Pontifícia Universidade Católica de Goiás

Alunos:

João Paulo Moreira Rosa

Igor Guimarães de Oliveira

Luiz Paulo T. Gonçalves

Thiago A. da Silva

Projeto Integrador II-B

Curso: Big Data e Inteligência Artificial

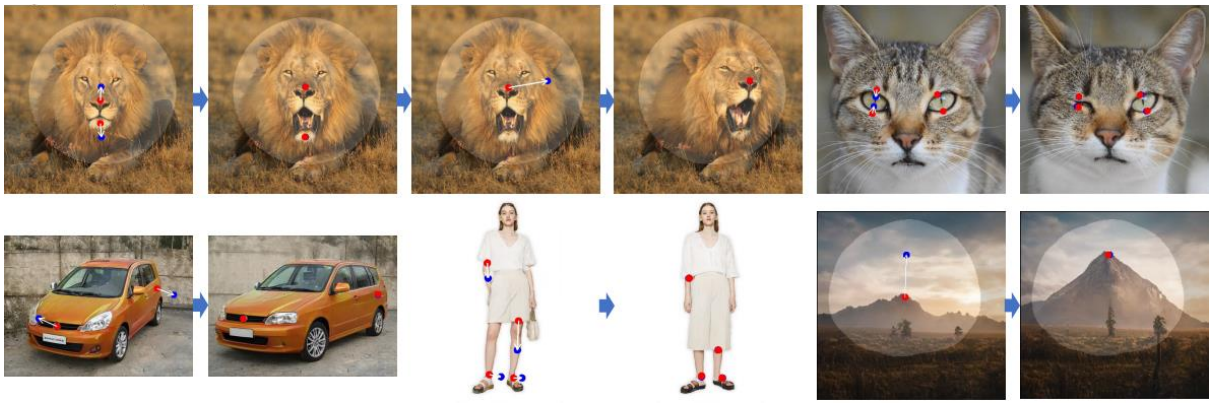
Goiânia,

2023

Manipulação interativa baseada em pontos no Manifold de Imagens Generativas

Fig. 1. Nossa abordagem DragGAN permite aos usuários "arrastarem" o conteúdo de qualquer imagem gerada por GAN. Os usuários só precisam clicar em alguns pontos de controle (vermelhos) e pontos de destino (azuis) na imagem, e nossa abordagem moverá os pontos de controle para alcançar precisamente seus pontos de destino correspondentes. Os usuários também podem opcionalmente desenhar uma máscara da região flexível (área mais clara), mantendo o restante da imagem fixo. Essa manipulação flexível baseada em pontos permite controlar muitos atributos espaciais, como pose, forma, expressão e layout em diversas categorias de objetos.

Imagem Resultado 2º Edição Resultado



Sintetizar conteúdo visual que atenda às necessidades dos usuários frequentemente requer controle flexível e preciso da pose, forma, expressão e layout dos objetos gerados. As abordagens existentes para obter controle sobre redes generativas adversariais (GANs) envolvem o uso de dados de treinamento anotados manualmente ou um modelo 3D prévio, o que geralmente apresenta falta de flexibilidade, precisão e generalidade. Neste trabalho, estudamos uma maneira poderosa, porém muito menos explorada, de controlar GANs, ou seja, "arrastar" quaisquer pontos da imagem para alcançar precisamente os pontos desejados de maneira interativa, conforme mostrado na Figura 1. Para isso, propomos o DragGAN, que consiste em dois componentes principais: 1) uma supervisão de movimento baseada em características que direciona o ponto de controle para se mover em direção à posição desejada, e 2) uma nova abordagem de rastreamento de pontos que utiliza as características discriminativas do gerador para localizar continuamente a posição dos pontos de controle. Através do DragGAN, qualquer pessoa pode deformar uma imagem com controle preciso sobre o destino dos pixels, manipulando assim a pose, forma, expressão e layout de diversas categorias, como animais, carros, humanos, paisagens, etc. Como essas manipulações são realizadas no manifold de imagens geradas por uma GAN, elas tendem a produzir saídas realistas, mesmo para cenários desafiadores, como gerar conteúdo oculto e deformar formas que seguem consistentemente a rigidez do objeto. Comparativos qualitativos e quantitativos demonstram a vantagem do DragGAN em relação às abordagens anteriores nas tarefas de manipulação de imagem e rastreamento de pontos. Também mostramos exemplos da manipulação de imagens reais por meio da inversão de GAN.

