

Pontifícia Universidade Católica de Goiás

Projeto Integrador II-B

Manipulação interativa baseada em pontos no Manifold de Imagens Generativas

João Paulo Moreira Rosa

Igor Guimarães de Oliveira

Luiz Paulo T. Gonçalves

Thiago A. da Silva

Curso: Big Data e Inteligência Artificial

Resumo: Busca-se no presente trabalho apresentar o novo modelo de nova inteligência artificial nomeado de DragGAN, o qual é uma abordagem que permite aos usuários manipular imagens geradas por redes adversárias generativas (GANs) de forma interativa. Através do "arrastar" de pontos de controle para pontos de destino, é possível controlar pose, forma, expressão e layout de objetos em diversas categorias. O DragGAN consiste em supervisão de movimento baseada em características e rastreamento de pontos, permitindo deformações precisas e realistas nas imagens. Comparativos mostram que o DragGAN supera abordagens anteriores em manipulação de imagens e rastreamento de pontos.

Goiânia

Junho de 2023

1. Introdução

Em maio de 2023, Xingang Pan, do Instituto Max Planck, publicou o artigo “Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold”. No qual busca apresentar juntamente com outros colaboradores, um novo modelo com redes adversárias generativas (GANs)¹.

O novo modelo, DragGAN, propõe uma abordagem que permite aos usuários manipular imagens geradas por GANs de forma interativa, controlando atributos como posição, forma, expressão e pose. Diferentemente de abordagens anteriores, o DragGAN utiliza uma manipulação baseada em pontos de referência e pontos alvo, garantindo flexibilidade, precisão e generalidade. O método consiste em supervisionar o movimento dos pontos de referência em direção aos pontos alvo e rastrear continuamente a posição dos pontos de referência. Experimentos mostram que o DragGAN é eficaz em diversas categorias de objetos, produzindo manipulações realistas e obedecendo às estruturas subjacentes dos objetos. A abordagem também supera técnicas anteriores de deformação de forma e rastreamento de pontos em quadros gerados por GANs. A interface gráfica do usuário (GUI) permite uma manipulação interativa e eficiente, tornando o DragGAN uma ferramenta poderosa para edição de imagens reais quando combinado com técnicas de inversão de GANs.

2. Referências no desenvolvimento da DragGAN

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. A maioria dos modelos GAN, como o StyleGAN, 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, além do vetor latente amostrado aleatoriamente, para gerar imagens fotorrealísticas. Em vez de modelar a distribuição condicional, o EditGAN 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. Outras abordagens calculam as direções semânticas importantes no espaço latente de forma não supervisionada. Recentemente, a capacidade de controle da posição grosseira do objeto é alcançada por meio da introdução de *blobs* –tipo de dado que funciona somente com informações binárias-intermediários ou mapas de calor. 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 denoizam iterativamente um ruído amostrado aleatoriamente para criar uma imagem fotorrealística. Modelos recentes têm mostrado síntese expressiva de imagens condicionadas a inputs de texto. 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.

