拡散モデルベースの３次元復元手法を用いた単一食事画像からのカロリー量推定

大岸 茉由[1,a)]　田邊 光[1,b)]　柳井 啓司[1,c)]

概要

食事のカロリー量や栄養情報を正確に把握することは健康維持に不可欠だが, 2 次元画像のみを用いた従来手法では, 奥行きや体積の推定が困難であり, 精度に限界があった. 本研究では, 拡散モデルを用いた 3 次元復元手法を導入し, 単一の食事画像から体積情報を考慮した高精度なカロリー量推定を実現する.

1. はじめに

健康を維持・向上させるためには, 食事のカロリー量や栄養素を正確に把握することが重要である. しかし, カロリー量推定には食材の種類や量, 調理方法などの詳細な情報が必要であり, 手動での収集・入力は利用者に大きな負担を強いる. さらに, 主観や計測誤差により推定精度にばらつきが生じる問題もある. このような課題を解決するため, 近年, 食事画像を用いたカロリー量推定が注目されている.

従来の手法では, 食事画像を 2 次元的に処理し, ピクセル情報や色, カテゴリ情報を活用してきたが, 2 次元画像のみでは食材の体積や高さといった 3 次元的情報を考慮できず, 特に形状が複雑な食事では推定精度が低下する.

本研究では, この課題を解決するため, 拡散モデルベースの 3 次元復元手法を導入し, 単一の食事画像から体積情報を考慮したカロリー量推定を行う. 拡散モデルの強力な事前知識を活用することで, 従来の 2 次元的アプローチを超える高精度な推定を可能にすることを目的とする.

2. 関連研究

2.1 単一画像からの 3 次元復元手法

ニューラルネットワークによる単一画像からの 3 次元復元は, これまで多くの研究が行われてきた. 従来の手法では, 直接 3 次元形状を学習するモデルが用いられてきたが, 3D データセット不足により, 未知カテゴリの形状生成に課題があった. そこで近年, 拡散モデルを用いた手法が注目されており, 特に Score Distillation Sampling (SDS) を活用した手法が高い生成品質を実現している. SDS を用いることで, 2 次元拡散モデルの知識を活用し, より多様な形状の生成が可能となったが, 計算コストの高さや視点間の一貫性の欠如が課題となっている. この問題を解決するため, ビュー毎の個別の最適化ではなく, マルチビューで一貫した生成を行う拡散モデルが提案された. 本手法で使用する Wonder3D [7] は, マルチビュー拡散を用いた手法の一つであり, 法線マップとカラー画像を統合的に扱うことで, 高品質で視点間の一貫性の取れた 3 次元復元を実現した.

2.2 カロリー量推定

現在, カロリー量の自動推定のために画像解析技術の活用が注目されている. 2 次元画像ベースの手法では, 食事カテゴリの分類や面積情報を用いた推定が主流であり, クレジットカードや箸などの基準物体を利用する方法が提案されてきた [1], [10]. また, 基準物体を用いずに, 食事カテゴリや食材情報を考慮するマルチタスク学習を活用した手法もある [16]. しかし, これらの手法では, 食事の奥行き情報を考慮できないため, 精度向上には限界がある.

そこで近年では, 3 次元情報を用いて食品の体積を推定し, カロリー量を求める手法が提案されている. Dehais らは, スマートフォンで撮影した複数の画像から 3 次元形状を再構成する手法を提案したが, 単一画像からの推定には対応していない [4]. 成冨らは, 皿の形状の一貫性を考慮した 3 次元復元手法を提案したが, カロリー量推定までには至っていない [9]. 本研究では, 拡散モデルベースの 3 次元復元モデルにより, 高品質な復元を行うことで, 2 次元画像のみを用いた従来手法よりも高精度なカロリー量推定を実現する.

3. 提案手法

本手法の全体的な流れを図 1 に示す. 入力は単一の食事画像であり, 食品カテゴリとカロリー量を推定する. 以下では, 各ステップの詳細について説明する.

