DS-06

DeepTasteDM: 拡散モデルを用いたリアルタイム食事変換ARシステム

唐澤 香梨菜[1]　中野 萌士[2]　柳井 啓司[1]
1.電気通信大学　2.東京大学

概要

- 食事の見た目を変更することで味覚に影響を与えることができる
- 拡散モデルを用いることで高品質かつプロンプト入力による柔軟な変換を行う
- 拡散モデルをARに適用する場合、レイテンシが課題となる

高速な拡散モデルを用いることで、
食事の見た目をリアルタイムに変換する
ARシステムDeepTasteDMを提案

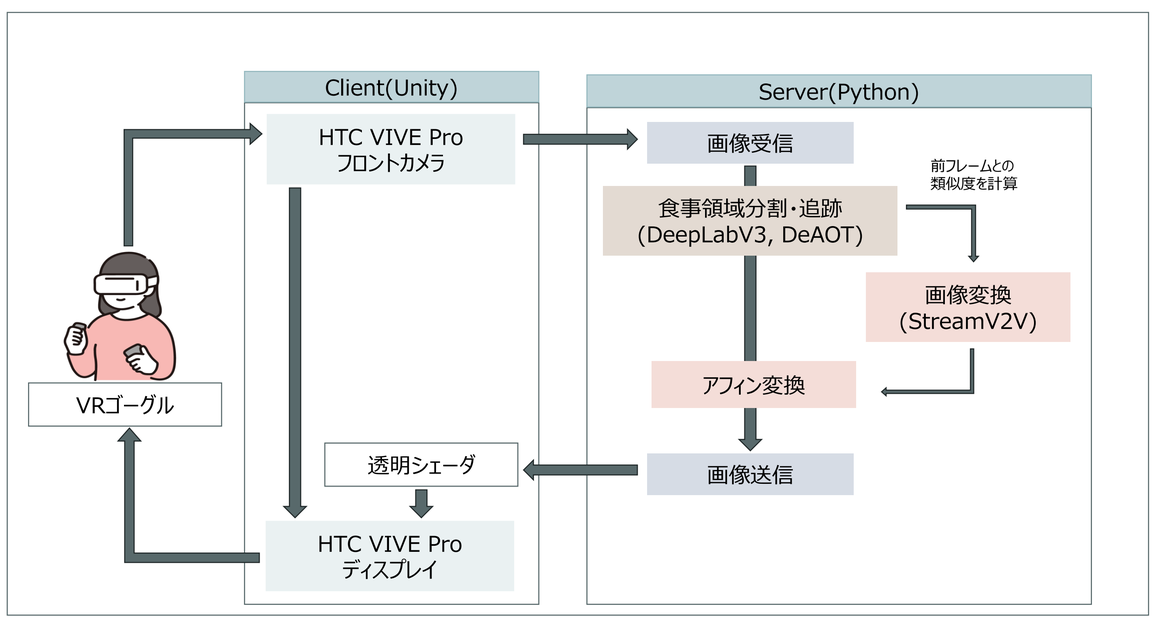
関連研究

DeepTaste[1]
- StarGAN[2]とARを用いて動的に食品の外観を変更できる味覚操作システムを作成した

StreamV2V[3]
- 拡散モデルは高品質な画像生成が可能だが生成速度が遅い
- 過去のキーとバリューを参照する拡張セルフアテンションと特徴バンクにより、高速性と時間的一貫性を両立

手法

システム概要図



食事領域セグメンテーションと追跡

- DeepLabV3[4]で食事領域を抽出し、DeAOT[5] でマスクを追跡・再利用することで滑らかな画像変換を実現
- 食事領域以外を白にした画像をStreamV2Vに入力し、変換結果の食事領域のみを重畳

LoRAによるドメイン適応

- 拡散モデルの学習データは西洋料理中心のため、和食・中華料理では外観が崩れやすい
- kohya-ss/sd-scripts[6]を用い、UECFood[7]でLoRA[8]を学習