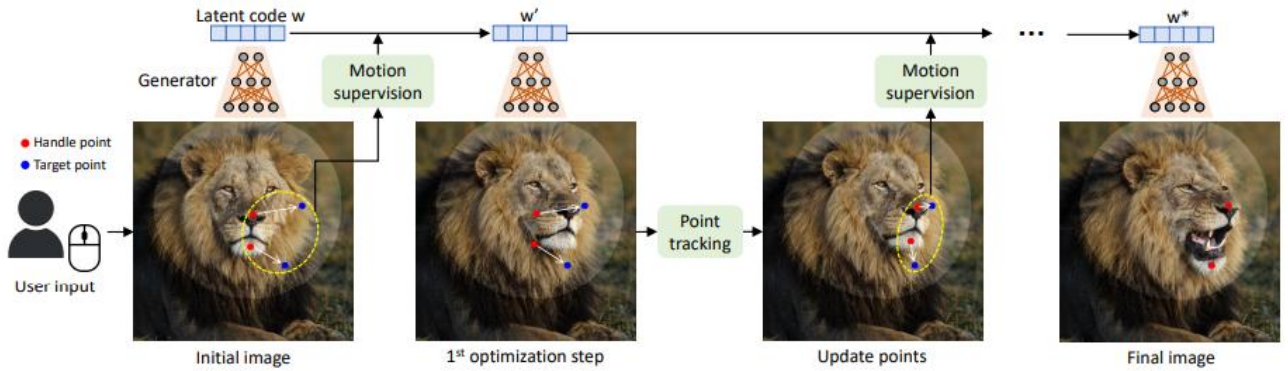
3. Método

Como mencionado anteriormente, o objetivo básico do modelo DragGAN é o deslocamento de pontos de uma determinada imagem, assim inicia-se em um ponto p_i até um ponto alvo t_i , respectivamente:

$$\begin{aligned} & (x_{p,i}, y_{p,i}) | i = 1, 2, \dots, n \} \\ & (x_{t,i}, y_{t,i}) | i = 1, 2, \dots, n \} \end{aligned}$$

Para isso o modelo é baseado na arquitetura StyleGAN2. Seguindo o pipeline da figura a seguir:

Figura 1 - Exemplo do Pipeline da DragGAN



Como pode ser observado na figura 1, o modelo permite aos usuários editar imagens geradas por GAN de forma controlada. Para isso, o usuário precisa definir pontos de controle (vermelhos) e pontos de destino (azuis) na imagem, além de uma máscara opcional que indica a região a ser editada (área mais clara). Método da DragGAN realiza um processo iterativo de supervisão do movimento, no qual os pontos de controle são movidos em direção aos pontos de destino, e um processo de seguimento dos pontos, que atualiza os pontos para acompanhar o objeto na imagem. Esse processo continua até que os pontos de controle alcancem os pontos de destino correspondentes. Dessa forma, permitimos ao usuário realizar edições precisas na imagem, guiando os pontos de controle para as posições desejadas

No qual $\Omega_1(p_i, r_1)$ para designar os pixels cuja distância a p_i é inferior a r_1 , então a perda de supervisão de movimento é:

$$\mathcal{L} = \sum_{i=0}^n \sum_{q_i \in \Omega_1(p_i, r_1)} \|F(q_i) - F(q_i + d_i)\|_1 + \lambda \| (F - F_0) \cdot (1 - M) \|_1$$

No qual, $F(q)$ representa os valores das características de F no pixel q , $d_i = t_i - p_i$ dividido por $\|t_i - p_i\|_2$ é um vetor normalizado que aponta de p_i para t_i ($d_i = 0$ se $t_i = p_i$), e F_0 é a imagem inicial. O primeiro termo é a somatória em todos os pontos de controle $\{p_i\}$. Como os componentes de $q_i + d_i$ não são inteiros, obtém-se $F(q_i + d_i)$ por interpolação bilinear. Assim, ao realizar a retropropagação utilizando esta perda, o gradiente não é retropropagado através de $F(q_i)$. Isso move p_i para $p_i + d_i$, mas não vice-versa. Deslocando a imagem entre o ponto inicial e o ponto escolhido como pode ser observado na figura 2 a seguir:

Figura 2 – Exemplo do modelo DragGAN



4. Experimentos

Conjuntos de dados. Avaliamos nossa abordagem com base no StyleGAN2 pré-treinado nos seguintes conjuntos de dados



Fig. 3. Manipulação de imagens reais. Dada uma imagem real, aplicamos a inversão da GAN para mapeá-la para o espaço latente do StyleGAN e, em seguida, editamos a pose, cabelo, forma e expressão.

