CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

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,a)]　中野 萌士[2,b)]　柳井 啓司[1,c)]

概要

本研究では，拡散モデルを用いてリアルタイムに食事変換を行う拡張現実（AR）システムである DeepTasteDM を提案する．高品質な画像生成が可能な拡散モデルとリアルタイムビデオ変換フレームワーク StreamV2V を採用し，プロンプト入力による柔軟な視覚提示が可能である．加えて，食事領域のセグメンテーションや LoRA によるドメイン適応，アフィン変換を併用し没入感の高い食事変換 AR システムを実現した．

1. はじめに

特定の食事（例：ラーメン）の味を再現し提示する場合，提示方法が確立している視覚情報とは異なり，単純な基本五味の組み合わせではラーメンの味を再現することは難しい．一方，味覚は視覚や嗅覚等からの影響を受け，異なった味のように感じられる．このクロスモーダル知覚を応用し，視覚情報を操作することで味覚を操作する研究が行われている．例えば，先行研究 [1] では StarGAN[2] を用いた画像生成と拡張現実（AR）により，味を操作可能であることを示している．我々は，StarGAN の代わりに目覚ましく発展する拡散モデルを用いることで，高精度かつプロンプト入力によって柔軟な画像変換を可能とし，味覚操作研究の応用性を拡大できるのではないかと考えた．しかし，拡散モデルを AR へ適用する場合，計算コストによるレイテンシが課題である．

本研究では, 食事の外観を高速に変換する AR システムの作成を目的とする．より本物に近い食品の見た目を提示するため，高品質な画像生成が可能な拡散モデルを使用し，さらにリアルタイムに近い画像変換を目指し StreamV2V を活用する．また食事領域のみに対し画像変換を行うことで, AR 食事体験による没入感を高め, 視覚提示による味覚操作の効果をより強く引き出すことを狙いとしている．

[1] 電気通信大学大学院
[2] 東京大学大学院
[a)] karasawa-k@mm.inf.uec.ac.jp
[b)] kizashi@cyber.t.u-tokyo.ac.jp
[c)] yanai@cs.uec.ac.jp



図 1: 本研究の概要図．ユーザは AR システムを用いて，実際の食事の見た目をプロンプトで入力した食事の見た目に変更することができる．入力画像：チャーハンの食品サンプル．出力画像：入力画像から生成し重畳表示されたカレーライス画像.

2. 関連研究

2.1 AR を用いた食事の外観変換

AR は現実世界にバーチャルな要素を重畳する技術であり，ユーザはスマートフォン・AR グラスなどのデバイスを通して，コンピュータで生成された情報・画像などと関わることができる．この技術を用いて，実際の食事を異なる食事の外観へと変化させる味覚ディスプレイの研究が行われている．視覚等の影響によって実際の味が変化するクロスモーダル知覚を利用したこれらの研究は，画像処理や生成技術の進歩に伴い，より大きな味覚操作効果が期待される．

FoodChangeLens[3] では，HoloLens と Conditional CycleGAN[4] を利用し，実際の食品形状を考慮した 10 種類の食事カテゴリ変換 AR システムを作成した．AR Food Changer[5] では Mask R-CNN を用いた物体抽出と二値化画像による食品の輪郭抽出を組み合わせることで，高精度なテクスチャ変換を実現した．本研究の基となった DeepTaste[1] では 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．StarGAN は UECFOOD-100[6] および X から収集した追

加画像で学習されており，そうめんから焼きそば，白米からカレーといった 10 種類の食事変換に対応している．本研究では，StreamV2V[7] を用いた高品質でプロンプト入力による柔軟な食事変換を実現し，DeepTaste の味覚操作効果の向上を目的とした．

2.2 食事画像生成

Stable Diffusion などの画像生成 AI は広範なデータセットから学習されており，多様な画像生成が可能である．その学習データには食事画像も含まれているが，あくまで一部に過ぎず食品領域に特化しているわけではない．そのため拡散モデルを用いてより高品質かつ魅力的な食事画像を生成する試みが多数なされている．FoodFusion[8] では LAION-5B データセットから高品質な食事画像を選び構築した LAIONFood データセットを学習することで，よりリアルな食事画像生成を実現した．また CW-Food[9] では画像から抽出されるプライベート特徴とテキストから得られるクラス内特徴を融合し，さらに LoRA を使用することでリアルで多様な食事画像生成を可能にした．

