Relatório do artigo: MVSplat efficient 3d gaussian splatting from sparse multi-view images

Revis@r Vitor Pereira Matias
ICMC - USP
vitorpmatias@usp.br

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

Hacker Davi Guimarães
UFF
davi.guima.castro@outlook.com

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

1. Revisão

Sugestão para o Revis@r: ler as orientações do CVPR. Tentem manter o relatório dentro de 8 páginas.

1.1. Resumo

O trabalho expõe uma nova arquitetura baseada em redes neurais para gerar modelos 3D Gaussian Splatting [5] a partir de imagens esparsas de uma cena. O modelo, chamado de MVSplat, consegue a partir de apenas duas imagens, com suas relativas poses e matrizes intrínsica, gerar um nuvem de gaussianas 3D. Esse é um problema que esta em alta, visto que diversas aplicações possuem apenas fotos esparsas de um objeto, por exemplo as fotos de um e-commerce.

Em resumo, o método consiste em 1) utilizar mapas de profundidade gerados por redes baseadas em transformers com feature matching e cost volume para gerar as posições das gaussians; 2) utilizando duas camadas convolucionais na distribuição de matching entre as imagens, é possível obter a opacidade da imagem; 3) as escalas e cores das gaussianas são obtidas também por duas camadas convolucionais que tem como entrada a concatenação das features das imagens, cost volume, e as imagens originais. A Figure 1 mostra a arquitetura geral do modelo.

Primeiramente, os mapas de profundidade são gerados por uma rede Multiview Swin-Transformer baseada na arquitetura gmflow [6, 11–13], a Figure 2 mostra a respectiva arquitetura. Porém, no caso do MVSplat é necessário encontrar correspondência entre as duas imagens de entrada (ou mais imagens), essa correspondência é definida por uma função de volume de custo baseada na abordagem *plane-sweep* [4], em que, dada uma imagem fonte e uma alvo, elas são projetas em planos que movem pelo espaço, essa projeção gera um custo, que é menor quando a projeção é igual a imagem alvo.

O volume de custo e os mapas de profundidade, podem ter problemas em seções que não possuem características suficientes para diferenciação, com isso, uma etapa adicional de refinamento é performada utilizando uma leve rede 2D-UNet [8, 9].

Após isso, como dito anteriormente, o modelo prediz uma nuvem de gaussianas 3D, para gerar as médias, ou posição central (x,y,z), é utilizado o próprio mapa de profundidade a partir dos parâmetros de câmera (pose e intrínsica); para gerar a opacidade, utilizamos a distribuição do cost volume feature matching; e para gerar as escalas e cores é utilizada uma rede convolucional de duas camadas. No fim, para treinamento, é utilizada as funções de perca $\updownarrow_2$ e LPIPS, com pesos 1 e 0.05, respectivamente.

1.2. Pontos positivos

No geral, o modelo contribuiu fortemente para o estado da arte, gerando resultado que ultrapassam o método anterior, pixelSplat [1], em todos os aspectos. Em velocidade que é duas vezes mais rápido, possuí 10 vezes menos parâmetros, obtém imagens com melhor qualidade, e generaliza muito melhor para K imagens. Além disso, não necessita de tanto pós-processamento igual o pixelSplat, que necessita de uma cuidadosa seleção dos planos *near* e *far*. O modelo também consegue generalizar melhor quando treinado em uma base de dados diferente do teste.

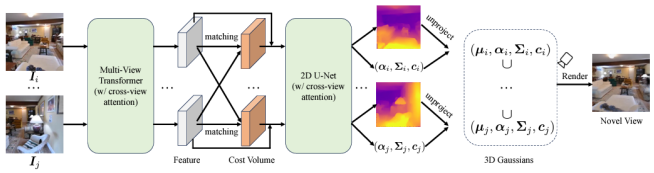

Figure 1. Arquitetura geral do modelo.

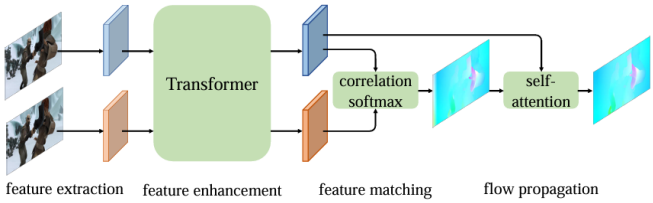

Figure 2. Arquitetura transformer de geração de mapas de profundidade.

1.3. Pontos negativos

O modelo não é capaz de generalizar superfícies não Lambertianas como vidros e espelhos. Além disso, assim como o pixelSplat, a necessidade das poses e matrizes intrínsica das câmeras causam uma grande dificuldade da utilização desses modelos em cenários reais. Ou seja, se quisermos rodar em duas fotos de uma cena qualquer, não iremos conseguir com facilidade.

1.4. Avaliação

Nota-se que pelo trabalho avançar no estado da arte em todos os aspectos, esse artigo deve ser aceito (5).

2. Arqueólog@

O MVSplat utiliza diversas técnicas como fundamentos, nota-se em especial que ele se baseia na técnica de Gaussian Splatting [5], técnicas de obtenção de mapas de profundidade, multi-view stero [11–13], transformers [6], cost volumes [4], e U-Nets [8, 9].

Além desses fundamentos é necessário conhecimento sobre a matemática entorno do método, listando algumas temos:

1. Domínio inverso da profundidade
2. Warp de features de CNN
3. Transformando mapas de profundidade em nuvem de pontos
4. Utilização da função softmax no cost volume feature matching entre as imagens para gerar a opacidade das gaussianas.
5. Aplicação de redes convolucionais para gerar as matrizes de covariância e as cores das gaussianas.
6. Funções de perca, LPIPS e $\updownarrow$.

