画像生成モデルによるカロリー量を考慮した食事画像編集

山本 耕平[1,a)]　大岸 茉由[1,b)]　柳井 啓司[1,c)]

概要

画像内の食事を指定されたカロリー量に変化させる、食事画像編集モデルを提案する。まず、入力画像を領域分割、カテゴリ・カロリー量認識を行う。その後、入力画像を線画に変換し、線画と食事領域を認識カロリー量と希望カロリー量を基にリサイズする。これらを条件付けとして、画像生成モデルに入力し、指定したカロリー量の画像を生成する。このモデルにより、視覚的に食事量を捉えることができ、食事の際に参考になる。

1. はじめに

人工知能の研究において、Transformer [11] を用いた大量のデータを学習させた言語モデル・言語視覚モデルが、急激な進化を遂げており、ChatGPT に代表される大規模言語モデルや、Stable Diffusion に代表される画像生成モデル、つまり生成 AI が、社会に浸透しつつある。しかし、想像する生成物が即時にできるわけではなく、画像生成 AI では大きさの操作や物理法則に即する画像を生成するのに時間や試行錯誤が必要となる。特に、人間の手や、食事の食べ方、数量や文字は、生成 AI が苦手とする対象である。食事画像も画像生成 AI が苦手とする対象であり、食事と数量を組み合わせたテキストを入力した場合、そのテキストに沿って画像が生成されることは稀である。

また、運動施設の充実や、新たな健康器具の開発、健康啓発活動により、社会全体での健康の意識が高まっている。健康管理の一環として食事記録を行うこともあり、自動でカロリー量の計算や栄養素の記録ができる機能が存在する。しかし、これらは食品成分表示に基づいた画一的な値になっている場合が多いため正確な記録ができず、標準量との視覚的な量の比較ができることが期待されている。

そこで本研究は、画像生成 AI と食事を結び付けたカロリー量を考慮した食事画像編集モデルを提案する。これにより、視覚的にカロリー量を捉えることができ、食べるときの参考とすることができる。

[1] 電気通信大学
[a)] yamamoto-k@mm.inf.uec.ac.jp
[b)] oghishi-m@mm.inf.uec.ac.jp
[c)] yanai@cs.uec.ac.jp

2. 関連研究

2.1 画像によるカロリー量測定

画像によるカロリー量測定には、主に 2 つの種類が存在する。基準物体を用いる方法と深層学習を用いて直接カロリー量を推定する方法である。

基準物体を用いる方法では、Smith ら [10] の研究がある。彼らの研究では、始めに基準となる紙を食品の前に配置し、紙の角を検出することによって三次元空間を認識する。次に手動で食品の 2D メッシュを構成し、三次元に投影することによって体積を算出する。最後に算出した体積をもとに、カロリー量を算出する手法である。

一方、深層学習を直接用いる方法では、會下ら [3] の研究や前田 [14] の研究がある。會下らの研究では、画像からカロリー量を推定する方法を 3 つ提案した。3 つの方法の中で、カロリー量、カテゴリ分類、食材推定、料理手順を同時学習することで、精度が向上することが分かった。

會下らの研究を基に、前田は、最新の深層学習モデルである Swin Transformer V2 と独自の出力関数である AutoBinning Softmax-Regression 関数を定義し、精度改善を図った。また、會下の研究と同様に、カロリー量測定と同時にカテゴリ推定を行う同時学習を行っている。

本研究では、基準物体が存在しない画像でもカロリー量の推定を行うため、深層学習を直接用いてカロリー量を予測する最新の手法である前田の手法を採用した。

2.2 画像生成 AI を用いた物体操作

本研究ではカロリー量の調整のため、画像内の物体の大きさを変更させる必要がある。画像生成 AI の研究には、画像内の物体の大きさや位置を操作するものもある。手法は主に、追加ネットワークを用いる方法と attention を操作する方法の二種類が存在する。

追加ネットワークを用いる方法では、ContorlNet [13] が代表的である。ContorlNet は、使用する diffusion モデルを固定し、新たに学習可能な diffusion ネットワークのエンコーダ部分を構築・学習する。これにより、diffusion モデルの生成能力を活かしながら、小さいデータセットでも過学習を避け、学習の早期収束を促すことができる。線画で