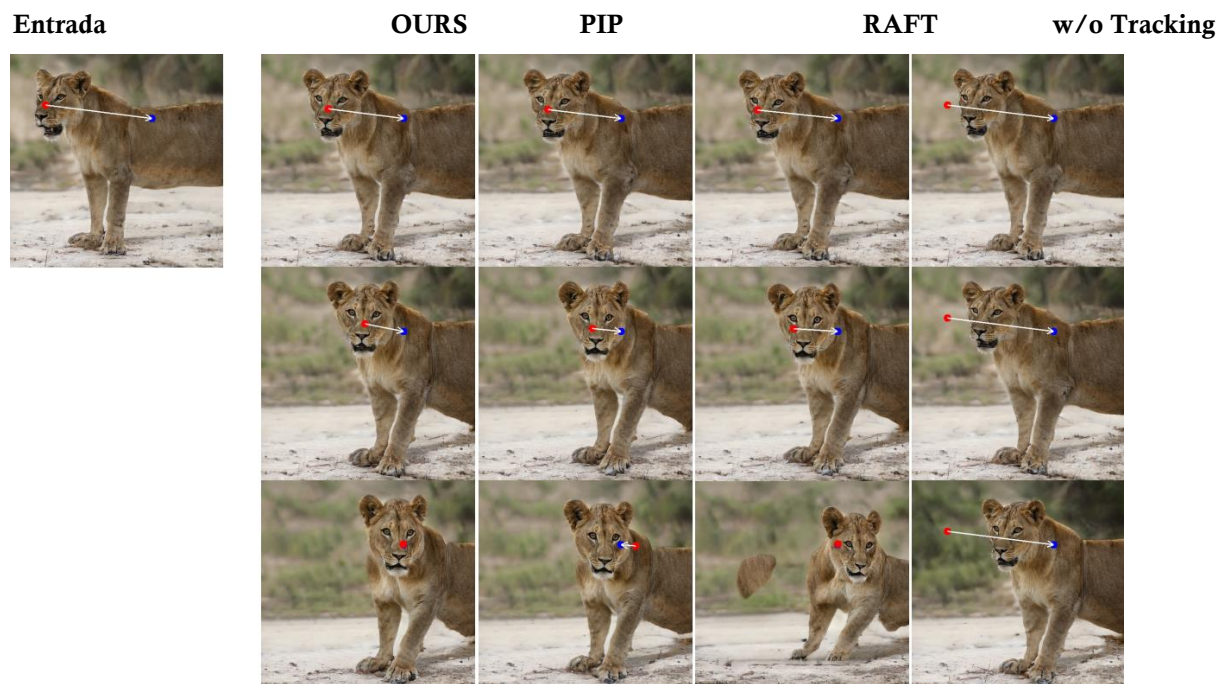


Fig. 4. Comparação qualitativa de rastreamento entre nossa abordagem, RAFT, PIPs e sem rastreamento. Nossa abordagem rastreia o ponto de referência de maneira mais precisa do que as abordagens comparativas, produzindo assim uma edição mais precisa.

4.1 Avaliação Qualitativa

A Figura 4 mostra a comparação qualitativa entre nosso método e o UserControllableLT. Apresentamos os resultados de manipulação de imagem para várias categorias diferentes de objetos e entradas do usuário. Nossa abordagem move com precisão os pontos de referência para atingir os pontos-alvo, alcançando efeitos de manipulação diversos e naturais, como a mudança de pose de animais, a forma de um carro e o layout de uma paisagem. Em contraste, o UserControllableLT não consegue mover fielmente os pontos de referência para os alvos e frequentemente resulta em mudanças indesejadas nas imagens, como as roupas do humano e o fundo do carro. Além disso, ele não mantém a região não mascarada fixa como o nosso método, como mostrado nas imagens dos gatos. Apresentamos mais comparações na Figura 10. Uma comparação entre nossa abordagem, PIPs e RAFT é fornecida na Figura 6. Nossa abordagem rastreia com precisão o ponto de referência acima do nariz do leão, conduzindo-o com sucesso ao alvo.