2.1. Trabalhos anteriores e concorrentes

O MVSplat conseguiu superar o estado da arte da época, pixelSplat, obtendo melhores resultados em todos os aspectos. As grandes melhorias foram na melhor generalização para K vistas, na melhora da qualidade da cena 3D, e a não necessidade de pós processamento. Nota-se que suas arquiteturas não são similares, o pixelSplat utiliza epipolar transformer (ET) para gerar features entre as duas entradas, após isso ele aprende uma distribuição de probabilidade que gera os posicionamentos das gaussianas. Enquanto que o MVSplat utiliza mapas de profundidade para gerar os posicionamentos das gaussianas. A utilização de ET acaba deixando o modelo com muito mais parâmetros para treino, menor velocidade de inferência e uma pior generalização para K vistas.

O trabalho se encontra na área de modelos feed-foward para geração de gaussian splattings. Outros modelos nessa área são o LaRa [2], GS-LRM [15] e GPS-Gaussian [16]. Todos esses modelos necessitam tanto da pose quanto da matriz intrínsica das cameras para gerar os modelos. Dentre esses, o GS-LRM é um modelo altamente escalável que prediz nuvens de gaussianas utilizando entre duas e quatro imagens. O modelo utiliza uma arquitetura transformers com a utilização das poses e da matriz intrínsica das câmera para gerar um encoding de raios de Plücker.

2.2. Trabalhos derivados (que citam)

Alguns trabalhos citam o MVSplat em seus corpos, como o HumanSplat [7] que gera modelos 3D de seres humanos, o V3D [3] que utiliza modelos baseados em difusão de geração de vídeos para gerar malhas 3D ou gaussian splattings, e o Flas3D [10] que reconstrói uma cena 3D com apenas uma imagem. Além desses, o NoPoSplat [14] é o único trabalho que não necessita das poses das câmera, visto que isso é um input extremamente caro de se conseguir, o NoPoSplat é um candidato para evoluir o estado da arte.

O método NoPoSplat também é baseado em transformers, nesse caso ViTs, além disso, ele utiliza uma camada linear para construir um enconding baseado na matriz intrínsica das câmeras. Os ViTs são utilizados em cada uma das imagens mas tem os seus pesos compartilhados entre si e também possuem uma camada de atenção cruzada. Após isso, para cada imagem é gerada uma nuvem de gaussianas que são concatenadas, o que gera um novo modelo 3D da cena, onde novas vistas podem ser renderizadas.

Em comparação com o MVSplat, o NoPoSplat consegue atingir um novo estado da arte, tanto pelo fato de que não há necessidade da pose da câmera, tanto pelo fato de uma melhor qualidade das imagens (PSNR, SSIM, LPIPS) e, também é cerca de 2 vezes mais rápido do que o MVSplat.

Em geral, o MVSplat não cita completamente de onde obteve as ideias do seu artigo, como: obtenção do cost-volume, feature matching warping, refinamento do cost-volume dos mapas de profundidade, geração das opacidades. Nota-se que no artigo gmflow [11], diversas das técnicas estão lá explicadas (não tenho total certeza disso). Uma importante referência de cost-volume está faltando: [4]. Em suma, o artigo está envolto num período curto de estado da arte em comparação com o pixelSplat e piora em comparação com o NoPoSplat.

3. Código e experimentos

Não tive tempo para realizar o relatório.

4. Projeto de doutorado

Problema: O MVSplat pode ser menos eficaz em superfícies não lambertianas e reflexivas.

Foi sugerido o projeto ”MVSplat com Funções de Sombreamento para Superfícies Reflexivas”. A ideia é treinar o modelo com conjuntos de dados mais diversificados e combiná-lo com o artigo *GaussianShader: 3D Gaussian Splatting with Shading Functions for Reflective Surfaces*. GaussianShader é um método inovador que aplica uma função de sombreamento simplificada em Gaussianas 3D para aprimorar a renderização neural em cenas com superfícies reflexivas, preservando a eficiência no treinamento e na renderização.

Plano de pesquisa:

1. Implementar o código do GaussianShader ao MVSplat
2. Treinar em vários conjuntos de dados:
   - NeRF Synthetic
   - Conjuntos de dados de objetos reflexivos: Shiny Blender e Glossy Synthetic
   - Cenas reais em grande escala: Tanks and Temples
   - Conjuntos de dados usados anteriormente
3. Comparar os resultados e, se necessário, ajustar os modelos
4. Verificar se o MVSplat perde significativamente a eficácia

5. Conclusões

References

[1] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction, 2024. 1

[2] Anpei Chen, Haofei Xu, Stefano Esposito, Siyu Tang, and Andreas Geiger. Lara: Efficient large-baseline radiance fields, 2024. 2

[3] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators, 2024. 3

[4] Robert T Collins. A space-sweep approach to true multi-image matching. In *Proceedings CVPR IEEE computer society conference on computer vision and pattern recognition*, pages 358–363. Ieee, 1996. 1, 2, 3

[5] 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. 1, 2

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 1, 2

[7] Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors, 2024. 3

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 2

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 1, 2

[10] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, João F. Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image, 2024. 3

[11] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching, 2022. 1, 2, 3

[12] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation, 2023.

[13] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo, 2018. 1, 2

[14] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images, 2024. 3

[15] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting, 2024. 2

[16] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis, 2024. 2