Finja até conseguir: análise de rosto na natureza usando apenas dados sintéticos

Erroll Wood* Tadas Baltrusaitis * ĊharlieHewitt Sebastian Dziadzio Matthew Johnson
Virginia Estellers Thomas J. Cashman Jamie Shotton
Microsoft

Abstrato

Demonstramos que é possível realizar visão computacional relacionada ao rosto na natureza usando apenas dados sintéticos. A comunidade há muito desfruta dos benefícios de sintetizar dados de treinamento com gráficos, mas a lacuna de domínio entre dados reais e sintéticos continua sendo um problema, especialmente para rostos humanos. Os pesquisadores tentaram preencher essa lacuna com mistura de dados, adaptação de domínio e treinamento de adversários de domínio, mas mostramos que é possível sintetizar dados com lacuna mínima de domínio, de modo que modelos treinados em dados sintéticos generalizem para real in-the- conjuntos de dados selvagens. Descrevemos como combinar um modelo de face 3D paramétrico gerado processualmente com uma biblioteca abrangente de recursos artesanais para renderizar imagens de treinamento com realismo e diversidade sem precedentes. Treinamos sistemas de aprendizado de máquina para tarefas relacionadas ao rosto, como localização de pontos de referência e análise facial, mostrando que os dados sintéticos podem corresponder com precisão aos dados reais, bem como abrir novas abordagens onde a rotulagem manual seria impossível.

Figura 1. Renderizamos imagens de treinamento de rostos com realismo e diversidade sem precedentes. O primeiro exemplo acima é mostrado junto com a geometria 3D e os rótulos que o acompanham para aprendizado de máquina.

1. Introdução

Quando confrontado com um problema de aprendizado de máquina, o desafio mais difícil geralmente não é escolher o modelo de aprendizado de máquina certo, mas encontrar os dados certos. Isso é especialmente difícil no domínio da visão computacional humana, onde as preocupações sobre a justiça dos modelos e a ética da emprego são fundamentais [31]. Em vez de coletar e rotular dados reais, que são lentos, caros e sujeitos a vieses, pode ser preferível sintetizar dados de treinamento usando computação gráfica [68]. Com dados sintéticos, você pode garantir rótulos perfeitos sem ruído de anotação, gerar rótulos avançados que seriam impossíveis de rotular manualmente e ter controle total sobre a variação e diversidade em um conjunto de dados.

Renderizar seres humanos convincentes é um dos problemas mais difíceis em computação gráfica. Filmes e videogames mostraram que humanos digitais realistas são possíveis, mas com esforço artístico significativo por indivíduo [22, 26]. Embora seja possível gerar infinitas novas imagens faciais com abordagens auto-supervisionadas recentes [27], rótulos correspondentes para aprendizado supervisionado não estão disponíveis. Como resultado, trabalhos anteriores recorreram à sintetização de dados de treinamento facial com simplificações, com resultados que estão longe de serem realistas. Vimos progresso em esforços que tentam cruzar a lacuna de domínio usando adaptação de domínio [60] refinando imagens sintéticas para parecerem mais reais e treinamento adversário de domínio [13] onde modelos de aprendizado de máquina são encorajados a ignorar diferenças entre o sintético e o domínios reais, mas menos trabalho tentou melhorar a qualidade dos próprios dados sintéticos. Sintetizar dados faciais realistas tem sido considerado tão difícil que encontramos a suposição de que os dados sintéticos não podem substituir totalmente os dados reais para problemas na natureza [60].

Neste artigo, demonstramos que as oportunidades para dados sintéticos são muito mais amplas do que imaginadas anteriormente e podem ser alcançadas hoje. Apresentamos um novo método de aquisição de dados de treinamento para rostos – renderizando modelos de rosto 3D com um nível sem precedentes de realismo e diversidade (consulte a Figura 1). Com um arcabouço sintético suficientemente bom, é possível

* Denota contribuição igual.
https://microsoft.github.io/FaceSynthetics

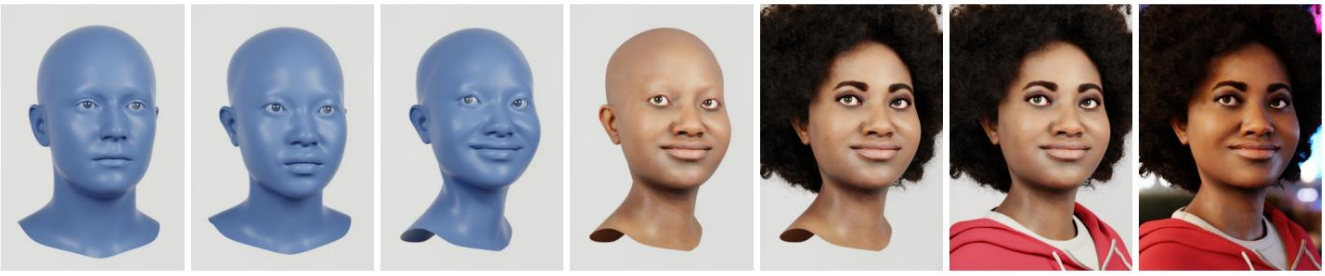

Figura 2. Construímos processualmente faces sintéticas realistas e expressivas. Começando com nosso modelo de rosto, randomizamos a identidade, escolhemos uma expressão aleatória, aplicamos uma textura aleatória, prendemos cabelos e roupas aleatórios e renderizamos o rosto em um ambiente aleatório.

para criar dados de treinamento que podem ser usados para resolver problemas do mundo real, sem usar nenhum dado real.

Requer conhecimento e investimento consideráveis para desenvolver uma estrutura sintética com lacuna mínima de domínio. No entanto, uma vez implementado, torna-se possível gerar uma ampla variedade de dados de treinamento com esforço incremental mínimo. Vamos considerar alguns exemplos; digamos que você tenha passado um tempo rotulando imagens de rosto com pontos de referência. No entanto, de repente você precisa de pontos de referência adicionais em cada imagem. Re-rotular e verificar levará muito tempo, mas com sintéticos, você pode regenerar rótulos limpos e consistentes a qualquer momento. Ou digamos que você esteja desenvolvendo algoritmos de visão computacional para uma nova câmera, por exemplo, uma câmera infravermelha de reconhecimento facial em um telefone celular. Podem existir poucos, se houver, protótipos de hardware, dificultando a coleta de um conjunto de dados. Synthetics permite que você renderize faces de um dispositivo simulado para desenvolver algoritmos e até mesmo guiar o próprio design de hardware.

Sintetizamos imagens de rosto combinando processualmente um modelo de rosto paramétrico com uma grande biblioteca de ativos de alta qualidade criados por artistas, incluindo texturas, cabelos e roupas (consulte a Figura 2). Com esses dados, treinamos modelos para tarefas comuns relacionadas à face: análise facial e localização de pontos de referência. Nossos experimentos mostram que os modelos treinados com um único conjunto de dados sintéticos genéricos podem ser tão precisos quanto aqueles treinados com conjuntos de dados reais específicos da tarefa, alcançando resultados de acordo com o estado da arte. Isso abre a porta para outras tarefas relacionadas ao rosto que podem ser abordadas com confiança com dados sintéticos em vez de reais.

Nossas contribuições são as seguintes. Primeiro, descrevemos como sintetizar dados de treinamento realistas e diversificados para análise facial na natureza, alcançando resultados de acordo com o estado da arte. Em segundo lugar, apresentamos estudos de ablação que validam os passos dados para alcançar o fotorrealismo. O terceiro é o próprio conjunto de dados sintético, que está disponível na página do nosso projeto: https://microsoft.github.io/FaceSynthetics.