アフィン変換

- フレーム間の食事領域マスクのIoUに基づき、画像生成とアフィン変換による高速追従を切り替え

スレッド並列処理

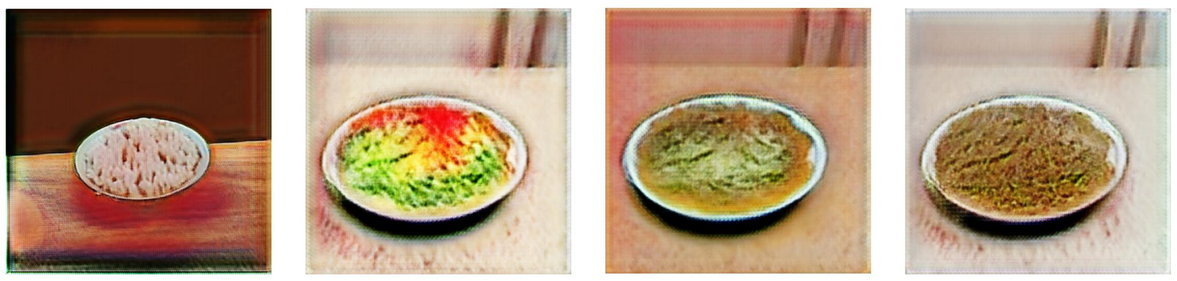
- 画像受信・画像生成・アフィン変換・画像送信の4スレッドで並列処理

実験・考察

- 拡散モデルにはJiali/stable-diffusion-1.5[9]、GPUにはNVIDIA RTX 6000 Adaを使用した

1. 変換結果

- 背景や食器の外観を維持しつつ具材まで高精細に再現



うな重　冷やし中華　ラーメン　焼きそば

DeepTaste(上段)と本研究(下段)の生成画像の比較

2. 時間的一貫性

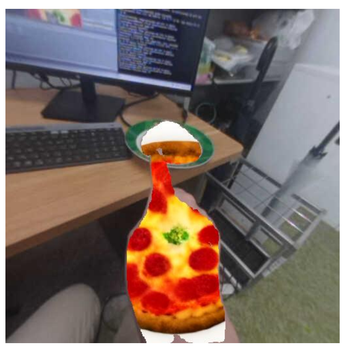
- 変化が少ない時は前フレームの結果を保持し自然な変換を実現



時間的一貫性の考慮による変換結果の比較

3. 限界

- DeepLabV3のセグメンテーション精度に依存し、ユーザの手を食事領域と誤認識することがある





結論

- 拡散モデルとLoRAの導入により、従来のGANでは困難だった高品質な画像生成とプロンプトによる柔軟な食事変換を実現
- セグメンテーション、アフィン変換の切替とスレッド並列処理により、時間的一貫性とリアルタイム性を両立
- 手の誤認識に対し、高精度かつ高速なセグメンテーションモデルやオクルージョンを考慮した変換方法を導入し、実用性の向上を目指す

[1] Nakano K et al.: DeepTaste : Augmented Reality Gustatory Manipulation with GAN-Based Real-Time Food-to-Food Translation, Proc. of 2019 IEEE International Symposium on Mixed and Augmented Reality, pp.212–223, 2019.
[2] Choi Y et al.: StarGAN : Unified generative adversarial networks for multi-domain image-to-image translation, Proc. of IEEE Computer Vision and Pattern Recognition, pp.8789–8797, 2018.
[3] Feng L et al.: Looking backward: Streaming video-to-video translation with feature banks. In International Conference on Learning Representations, Vol. 2025, pp.46425–46445, 2025.
[4] Chen L-C: Rethinking atrous convolution for semantic image segmentation, arXiv:1706.05587, 2017.
[5] Aongxin Y et al.: Decoupling featuresin hierarchical propagation for video object segmentation. Advances in Neural Information Processing Systems, Vol. 35, pp. 36324–36336, 2022.
[6] https://github.com/kohya-ss/sd-scripts
[7] Yuji M et al.: Recognition of multiple-food images by detecting candidate regions. In IEEE International Conference on Multimedia and Expo, pp. 25–30, 2012.
[8] Edward J Haw et al.: Lora: Low-rank adaptation of large language models. Iclr, Vol. 1, No. 2, p. 3, 2022.
[9] https://huggingface.co/Jiali/stable-diffusion-1.5