このように食事という特定のドメインに特化した高品質な画像を生成するためには，学習画像データの調整や，軽量なパラメータを追加して学習を行う LoRA の活用などが一般的である．特に LoRA[10] はモデルの重みを固定したまま低ランク行列のみ学習させることで，計算資源やメモリ消費を大幅に削減しつつ，ターゲットとなるドメインを効率的に学習でき，特定のドメインに特化させる際は有効な手法として広く用いられている．

2.3 StreamV2V

拡散モデルは高品質な画像を生成できる点で従来の生成モデルより優れているが，逆拡散過程において反復的なノイズ処理を行うために生成速度が遅いという課題がある．この生成を高速化する試みとして Latent Consistency Models[11] や LDDGAN[12]，また動画領域においては TurboDiffusion[13]，StreamDiffusion[14] などの研究が盛んに行われている．これらを踏まえ本研究ではユーザプロンプトを用いたリアルタイムビデオ間変換が可能な StreamV2V[7] を活用する．

StreamDiffusion では StreamBatch により待ち時間を最小限に抑え高速な処理を可能にした．それを踏まえ，StreamV2V ではストリーミング方式によりフレーム数を無限にサポートし，過去のキー・バリューを参照する拡張 Self-Attention 機構や，過去フレームの情報を維持する特徴バンクを導入している．これにより処理の高速性を維持したまま，時間的一貫性を維持する能力が特徴となっている．

3. 手法

本システムの概要図を図 2 に示す．本研究では，拡散モデルがもつ高い表現力を維持しつつ，リアルタイムで動作する食事変換 AR システムを提案する．本システムでは高速な拡散モデルである StreamV2V をベースとし，毎フレームの画像生成を行うのではなく，カメラのフレーム間の動き量に応じて処理を切り替えるパイプラインを構築した．すなわち，ユーザの視点変化や食事領域の変化が少ない場合は，時間のかかる拡散モデルの推論をスキップし，前フレームの変換結果をアフィン変換によって高速に変形する．また食事領域の認識と高速追跡を両立するため，DeepLabV3 によるセグメンテーションと DeAOT[15] によるリアルタイム追跡を組み合わせた．

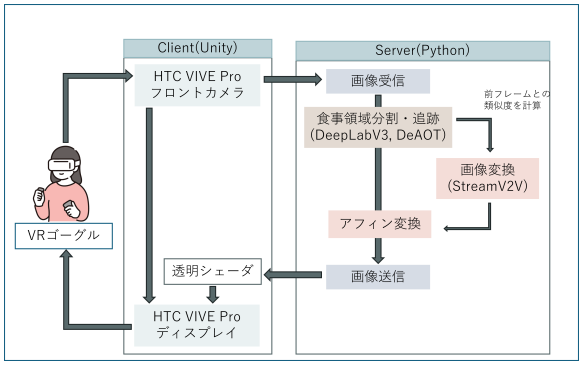


図 2: 本システムの概要図

本研究のシステムは，Unity 2018.2.9f1 を用いたクライアントと Python 3.10.16 を用いた Flask サーバで構築される．最初に，HTC VIVE Pro 2 のフロントカメラから取得した入力画像の中心から 512 × 512 の範囲でクロップした画像を，共有メモリによるプロセス間通信を用いてサーバに送信する．同時に，プロンプトとメタデータもクライアントからサーバに送信する．次にサーバは，送られた画像に対して画像処理を行った後，変換画像を生成し，クライアントに送り返す．最後にクライアントは，受け取った生成画像をフロントカメラの映像に重畳表示する．サーバでの画像処理では，DeepLabV3[16] による食事領域のセグメンテーション，変換画像生成，およびレイテンシ短縮のためのアフィン変換を行う．HTC VIVE Pro 2 と Intel Core i7-9700K，NVIDIA RTX 6000 Ada を用いた環境で計測した結果，フレームレート：約 14 [fps]，レイテンシ：約 130 [ms] で生成画像の重畳表示に成功した．