2. Trabalho relacionado

Diversos conjuntos de dados faciais são muito difíceis de coletar e anotar. Técnicas de coleta, como rastreamento da Web, representam preocupações significativas de privacidade e direitos autorais. anotação manual ção é propensa a erros e muitas vezes pode resultar em rótulos inconsistentes. Portanto, a comunidade de pesquisa está cada vez mais procurando aumentar ou substituir dados reais por sintéticos.

2.1. Dados faciais sintéticos

A comunidade de visão computacional tem usado dados sintéticos para muitas tarefas, incluindo reconhecimento de objetos [23, 44, 51, 73], compreensão de cena [12, 25, 47, 50], rastreamento ocular [63, 68], rastreamento manual [40, 61] e análise de corpo inteiro [41, 59, 65]. No entanto, relativamente poucos trabalhos anteriores tentaram gerar sintéticos de rosto inteiro usando computação gráfica, devido à complexidade da modelagem da cabeça humana.

Uma abordagem comum é usar um Modelo Morfável 3D (3DMM) [5], uma vez que estes podem fornecer rótulos consistentes para diferentes faces. Trabalhos anteriores focaram em partes do rosto, como a região dos olhos [62] ou a máscara de hóquei [45, 76]. Zeng et ai. [76], Richardson et al. [46], e Sela et al. [58] usaram 3DMMs para renderizar dados de treinamento para reconstruir a geometria facial detalhada. Da mesma forma, Wood et al. [69] renderizou uma região ocular 3DMM para estimativa do olhar. No entanto, como essas abordagens renderizam apenas parte da face, os dados resultantes têm uso limitado para tarefas que consideram toda a face.

Construir modelos paramétricos é desafiador, então uma alternativa é renderizar digitalizações 3D diretamente [4, 55, 62, 68]. Jeni et al. [24] renderizou o conjunto de dados BU-4DFE [74] para alinhamento facial 3D denso, e Kuhnke e Ostermann [30] renderizaram varreduras de cabeça 3D comercialmente disponíveis para estimativa de pose da cabeça . Embora muitas vezes realistas, essas abordagens são limitadas pela diversidade expressa nas próprias varreduras e não podem fornecer rótulos semânticos ricos para aprendizado de máquina.

A manipulação de imagens 2D pode ser uma alternativa ao uso de um pipeline de gráficos 3D. Zhu et al. [79] encaixou um 3DMM nas imagens do rosto e as distorceu para aumentar a pose da cabeça. Noja vanasghari et al. [42] compôs imagens de mãos em rostos para melhorar a detecção de rostos. Essas abordagens podem apenas fazer pequenos ajustes nas imagens existentes, limitando seu uso.

2.2. Treinamento com dados sintéticos

Embora seja comum confiar apenas em dados sintéticos para tarefas de corpo inteiro [54, 59], os dados sintéticos raramente são usados em sua

próprio para aprendizado de máquina relacionado ao rosto. Em vez disso, é primeiro adaptado para torná-lo mais parecido com algum domínio de destino ou usado juntamente com dados reais para pré-treinamento [76] ou modelos de regularização [16, 29]. A razão para isso é a lacuna de domínio - uma diferença nas distribuições entre dados reais e sintéticos que dificulta a generalização [25].

A adaptação de domínio aprendida modifica imagens sintéticas para melhor corresponder à aparência de imagens reais. Shrivastava et ai. [60] usam uma rede de refino adversária para adaptar imagens oculares sintéticas com regularização para preservar as anotações.
Da mesma forma, Bak et al. [3] adaptar dados sintéticos usando um Ciclo GAN [77] com um termo de regularização para preservação de identidades. Uma limitação da adaptação do domínio aprendido é a tendência da semântica da imagem mudar durante a adaptação [15], daí a necessidade de regularização [3, 40, 60]. Essas técnicas são , portanto, inadequadas para anotações refinadas, como rótulos por pixel ou coordenadas precisas de pontos de referência.

Em vez de adaptar os dados, é possível aprender características que são resistentes às diferenças entre os domínios [13, 57].
Wu e outros. [71] misturam dados reais e sintéticos por meio de um classificador de domínio para aprender recursos invariantes de domínio para detecção de texto, e Saleh et al. [56] exploram a observação de que a forma é menos afetada pela lacuna de domínio do que a aparência para a segmentação semântica da cena.

Em nosso trabalho, não realizamos nenhuma dessas técnicas e, em vez disso, minimizamos a lacuna de domínio na fonte, gerando dados sintéticos altamente realistas.

3. Sintetizando imagens faciais

A indústria de Efeitos Visuais (VFX) desenvolveu muitas técnicas para convencer o público de que os rostos 3D são reais, e desenvolvemos essas técnicas em nossa abordagem. No entanto, uma diferença fundamental é a escala: embora VFX possa ser usado para um punhado de atores, exigimos diversos dados de treinamento de milhares de indivíduos sintéticos. Para resolver isso, usamos a geração procedural para criar e renderizar aleatoriamente novas faces 3D sem nenhuma intervenção manual.

Começamos por amostrar um modelo de face 3D generativo que captura a diversidade da população humana. Em seguida, 'vestimos' aleatoriamente cada rosto com amostras de grandes coleções de cabelos, roupas e acessórios. Todas as coleções são amostradas independentemente para criar indivíduos sintéticos tão diversos quanto possível uns dos outros. Esta seção descreve os componentes técnicos que construímos para permitir coleções de ativos que podem ser misturadas e combinadas sobre faces 3D de maneira aleatória, mas plausível.

3.1. modelo de rosto 3D

Nosso modelo generativo de rosto 3D captura como o formato do rosto varia na população humana e muda durante as expressões faciais. É um face rig baseado em blendshape semelhante ao trabalho anterior [17, 34], e compreende uma malha de N = 7, 667

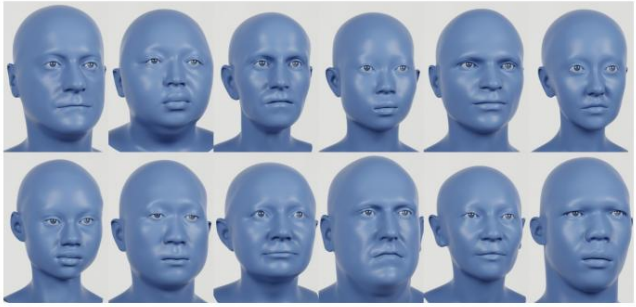

Figura 3. Faces 3D amostradas de nosso modelo generativo, demonstrando como nosso modelo captura a diversidade da população humana.

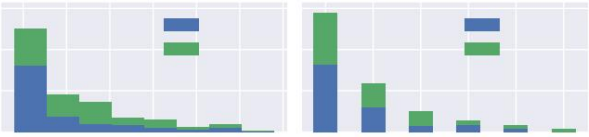

Figura 4. Histogramas de idade, gênero e etnia autorrelatados em nossa coleção de varreduras, que foi usada para construir nosso modelo facial e biblioteca de texturas. Nossa coleção abrange uma variedade de idades e etnias.

vértices e 7.414 polígonos, e um esqueleto mínimo de K = 4 articulações: cabeça, pescoço e dois olhos.

