Relatório do artigo: COLMAP-Free 3D Gaussian Splatting

Revis@r Veronika Treumova
IMPA
veronika.treumova@impa.br

Arqueólog@ Veronika Treumova
IMPA
veronika.treumova@impa.br

Hacker Mateus Barbosa
IMPA
secondauthor@i3.org

Doutorand@ Vitor Pereira Matias
ICMC - USP
vitorpmatias@usp.br

1. Revisão

1.1. Resumo

O artigo aborda o problema de reconstrução de cenas 3D e síntese de vistas novas a partir de vídeos ou imagens não calibradas, sem a necessidade de estimar poses de câmera previamente (usualmente realizadas com ferramentas como COLMAP). Apesar de sua popularidade, o COLMAP apresenta limitações significativas:

- Sensibilidade a Erros de Extração de Recursos: Cenas com texturas pobres ou padrões repetitivos podem levar a falhas no cálculo das poses.
- Custo Computacional Elevado: O processo de SfM pode ser demorado, tornando-o impraticável para aplicações em larga escala ou em cenários dinâmicos.

As abordagens convencionais de campos de radiância neural (NeRF) necessitam de poses precisas das câmeras para otimizar os parâmetros da cena. Isso cria um gargalo quando o pré-processamento falha ou é impraticável. O uso recente de splatting gaussiano 3D, que representa explicitamente a cena como nuvens de pontos, abre oportunidades para lidar com esse problema de maneira eficiente e robusta.

O método, chamado COLMAP-Free 3D Gaussian Splatting (CF-3DGS), propõe a otimização conjunta de poses de câmera e uma representação 3D explícita baseada em splatting gaussiano. Ele constrói e ajusta progressivamente conjuntos de pontos gaussianos a partir de sequências de vídeo, explorando a continuidade temporal para estimar transformações relativas entre quadros consecutivos.

O método que combina:

- Representação Explícita Baseada em Pontos Gaussianos: Utilizando o método recente de splatting gaussiano 3D, que representa cenas como uma coleção de pontos gaussianos, o modelo facilita a estimativa direta das poses de câmera
- Exploração da Continuidade Temporal de Vídeos: O método processa frames sequencialmente, usando transformações afins para estimar movimentos relativos entre quadros consecutivos.

O pipeline é dividido em:

- 3DGS local: para estimar poses relativas entre pares de quadros adjacentes;
- 3DGS global: para agregar essas informações e ajustar a nuvem de pontos para reconstruir a cena de forma progressiva.

Lista de Contribuições:

- Introdução de um método robusto para síntese de vistas e estimativa de poses sem pré-processamento de estrutura a partir do movimento (SfM);
- Aproveitamento da continuidade temporal de vídeos e representações explícitas (pontos gaussianos) para estimativas mais robustas em cenários com movimento complexo;
- Demonstração de superioridade em relação a métodos existentes em termos de qualidade de renderização e precisão de poses, especialmente em cenários de movimento de câmera 360°;
- Redução significativa no tempo de treinamento em comparação com métodos baseados em NeRF.

1.2. Pontos positivos

- O método tem um desempenho significativamente melhor do que as abordagens anteriores sem poses de câmera pré-computadas

- Eficácia e robustez da abordagem em cenas desafiadoras, como vídeos em 360°
- Graças às vantagens do splatting gaussiano, a abordagem atinge rápidas velocidades de treinamento e inferência

1.3. Pontos negativos

A maior desvantagem é que o método otimiza a pose da câmera e o 3DGS em conjunto de maneira sequencial, restringindo assim sua aplicação principalmente a fluxos de vídeo ou coleções ordenadas de imagens.

A qualidade decresce nas estimações conforme mais quadros são treinados, de forma que enquanto o primeiro quadro é reproduzido fielmente, a o modelo com a visão do último quadro apresenta distorções significativas.

Depende de intrínsecos pré-calculados.

1.4. Avaliação

Tem mais vantagens do que desvantagens, material foi apresentado com clareza, nota 4.

2. Arqueólog@

3. Código e experimentos

O artigo apresenta inconsistências em sua implementação. Apesar dos autores proporem a utilização de diferentes detectores de profundidade, a depender do tipo de dataset utilizado, na prática eles utilizam em todos os casos o MiDas.

Além disso, mesmo permitindo o uso de datasets sem intrínsecos dados, nesses casos os intrínsecos são sempre tomados como um parâmetro fixo. Configura-se então o FoV para 79 graus, e os pontos principais são definidos para o centro da imagem.