[1] 電気通信大学
[a)] ogishi-m@mm.inf.uec.ac.jp
[b)] tanabe-h@mm.inf.uec.ac.jp
[c)] yanai@cs.uec.ac.jp

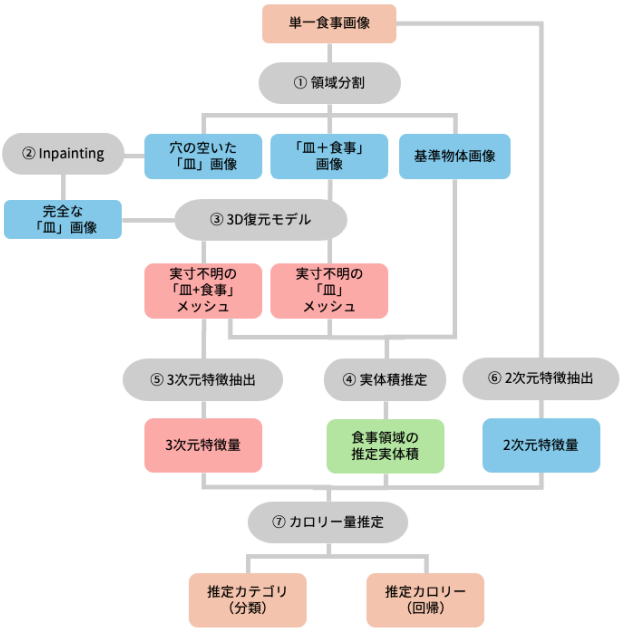

図 1　提案手法の概要

3.1　マスク抽出

提案手法の最初のステップでは, 食事画像から食事・皿・基準物体の領域を抽出する. 本研究では, 自然言語プロンプトによる柔軟な物体検出を行う GroundingDINO1.5 [12] と高精度な領域分割を実現する SAM-2 [11] を組み合わせた Grounded-SAM [13] を使用し, 自然言語に基づく柔軟な物体検出を行う. 本研究ではデータ特性に基づき “food . plate . chessboard” を使用した. 図 2 に抽出結果を示す.

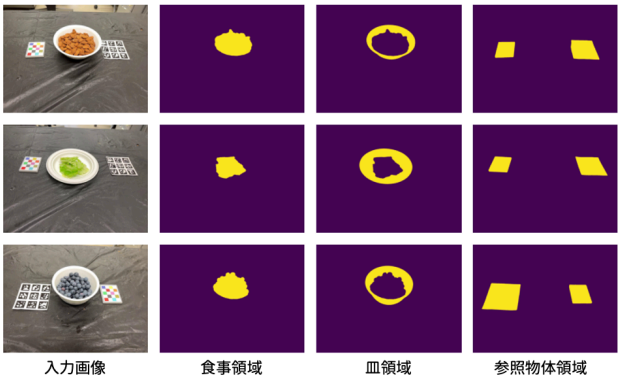

図 2　Grounded-SAM による領域抽出結果

このようにして柔軟な領域抽出を実現し, 得られたマスクを Inpainting や 3 次元復元に活用する.

3.2　Inpainting による皿画像の生成

皿の 3 次元復元には, 食事領域を補完した完全な皿画像が必要である. 本手法では, Stable Diffusion v2 [14] を基盤とする事前学習済みモデルを使用し, プロンプトに “plain plate” を指定して皿画像を生成する. しかし, デフォルト設定では図 3 のように食事を補完する傾向が確認された. これは, 学習データの多くに食事が載っていたためと考えられる.

図 3　デフォルト設定による Inpainting 結果

この課題を解決するため, モデルを, 食事の載っていない皿画像でファインチューニングし, 改善を図った. まず, インターネットから “plain plate” の検索クエリで画像を収集し, 食事なしの 1000 枚を選定した. （図 4）.

図 4　収集した皿画像例

しかし, 実画像では皿の向きや背景などに偏りがあるため, Stable Diffusion v2 で追加で画像生成を行い, 皿の厚み・色・背景をランダムに生成した. ネガティブプロンプトに “multiple plates, food, text, watermark” を指定し, 不適切な画像を排除して, 最終的に 3000 枚のデータセットを作成した（図 5）.

図 5　Stable Diffusion により生成された皿画像

次に, この皿データセットを用いて, モデルを LoRA (Low-Rank Adaptation) [6] でファインチューニングした. LoRA は, 既存のネットワークの重みを固定し, 新たなパラメータのみを学習する手法である. ファインチューニング後のモデルでは, 食事の補完問題が解消され, 穴のない皿画像を高精度に生成可能となった（図 6）.

図 6　ファインチューニング後の Inpainting 結果

本手法により, 食事領域を正確に補完した皿画像を得ることができた.

3.3 3 次元復元

マスク画像の取得と皿画像の生成後,「食事+皿」および「皿」単体の 3 次元復元を行う. 本手法では, 高品質な 3 次元形状を復元するためにマルチビュー拡散モデルを活用した Wonder3D [7] を採用した.