As posições dos vértices da malha de face são definidas pela função geradora de malha M(ÿ, ÿ, ÿ): R |ÿ|×|ÿ|×| 3 ÿ R que N × ÿ| toma parâmetros ÿ ÿ R |ÿ| para identidade, ÿ ÿ R |ÿ| para expressão , e ÿ ÿ R para pose esquelética. Os parâmetros de pose ÿ são rotações locais por junta representadas como ângulos de Euler. K × 3 M é definido como

$$M(\ddot{y}, \ddot{y}, \ddot{y}) = L(T(\ddot{y}, \ddot{y}), \ddot{y}, J(\ddot{y}); W)$$

onde L(X, ÿ, J;W) é uma pele de mistura linear padrão (LBS) [33] que gira as posições dos vértices X ÿ R sobre N×3 as localizações das juntas J K × 3 por rotações de juntas locais ÿ, ÿ R com pesos por vértice W ÿ K×N determinando como rota R ções são interpoladas através da malha. T (ÿ, ÿ):R |ÿ|×|ÿ|ÿ R N×3 constrói uma malha de face na pose de ligação adicionando qual deslocamentos para a malha do modelo TÿR envia N×3 , representação a face média com expressão neutra:

$$T(\ddot{y}, \ddot{y}) = \bar{T}^{ij}_{k} + \ddot{y}^{i}S^{ij}_{k} + \ddot{y}^{i}E^{ij}_{k}$$

base de identidade linear dada S ÿ R |ÿ|×N×3 e base de expressão E ÿ R | ÿ|×N×3 . Observe o uso da notação de soma de Einstein nesta definição e abaixo. Finalmente, J (ÿ):R |ÿ| move as localizações das juntas modelo ÿ R K × 3 J ÿ R conta para mudanças na identidade: K × 3 para ac

$$J(\ddot{y})^{j}_{k} = \bar{J}^{ij}_{k} + W^{il}_{k}\ddot{y}^{i}S^{jl}_{k}.$$

Aprendemos a base de identidade S a partir de varreduras 3D de alta qualidade de M = 511 indivíduos com expressão neutra. Cada varredura

Figura 5. Nós “limpamos” manualmente as varreduras de cabeça 3D de alta resolução brutas para remover ruídos e cabelos. Usamos as varreduras limpas resultantes para construir nosso modelo de geometria generativa e biblioteca de texturas.

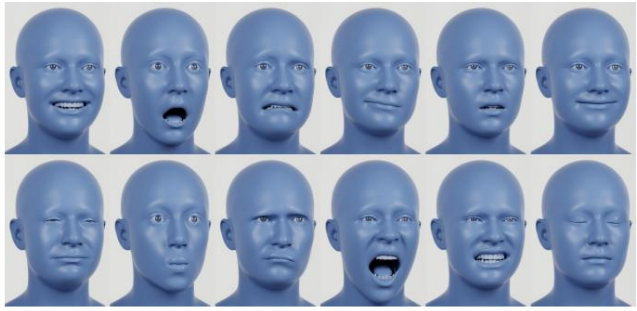

Figura 6. Exemplos de nossa biblioteca de expressões orientadas por dados e sequências animadas manualmente, visualizadas em nossa face de modelo.

foi limpo (ver Figura 5) e registrado na topologia de T usando software comercial [52], resultando no conjunto de dados de treinamento V ÿ RM×3N . Em seguida, ajustamos conjuntamente a base de identidade S e os parâmetros [ÿ 1, . . . , ÿM] para V. Para gerar novas formas de rosto, ajustamos uma distribuição normal multivariada aos parâmetros de identidade ajustados e extraímos dela (consulte a Figura 3).
Como é comum em animação por computador, tanto a base de expressão E quanto os pesos de esfola W foram criados por um artista e são mantidos fixos durante o aprendizado de S.

3.2. Expressão

Aplicamos expressões aleatórias a cada rosto para que nossos modelos de aprendizado de máquina downstream sejam robustos ao movimento facial . Usamos duas fontes de expressão facial. Nossa fonte primária é uma biblioteca de 27.000 parâmetros de expressão {ÿ i} construídos ajustando um modelo de rosto 3D a um corpus de imagens 2D com pontos de referência faciais anotados. No entanto, como os pontos de referência anotados são esparsos, não é possível recuperar todos os tipos de expressão apenas desses pontos de referência, por exemplo, bochechas inchadas. Portanto, também amostramos expressões de uma sequência animada manualmente que foi projetada para preencher as lacunas em nossa biblioteca de expressões, exercitando o rosto de maneiras realistas, mas extremas. A Figura 6 mostra amostras de nossa coleção de expressões. Além da expressão facial, colocamos em camadas direções aleatórias do olhar sobre as expressões amostradas e usamos a lógica processual para posicionar as pálpebras de acordo.

3.3. Textura

Rostos sintéticos devem parecer realistas mesmo quando vistos de uma distância extremamente próxima, por exemplo, por uma câmera de rastreamento ocular em um dispositivo montado na cabeça. Para conseguir isso, coletamos 200 conjuntos de texturas de alta resolução (8192 × 8192 px)

Figura 7. Aplicamos deslocamento aproximado e meso ao nosso modelo de face 3D para garantir que as faces pareçam realistas mesmo quando vistas de perto.

Figura 8. Nossa biblioteca de cabelos contém uma gama diversificada de cabelos, sobrancelhas e barbas. Ao montar um rosto 3D, escolhemos o estilo de cabelo e a aparência aleatoriamente.

de nossas varreduras faciais limpas. Para cada varredura, extraímos uma textura de albedo para a cor da pele e dois mapas de deslocamento (consulte a Figura 7). O mapa de deslocamento grosseiro codifica a geometria de varredura que não é capturada pela natureza esparsa de nosso modelo de identidade em nível de vértice. O mapa de meso-deslocamento aproxima os detalhes do nível dos poros da pele e é construído pela filtragem passa-alta da textura albedo, assumindo que os pixels escuros correspondem a partes ligeiramente recuadas da pele.

Ao contrário do trabalho anterior [45, 76], não construímos um modelo generativo de textura, pois esses modelos lutam para produzir fielmente detalhes de alta frequência, como rugas e poros. Em vez disso, simplesmente escolhemos um conjunto correspondente de albedo e texturas de deslocamento de cada varredura. As texturas são combinadas em um material de pele de base física com dispersão subsuperficial [9]. Finalmente, opcionalmente, aplicamos efeitos de maquiagem para simular sombra, delineador e rímel.

3.4. Cabelo

Em contraste com outros trabalhos que aproximam o cabelo com texturas ou geometria grosseira [17, 55], representamos o cabelo como fios 3D individuais, com uma cabeça cheia de cabelo compreendendo mais de 100.000 fios. Modelar o cabelo no nível do fio nos permite capturar efeitos realistas de iluminação de vários caminhos.
Mostrado na Figura 8, nossa biblioteca de cabelo inclui 512 estilos de cabelo, 162 sobrancelhas, 142 barbas e 42 conjuntos de cílios. Cada recurso foi criado por um artista especializado em criar cabelos digitais. Na hora da renderização, combinamos aleatoriamente o couro cabeludo, a sobrancelha, a barba e os cílios.

Usamos um sombreador de cabelo processual com base física para

Figura 9. Cada rosto está vestido com uma roupa aleatória montada a partir de nosso guarda-roupa digital – uma coleção de diversas roupas 3D e ativos acessórios que podem ser ajustados em torno de nosso modelo de cabeça 3D.