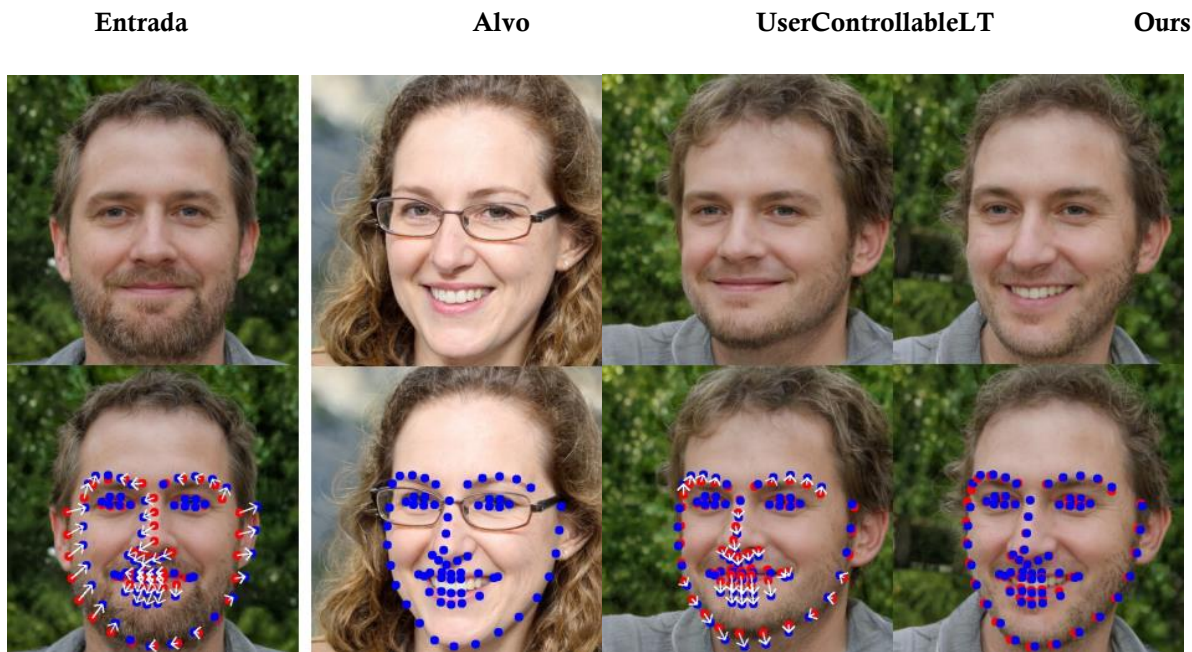


Fig. 7. Manipulação de marcos faciais. Em comparação com o UserControllableLT, este método pode manipular os marcos detectados na imagem de entrada para corresponder aos marcos detectados na imagem de destino com menos erros de correspondência.

Tabela 1. Avaliação quantitativa da manipulação de pontos-chave faciais. Calculamos a distância média entre os pontos editados e os pontos-alvo. O FID e o Tempo são relatados com base na configuração de '1 ponto'.

Método	1 ponto	5 pontos	68 pontos	FID	Tempo(s)
Sem edição	12.93	11.66	16.02	-	-
UserControllableLT	11.64	10.41	10.15	25.32	0.03
Ours w. RAFT tracking	13.43	13.59	15.92	51.37	15.4
Ours w. PIPs tracking	2.98	4.83	5.30	31.87	6.6
Ours	2.44	3.18	4.73	9.28	2.0

No PIPs e no RAFT, o ponto rastreado começa a se desviar do nariz durante o processo de manipulação. Consequentemente, eles movem a parte errada para a posição de destino. Quando nenhum rastreamento é realizado, o ponto fixo de controle logo começa a afetar outra parte da imagem (por exemplo, o fundo) após alguns passos e nunca sabe quando parar, o que falha em alcançar o objetivo de edição. Edição de imagem real. Usando técnicas de inversão de GAN que incorporam uma imagem real no espaço latente do StyleGAN, também podemos aplicar nossa abordagem para manipular imagens reais. A Figura 5 mostra um exemplo em que aplicamos a inversão PTI à imagem real e, em seguida, realizamos uma série de manipulações para editar a pose, cabelo, forma e expressão do rosto na imagem. Mostramos mais exemplos de edição de imagem real na Figura 13.

4.2 Avaliação Quantitativa

Avaliamos quantitativamente nosso método em duas configurações, incluindo manipulação de marcos faciais e reconstrução de imagens em pares.