Wonder3D は, 2 次元拡散モデルの事前知識を利用し, 法線マップとカラー画像を同時生成するクロスドメイン手法であり, 高精度かつ一貫性のある 3 次元形状復元を実現する. 本研究では, 以下の 2 種類の 3D メッシュを生成した.

- 皿のみの 3D メッシュ
- 食事と皿を統合した 3D メッシュ

これらの 3D メッシュは実スケール情報を持たないため, 後続のスケール推定で参照物体を用いて補正する. Wonder3D の適用により, 視点間の整合性を維持しつつ高忠実度な形状復元が可能となり, 食事領域の体積推定やカロリー量推定に必要なデータが得られる. また, 本手法は従来困難だった複雑な形状の食事の 3 次元復元にも対応可能である.

図 7 に復元結果を示す. 上段は「食事+皿」, 下段は「皿のみ」の 3D メッシュである.

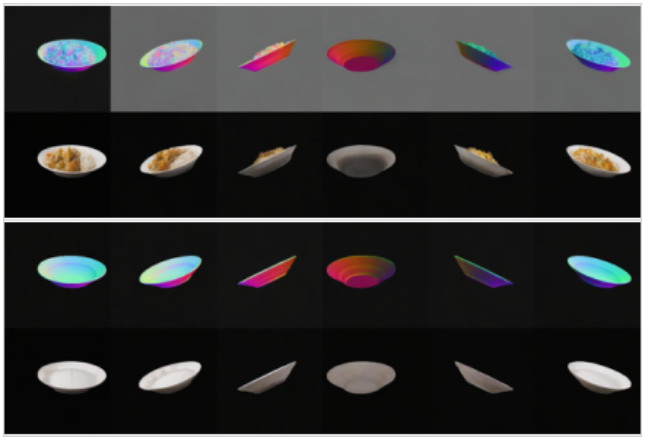

図 7 Wonder3D を用いた 3 次元形状の復元例

3.4 実体積推定

復元した 3 次元形状から食事領域の体積を正確に計算するため, スケール推定を行う. 本研究では, MetaFood3D [3] に含まれるキャリブレーションカードを参照物体とし, 基準サイズを用いたスケール変換を実施した.

提案手法は Grounded-SAM を用いて任意の参照物体を検出できるため, キャリブレーションカードに限らず, サイズが既知の箸や皿, コップなどの一般的な食事シーンで用いられるものを使用することも可能である.

スケール推定では, 基準物体の既知の物理的寸法と画像内のピクセル数を対応付けることで, 1 ピクセルあたりの実際のサイズを算出する. これに基づいて 3 次元形状全体をスケール変換し, 各メッシュの正確な実体積を取得する. 「皿+食事の 3D メッシュ」と「皿のみの 3D メッシュ」の体積差分を取ることで, 食事領域の体積を高精度に計算する.

3.5 3 次元形状特徴抽出

体積情報に加え,「食事+皿」領域の 3 次元形状から高精度な特徴を抽出する. 具体的には, 3D メッシュを点群に変換し, PointGPT [2] を用いた 3 次元形状エンコーダーを適用した. PointGPT は, 3 次元点群を Generative Pre-trained Transformer（GPT）のように自己回帰的に生成・学習することで, 情報密度の低さ, タスク間のギャップといった課題に対処する革新的な手法である. 本研究では, ModelNet40 [15] で事前学習した PointGPT を特徴抽出器として用い, 最終層を凍結し 2048 次元の埋め込みベクトルを取得した. これにより, 体積情報のみならず, 3 次元形状を考慮した高精度な特徴抽出を実現した.

3.6 2 次元画像特徴抽出

3 次元形状特徴に加え, 2 次元画像からの特徴量も抽出することで, 3 次元形状特徴からは得ることができない食品の種類や表面的特性（例: 油分量や焼き加減）等を補完し, カロリー量推定精度を向上させる. 本研究では, 2 次元画像からの特徴量抽出器として VisionTransformer [5] を用いた. VisionTransformer は, 画像をパッチに分割し, それぞれのパッチをトランスフォーマーに入力することで, 高性能な画像認識を実現するモデルである. 事前学習済みの VisionTransformer を使用し, 1024 次元の特徴ベクトルを抽出した. これにより, 食品の色やテクスチャなど視覚的特徴をモデルに提供する.

3.7 カロリー量推定

最後に, 図 8 のように, 推定体積を基盤とし, 2 次元画像特徴と 3 次元形状特徴を統合してカロリー量を推定する. この統合により, 食品の種類や表面特性を考慮した高精度な推定を実現する.