Figura 10. Usamos HDRIs para iluminar o rosto. O mesmo rosto pode parecer muito diferente sob diferentes iluminações.

modelar com precisão as propriedades materiais complexas do cabelo [8]. Este shader nos permite controlar a cor do cabelo com parâmetros de melanina [38] e cinza, e ainda nos permite pintar ou descolorir o cabelo para penteados menos comuns.

3.5. Roupas

As imagens de rostos geralmente incluem o que alguém está vestindo, então vestimos nossos rostos com roupas 3D. Nosso guarda-roupa digital contém 30 roupas de parte superior do corpo que foram criadas manualmente usando design de roupas e software de simulação [10]. Conforme mostrado na Figura 9, essas roupas incluem roupas formais, casuais e esportivas. Além das roupas para a parte superior do corpo, vestimos nossos rostos com chapelaria (36 itens), facewear (7 itens) e óculos (11 itens), incluindo capacetes, lenços de cabeça, máscaras faciais e óculos. Todos os itens de vestuário foram criados em uma malha de corpo nu com as proporções médias do corpo masculino ou feminino [37] em uma postura relaxada.

Deformamos as roupas com uma técnica de deformação baseada em gaiola não rígida [2] para que se ajustem perfeitamente em diferentes formatos de faces. Os óculos são montados com um esqueleto e posicionados usando cinemática inversa, de modo que as têmporas e a ponte do nariz fiquem nas partes correspondentes do rosto.

3.6. Renderização

Renderizamos imagens faciais com Cycles, um renderizador fotorrealista de traçado de raios [6]. Posicionamos aleatoriamente uma câmera ao redor da cabeça e apontamos para o rosto. A distância focal e a profundidade de campo são variadas para simular diferentes câmeras e lentes. Empregamos iluminação baseada em imagem [11] com alto

Figura 11. Exemplos de faces sintéticas que geramos e renderizamos aleatoriamente para uso como dados de treinamento.

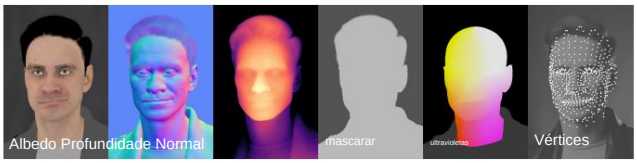

Figura 12. Também sintetizamos rótulos para aprendizado de máquina. Acima estão os tipos de rótulos adicionais além daqueles mostrados na Figura 1.

imagens de faixa dinâmica (HDRI) para iluminar o rosto e fornecer um plano de fundo (consulte a Figura 10). Para cada imagem, escolhemos aleatoriamente de uma coleção de 448 HDRIs que incluem uma variedade de ambientes diferentes [75]. Consulte a Figura 11 para obter exemplos de faces renderizadas com nossa estrutura.

Além de renderizar imagens coloridas, geramos rótulos de informações básicas (consulte a Figura 12). Enquanto nossos experimentos na seção 4 se concentram em anotações de referência e segmentação, a síntese nos permite criar facilmente uma variedade de rótulos ricos e precisos que permitem novas tarefas relacionadas à face (consulte a subseção 4.5).

4. Análise facial

Avaliamos nossos dados sintéticos em duas tarefas comuns de análise de faces: análise de faces e localização de pontos de referência. Mostramos que os modelos treinados em nossos dados sintéticos demonstram desempenho competitivo ao estado da arte. Observe que todas as avaliações que usam nossos modelos são conjuntos de dados cruzados – treinamos puramente em dados sintéticos e testamos em dados reais, enquanto o estado da arte avalia dentro do conjunto de dados, permitindo que os modelos aprendam possíveis vieses nos dados.

4.1. Metodologia de treinamento

Renderizamos um **único** conjunto de dados de treinamento para localização de pontos de referência e análise facial, compreendendo 100.000 imagens com resolução de 512 × 512 . Demorou 48 horas para renderizar usando 150 GPUs NVIDIA M60.

Durante o treinamento, realizamos aumento de dados, incluindo rotações, distorções de perspectiva, desfoques, modulações de brilho e contraste, adição de ruído e conversão para escala de cinza. Esses aumentos são especialmente importantes para imagens sintéticas que, de outra forma, são livres de imperfeições (consulte a subseção 4.4). Enquanto alguns deles podem ser feitos no tempo de renderização, nós os executamos no tempo de treinamento para aleatoriamente

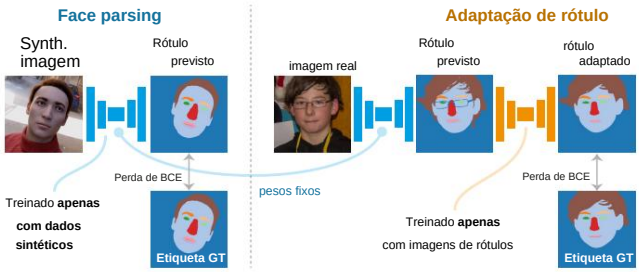

Figura 13. Treinamos uma rede de análise facial (usando apenas dados sintéticos) seguida por uma rede de adaptação de rótulos para lidar com diferenças sistemáticas entre rótulos sintéticos e anotados por humanos.

aplicar diferentes ampliações à mesma imagem de treinamento. Implementamos redes neurais com PyTorch [43] e as treinamos com o otimizador Adam [28].

4.2. Análise facial

A análise facial atribui um rótulo de classe a cada pixel em uma imagem, por exemplo, pele, olhos, boca ou nariz. Avaliamos nossos dados de treinamento sintético em dois conjuntos de dados de análise de faces: **He len** [32] é o benchmark mais conhecido na literatura. Ele contém 2.000 imagens de treinamento, 230 imagens de validação e 100 imagens de teste, cada uma com 11 classes. Devido a erros de rotulagem no conjunto de dados original, usamos Helen* [35], uma versão retificada popular do conjunto de dados que apresenta rótulos de treinamento corrigidos, mas deixa os rótulos de teste inalterados para uma comparação justa. **LaPa** [36] é um conjunto de dados lançado recentemente que usa os mesmos rótulos de Helen, mas tem mais imagens e exibe expressões, poses e oclusões mais desafiadoras.

Ele contém 18.176 imagens de treinamento, 2.000 imagens de validação e 2.000 imagens de teste.

Como é comum [35, 36], usamos as marcas de terreno 2D fornecidas para alinhar as faces antes do processamento. Nós dimensionamos e cortamos cada imagem para que os pontos de referência fiquem centralizados em uma região de interesse de 512 × 512px . Após a previsão, desfazemos essa transformação para calcular os resultados em relação à anotação do rótulo original, sem qualquer redimensionamento ou corte.

Método Tratamos a análise facial como tradução de imagem para imagem. Dada uma imagem de cor de entrada x contendo classes C , desejamos prever uma imagem de rótulo de canal C yˆ das mesmas dimensões espaciais que corresponda à imagem de rótulo de verdade y. Os pixels em y são codificados one-hot com o índice da classe verdadeira. Para isso, utilizamos uma UNet [49] com encoder ResNet-18 [21, 72]. Treinamos essa rede apenas com dados sintéticos, minimizando uma perda de entropia cruzada binária (BCE) entre imagens de rótulos preditas e de verdade. Observe que não há nada de novo em nossa escolha de arquitetura ou função de perda, esta é uma abordagem bem compreendida para esta tarefa.