O código disponibilizado pelos autores funciona sem problemas. Ao treinar o modelo para o dataset Tanks, percebemos que o modelo piora significativamente de qualidade o treinamento avança nos quadros. Atribuímos isso à natureza sequencial do método, que pode ir somando os erros de cada passagem de quadro.

Tendo em vista a definição heurística dos intrínsecos para datasets compostos somente por imagem, realizamos um experimento para verificar qualitativamente a diferença entre um mesmo conjunto de imagens com e sem informações prévias de intrínsecos.

Figure 1. Imagem obtida pelo CF-3DGS com e sem dados prévios de intrínsecos, respectivamente. Fonte: Elaboração própria

A figura 1 mostra que a definição heurística amplifica consideravelmente as distorções obtidas no treinamento dos últimos quadros

4. Projeto de doutorado

Em suma, o artigo que estamos revisando tenta remover a necessidade do uso de softwares Structure-From-Motion como o COLMAP [3] do pipeline de reconstrução de cenas 3D por meio da técnica 3D Gaussian Splatting [2]. Para isso, apresentam uma forma de auto regressão das posições das câmeras que consegue obter resultados similares ao COLMAP. Porém, ainda há a necessidade de obtenção dos parâmetros intrínsicos da câmera como: centro óptico, distância focal e cisalhamento.

Nos dias de hoje ainda não é tão comum encontrar câmeras e aparelhos celulares com acesso aos parâmetros intrínsicos da câmera, além disso, durante a gravação de um video, esses parâmetros podem sofrerem alterações. Então, existe a necessidade do uso de algum software a priori que faça adquira, a partir das imagens, os parâmetros. Entretanto, podemos efetuar simplificações que nos direcionaram para o encontro da matriz intrínsica K de um modo mais rápido.

Aqui, utilizaremos a matriz intrínsica como feito FlowMap [4] em que eles removem o termo de cisalhamento, igualam as distâncias focais $f_x = f_y$ e o centro óptico é o pixel central da imagem. Além disso, a matriz K é a mesma para todas as imagens inseridas no software.

Com essas simplificações é possível rodar a técnica COLMAP-Free com diferentes distâncias focais f e comparando a perca gerada para cada uma das distâncias. Essa comparação pode ser feita de modo ponderado ou de modo a selecionar a melhor. Nota-se que no FlowMap eles utilizam uma ponderação por meio da softmin das percas que é multiplicada para gerar a matriz K final, em termos temos

$$\mathbf{K} = \sum_k w_k \mathbf{K}_k \qquad w_k = \frac{\exp(-\mathcal{L})}{\sum_k \exp(-\mathcal{L}_k)}, \tag{1}$$

em que $\mathcal{L}$ é a perca de uma reconstrução renderizada.

Com essas duas possibilidades, nosso esquema se resume da seguinte forma:

- As poses das câmera são estimadas pelo método original COLMAP-Free
- Intrinsicas são obtidas da seguinte forma:
  - Para cada K em uma amostra de tamanho k reconstrua a cena usando o método 3DGS
  - Calcule a perca.
- Compare as percas de modo singular ou ponderado.
- Escolha um K.

Por fim, para serem definidas as percas, existem duas. (1) A perca fotométrica utilizada pelo COLMAP-Free ou (2) a *Camera-Induced Flow Loss* do FlowMap.

$$\mathcal{L}_{rgb} = (1 - \lambda)\mathcal{L}_1 + \lambda(\mathcal{L}_{D-SSIM}) \tag{2}$$

$$\mathcal{L} = \|\hat{u}_{ij} - u_{ij}\| \tag{3}$$

No geral, essa arquitetura pode ser chamada de RGB-splat por ter apenas com entrada imagens RGB, a Figura 3 exemplifica a arquitetura do modelo. Nota-se que na figura temos apenas a primeira iteração do modelo, mas isso pode ser feito após várias iterações, ou em paralelo.

Camera-Induced Flow Loss: Dado duas imagens de entrada i e j, e uma pixel u_i em i, podemos utilizar o mapa de profundidade e a matriz K para induzir uma posição x_i no espaço 3D. Após isso, podemos utilizar uma a pose estimada P_{ij} entre as imagens e projetar a posição x_i como $P_{ij}x_i$ no plano da imagem j que nos gerará uma correspondência $\hat{u}_{ij}$. A correspondência u_{ij} é conhecida e derivada do fluxo óptico entre imagens em sequência e, pontos esparsos que aparecem por uma longa janela. A Figura 2 exemplifica o funcionamento da função de perca.