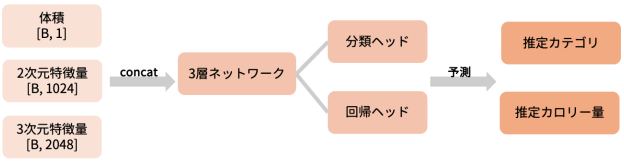

図 8 カロリー量推定の概要

提案手法では, 以下の特徴量を使用する.

- **実体積の推定値**: 3 次元復元で得られる「皿+食事メッシュ」と「皿メッシュ」の差分を計算.
- **3 次元形状特徴**: PointGPT を用いた 2048 次元の特徴ベクトル. 表面形状や空間構造を表現.

- **2 次元画像特徴**: VisionTransformer により抽出された 1024 次元の特徴ベクトル．食品の色やテクスチャを捉える．

これらの特徴量は結合され，ニューラルネットワークに入力される．カロリー量推定（回帰タスク）と食品カテゴリ分類（分類タスク）を同時に学習するマルチタスク学習を採用し，両タスクの相互作用により精度を向上させた．モデルは 3 層の MLP と，カロリー量推定・カテゴリ分類の 2 つの出力層から構成される．

4. 実験

4.1 データセット

本研究では，学習に MetaFood3D データセット [3] を使用した．MetaFood3D は 108 カテゴリ・637 個の食品オブジェクトを含み，3D メッシュ，2D 画像，点群，栄養情報など多様なデータを提供する．データを学習用とテスト用に 8:2 で分割し，評価を行った．

4.2 体積・カロリー量・カテゴリ推定の評価

提案手法と MFP3D [8] の比較を行い，体積推定，カロリー量推定，食品カテゴリ分類の精度を評価した．結果を表 1 に示す．

表 1 提案手法と MFP3D の評価結果

指標	MFP3D	提案手法
体積推定 MAE (ml)	62.60	**57.53**
体積推定 MAPE (%)	41.43	269.31
カロリー量推定 MAE (kCal)	**77.98**	92.74
カロリー量推定 MAPE (%)	**68.05**	212.78
カテゴリ推定 Accuracy	-	0.9641

提案手法は体積推定において MFP3D と比べ MAE が低く，高精度な推定が可能であることが確認された．一方で MAPE は大きく，特にナッツ類やブルーベリーなどの小型食品に対するセグメンテーションの失敗に起因し，相対的な誤差が増加する傾向が見られた．カロリー量推定では MFP3D に劣る結果となったが，体積情報や 3D 特徴の有用性が示唆された．カテゴリ推定では高い精度を示し，食品分類タスクへの応用可能性が示された．

4.3 アブレーション研究

体積情報の有無がカロリー量推定に与える影響を検証した結果を表 2 に示す．

表 2 アブレーション研究の結果

条件	MAE (kCal)	MAPE (%)
2 次元特徴量のみ	133.27	373.23
2 次元特徴 + 体積	101.12	254.09
提案手法 (3 次元特徴含む)	92.74	212.78

体積情報の追加により，推定精度が向上することが確認された．

4.4 定性評価

提案手法の各処理ステップとそれによって得られた結果を図 9 に示す．従来の 2 次元的アプローチでは対応が困難であったお椀に盛られた料理などの複雑な形状や構造を持つ食事に対しても，本手法のプロセスを通じて高精度な 3 次元復元が可能であることが確認された．