Adaptação de rótulos. É provável que existam pequenas diferenças sistemáticas entre rótulos sintéticos e rótulos anotados por humanos. Por exemplo, onde exatamente está o limite entre o nariz e o resto do rosto? Avaliar

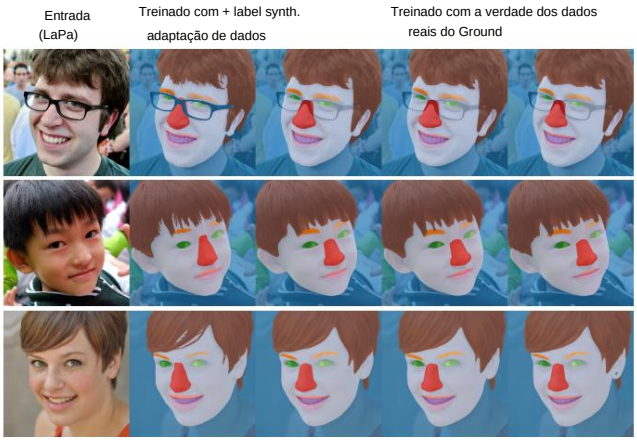

Figura 14. Resultados da análise de faces por redes treinadas com dados sintéticos (com e sem adaptação de rótulos) e dados reais. A adaptação de rótulos aborda diferenças sistemáticas entre rótulos sintéticos e reais, por exemplo, o formato da classe do nariz ou a granularidade do cabelo.

nossos dados sintéticos sem precisar ajustar cuidadosamente nosso processo de geração de rótulos sintéticos para um conjunto de dados real específico, usamos a adaptação de rótulos. A adaptação de rótulo transforma os rótulos previstos por nossa rede de análise facial (treinada apenas com dados sintéticos) em rótulos mais próximos da distribuição no conjunto de dados real (consulte a Figura 13). Tratamos a adaptação de rótulos como outra tarefa de tradução de imagem para imagem e usamos uma UNet com codificador ResNet18 [72]. Para garantir que este estágio não seja capaz de 'trapacear', ele é treinado apenas em pares de rótulos previstos yˆ e rótulos de verdade y. Ele é treinado inteiramente separadamente da rede de análise facial e nunca vê nenhuma imagem real.

Resultados Consulte as Tabelas 1 e 2 para comparações com o estado da arte e a Figura 14 para alguns exemplos de previsões . Embora as redes treinadas com nossos dados sintéticos genéricos não superem o estado da arte, é notável que elas alcancem resultados semelhantes ao trabalho anterior treinado no conjunto de dados em dados específicos da tarefa.

Comparação com dados reais. Também treinamos uma rede na parte de treinamento de cada conjunto de dados real para separar nossa metodologia de treinamento de nossos dados sintéticos, apresentados como “Nossos (reais)” nas Tabelas 1 e 2. Pode-se ver que o treinamento apenas com dados sintéticos produz resultados comparáveis ao treinamento com dados reais.

4.3. Localização de ponto de referência

A localização de marco encontra a posição dos pontos faciais de interesse em 2D. Avaliamos nossa abordagem no conjunto de dados **de 300 W** [53] , que é dividido em subconjuntos comum (554 imagens), desafiador (135 imagens) e privado (600 imagens) .

Método Nós treinamos um ResNet34 [21] com perda de erro quadrático médio para prever diretamente 68 coordenadas de marco 2D por imagem. Usamos as caixas delimitadoras fornecidas para extrair uma região de interesse de 256 × 256 pixels de cada imagem. O

Tabela 1. Uma comparação com o estado da arte no conjunto de dados Helen, usando pontuação F1 . Como é comum, as pontuações para cabelos e outras categorias refinadas são omitidas para ajudar na comparação com trabalhos anteriores. A pontuação geral é calculada mesclando as categorias de nariz, sobrancelhas, olhos e boca. O treinamento com nossos dados sintéticos alcança resultados em linha com o estado da arte, treinados com dados reais.

Método	Pele	Nariz	Lábio superior	Boca interna	Lábio inferior	Sobrancelhas	Olhos	Boca	Geral
Guo et al. [19] AAAI'18	93,8	94,1	89,8	89,6	83,7	91,4	83,7	91,0	84,9
Wei et al. [67] TIP'19	95,6	95,2	80,4	92,4	90,5				
Lin et al. [35] CVPR'19	94,5	95,6	86,7	82,6	93,6	91,6			
Liu et al. [36] AAAI'20	94,9	95,8	86,7	83,1	95,0	92,4			
Te et al. [64] ECCV'20	94,6	96,1	89,1	83,5	96,1	93,1			
Nosso (real)	95,1	94,7	95,1	94,5	83,6	89,8	90.2	95,5	93,2
Nosso (sintético)	81,6	87,0	88,9	81,5	87,6	94,8	91,6		
	82.3	89.1	89,9	83,5	87,3	95,1	92,0		

Tabela 2. Comparação com o estado da arte no LaPa, usando o escore F1 . Para olhos e sobrancelhas, L e R são esquerda e direita. Para os lábios, U, I e L são superiores, internos e inferiores. O treinamento com nossos dados sintéticos alcança resultados em linha com o estado da arte, treinados com dados reais.

Método	Pele	Cabelo	Olho D	Olho D	Lábio em U	Boca em D	Lábio em D	Nariz	Sobrancelha em D	Sobrancelha	Média
Liu et al. [36] AAAI'20	97,2	96,3	88.1	88,0	84,4	85,7	95,5	87,6	89,8	87,6	87,7
Te et al. [64] ECCV'20	97,3	96,2	89,5	90,0	88.1	90,0	89,0	97,1	86,5	87,0	91.1
Nosso (real)	97,5	86,9	91,4	91,5	87,3	89,8	89,4	96,9	89,3	89,3	90,9
Nosso (sintético)	97,1	85,7	90,6	90.1	85,9	88,8	88,4	96,7	88,6	88,5	90.1

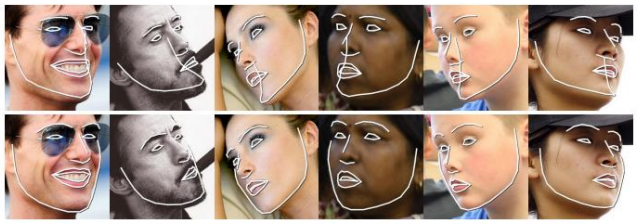

Figura 15. Previsões antes (em cima) e depois (em baixo) da adaptação do rótulo . A principal diferença é mudar o queixo de uma projeção 3D para 2D para seguir o contorno facial na imagem.

Figura 16. Previsões por redes treinadas com dados reais (superior) e sintéticos (inferior). Observe como a rede de dados sintéticos generaliza melhor em expressão, iluminação, pose e oclusão.

Tabela 3. Resultados de localização de referência nos subconjuntos comuns, desafiadores e privados de 300 W. Mais baixo é melhor em todos os casos. Observe que a taxa de 0,5 FR se traduz em 3 imagens, enquanto 0,17 corresponde a 1.

Método		Comum NME	Desafiador NME	Particular NME	FR10%
DenseReg [20]	CVPR'17	-	-	3,67	
LAB [70]	CVPR'18	2,98	5.19	0,83	
AWING [66]	ICCV'19	2.72	4.52	0,33	
ODN [78]	CVPR'19	3,56	6.67	-	
Laplace KL [48]	ICCV'19	3.19	6.87	-	
3FabRec [7]	CVPR'20	3.36	5.74	0,17	
Nossa (verdadeiro)		3,37	5,77	1,17	
Nosso (sintético)		3,09	4,86	0,50	
Estudos de ablação					
Sem aumento		4.25	7.87	4,00	
Aumento da aparência		3,93	6,80	1,83	
Sem cabelo ou roupas		3.36	5.37	2.17	
Sem roupas		3.20	5.09	1,00	
adaptação de rótulo (synth.)		5,61	8,43	4,67	
Sem adaptação de rótulo (real)		3,44	5,71	1,17	

o conjunto privado não tem caixas delimitadoras, por isso usamos um recorte apertado em torno dos pontos de referência.

