhores valores por linha estão destacados em vermelho. De modo geral, observa-se que o método proposto (Ours) apresenta desempenho superior em todas as combinações de teste, com destaque para a cena MIC, onde as diferenças em PSNR e LPIPS são particularmente acentuadas.

Para facilitar a interpretação dos dados, os resultados são apresentados também por meio de gráficos:

- As Figuras 5.8, 5.9 e 5.10 mostram os valores de cada métrica, organizados por cenário (*Aparência / Distorção*).
- As Figuras 5.11, 5.12 e 5.13 apresentam as médias gerais de cada métrica ao longo de todos os testes.

Tabela 5.1: Comparação entre os métodos Ours e 3DGS para cenas LEGO e MIC. Em vermelho os melhores valores por par.

			Método	Ours			3DGS		
			Métrica	SSIM	PSNR	LPIPS	SSIM	PSNR	LPIPS
Cena	Set	Aparência	Distorção						
LEGO	Test	Base	SemDist	**0.914**	**23.903**	**0.090**	0.879	20.690	0.121
		Var.	Cor	**0.895**	**21.586**	**0.102**	0.867	19.383	0.134
	Train	Base	SemDist	**0.926**	**24.485**	**0.080**	0.891	21.058	0.110
		Var.	Cor	**0.930**	**27.554**	**0.061**	0.870	18.883	0.119
MIC	Test	Base	SemDist	**0.980**	**30.747**	**0.021**	0.957	22.857	0.054
		Var.	Cor	**0.972**	**28.207**	**0.027**	0.950	21.740	0.062
	Train	Base	SemDist	**0.985**	**31.845**	**0.017**	0.967	24.216	0.044
		Var.	Cor	**0.984**	**32.270**	**0.014**	0.964	23.460	0.047
CHAIR	Test	Base	SemDist	**0.961**	**29.368**	**0.039**	0.936	23.664	0.076
		Var.	Cor	**0.953**	**26.164**	**0.052**	0.929	21.507	0.087
	Train	Base	SemDist	**0.967**	**29.842**	**0.035**	0.947	24.944	0.064
		Var.	Cor	**0.966**	**31.144**	**0.035**	0.943	23.014	0.070
FICUS	Test	Base	SemDist	**0.972**	**29.664**	**0.027**	0.944	22.806	0.065
		Var.	Cor	**0.964**	**26.945**	**0.037**	0.937	21.651	0.074
	Train	Base	SemDist	**0.978**	**30.153**	**0.024**	0.956	23.850	0.053
		Var.	Cor	**0.977**	**30.461**	**0.024**	0.954	23.319	0.055
HOTDOG	Test	Base	SemDist	**0.963**	**30.587**	**0.056**	0.925	22.661	0.107
		Var.	Cor	**0.941**	**23.617**	**0.093**	0.911	20.701	0.130
	Train	Base	SemDist	**0.965**	**31.078**	**0.054**	0.931	23.133	0.107
		Var.	Cor	**0.938**	**23.256**	**0.104**	0.922	21.938	0.111

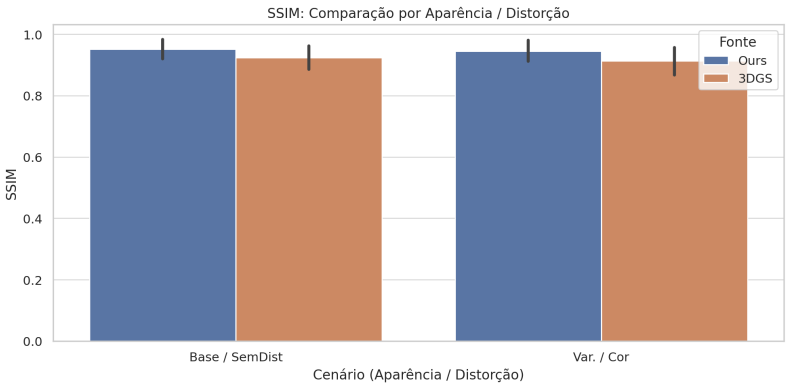

Figura 5.8: Comparação da métrica SSIM entre os métodos, agrupada por cenário (*Aparência / Distorção*).