3.1 食事領域セグメンテーションと追跡

自然な食事変換を達成するため，背景や食器は入力画像のまま維持することが望ましい．そこで本システムではセマンティックセグメンテーションモデルである DeepLabV3 を用いて，食事領域を抽出する．また動画物体領域追跡モデルである DeAOT により食事領域マスクを追跡・再利用する．このように食事領域マスクを保持することで，滑らかな画像変換を可能にしている．

実際の処理ではマスクを基に食事領域以外を白背景に置き換えた画像を StreamV2V に入力し，食事のみの変換画像を生成する．最終的に，変換画像の食事領域のみをフロントカメラからの入力画像に重畳することで，背景や皿は変換せず食事のみの変換を行い，没入感の高い AR 表示を可能にした．セグメンテーション処理の例を図 3 に示す．

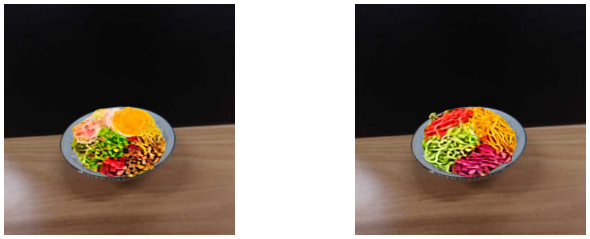


(a) 入力画像　(b) セグメンテーション　(c) 画像変換　(d) アフィン変換

図 3: セグメンテーションとカレーへの画像変換の例

3.2 LoRA

ベースとなる拡散モデルの学習データは西洋料理が中心であり，和食や中華料理を対象とした場合，本来とは異なる外観の画像が生成される．また入力画像によるブレやノイズに対し，生成品質が不安定になる問題も存在する．そこで本研究では LoRA を使用することで多様な入力環境下でも安定した食事画像変換を行う．



(a) LoRA なし　(b) LoRA あり

図 4: 冷やし中華画像変換における LoRA 有無の比較

図 4 に示すように LoRA を使用することでより理想的な食品画像が生成可能になる．学習には kohya-ss/sd-scripts[*1]を利用し，学習データには UECFood の解像度 512 × 512，食事以外を白背景にした画像を使用した．

3.3 入力画像に対する食事領域の大きさ

StreamV2V への入力画像に対する食事領域の大きさと生成画像の品質の関係を以下の図 5 に示す．フロントカメラと食事が離れると StreamV2V への入力画像に対する食事領域も小さくなり，白背景に影響されリアルな食事画像が生成されなかった．また入力画像に対する食事が大きい場合でも粗い画像が生成されてしまうことが分かった．そのため，本システムでは入力画像に対する食事領域の大きさを 60%と固定している．

[*1] https://github.com/kohya-ss/sd-scripts

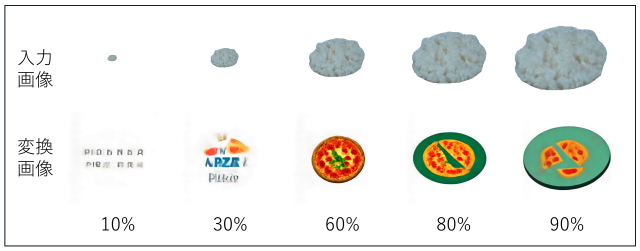


図 5: 入力画像に対する食事領域の割合の比較

3.4 アフィン変換

拡散モデルは推論コストが大きいため，毎フレーム推論を行うと大きなレイテンシが生じる．一方で食事中のユーザの頭部運動は比較的緩やかであり，視点変化が小さいシーンも多い．このような時は前フレームの生成結果を表示し追従させる方が望ましい．そこで本研究では毎フレーム変換をするのではなく，フレーム間の食事領域での類似度に基づいて処理を切り替える．前フレームで追跡していた食事領域マスクと現フレームのマスク間の IoU が 0.3 以下の場合は新たな画像を生成する．IoU が 0.3 より大きい場合は，保持している変換画像と入力画像の食事領域の輪郭からアフィン変換を行う．これにより，重い画像生成処理を効果的にスキップし，リアルタイム性を担保している．