5. Conclusões

O presente trabalho introduz uma técnica que remove a necessidade prévia do uso de softwares como o COLMAP, porém, pela necessidade priori da matriz intrínsica da câmera isso é visto de maneira dúbia, desse modo, o artigo consegue com êxito estimar as poses das câmeras, mas não consegue ficar livre do COLMAP. Portanto, o título do trabalho deveria ser revisto, como por exemplo *Pose-Free 3D Gaussian Splatting*.

References

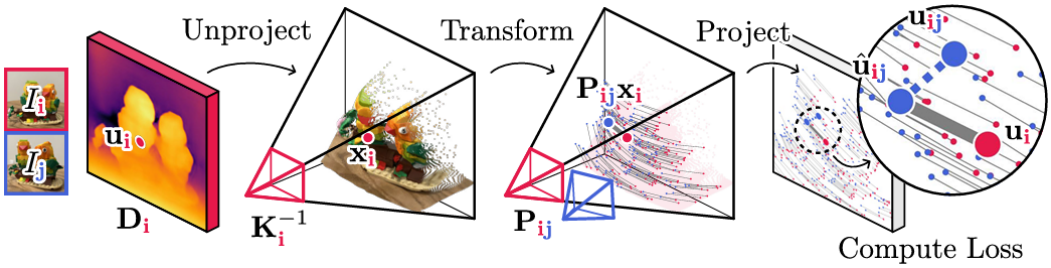
[1] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting, 2024. 4

[2] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[3] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2

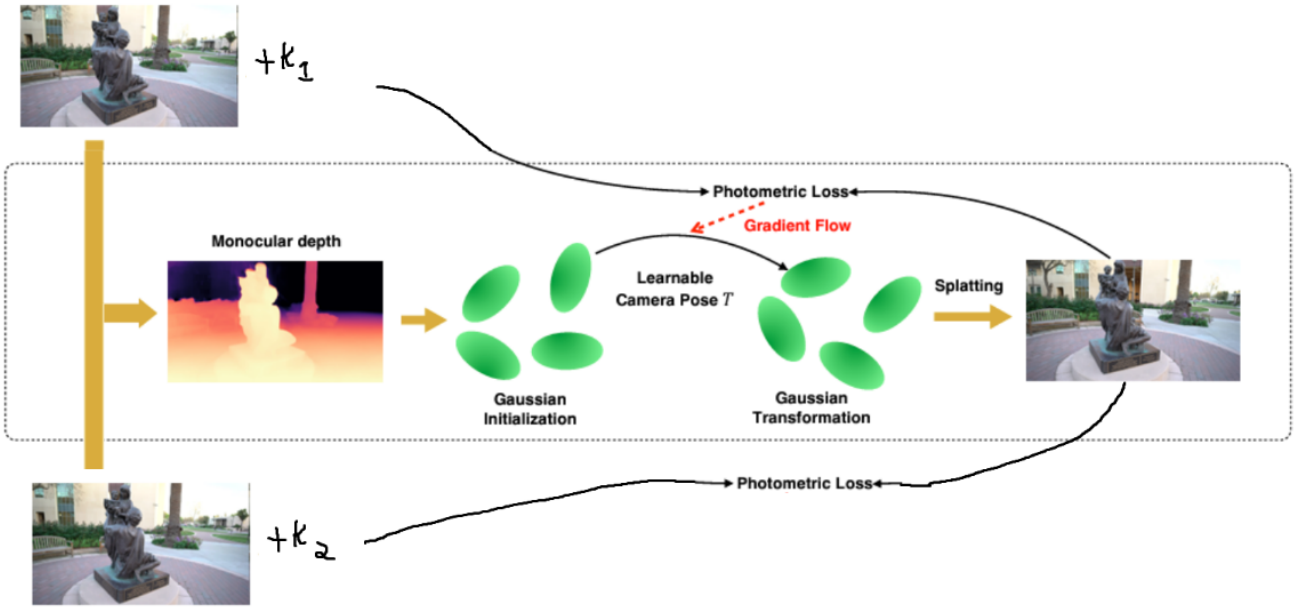
[4] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent, 2024. 2, 4

Figure 2. *Camera-Induced Flow Loss*



Fonte: [4].

Figure 3. Arquitetura do RGBsplat.



Fonte: Modificado de [1].