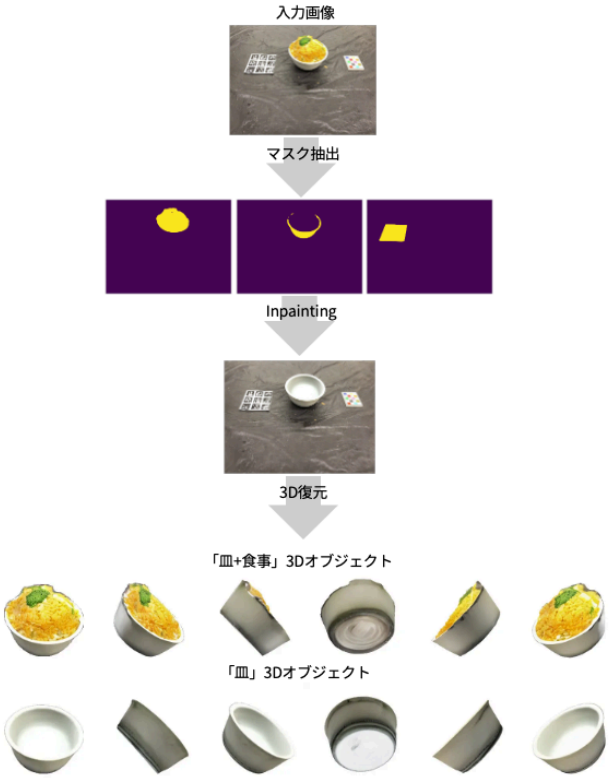

図 9 提案手法を用いた生成結果の例

5. おわりに

本研究では，食事の単一画像から三次元形状を復元し，体積を基にカロリー量を推定する手法を提案した．実験では，高精度な体積推定が可能であることを確認した．カロリー量推定では既存手法を上回る結果には至らなかったものの，アブレーション研究により，体積情報に加えて画像・形状特徴量が有用であることが示唆された．また，カテゴリ推定の精度も高く，食品分類タスクへの応用可能性を示した．今後の課題として，参照物体検出の精度向上，体積推定の改良，3 次元復元の高速化が挙げられる．現状，3 次元復元の処理時間は A4000 GPU で約 1 分程度であり，さらに前処理やカロリー量推定を含めると全体で約 2 分を要する．実用化に向けては，より高速な処理の実現が課題となる．これらの改善を通じて，提案手法の実用性が一層高まることが期待される．

参考文献

[1] Akpa, E., Suwa, H., Arakawa, Y. and Yasumoto, K.: Smartphone-Based Food Weight and Calorie Estimation Method for Effective Food Journaling, *In SICE Journal of Control, Measurement, and System Integration* (2017).

[2] Chen, G., Wang, M., Yang, Y., Yu, K., Yuan, L. and Yue, Y.: PointGPT: Auto-regressively generative pre-training from point clouds, *Advances in Neural Information Processing Systems* (2024).

[3] Chen, Y., He, J., Czarnecki, C., Vinod, G., Mahmud, T. I., Raghavan, S., Ma, J., Mao, D., Nair, S., Xi, P., Wong, A., Delp, E. and Zhu, F.: MetaFood3D: Large 3D Food Object Dataset with Nutrition Values, *Proc. of International Conference on Learning Representations* (2024).

[4] Dehais, J., Anthimopoulos, M., Shevchik, S. and Mougiakakou, S.: Two-View 3D Reconstruction for Food Volume Estimation, *IEEE Transactions on Multimedia*, Vol. 19, No. 5, pp. 1090–1099 (2017).

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, *Proc. of International Conference on Learning Representations* (2021).

[6] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models, *arXiv preprint arXiv:2106.09685* (2021).

[7] Long, X., Guo, Y.-C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.-H., Habermann, M., Theobalt, C. and Wang, W.: Wonder3D: Single Image to 3D using Cross-Domain Diffusion, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 9970–9980 (2024).

[8] Ma, J., Zhang, X., Vinod, G., Raghavan, S., He, J. and Zhu, F.: MFP3D: Monocular Food Portion Estimation Leveraging 3D Point Clouds, *arXiv preprint arXiv:2411.10492* (2024).

[9] Naritomi, S. and Yanai, K.: Hungry networks: 3D mesh reconstruction of a dish and a plate from a single dish image for estimating food volume, *Proc. of ACM International Conference Multimedia* (2021).

[10] Okamoto, K. and Yanai, K.: An Automatic Calorie Estimation System of Food Images on a Smartphone, *Proc. of 2nd International Workshop on Multimedia Assisted Dietary Management*, p. 63–70 (2016).

[11] Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P. and Feichtenhofer, C.: SAM 2: Segment Anything in Images and Videos, *arXiv preprint arXiv:2408.00714* (2024).

[12] Ren, T., Jiang, Q., Liu, S., Zeng, Z., Liu, W., Gao, H., Huang, H., Ma, Z., Jiang, X., Chen, Y., Xiong, Y., Zhang, H., Li, F., Tang, P., Yu, K. and Zhang, L.: Grounding DINO 1.5: Advance the "Edge" of Open-Set Object Detection, *arXiv preprint arXiv:2405.10300* (2024).

[13] Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q. and Zhang, L.: Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks, *arXiv preprint arXiv:2401.14159* (2024).

[14] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B.: High-Resolution Image Synthesis With Latent Diffusion Models, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 10684–10695 (2022).

[15] Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X. and Xiao, J.: 3D ShapeNets: A Deep Representation for Volumetric Shapes, *Proc. of IEEE Computer Vision and Pattern Recognition* (2015).

[16] 前田将貴：Vision transformer を用いた食事画像からのカロリー量推定，電気通信大学修士論文 (2023).