3.5 スレッド

画像受信，セグメンテーション，画像変換，アフィン変換，画像送信の一連の処理を単一スレッドで実行すると，高負荷な画像変換中は他の処理を行うことができず，クライアント側の HMD (Head-Mounted Display) 表示映像がフリーズしてしまう．時間を要する画像生成のバックグラウンド実行中であっても，クライアント側はアフィン変換による追従処理を継続し，滑らかな描画を維持する必要がある．そこで本研究では画像受信スレッド，画像生成スレッド，アフィン変換スレッド，画像送信スレッドの 4 つのスレッドを立ち上げる並列処理構造を構築した．これにより各処理間の待機時間を減らし，重い画像生成タスクと軽量な追従タスクを並列に動作させることで，描画の低遅延化と滑らかな映像出力を実現している．

3.6 デモ展示内容

デモ発表では，提案システムの実機デモを展示する．会場には，フロントカメラを搭載した HMD と，画像変換サーバとして機能するノート PC を配置する．ユーザは HMD を装着し，ブース内に用意された食品を観察する．カレーやピザなどのプロンプトを入力すると，リアルタイムに別料理へと変換される体験が可能である．

4. 実験・考察

本システムによる食事変換の視覚的品質および時間的一貫性を検証するため，変換実験を行った．なお拡散モデル

には”Jiali/stable-diffusion-1.5”を使用し，ハードウェア環境には NVIDIA RTX 6000 Ada Generation を用いた．

4.1 変換結果

従来システムである DeepTaste（StarGAN）と提案システムの変換結果の比較を図 6 に示す．上段は DeepTaste, StarGAN での変換結果，中段は本システムの変換結果であり，下段には画像変換時に使用したテキストプロンプトを示している．



(a) うな重　(b) 冷やし中華　(c) ラーメン　(d) 焼きそば

(e)　(f)　(g)　(h)

(e) “unadon, grilled eel with sweet soy glaze, lacquered bowl, sansho pepper, char-grilled marks”

(f) “hiyachu, hiyashi chuka, colorful toppings, shredded ham, cucumber, egg, cold noodles, ice cubes”

(g) “ramen, rich tonkotsu broth, soft boiled egg, nori, chashu pork, oil droplets on surface”

(h) “yakisoba, brown sauce glaze, red pickled ginger, aonori seaweed, steaming hot”

図 6: DeepTaste と本システムの比較結果

上段の StarGAN では机や食器などの背景を含めた入力画像全体を変換していた．またうな重や焼きそばの変換結果においては，白米や麺に対し簡易的な着色に留まる傾向が見られた．一方で提案システムでは机や食器の外観が維持されており，よりリアルで没入感が高い結果となっている．またうな重のうなぎやラーメンののりなど具材が現れたり，麺の質感が表現されていたりなど高品質でディティールが鮮明な変換結果となっており，拡散モデルと LoRA の有効性が確認できる．

4.2 時間的一貫性

以前に採用していた StreamDiffusion と時間的一貫性を考慮した提案システムの StreamV2V およびアフィン変換との時系列比較を図 7 に示す．毎フレーム独立して画像変換を行う図 7(a) では，カメラの微小な動きに伴って 1 フレームごとにカレールゥの色が激しく変化しており，食事変換 AR として適していなかった．これに対し本手法の図 7(b) では変化が少ないシーンでは前フレームの結果を保持するため，画面のちらつきが少なく自然な変換結果となった．



(a) フレームごとに変換した結果



(b) 時間的一貫性を考慮した変換結果

図 7: 時間的一貫性を考慮した前後の比較

4.3 限界と今後の展望

本システムでは DeepLabV3 によるセグメンテーションに依存しているため，図 8 に示すようにユーザの手を食事領域の一部として誤認識してしまうことがある．この場合，手を含んだマスクの形状変化に合わせてアフィン変換が行われるため，変換結果が不自然に引き延ばされてしまう．今後はより高速かつ高精度なセグメンテーションモデルの導入や手の遮蔽を考慮したパイプラインへの拡張を予定している．