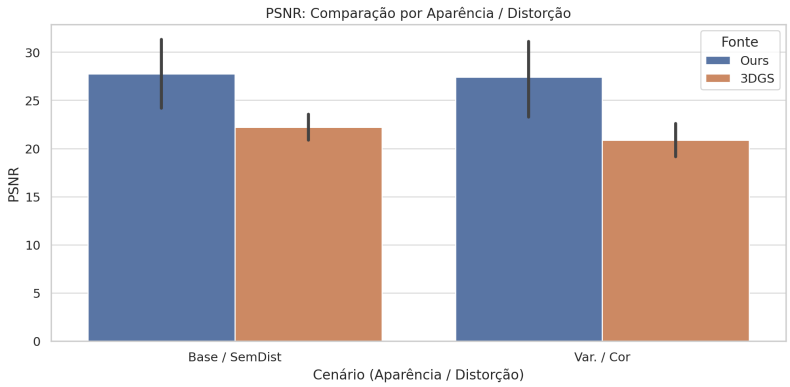

Figura 5.9: Comparação da métrica PSNR entre os métodos, agrupada por cenário (*Aparência / Distorção*).

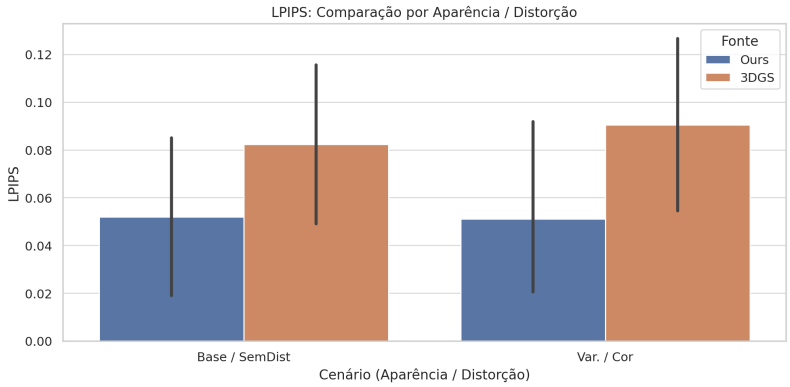

Figura 5.10: Comparação da métrica LPIPS entre os métodos, agrupada por cenário (*Aparência / Distorção*).

A Figura 5.11 revela que, em termos de similaridade estrutural (SSIM), o método Ours supera o 3DGS de forma consistente. A diferença de desempenho é notável mesmo considerando a escala restrita da métrica (valores próximos de 1), indicando maior preservação estrutural nas imagens reconstruídas.

Na Figura 5.12, observa-se um ganho expressivo do método proposto sobre o 3DGS na métrica PSNR. A diferença de diversos decibéis reforça a fidelidade da reconstrução e a menor presença de ruído.

Já a Figura 5.13 mostra que o método Ours também se destaca na métrica perceptual LPIPS, com valores visivelmente menores. Isso sugere uma maior proximidade visual com as imagens de referência, segundo modelos treinados com avaliações humanas.

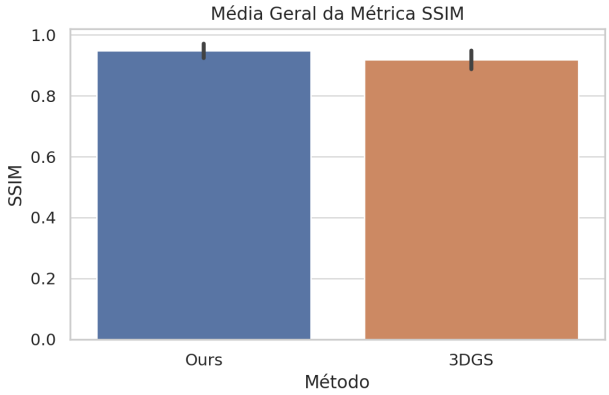

Figura 5.11: Média geral da métrica SSIM para os métodos Ours e 3DGS, considerando todas as cenas e configurações.