Manipulação de marcos faciais: Como a detecção de marcos faciais é muito confiável usando uma ferramenta pronta, usamos suas previsões como marcos de referência verdadeiros. Especificamente, geramos aleatoriamente duas imagens faciais usando o StyleGAN treinado no conjunto de dados FFHQ e detectamos seus marcos. O objetivo é manipular os marcos para corresponderem aos marcos de destino.

Tabela 2. Avaliação quantitativa na reconstrução de imagens em pares. Seguimos a avaliação em e relatamos as pontuações de MSE ($\times 102$) \downarrow e LPIPS ($\times 10$) \downarrow .

Dataset	LEÃO		LSUN Gato		Cachorro		LSUN Carro	
Métrica	MSE	LPIPS	MSE	LPIPS	MSE	LPIPS	MSE	LPIPS
UserControllableLT	1.82	1.14	1.25	0.87	1.23	0.92	1.98	0.85
Ours w. RAFT tracking	1.09	0.99	1.84	1.15	0.91	0.76	2.37	0.94
Ours w. PIPs tracking	0.80	0.82	1.11	0.85	0.78	0.63	1.81	0.79
Ours	0.66	0.72	1.04	0.82	0.48	0.44	1.67	0.74



Figura 8. Efeitos da máscara. Nossa abordagem permite mascarar a região móvel. Após mascarar a região da cabeça do cachorro, a parte restante permanecerá praticamente inalterada.

Da primeira imagem para corresponder aos marcos da segunda imagem. Após a manipulação, detectamos os marcos da imagem final e calculamos a distância média (MD) em relação aos marcos de destino. Os resultados são calculados em média ao longo de 1000 testes. O mesmo conjunto de amostras de teste é usado para avaliar todos os métodos. Dessa forma, a pontuação final de MD reflete o quão bem o método pode mover os marcos para as posições de destino. Realizamos a avaliação em três configurações com diferentes números de marcos, incluindo 1, 5 e 68, para mostrar a robustez de nossa abordagem em diferentes números de pontos de referência. Também relatamos a pontuação FID entre as imagens editadas e as imagens iniciais como indicação da qualidade da imagem. Em nossa abordagem e suas variantes, o número máximo de etapas de otimização é definido como 300.

Os resultados são fornecidos na Tabela 1. Nossa abordagem supera significativamente o UserControllableLT em diferentes números de pontos. Uma comparação qualitativa é mostrada na Figura 7, onde nosso método abre a boca e ajusta a forma da mandíbula para corresponder ao rosto de destino, enquanto o UserControllableLT não consegue fazê-lo. Além disso, nossa abordagem preserva uma melhor qualidade de imagem, como indicado pelas pontuações FID. Graças a uma capacidade de rastreamento melhorada, também alcançamos uma manipulação mais precisa do que o RAFT e o PIPs. O rastreamento impreciso também leva a uma manipulação excessiva, o que deteriora a qualidade da imagem, como mostrado nas pontuações FID. Embora o UserControllableLT seja mais rápido, nossa abordagem empurra significativamente o limite desse desafio, alcançando uma manipulação muito mais fiel, mantendo um tempo de execução confortável para os usuários.

Reconstrução de imagens em pares. Nesta avaliação, seguimos a mesma configuração do UserControllableLT. Especificamente, amostramos um código latente w_1 e o perturbamos aleatoriamente para obter w_2 da mesma maneira que em. Sejam I_1 e I_2 as imagens do StyleGAN geradas a partir dos dois códigos latentes. Em seguida, calculamos o fluxo óptico entre I_1 e I_2 e amostramos aleatoriamente 32 pixels do campo de fluxo como entrada do usuário U . O objetivo é reconstruir I_2 a partir de I_1 e U . Relatamos o MSE e o LPIPS e calculamos a média dos resultados ao longo de 1000 amostras. O número máximo de etapas de otimização é definido como 100 em nossa abordagem e suas variantes. Conforme mostrado na Tabela 2, nossa abordagem supera todos os baselines em diferentes categorias de objetos, o que é consistente com resultados anteriores.