A adaptação do rótulo é realizada usando um perceptron de duas camadas para abordar diferenças sistemáticas entre rótulos de referência sintéticos e reais (Figura 15). Esta rede nunca é exposta a nenhuma imagem real durante o treinamento.

Resultados Como métricas de avaliação utilizamos: Erro Médio Normalizado (NME) [53] – normalizado pela distância interocular externa do olho; e Taxa de Falha abaixo de um limite de erro de 10% (FR10%).

Consulte a Tabela 3 para comparações com o estado da arte no conjunto de dados de 300 W. É claro que a rede treinada com nossos dados sintéticos pode detectar pontos de referência com precisão comparável a métodos recentes treinados com dados reais.

Comparação com dados reais Aplicamos nossa metodologia de treinamento (incluindo aumentos de dados e adaptação de rótulos) às partes de treinamento e validação do conjunto de dados de 300 W, para comparar dados reais e sintéticos mais diretamente. A Tabela 3 mostra claramente que o treinamento com dados sintéticos leva a melhores resultados, mesmo quando comparado a um modelo treinado com dados reais e avaliado dentro do conjunto de dados.

4.4. estudos de ablação

Investigamos o efeito do **tamanho do conjunto de dados** sintético na precisão do ponto de referência. A Figura 17 mostra que a localização do ponto de referência melhora à medida que aumentamos o número de imagens de treinamento,

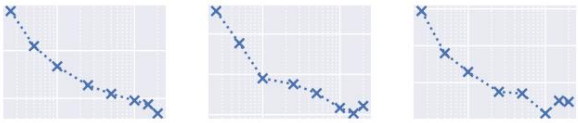

Figura 17. A precisão da localização do ponto de referência melhora à medida que usamos mais e mais dados de treinamento sintéticos.

Figura 18. É fácil gerar dados de treinamento sintéticos para rastreamento ocular (à esquerda) que se generalizam bem para imagens do mundo real (à direita).

antes de começar a estabilizar em 100.000 imagens.

Estudamos a importância do **aumento de dados** ao treinar modelos em dados sintéticos. Treinamos modelos com: 1) sem aumento; 2) apenas aumento da aparência (por exemplo, mudanças de cor, brilho e contraste); 3) aumento total, variando tanto a aparência quanto a geometria (por exemplo, rotação e deformação). A Tabela 3 mostra a importância do aumento, sem o qual os dados sintéticos não superam os reais.

A Tabela 3 também mostra a importância da **adaptação do rótulo** ao avaliar modelos treinados em dados sintéticos – usar a adaptação do rótulo para melhorar a consistência do rótulo reduz o erro.
Adicionar adaptação de rótulo a um modelo treinado em dados reais resulta em pouca mudança no desempenho, mostrando que não beneficia rótulos já consistentes dentro do conjunto de dados.

Se removermos **roupas e cabelos,** a precisão do ponto de referência é prejudicada (Tabela 3). Isso comprova a importância de nossa biblioteca de cabelos e guarda-roupa digital, que melhoram o realismo de nossos dados.

Estudos de ablação adicionais que analisam o impacto da qualidade da renderização e a variação de pose, expressão e identidade podem ser encontrados no material complementar.

4.5. Outros exemplos

Além dos resultados quantitativos acima, esta seção demonstra qualitativamente como podemos resolver problemas adicionais usando nossa estrutura de face sintética.

O rastreamento ocular pode ser um recurso fundamental para dispositivos de realidade virtual ou aumentada, mas dados de treinamento reais podem ser difíceis de adquirir [14]. Como nossos rostos parecem realistas em close, é fácil para nós configurar uma câmera de rastreamento ocular sintético e renderizar diversas imagens de treinamento, juntamente com a verdade do terreno. A Figura 18 mostra exemplos de dados de treinamento sintético para tal câmera, junto com resultados para segmentação semântica.

Marcos densos. Na subseção 4.3, apresentamos os resultados da localização de 68 marcos faciais. E se quiséssemos prever dez vezes mais marcos? Seria impossível

Figura 19. Com dados sintéticos, podemos facilmente treinar modelos que prevêem com precisão dez vezes mais pontos de referência do que o normal. Aqui estão alguns exemplos de previsões de marcos densos no conjunto de dados de 300 W.

ble para um ser humano anotar tantos pontos de referência de forma consistente e correta. No entanto, nossa abordagem nos permite gerar facilmente rótulos de pontos de referência densos e precisos. A Figura 19 mostra os resultados da modificação de nossa rede de referência para regredir 679 coordenadas em vez de 68 e treiná-la com dados sintéticos.

4.6. Discussão

Mostramos que é possível obter resultados comparáveis com o estado da arte para duas tarefas bem conhecidas: análise facial e localização de pontos de referência, sem usar uma única imagem real durante o treinamento. Isso é importante, pois abre a porta para muitas outras tarefas relacionadas à face que podem ser abordadas usando dados sintéticos no lugar de dados reais.

Limitações permanecem. Como nosso modelo de rosto paramétrico inclui apenas a cabeça e o pescoço, não podemos simular roupas com decotes baixos. Não incluímos efeitos de enrugamento dependentes de expressão, então o realismo sofre durante certas expressões . Como amostramos partes de nosso modelo de forma independente, às vezes obtemos combinações incomuns (mas não impossíveis), como rostos femininos com barba. Planejamos abordar essas limitações com trabalhos futuros.

A renderização fotorrealista é computacionalmente cara, então devemos considerar o custo ambiental. Para gerar o conjunto de dados usado neste artigo, nosso cluster de GPU usou aproximadamente 3.000 kWh de eletricidade, equivalente a cerca de 1,37 toneladas métricas de CO_2, 100% do qual foi compensado por nosso provedor de computação em nuvem. Esse impacto é mitigado pelo progresso contínuo dos provedores de computação em nuvem para se tornarem negativos em carbono e usarem fontes de energia renováveis [1, 18, 39].
Há também o custo financeiro a considerar. Considerando US$ 1 por hora para uma GPU M60 (preço médio entre provedores de nuvem), custaria US$ 7.200 para renderizar 100.000 imagens. Embora isso pareça caro, os custos reais de coleta de dados podem ser muito maiores, especialmente se levarmos em consideração a anotação.

Agradecimentos Agradecemos a Pedro Urbina, Jon Hanzelka, Rodney Brunet e Panagiotis Giannakopoulos por suas contribuições artísticas. Este trabalho foi publicado no ICCV 2021. A lista de autores na biblioteca digital do IEEE está faltando VE e MJ devido a um erro nosso durante o processo de publicação.