1. Introdução

Modelos generativos profundos, como as redes generativas adversariais (GANs) [Goodfellow et al. 2014], têm alcançado um sucesso sem precedentes na síntese de imagens fotorrealistas aleatórias. Em aplicações do mundo real, um requisito funcional crítico desses métodos de síntese de imagens baseados em aprendizado é a capacidade de controlar o conteúdo visual sintetizado. Por exemplo, usuários de redes sociais podem querer ajustar a posição, forma, expressão e pose corporal de um humano ou animal em uma foto casualmente capturada; profissionais de pré-visualização de filmes e edição de mídia podem precisar criar eficientemente esboços de cenas com determinados layouts; e designers de carros podem querer modificar interativamente a forma de suas criações. Para atender a esses diversos requisitos dos usuários, uma abordagem ideal de síntese de imagens controláveis deve possuir as seguintes propriedades: 1) Flexibilidade: deve ser capaz de controlar diferentes atributos espaciais, incluindo posição, pose, forma, expressão e layout dos objetos ou animais gerados; 2) Precisão: deve ser capaz de controlar os atributos espaciais com alta precisão; 3) Generalidade: deve ser aplicável a diferentes categorias de objetos, mas não limitada a uma categoria específica. Enquanto trabalhos anteriores satisfazem apenas uma ou duas dessas propriedades, nosso objetivo neste trabalho é alcançar todas elas.

A maioria das abordagens anteriores obtém a capacidade de controle das GANs por meio de modelos 3D pré-existentes [Deng et al. 2020; Ghosh et al. 2020; Tewari et al. 2020] ou aprendizado supervisionado que depende de dados manualmente anotados [Abdal et al. 2021; Isola et al. 2017; Ling et al. 2021; Park et al. 2019; Shen et al. 2020]. Assim, essas abordagens falham em generalizar para novas categorias de objetos, frequentemente controlam uma faixa limitada de atributos espaciais ou fornecem pouco controle sobre o processo de edição. Recentemente, a síntese de imagens guiada por texto tem atraído atenção [Ramesh et al. 2022; Rombach et al. 2021; Saharia et al. 2022]. No entanto, a orientação por texto carece de precisão e flexibilidade em termos de edição de atributos espaciais. Por exemplo, ela não pode ser usada para mover um objeto por um número específico de pixels.

Para alcançar uma controllabilidade flexível, precisa e genérica das GANs, neste trabalho, exploramos uma poderosa, porém menos explorada, manipulação interativa baseada em pontos. Especificamente, permitimos que os usuários cliquem em qualquer número de pontos de referência (handle points) e pontos alvo (target points) na imagem, e o objetivo é fazer com que os pontos de referência alcancem seus respectivos pontos alvo. Como mostrado na Fig. 1, essa manipulação baseada em pontos permite aos usuários controlar diversos atributos espaciais e é independente de categorias de objetos. A abordagem com configuração mais próxima da nossa é a UserControllableLT [Endo 2022], que também estuda a manipulação baseada em arrastar. Comparado a ela, o problema estudado neste artigo apresenta dois desafios adicionais: 1) consideramos o controle de mais de um ponto, que a abordagem deles não lida adequadamente; 2) exigimos que os pontos de referência alcancem precisamente os pontos alvo, enquanto a abordagem deles não o faz. Conforme mostraremos nos experimentos, lidar com mais de um ponto com controle preciso de posição permite uma manipulação de imagem muito mais diversa e precisa.

Para realizar essa manipulação interativa baseada em pontos, propomos o DragGAN, que aborda dois subproblemas, incluindo: 1) supervisionar os pontos de referência para se moverem em direção aos alvos e 2) rastrear os pontos de referência para que suas posições sejam conhecidas em cada etapa de edição. Nossa técnica é baseada na ideia fundamental de que o espaço de características de uma GAN é suficientemente discriminativo para permitir tanto a supervisão de movimento quanto o rastreamento preciso dos pontos. Especificamente, a supervisão de movimento é alcançada por meio de uma perda de remendo de características deslocadas que otimiza o código latente. Cada etapa de

otimização faz com que os pontos de referência se aproximem dos alvos; assim, o rastreamento de pontos é realizado por meio de uma busca pelo vizinho mais próximo no espaço de características. Esse processo de otimização é repetido até que os pontos de referência alcancem os alvos. O DragGAN também permite que os usuários desenhem opcionalmente uma região de interesse para realizar edições específicas da região. Como o DragGAN não depende de redes adicionais como o RAFT [Teed and Deng 2020], ele realiza manipulações eficientes, levando apenas alguns segundos em uma única GPU RTX 3090 na maioria dos casos. Isso permite sessões de edição ao vivo e interativas, nas quais o usuário pode iterar rapidamente em diferentes layouts até obter o resultado desejado.