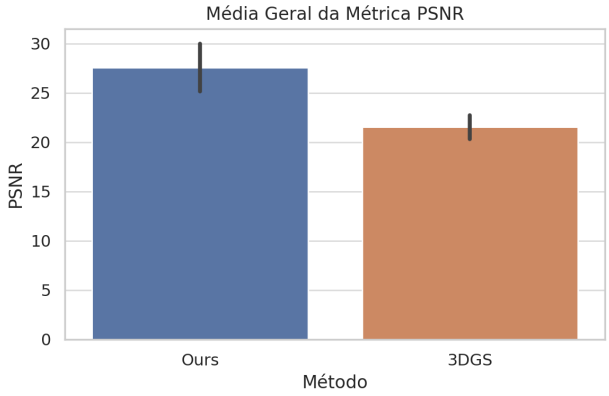

Figura 5.12: Média geral da métrica PSNR para os métodos Ours e 3DGS. A métrica avalia a fidelidade de reconstrução em decibéis (dB).

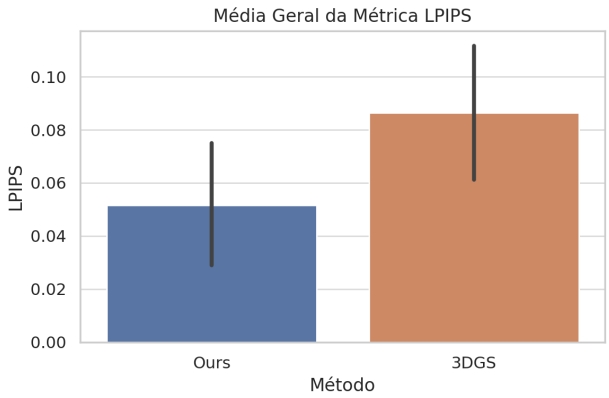

Figura 5.13: Média geral da métrica LPIPS para os métodos Ours e 3DGS. Valores menores indicam maior similaridade perceptual.

5.5 Ablations

Nesta seção, iremos destrinchar nossa solução, e os módulos existentes. Pretendemos fazer uma análise para demonstrar que cada parte no processo é necessária para a obtenção dos resultados encontrados, e como cada elemento se relaciona com os problemas vistos.

5.5.1 Sem nenhum Módulo

Figura 5.14: Resultado obtido com ambos os módulos de aparência e oclusão desativados, equivalente ao Gaussian Splatting tradicional. Observa-se instabilidade cromática, artefatos de sobreposição e perda de detalhes estruturais.

Na ausência de mecanismos explícitos de codificação de aparência e de tratamento de oclusão, o modelo equivale diretamente ao *Gaussian Splatting* original. Como ilustrado nas imagens, essa configuração básica apresenta limitações importantes na qualidade da reconstrução visual

Primeiramente, observa-se uma variação intensa e inconsistente de cor ao redor da base do microfone, causada pela ausência de um modelo de aparência que estabilize a distribuição cromática entre pontos de vista. Esse artefato visual é fortemente sensível ao posicionamento da câmera, resultando em *flickering* ou ruídos cromáticos, especialmente em regiões de oclusão parcial. O modelo, sem informação contextual de visibilidade, tenta compensar essas regiões com efeitos de *fog* ou borrões coloridos, como uma tentativa implícita de "esconder"a incerteza geométrica.

Além disso, há falhas evidentes na reconstrução de detalhes finos, como a grade metálica do microfone, que se apresenta borrada ou distorcida. Essa perda de definição ocorre tanto pela imprecisão na modelagem de profundidade quanto pela incapacidade do modelo de discernir entre superfícies altamente detalhadas e regiões de fundo. Sem a máscara de oclusão, o modelo tende a acumular splats mesmo em áreas que não deveriam contribuir para a imagem final, prejudicando a fidelidade estrutural.

Outros pontos relevantes incluem:

- *Bleeding de cor* entre partes não contíguas da geometria;
- Falta de coerência nas bordas, especialmente entre o cabo e o tripé;
- Mistura de camadas visuais, resultando em transparências artificiais.