Referências

[1] Amazônia. Compromisso climático da Amazônia. https://www. aboutamazon.com/planet/climate-pledge, 2021. 8 See More

[2] GR Anderson, MJ Aftosmis e M. Nemec. Deformação Paramétrica de Geometria Discreta para Projeto de Forma Aerodinâmica. Journal of Aircraft, 2012. 5 [3] S. Bak, P. Carr e J.-F. Lalonde. Adaptação de Domínio através de Síntese para Reidentificação de Pessoa Não Supervisionada. In ECCV, 2018. 3

[4] T. Baltrusaitis, P. Robinson e L.-P. Morency. Modelo local restrito 3D para rastreamento facial rígido e não rígido . In CVPR, 2012. 2 [5] V. Blanz e T. Vetter. Um modelo morfável para a síntese de faces 3D. Em Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, páginas 187–194 , 1999. 2 [6] Blender Foundation. Renderizador de ciclos. https:// www. cycles-renderer.org/, 2021. 5 See More

[7] B. Browatzki e C. Wallraven. 3FabRec: Alinhamento facial rápido em poucos disparos por reconstrução. In CVPR, 2020. 7 [8] MJ-Y. Chiang, B. Bitterli, C. Tappan e B. Burley. Um modelo de cabelo e pêlo prático e controlável para rastreamento de caminho de produção. In Computer Graphics Forum, 2016. 5 [9] PH Christensen. Um perfil de refletância aproximado para dispersão subsuperficial eficiente. In SIGGRAPH Talks, 2015. 4 [10] CLO Virtual Fashion Inc. Designer maravilhoso. https: //www.marvelousdesigner.com/, 2021. 5 See More

[11] P. Debevec. Iluminação baseada em imagem. Nos Cursos SIGGRAPH, 2006. 5

[12] A. Gaidon, Q. Wang, Y. Cabon e E. Vig. VirtualWorlds como proxy para análise de rastreamento de vários objetos. In CVPR, 2016. 2 [13] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand e V. Lempitsky. Treinamento adversário de domínio de redes neurais. JMLR, 2016. 1, 3 [14] SJ Garbin, Y. Shen, I. Schuetz, R. Cavin, G. Hughes e SS Talathi. OpenEDS: conjunto de dados de olho aberto. arXiv preprint arXiv:1905.03702, 2019. 8 [15] SJ Garbin, M. Kowalski, M. Johnson e J. Shotton. Adaptação de domínio zero-shot de alta resolução de imagens faciais renderizadas sinteticamente. In ECCV, 2020. 3 [16] B. Gecer, B. Bhattarai, J. Kittler e T.-K. Kim. Aprendizagem adversarial semi-supervisionada para gerar imagens faciais fotorrealistas de novas identidades a partir de modelos 3D morfáveis. In ECCV, 2018. 3

[17] T. Gerig, A. Forster, C. Blumer, B. Egger, M. Luthi, ¨ S. Schonborn e T. Vetter. Modelos faciais morfáveis - uma estrutura aberta. Reconhecimento automático de rostos e gestos, 2017. 3, 4 [18] Google. Sustentabilidade da nuvem do Google. https://cloud. google.com/sustainability/, 2021. 8 [19] T. Guo, Y. Kim, H. Zhang, D. Qian, B. Yoo, J. Xu, D. Zou, J.-J. Han e C. Choi. Rede de decodificador de codificador residual e antes adaptável para análise de face. In AAAI, 2018. 7 [20] RA Guler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou e I. Kokkinos. DenseReg: Totalmente Convolucional Dense Shape Regression In-the-Wild, 2017. 7

[21] K. He, X. Zhang, S. Ren e J. Sun. Aprendizado Residual Profundo para Reconhecimento de Imagem. In CVPR, 2016. 6

[22] D. Hendler, L. Moser, R. Battulwar, D. Corral, P. Cramer, R. Miller, R. Cloudsdale e D. Roble. Vingadores: Capturando o rosto complexo de Thanos. Em SIGGRAPH Talks, 2018. 1 [23] T. Hodan, V. Vineet, R. Gal, E. Shalev, J. Hanzelka, T. Connell, P. Urbina, SN Sinha e B. Guenter. Síntese de imagens fotorrealistas para detecção de instâncias de objetos. Em 2019 IEEE International Conference on Image Processing (ICIP), páginas 66–70. IEEE, 2019. 2 [24] LA Jeni, JF Cohn e T. Kanade. Alinhamento facial 3D denso a partir de vídeos 2D em tempo real. No 11º IEEE Internacional Conferência e Workshops sobre Reconhecimento Automático de Gestos e Faces (FG), 2015. 2 [25] A. Kar, A. Prakash, M.-Y. Liu, E. Cameracci, J. Yuan, M. Rusiniak, D. Acuna, A. Torralba e S. Fidler. Meta Sim: Aprendendo a Gerar Conjuntos de Dados Sintéticos. In ICCV, 2019. 2, 3 [26] B. Karis, T. Antoniades, S. Caulkin e V. Mastilovic. Humanos digitais : cruzando o vale misterioso em ue4. Conferência de desenvolvedores de jogos , 2016. 1

[27] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen e T. Aila. Analisar e melhorar a qualidade de imagem do Style GAN. Em CVPR, 2020. 1

[28] DP Kingma e J. Ba. Adam: Um método para otimização estocástica. arXiv preprint arXiv:1412.6980, 2014. 6 [29] M. Kowalski, SJ Garbin, V. Estellers, T. Baltrusaitis, M. Johnson e J. Shotton. Config: Geração de imagem facial neural controlável. In ECCV, 2020. 3 [30] F. Kuhnke e J. Ostermann. Estimativa de pose de cabeça profunda usando imagens sintéticas e adaptação de domínio adversário parcial para espaços de rótulos contínuos. In CVPR, 2019. 2 [31] K. Karkk ¨ ainen e J. Joo. Fairface: conjunto de dados de atributos faciais para raça, gênero e idade equilibrados. Em WACV, 2021. 1

[32] V. Le, J. Brandt, Z. Lin, L. Bourdev e TS Huang. Inter localização de características faciais ativas. In ECCV, 2012. 6

[33] JP Lewis, M. Cordner e N. Fong. Pose Space Deformation: Uma Abordagem Unificada para Interpolação de Forma e Deformação Acionada por Esqueleto. Em SIGGRAPH, 2000. 3 [34] T. Li, T. Bolkart, MJ Black, H. Li e J. Romero. Aprender um modelo de forma e expressão facial a partir de digitalizações 4D. SIGGRAPH Ásia, 2017. 3

[35] J. Lin, H. Yang, D. Chen, M. Zeng, F. Wen e L. Yuan. Face Parsing com RoI Tanh-Warping. In CVPR, 2019. 6, 7 [36] Y. Liu, H. Shi, H. Shen, Y. Si, X. Wang e T. Mei. Um novo conjunto de dados e segmentação semântica de atenção de limite para análise facial. Em AAAI, 2020. 6, 7 [37] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll e MJ Preto. SMPL: Um modelo linear de várias pessoas com capa. SIG GRAPH Asia, 2015. 5 [38] I. Lozano, J. Saunier, S. Panhard e G. Loussouarn. A diversidade da cor do cabelo humano avaliada por escalas visuais e medições instrumentais. uma pesquisa mundial. Jornal internacional de ciência cosmética, 39:101–107, 2017. 5 [39] Microsoft. negativo até 2030. https://blogs. microsoft.com/blog/ Microsoft será carbono 2020/01/16/ microsoft-will-be-carbon-negative-by-2030/,

2021. 8 See More

[40] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas e C. Theobalt. Mãos GANificadas para Rastreamento de Mãos 3D em Tempo Real de Monocular RGB. In CVPR, 2018. 2, 3