Realizamos uma extensa avaliação do DragGAN em diversos conjuntos de dados, incluindo animais (leões, cães, gatos e cavalos), humanos (rosto e corpo inteiro), carros e paisagens. Como mostrado na Fig. 1, nossa abordagem move efetivamente os pontos de referência definidos pelo usuário para os pontos alvo, alcançando efeitos de manipulação diversos em várias categorias de objetos. Ao contrário das abordagens convencionais de deformação de forma que simplesmente aplicam deformações [Igarashi et al. 2005], nossa deformação é realizada no espaço de imagem aprendido de uma GAN, que tende a obedecer às estruturas subjacentes dos objetos. Por exemplo, nossa abordagem pode gerar conteúdo oculto, como os dentes dentro da boca de um leão, e pode deformar seguindo a rigidez do objeto, como a flexão de uma perna de cavalo. Também desenvolvemos uma interface gráfica do usuário (GUI) para que os usuários realizem interativamente a manipulação, simplesmente clicando na imagem. Tanto a comparação qualitativa quanto a quantitativa confirmam a vantagem de nossa abordagem em relação ao UserControllableLT. Além disso, nosso algoritmo de rastreamento de pontos baseado em GAN também supera abordagens de rastreamento de pontos existentes, como o RAFT [Teed and Deng 2020] e o PIPs [Harley et al. 2022] para quadros gerados por GANs. Além disso, ao combinar com técnicas de inversão de GAN, nossa abordagem também se torna uma poderosa ferramenta para edição de imagens reais.

2. Trabalho Relacionado

2.1 Modelos Generativos para Criação Interativa de Conteúdo

A maioria dos métodos atuais utiliza redes generativas adversariais (GANs) ou modelos de difusão para síntese controlável de imagens. GANs incondicionais são modelos generativos que transformam vetores latentes de baixa dimensão, amostrados aleatoriamente, em imagens fotorrealísticas. Eles são treinados utilizando aprendizado adversarial e podem ser utilizados para gerar imagens fotorrealísticas de alta resolução [Creswell et al. 2018; Goodfellow et al. 2014; Karras et al. 2021, 2019]. A maioria dos modelos GAN, como o StyleGAN [Karras et al. 2019], não permite a edição controlável das imagens geradas.

GANs condicionais. Vários métodos propuseram GANs condicionais para abordar essa limitação. Nesses casos, a rede recebe uma entrada condicional, como um mapa de segmentação [Isola et al. 2017; Park et al. 2019] ou variáveis 3D [Deng et al. 2020; Ghosh et al. 2020], além do vetor latente amostrado aleatoriamente, para gerar imagens fotorrealísticas. Em vez de modelar a distribuição condicional, o EditGAN [Ling et al. 2021] permite a edição ao primeiro modelar uma distribuição conjunta de imagens e mapas de segmentação e, em seguida, calcular novas imagens correspondentes a mapas de segmentação editados.

Controle usando GANs incondicionais. Vários métodos foram propostos para a edição de GANs incondicionais por meio da manipulação dos vetores latentes de entrada. Algumas abordagens encontram direções latentes significativas por meio de aprendizado supervisionado a partir de

anotações manuais ou modelos 3D prévios [Abdal et al. 2021; Leimkühler e Drettakis 2021; Patashnik et al. 2021; Shen et al. 2020; Tewari et al. 2020]. Outras abordagens calculam as direções semânticas importantes no espaço latente de forma não supervisionada [Härkönen et al. 2020; Shen e Zhou 2020; Zhu et al. 2023]. Recentemente, a capacidade de controle da posição grosseira do objeto é alcançada por meio da introdução de "blobs" intermediários [Epstein et al. 2022] ou mapas de calor [Wang et al. 2022b]. Todas essas abordagens permitem a edição de atributos semânticos alinhados à imagem, como aparência, ou atributos geométricos grosseiros, como posição e pose do objeto. Embora o Editing-in-Style [Collins et al. 2020] demonstre alguma capacidade de edição de atributos espaciais, isso só é possível transferindo semântica local entre amostras diferentes. Ao contrário desses métodos, nossa abordagem permite aos usuários um controle refinado sobre os atributos espaciais por meio de edição baseada em pontos.