Os resultados reforçam a necessidade de mecanismos adicionais, como *appearance embeddings* e máscaras de oclusão, para lidar com as limitações observadas.

5.5.2 Testes somente com o Módulo de Aparência

Ao introduzirmos o módulo de *appearance embedding*, observa-se uma melhora significativa em diversos aspectos em comparação à versão original do *Gaussian Splatting*, conforme discutido anteriormente.

A principal mudança está na estabilização da coloração dos pontos Gaussianos ao longo dos diferentes pontos de vistas. Enquanto a versão tradicional sofria com variações cromáticas abruptas e efeitos fantasmas coloridos — principalmente nas regiões de sobreposição e oclusão parcial —, o uso de embeddings de aparência permite uma representação mais coerente do conteúdo visual, minimizando flutuações sensíveis à posição da câmera.

Figura 5.15: Resultado com o módulo de aparência ativado e o módulo de oclusão desativado. Há melhoria na estabilidade cromática e suavização dos borrões, mas persistem artefatos espaciais em regiões ocluídas.

Nota-se também uma redução considerável de ruídos cromáticos que anteriormente se manifestavam como borrões multicoloridos em torno da haste e da base do microfone. Com o embedding de aparência, esses artefatos são substituídos por uma névoa acinzentada, mais neutra, que reflete uma tentativa mais informada do modelo em lidar com regiões ambíguas ou parcialmente ocluídas. Ainda que o *fog* permaneça visível — o que indica que o módulo de oclusão ainda não está ativo —, sua natureza é mais suave e visualmente consistente.

Em termos de detalhamento fino, como a malha da grade do microfone, há uma recuperação parcial da estrutura. A regularidade do padrão é melhor preservada, embora ainda apresente certa suavização ou borramento em áreas onde há confusão entre camadas frontais e de fundo.

Outros pontos relevantes:

- A aparência contribui para preservar texturas locais, mesmo em regiões com profundidade variável, como a conexão entre o cabo e o suporte;
- Redução do "vazamento"de cor entre superfícies desconectadas;
- A percepção volumétrica do objeto melhora, ainda que sombras, reflexos ou transparências complexas não sejam plenamente modeladas nesta etapa.

Em resumo, a ativação do módulo de aparência representa um avanço claro sobre o modelo base, reduzindo instabilidades e proporcionando maior fidelidade perceptual, ainda que sem resolver por completo os problemas oriundos da oclusão.

5.5.3 Teses somente com o Módulo de Oclusão

Figura 5.16: Resultado com o módulo de oclusão ativado e aparência desativada. A organização espacial é significativamente melhorada, com redução de sobreposições indevidas, embora persistam variações cromáticas entre pontos de vista.

Com a ativação isolada do módulo de **máscara de oclusão**, já se observa uma melhora considerável na organização espacial da cena e na supressão de artefatos causados por sobreposição indevida de pontos Gaussianos. Ainda sem o uso de appearance embeddings, os benefícios concentram-se principalmente na estrutura geométrica e na coerência visual da profundidade.

A máscara de oclusão permite ao modelo descartar splats que deveriam estar ocultos sob outros elementos da geometria, o que resulta na eliminação

quase total dos efeitos de borrão multicolorido anteriormente visíveis ao redor do suporte do microfone. Isso se traduz em uma composição visual mais limpa e realista, especialmente em regiões com interseções complexas, como a base do tripé e os segmentos do cabo.

Entretanto, como o componente de aparência permanece desativado, persistem variações cromáticas abruptas e instabilidades de cor em função do ponto de vista. Essas variações são menos intensas do que no modelo puramente tradicional, mas ainda presentes, especialmente em superfícies com menor textura. Isso indica que, embora a oclusão organize bem a contribuição dos pontos no espaço, ela não estabiliza seu conteúdo visual.