[41] H. Ning, W. Xu, Y. Gong e T. Huang. Aprendizagem discriminativa de palavras visuais para estimativa de pose humana 3D. In CVPR, 2003. 2 [42] B. Nojavanasghari, T. Baltrusaitis, CE Hughes, e L.-P. Morency. Hand2face: Síntese automática e reconhecimento de oclusões de mão sobre face. In ACII, 2017. 2 [43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. PyTorch: um estilo imperativo, biblioteca de aprendizado profundo de alto desempenho. NeurIPS, 2019. 6 [44] W. Qiu, F. Zhong, Y. Zhang, S. Qiao, Z. Xiao, TS Kim, Y. Wang e A. Yuille. Unrealcv: mundos virtuais para visão computacional. Petição ACM Multimedia Open Source Software , 2017. 2

[45] E. Richardson, M. Sela e R. Kimmel. Reconstrução facial 3D aprendendo com dados sintéticos. Conferência Internacional sobre Visão 3D, 2016. 2, 4 [46] E. Richardson, M. Sela, R. Or-El e R. Kimmel. Aprendendo reconstrução facial detalhada a partir de uma única imagem. In CVPR, 2017. 2

[47] SR Richter, V. Vineet, S. Roth e V. Koltun. Jogando por dados: verdade fundamental de jogos de computador. Na Conferência Europeia sobre Visão Computacional, páginas 102–118. Springer, 2016. 2

[48] JP Robinson, Y. Li, N. Zhang, Y. Fu, e S. Tulyakov. Localização de marcos de Laplace. In ICCV, 2019. 7

[49] O. Ronneberger, P. Fischer e T. Brox. U-net: redes convolucionais para segmentação de imagens biomédicas. Na Conferência Internacional sobre Computação de Imagens Médicas e Intervenção Assistida por Computador, 2015. 6 [50] G. Ros, L. Sellart, J. Materzynska, D. Vazquez e AM Lopes. O Conjunto de Dados SYNTHIA: Uma Grande Coleção de Imagens Sintéticas para Segmentação Semântica de Cenas Urbanas. In CVPR, 2016. 2

[51] A. Rozantsev, V. Lepetit e P. Fua. Na renderização de imagens sintéticas para treinar um detector de objetos. CVIU, 2014. 2 [52] Russian3DScanner. Enrole3. https://www.russian3dscanner.com/, 2021. 4 [53] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou e M. Pantic. 300 faces Desafio selvagem: banco de dados e resultados. Image and Vision Computing (IMAVIS), 2016. 6, 7 [54] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa e H. Li. PIFu: Função implícita alinhada a pixels para digitalização humana vestida de alta resolução. In ICCV, outubro de 2019. 2

[55] S. Saito, T. Simon, J. Saragih e H. Joo. PIFuHD: função implícita alinhada a pixels de vários níveis para digitalização humana 3D de alta resolução. In CVPR, 2020. 2, 4 [56] FS Saleh, M. Sadegh Aliakbarian, M. Salzmann, L. Peters son e JM Alvarez. Uso Efetivo de Dados Sintéticos para Segmentação Semântica de Cena Urbana. In ECCV, 2018. 3 [57] S. Sankaranarayanan, Y. Balaji, A. Jain, SN Lim e R. Chellappa. Aprendendo com dados sintéticos: abordando Deslocamento de domínio para segmentação semântica. In CVPR, 2018. 3 [58] M. Sela, E. Richardson e R. Kimmel. Reconstrução irrestrita da geometria facial usando tradução de imagem para imagem. ICCV, 2017. 2

[59] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman e A. Blake. Reconhecimento de pose humana em tempo real em partes de imagens de profundidade única. In CVPR, 2011. 2

[60] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang e R. Webb. Aprendendo com imagens simuladas e não supervisionadas por meio de treinamento adversário. In CVPR, 2017. 1, 3 [61] T. Simon, H. Joo, I. Matthews e Y. Sheikh. Detecção de ponto-chave manual em imagens únicas usando bootstrapping de múltiplas visualizações. Em Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, páginas 1145–1153, 2017. 2

[62] Y. Sugano, Y. Matsushita e Y. Sato. Aprendizado por Síntese para Estimativa de Olhar 3D Baseada na Aparência. In CVPR, 2014. 2

[63] L. Swirski e NA Dodgson. Renderização de imagens de verdade sintéticas para avaliação do rastreador ocular. In ETRA, 2014. 2

[64] G. Te, Y. Liu, W. Hu, H. Shi e T. Mei. Aprendizagem de representação gráfica com reconhecimento de arestas e raciocínio para análise de faces. In ECCV, 2020. 7 [65] G. Varol, J. Romero, X. Martin, N. Mahmood, MJ Black, I. Laptev e C. Schmid. Aprendendo com humanos sintéticos. In CVPR, 2017. 2

[66] X. Wang, L. Bo e L. Fuxin. Perda de asa adaptável para alinhamento robusto da face via regressão de mapa de calor. In ICCV, 2010. 7

[67] Z. Wei, S. Liu, Y. Sun e H. Ling. Análise precisa da imagem facial em velocidade em tempo real. IEEE Transactions on Image Processing, 28(9):4659–4670, 2019. 7 [68] E. Wood, T. Baltrusaitis, X. Zhang, Y. Sugano, P. Robinson e A. Bulling. Renderização de olhos para registro da forma do olho e estimativa do olhar. In ICCV, 2015. 1, 2 [69] E. Wood, T. Baltrusaitis, L.-P. Morency, P. Robinson e A. Bulling. Aprendendo um estimador de olhar baseado em aparência a partir de um milhão de imagens sintetizadas. In ETRA, 2016. 2 [70] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai e Q. Zhou. Veja Boundary: um algoritmo de alinhamento facial com reconhecimento de limites. In CVPR, 2018. 7

[71] W. Wu, N. Lu e E. Xie. Adaptação de domínio não supervisionado sintético para real para detecção de texto de cena na natureza. In ACCV, 2020. 3

[72] P. Yakubovskiy. Modelos de segmentação pytorch. https://github.com/qubvel/segmentation_ models.pytorch, 2020. 6

[73] Y. Yao, L. Zheng, X. Yang, M. Naphade e T. Gedeon. Simulando conjuntos de dados de veículos consistentes de conteúdo com descida de atributos. In ECCV, 2020. 2 [74] L. Yin, X. Chen, Y. Sun, T. Worm e M. Reale. Um banco de dados de expressões faciais dinâmicas em 3D de alta resolução. Em 2008, 8ª Conferência Internacional IEEE sobre Reconhecimento Automático de Gestos Faciais, páginas 1–6, 2008. doi: 10.1109/AFGR.2008. 4813324. 2

[75] G. Zaal, S. Majboroda e A. Mischok. Refúgio HDRI. https://hdrihaven.com, 2020. 5 [76] X. Zeng, X. Peng e Y. Qiao. Df2net: Um denso-fino-fino

rede para reconstrução facial 3D detalhada. In CVPR, 2019. 2, 3, 4

[77] J.-Y. Zhu, T. Park, P. Isola e AA Efros. Tradução não emparelhada de imagem para imagem usando redes adversárias consistentes em ciclo . In ICCV, 2017. 3

[78] M. Zhu, D. Shi, M. Zheng e M. Sadiq. Detecção robusta de marcos faciais por meio de redes profundas adaptáveis à oclusão. In CVPR, 2019. 7

[79] X. Zhu, X. Liu, Z. Lei e SZ Li. Alinhamento facial em toda a gama de poses: uma solução total em 3D. TPAMI, 2017. 2