Modelos de Difusão. Mais recentemente, os modelos de difusão [Sohl-Dickstein et al. 2015] possibilitaram a síntese de imagens de alta qualidade [Ho et al. 2020; Song et al. 2020, 2021]. Esses modelos denoisam iterativamente um ruído amostrado aleatoriamente para criar uma imagem fotorrealística. Modelos recentes têm mostrado síntese expressiva de imagens condicionadas a inputs de texto [Ramesh et al. 2022; Rombach et al. 2021; Saharia et al. 2022]. No entanto, a linguagem natural não permite um controle refinado sobre os atributos espaciais das imagens, e, portanto, todos os métodos condicionados a texto estão restritos à edição semântica de alto nível. Além disso, os modelos de difusão atuais são lentos, pois requerem várias etapas de denoising. Embora tenham ocorrido avanços em relação à amostragem eficiente, as GANs ainda são significativamente mais eficientes.

2.2 Rastreamento de Pontos

Para rastrear pontos em vídeos, uma abordagem óbvia é através da estimativa de fluxo óptico entre quadros consecutivos. A estimativa de fluxo óptico é um problema clássico que estima campos de movimento entre duas imagens.

Abordagens convencionais resolvem problemas de otimização com critérios feitos à mão [Brox e Malik 2010; Sundaram et al. 2010], enquanto abordagens baseadas em aprendizado profundo começaram a dominar o campo nos últimos anos devido ao melhor desempenho [Dosovitskiy et al. 2015; Ilg et al. 2017; Teed e Deng 2020]. Essas abordagens baseadas em aprendizado profundo geralmente usam dados sintéticos com fluxo óptico verdadeiro para treinar as redes neurais profundas. Entre eles, o método mais amplamente usados atualmente é o RAFT [Teed e Deng 2020], que estima o fluxo óptico por meio de um algoritmo iterativo. Recentemente, Harley et al. [2022] combinaram esse algoritmo iterativo com uma abordagem convencional de "vídeo de partículas", dando origem a um novo método de rastreamento de pontos chamado PIPs. O PIPs considera informações em vários quadros e, portanto, lida melhor com o rastreamento de longo alcance do que abordagens anteriores. Neste trabalho, mostramos que o rastreamento de pontos em imagens geradas por GANs pode ser realizado sem o uso de nenhuma das abordagens mencionadas anteriormente ou redes neurais adicionais. Revelamos que os espaços de características das GANs são suficientemente discriminativos para que o rastreamento possa ser alcançado simplesmente através da correspondência de características. Embora alguns trabalhos anteriores também aproveitem a característica discriminativa na segmentação semântica [Tritrong et al. 2021; Zhang et al. 2021], somos os primeiros a conectar o problema de edição baseada em pontos à intuição das características discriminativas das GANs e projetar um método concreto. Eliminar modelos de rastreamento adicionais permite que nossa abordagem seja executada de maneira muito mais eficiente para suportar edição interativa. Apesar

da simplicidade de nossa abordagem, mostramos que ela supera as abordagens de rastreamento de pontos de ponta, incluindo o RAFT e o PIPs, em nossos experimentos.

3. Método

LUIZ

4.

5. Conclusão

Apresentamos o DragGAN, uma abordagem interativa para edição intuitiva de imagens baseada em pontos. Nosso método utiliza uma GAN pré-treinada para sintetizar imagens que não apenas seguem precisamente a entrada do usuário, mas também permanecem na manifolds de imagens realistas. Ao contrário de muitas abordagens anteriores, apresentamos uma estrutura geral que não depende de modelagem específica de domínio ou redes auxiliares. Isso é alcançado usando dois elementos novos: uma otimização de códigos latentes que move incrementalmente múltiplos pontos de referência em direção às suas localizações alvo, e um procedimento de rastreamento de pontos para traçar fielmente a trajetória dos pontos de referência. Ambos os componentes utilizam a qualidade discriminativa dos mapas de recursos intermediários da GAN para obter deformações precisas em nível de pixel e desempenho interativo. Demonstramos que nossa abordagem supera o estado da arte na manipulação baseada em GAN e abre novas direções para a edição poderosa de imagens usando priors generativas. Quanto a trabalhos futuros, planejamos estender a edição baseada em pontos para modelos generativos 3D.