Outros pontos observáveis:

- A **grade do microfone** é renderizada com nitidez superior à do modelo tradicional, com melhor separação entre os fios da malha, graças à correta visibilidade frontal;
- A geometria do **cabo e tripé** é consistente e não sofre mais com vazamentos ou sobreposição incorreta;
- Ainda há **inconsistência cromática local**, com trechos mais esverdeados ou avermelhados, reflexo da ausência de controle explícito de aparência.

Em resumo, o uso exclusivo da máscara de oclusão soluciona grande parte dos problemas estruturais, reduzindo drasticamente a interferência de splats incorretos e permitindo uma renderização mais fiel em termos de profundidade e forma. No entanto, a **instabilidade cromática residual** indica que o embedding de aparência continua sendo necessário para uma composição visual mais estável e uniforme.

5.5.4 Modelo Completo

A ativação simultânea dos módulos de *appearance embedding* e máscara de oclusão resulta em uma composição visual significativamente superior às versões anteriores, tanto em termos de fidelidade estrutural quanto de consistência perceptual.

A aparência é estabilizada globalmente, o que elimina variações cromáticas sensíveis ao ponto de vista, enquanto a máscara de oclusão garante que apenas os splats visíveis contribuam para a imagem final. Com isso, evitam-se artefatos típicos como *fog*, transparências espúrias, sobreposição de geometrias e sobreposição de camadas indevidas.

A grade do microfone, anteriormente borrada ou deformada, é agora renderizada com clareza, mantendo a regularidade do padrão metálico mesmo em posições de câmera desfavoráveis. O cabo e o suporte mantêm suas formas e posições corretas, sem ruídos visuais ou mistura de planos. A transição entre superfícies é suave, e os detalhes finos são preservados sem perder coesão com o volume geral da cena.

Além disso, a cena como um todo se torna visualmente natural e coerente, mesmo com movimento ou variações de câmera. A profundidade é respeitada, as cores são consistentes, e a imagem final se aproxima de uma renderização realista.

5.5.5 Comparativo dos Resultados

As avaliações qualitativas indicam que o modelo proposto é capaz de preservar e reproduzir variações de aparência presentes nas imagens de referência. Diferentes tonalidades (como azuladas ou lilás) foram mantidas de forma consistente em múltiplas poses da cena, demonstrando que os embeddings aprendidos representam com precisão características visuais relevantes.

Além disso, ao manipular o espaço latente — por exemplo, ao remover componentes principais de menor variância — foi possível gerar renderizações que se aproximam da média visual da cena. Isso mostra que a representação aprendida é não apenas robusta, mas também interpretável e manipulável.

Esses resultados reforçam a importância da modulação explícita da aparência, especialmente em cenários com iluminação variável, sombras móveis ou outras formas de inconsistência visual.

Tabela 5.2: Comparação entre as diferentes configurações de módulos de aparência e oclusão.

Critério	**Sem Aparência Sem Oclusão**	**Com Aparência Sem Oclusão**	**Sem Aparência Com Oclusão**	**Com Aparência Com Oclusão**
Cor / Aparência	Variações cromáticas instáveis, sensíveis à câmera	Cores suavizadas, mais naturais	Inconsistência cromática residual	Cor consistente sob múltiplas perspectivas
Regiões Ocluídas	Fog colorido, sobreposição indevida	Fog neutro e mais suave	Remoção correta de splats ocultos	Oclusão corretamente tratada, sem artefatos
Geometria	Confusão espacial, partes se sobrepõem	Estrutura parcialmente organizada	Geometria coerente, bem separada	Estrutura fiel à cena, sem conflitos
Detalhes Finos	Grade borrada ou ausente	Grade parcialmente recuperada	Grade bem definida, mas cromaticamente instável	Grade nítida e bem renderizada
Consistência Angular	Ruídos visuais variam com o ângulo	Menos sensível ao ponto de vista	Geométrica estável, mas com cor instável	Totalmente estável em todas as perspectivas
Percepção Visual Geral	Instável e artificial	Visualmente mais agradável	Estruturalmente correta, mas com cor incoerente	Natural, limpa e próxima da renderização real
Vantagem Principal	Referência base (3DGS tradicional)	Estabilidade visual e suavidade perceptual	Correção espacial e separação de planos	Equilíbrio completo entre estrutura e aparência
Limitação Principal	Alta instabilidade e artefatos	Confusão estrutural nas regiões ocluídas	Cor instável entre ângulos	Custo computacional e necessidade de ambos os módulos