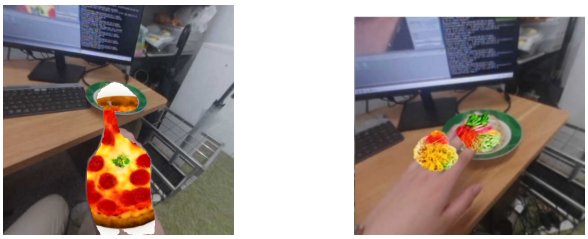


図 8: 手を誤認識した結果

5. 結論

本研究では，拡散モデルを用いたリアルタイム食事変換 AR システムである DeepTasteDM を提案した．拡散モデルや LoRA を導入することで，従来の GAN では困難であった高品質な画像生成と，テキストプロンプトによる柔軟な食事変換が可能になった．さらに食事領域セグメンテーションやアフィン変換への適応的切り替え，スレッド並列処理を構築することで，時間的一貫性とリアルタイム性を両立した．今後は手の誤認識に対し，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の導入により実用性の向上を目指す．

参考文献

[1] Kizashi Nakano, Daichi Horita, Nobuchika Sakata, Kiyoshi Kiyokawa, Keiji Yanai, and Takuji Narumi. DeepTaste : Augmented reality gustatory manipulation with GAN-based real-time food-to-food translation. In *Proc. of 2019 IEEE International Symposium on Mixed and Augmented Reality*, pp. 212–223, 2019.

[2] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN : Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 8789–8797, 2018.

[3] Shu Naritomi, Ryosuke Tanno, Takumi Ege, and Keiji Yanai. FoodChangeLens: CNN -based food transformation on hololens. In *Proc. of IEEE International Conference on Artificial Intelligence and Virtual Reality*, pp. 197–199, 2018.

[4] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of IEEE International Conference on Computer Vision*, pp. 2242–2251, 2017.

[5] Junya Ueda and Katsunori Okajima. Ar food changer using deep learning and cross-modal effects. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pp. 110–1107. IEEE, 2019.

[6] Yuji Matsuda, Hajime Hoashi, and Keiji Yanai. Recognition of multiple-food images by detecting candidate regions. In *IEEE International Conference on Multimedia and Expo*, pp. 25–30, 2012.

[7] Feng Liang, Akio Kodaira, Chenfeng Xu, Masayoshi Tomizuka, Kurt Keutzer, and Diana Marculescu. Looking backward: Streaming video-to-video translation with feature banks. In *International Conference on Learning Representations*, Vol. 2025, pp. 46425–46445, 2025.

[8] Olivia Markham, Yuhao Chen, Chi-en Amy Tai, and Alexander Wong. Foodfusion: A latent diffusion model for realistic food image generation. *arXiv preprint arXiv:2312.03540*, 2023.

[9] Dongjian Yu, Weiqing Min, Xin Jin, Qian Jiang, Ying Jin, and Shuqiang Jiang. Diverse and high-quality food image generation from only food names. *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 21, No. 5, pp. 1–22, 2025.

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, Vol. 1, No. 2, p. 3, 2022.

[11] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. In *arXiv:2310.04378*, 2023.

[12] Luan Thanh Trinh and Tomoki Hamagami. Latent Denoising Diffusion GAN: Faster sampling, higher image quality. *IEEE Access*, Vol. 12, pp. 78161–78172, 2024.

[13] Jintao Zhang, Kaiwen Zheng, Kai Jiang, Haoxu Wang, Ion Stoica, Joseph E Gonzalez, Jianfei Chen, and Jun Zhu. Turbodiffusion: Accelerating video diffusion models by 100-200 times. *arXiv preprint arXiv:2512.16093*, 2025.

[14] Akio Kodaira, Chenfeng Xu, Toshiki Hazama, Takanori Yoshimoto, Kohei Ohno, Shogo Mitsuhori, Soichi Sugano, Hanying Cho, Zhijian Liu, and Kurt Keutzer. StreamDiffusion : A pipeline-level solution for real-time interactive generation. In *arXiv:2312.12491*, 2023.

[15] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *Advances in Neural Information Processing Systems*, Vol. 35, pp. 36324–36336, 2022.

[16] Liang-Chieh Chen. Rethinking atrous convolution for semantic image segmentation. In *arXiv:1706.05587*, 2017.