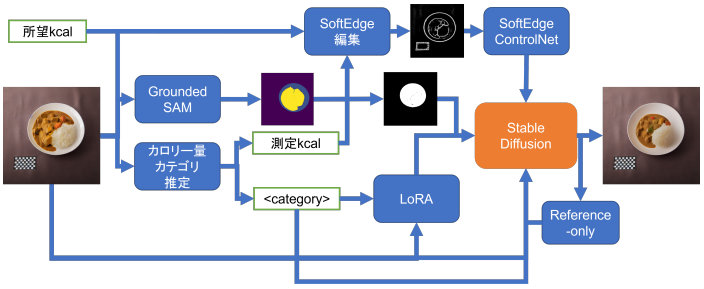

図 1 提案手法の処理の流れ。

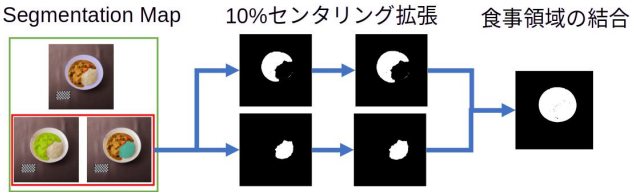

図 2 segmentation mask の微調整。

ある Canny 画像や、人間の骨格を表す人物ポーズ画像など、様々な画像を条件として画像を生成できる。制御性の高さから、画像生成 AI を用いた人物画像や漫画の制作では、よく使われている[*1] [*2]。

Attention を操作する方法では、self-guidance[4] がある。これは、diffusion における attention と activation を用いて最適化を行うことで、物体の移動、縮小拡大、外観の保存・変更を行うことができるモデルである。事前学習が必要なくシンプルであるため、どの画像生成 AI モデルでも適用することができる。しかし、編集の都度最適化を行っているため、その重みの調整や実画像との忠実性と編集能力の調整が困難である。

本研究では、不安定である attention を用いて画像を操作する手法は用いず、追加ネットワークである ContorlNet を用いて、画像の大小の編集を行った。

3. 手法

3.1 手法概要

提案手法は、図 1 のような構造とした。入力は、入力画像と変換後に欲しいカロリー量である。最初に、入力画像をカロリー量・カテゴリ推定器に入力し、カロリー量とカテゴリを得る。また、Grounded-Segment-Anything(Grounded-SAM)[*3]により食事と皿の矩形領域と segmentation mask を得る。次に、Grounded-SAM から得た segmentation mask を微調整する。微調整された segmentation mask と所望するカロリー量と測定したカロリー量の比から、入力画像の SoftEdge 画像を編集する。この編集された SoftEdge 画像が物体の形状を定める。さらに、元画像の外観を保存するため、一枚の画像から LoRA [5] の学習を行い、事前に Stable Diffusion の入力画像の特徴を得る Reference-only[*4]を用いる。最後に、入力画像、segmentation mask、編集 SoftEdge 画像、LoRA、Reference-only を用いて、カロリーの調整された画像を生成する。

[*1] https://note.com/takahiroanno/n/n649cac118429
[*2] https://note.com/maruhidd/n/n40c74c02f9a5
[*3] https://github.com/IDEA-Research/Grounded-Segment-Anything
[*4] https://github.com/Mikubill/sd-webui-controlnet/discussions/1236

3.2 カロリー量認識

カロリー量を考慮する画像生成を行うにあたり、入力された画像の現状のカロリー量がいくらか推定する必要がある。よって、画像からカロリー量を推定する、カロリー量推定器が必要になる。

本研究のカロリー量推定器には、前田 [14] のモデルを用いた。前田のモデルは、會下ら [2] のカロリー量情報付き食事画像データセットで学習されている。會下らが作成したカロリー量情報付き食事画像データセットのカテゴリ数は 15 で、画像枚数は総計 4,877 枚である。データセットは、カロリー量が記載されている 6 つのレシピサイトから収集されたものである。