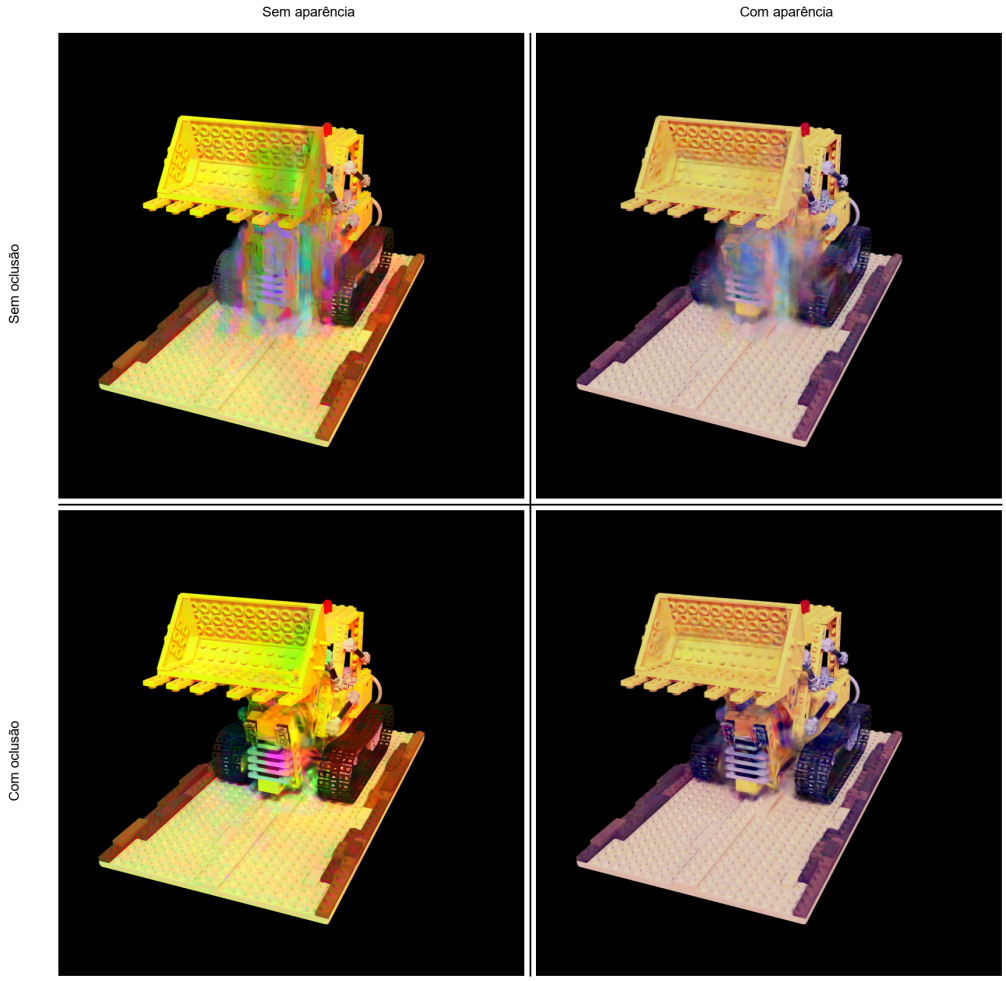

Figura 5.17: Comparação entre diferentes configurações do método proposto aplicadas a uma cena com movimento. À esquerda, modelos sem uso de appearance embeddings; à direita, com uso de aparência. Na linha superior, sem aplicação de máscara de oclusão; na inferior, com oclusão. Nota-se que a combinação de aparência e oclusão (canto inferior direito) produz uma renderização mais coerente, com menor distorção visual e melhor definição estrutural.