Tabela 3. Efeitos de qual recurso usar. $x+y$ significa a concatenação de dois recursos. Relatamos o desempenho (MD) da manipulação de marcos faciais (1 ponto).

Block No	4	5	6	7	5+6	6+7
Motion Sup	2.73	2.50	2.44	2.51	2.47	2.45
Monitorando	3.61	2.55	2.44	2.58	2.47	2.45

Tabela 4. Efeitos de $r1$.

$r1$	1	2	3	4	5
MD	2.49	2.51	2.44	2.45	2.46




Figura 9. Manipulações fora da distribuição. Nossa abordagem possui a capacidade de extrapolação para criar imagens fora da distribuição de imagens de treinamento, por exemplo, uma boca extremamente aberta e uma roda muito ampliada.

Estudo de ablação. Aqui estudamos os efeitos de qual recurso usar na supervisão de movimento e rastreamento de pontos. Relatamos o desempenho (MD) da manipulação de marcos faciais usando diferentes recursos. Conforme mostra a Tabela 3, tanto na supervisão de movimento quanto no rastreamento de pontos, os mapas de recursos após o 6º bloco do StyleGAN apresentam o melhor desempenho, mostrando o melhor equilíbrio entre resolução e discriminabilidade. Também fornecemos os efeitos de $r1$ na Tabela 4. Pode-se observar que o desempenho não é muito sensível à escolha de $r1$ e que $r1 = 3$ apresenta um desempenho ligeiramente melhor.

4.3 Discussões

Efeitos da máscara. Nossa abordagem permite que os usuários insiram uma máscara binária que denota a região móvel. Mostramos seus efeitos na Figura 8. Quando uma máscara sobre a cabeça do cachorro é fornecida, as outras regiões ficam praticamente fixas e apenas a cabeça se move. Sem a máscara, a manipulação move todo o corpo do cachorro. Isso também mostra que a manipulação baseada em pontos frequentemente possui várias soluções possíveis e a GAN tenderá a encontrar a solução mais próxima no espaço de imagens aprendido a partir dos dados de treinamento. A função de máscara pode ajudar a reduzir a ambiguidade e manter certas regiões fixas.

Manipulação fora da distribuição. Até agora, as manipulações baseadas em pontos que mostramos são manipulações "dentro da distribuição", ou seja, é possível satisfazer os requisitos de manipulação com uma

imagem natural dentro da distribuição de imagens do conjunto de dados de treinamento. Aqui, mostramos algumas manipulações fora da distribuição na Figura 9. Pode-se ver que nossa abordagem tem alguma capacidade de extrapolação, criando imagens fora da distribuição de imagens de treinamento, por exemplo, uma boca extremamente aberta e uma roda grande. Em alguns casos, os usuários podem sempre querer manter a imagem dentro da distribuição de treinamento e impedir que ela atinja tais manipulações fora da distribuição. Uma maneira potencial de alcançar isso é adicionar regularização adicional ao código latente w , que não é o foco principal deste artigo.

Limitações. Apesar de ter alguma capacidade de extrapolação, nossa qualidade de edição ainda é afetada pela diversidade dos dados de treinamento. Como exemplificado na Figura 14 (a), criar uma pose humana que se desvia da distribuição de treinamento pode resultar em artefatos. Além disso, pontos de referência em regiões sem textura às vezes sofrem mais desvio no rastreamento, como mostrado na Figura 14 (b)(c). Sugerimos, portanto, escolher pontos de referência ricos em textura, se possível.

Impactos sociais. Como este método pode alterar os atributos espaciais das imagens, ele poderia ser mal utilizado para criar imagens de uma pessoa real com uma pose, expressão ou forma falsa. Assim, qualquer aplicação ou pesquisa que utilize esta abordagem deve respeitar estritamente os direitos de personalidade e regulamentos de privacidade.

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 manifold 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, espera-se estender a edição baseada em pontos para modelos generativos 3D.

Referências Bibliográficas:

KARRAS,R.; SAMULI,L. 2019. A style-based generator architecture for generative adversarial networks. In CVPR. 4401–4410