6
Conclusões

Reconstruir cenas tridimensionais a partir de imagens capturadas em ambientes não controlados envolve lidar com uma série de imperfeições: variações de aparência, objetos transitórios, oclusões parciais e inconsistências geométricas. Este trabalho apresentou uma extensão ao método *Gaussian Splatting*, com o objetivo de torná-lo mais robusto frente a esses desafios, mantendo aplicabilidade prática e fidelidade visual.

A proposta combinou dois mecanismos complementares: a utilização de *appearance embeddings* para regularizar a aparência das imagens ao longo de múltiplas posições, e a aplicação de máscaras de oclusão geradas automaticamente para excluir regiões inconsistentes da reconstrução. Esses componentes foram integrados ao pipeline do *Gaussian Splatting* sem comprometer sua eficiência, permitindo uma reconstrução mais estável e limpa em contextos visuais diversos.

Os experimentos conduzidos com dados sintéticos e reais demonstraram que a abordagem proposta é capaz de lidar com variações cromáticas, iluminações distintas e objetos transitórios, resultando em representações tridimensionais visualmente mais coerentes. As imagens geradas mantiveram fidelidade estrutural e consistência perceptual, mesmo em condições não ideais de captura.

Além disso, as análises de ablação permitiram compreender o papel individual de cada módulo. Observou-se que a ausência dos componentes propostos leva a artefatos visuais e instabilidades cromáticas, o que reforça sua importância no contexto do modelo. Essa avaliação qualitativa, combinada com métricas perceptuais, sustentou a eficácia da abordagem.

Mais do que aprimorar um método existente, este trabalho buscou investigar formas de tornar a reconstrução 3D mais tolerante à diversidade e imperfeição dos dados reais — um aspecto central para a adoção dessas técnicas em aplicações concretas, como reconstrução urbana, preservação digital, visualização científica ou realidade aumentada.

Limitações e Trabalhos Futuros

Embora os resultados tenham sido satisfatórios, algumas limitações permanecem. A qualidade das máscaras geradas automaticamente pode impactar negativamente a reconstrução, especialmente em casos em que o modelo

de segmentação falha em capturar bordas complexas ou objetos parcialmente ocultos. Além disso, a seleção de pontos de interesse para iniciar o processo de segmentação, embora parcialmente automatizada, ainda exige intervenção manual em alguns cenários.

Outro ponto relevante refere-se ao espaço latente de aparência. Apesar de eficaz na prática, sua estrutura interna ainda carece de mecanismos que favoreçam a interpretação e o controle semântico das variações aprendidas. Trabalhos futuros podem investigar formas de explorar esse espaço com maior profundidade, por meio de visualizações, agrupamentos ou técnicas de análise de variância, permitindo compreender como atributos visuais estão sendo codificados e utilizados pelo modelo.

Em relação ao tratamento de oclusões, uma direção promissora é a integração de segmentação semântica com análise estatística do conjunto de imagens. Em vez de considerar cada imagem de forma isolada, o modelo poderia utilizar o contexto global da cena para identificar quais objetos são persistentes, estruturais e semanticamente relevantes. Elementos inconsistentes — como pessoas, veículos ou objetos móveis — que aparecem esporadicamente poderiam ser filtrados com mais precisão, utilizando critérios baseados em frequência, co-ocorrência ou compatibilidade com padrões dominantes. Essa abordagem permitiria um refinamento mais robusto das máscaras e maior fidelidade na reconstrução final.

Essas investigações podem não apenas aprofundar a compreensão dos mecanismos aqui propostos, como também abrir caminho para reconstruções mais precisas, interpretáveis e adaptáveis a contextos diversos.

Outras possibilidades incluem:

- Aplicação do modelo a sequências temporais, como vídeos ou varreduras em tempo real, avaliando sua estabilidade em cenários dinâmicos;
- Integração com modelos generativos mais expressivos, como redes baseadas em difusão, que poderiam preencher lacunas estruturais ou visuais na cena;
- Emprego de métricas de avaliação perceptual mais alinhadas à experiência humana, incorporando critérios como continuidade local, nitidez e plausibilidade visual.

7
Referências bibliográficas

ALI, M. S. et al. Compression in 3d gaussian splatting: A survey of methods, trends, and future directions. **arXiv e-prints**, p. arXiv–2502, 2025.

BARRON, J. T. et al. **Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields**. 2021.

CARON, M. et al. Emerging properties in self-supervised vision transformers. **arXiv preprint arXiv:2104.14294**, 2021. Disponível em: https://arxiv.org/abs/2104.14294.

CHEN, G.; WANG, W. A survey on 3d gaussian splatting. **arXiv preprint arXiv:2401.03890**, 2024.

CHEN, L.-C. et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 40, n. 4, p. 834–848, 2017.

GOESELE, M. et al. Multi-view stereo for community photo collections. In: IEEE. **2007 IEEE 11th international conference on computer vision**. [S.l.], 2007. p. 1–8.

JIN, X. et al. Lighting every darkness with 3dgs: Fast training and real-time rendering for hdr view synthesis. **Advances in Neural Information Processing Systems**, v. 37, p. 80191–80219, 2024.

KERBL, B. et al. 3d gaussian splatting for real-time radiance field rendering. **ACM Transactions on Graphics**, v. 42, n. 4, July 2023.

KIRILLOV, A. et al. Segment anything. **arXiv preprint arXiv:2304.02643**, 2023.

KNAPITSCH, A. et al. Tanks and temples. **ACM Transactions on Graphics (TOG)**, v. 36, p. 1 – 13, 2017.

KULHANEK, J. et al. Wildgaussians: 3d gaussian splatting in the wild. 2024. Disponível em: https://arxiv.org/abs/2407.08447.

LI, Z. et al. Neuralangelo: High-fidelity neural surface reconstruction. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2023. p. 8456–8465.

MARTIN-BRUALLA, R. et al. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In: **CVPR**. [S.l.: s.n.], 2021.

MILDENHALL, B. et al. Nerf: Representing scenes as neural radiance fields for view synthesis. In: **ECCV**. [S.l.: s.n.], 2020.

MÜLLER, T. et al. Instant neural graphics primitives with a multiresolution hash encoding. **ACM Trans. Graph.**, ACM, New York, NY, USA, v. 41, n. 4, p. 102:1–102:15, jul. 2022.

PFISTER, H. et al. Surfels: Surface elements as rendering primitives. **Proceedings of the 27th annual conference on Computer graphics and interactive techniques**, p. 335–342, 2000.

RADFORD, A. et al. Learning transferable visual models from natural language supervision. In: **Proceedings of the 38th International Conference on Machine Learning (ICML)**. [s.n.], 2021. Disponível em: https://arxiv.org/abs/2103.00020.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. **Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18**. [S.l.], 2015. p. 234–241.

SABOUR, S. et al. Spotlesssplats: Ignoring distractors in 3d gaussian splatting. 2024. Disponível em: https://arxiv.org/abs/2406.20055.

SHEN, L. et al. Gaussian time machine: A real-time rendering methodology for time-variant appearances. **arXiv preprint arXiv:2405.13694**, 2024.

SNAVELY, N.; SEITZ, S. M.; SZELISKI, R. Photo tourism: exploring photo collections in 3d. In: **ACM siggraph 2006 papers**. [S.l.: s.n.], 2006. p. 835–846.

TURKI, H.; RAMANAN, D.; SATYANARAYANAN, M. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2022. p. 12922–12931.

WANG, Y.; WANG, J.; QI, Y. We-gs: An in-the-wild efficient 3d gaussian representation for unconstrained photo collections. 2024. Disponível em: https://arxiv.org/abs/2406.02407.

XU, Q. et al. Point-nerf: Point-based neural radiance fields. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2022. p. 5438–5448.

ZHU, S. et al. 3d gaussian splatting in robotics: A survey. **arXiv preprint arXiv:2410.12262**, 2024.

ZWICKER, M. et al. Ewa splatting. **IEEE Transactions on Visualization and Computer Graphics**, v. 8, n. 3, p. 223–238, 2002.