本研究においては、より高いカロリー量の検出精度を求めるため、會下らと同様にカロリー量付きレシピサイトを用いて、データセットの収集を行い、前田のモデルの再学習を行った。

3.3 Segmentation Mask の微調整と SoftEdge 編集

ここでは、実際に量を調節するための SoftEdge 画像を編集する。Grounded-SAM で得た segmentation mask は、SoftEdge 画像の編集を行うには狭く、食事領域が重複して存在する場合があるため、線画が編集が行える 10%の中央拡大と、重複食事領域の結合を行う。その様子を、図 2 に示す。

最初に、同じ皿の中に存在している食事ラベルを検出する。皿の矩形領域の左上頂点と右下頂点にそれぞれ余白の 40 ピクセルを引き足しした矩形内に存在する食事領域を検出する。検出した食事領域は、食事領域の結合のため、その食事の矩形領域を中心に 10%の中央拡大処理を行う。この中央拡大処理を行うことで、Grounded-SAM で得た余白が少ない segmentation mask をうまく結合させることができ、SoftEdge 画像編集時も余裕のある線画の切り取りを行うことができる。

最後に同じ皿の中に複数食事領域が存在する場合は、segmentation mask を結合し、新たに結合した segmentation mask を作成する。

Segmentation mask の微調整によって、1 つの皿領域に対して 1 つの食事領域を持つようになる。これを元に図 3 のように欲しいカロリー量に食事領域を変化させるための SoftEdge 画像の編集を行う。

SoftEdge 画像の編集では、まず入力画像を SoftEdge 画

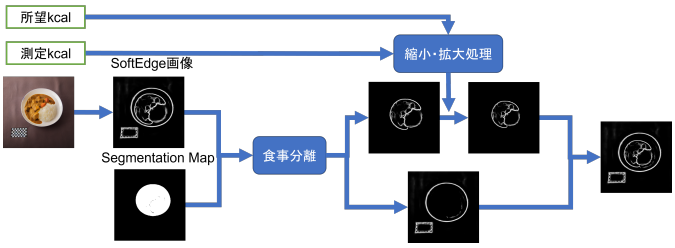

図 3 SoftEdge 画像の編集。図は、所望カロリー量 400kcal、計測カロリー量 600kcal として、縦と横サイズを $\frac{2}{3}^{\frac{1}{3}} \fallingdotseq 0.874$ 倍に SoftEdge 画像をリサイズした場合である。

表 1 追加で収集したカロリー量付きデータセット

料理名	レシピ数	画像数
カレー	149	3,443
チャーハン	107	3,334
炊き込みご飯	217	4782
ピラフ	78	1,632
ポテトサラダ	83	1,703
スパゲッティ	93	3,399
肉じゃが	153	3,872
ちらし寿司	44	229
グラタン	83	1,521
ハンバーグ	144	5,011
味噌汁	34	440
オムライス	10	128
シチュー	236	6,242
焼きそば	65	1,299

像に変換する。SoftEdge 画像は、Holistically-nested Edge Detection(HED) [12] によって生成される。segmentation mask の微調整で得た食事マスクを使い、食事領域と背景の分離を行う。次に、食事領域の SoftEdge 画像を編集する。編集は、欲しいカロリー量と測定したカロリー量を体積とみなして、縮小・拡大処理を行う。所望するカロリー量が 400kcal、測定したカロリー量が 600kcal の場合、体積比 (カロリー量比) は $4:6=\frac{2}{3}:1$ となる。これを縦・横それぞれ $\left(\frac{2}{3}\right)^{\frac{1}{3}} \fallingdotseq 0.874$ 倍することで、所望する体積 (カロリー量) になるとみなし、中央縮小処理を行う。最後に、編集した食事領域 SoftEdge 画像を背景領域 SoftEdge 画像に合成することで、カロリー量を考慮した食事画像を生成するための条件画像を得る。

なお背景以外の segmentation map は、画像生成 AI の inpainting マスクとして用いるため、カロリー量を増加させる場合には、ここで皿のマスクもその比率に合わせて拡大を行う。

4. 実験

4.1 カロリー量認識

前田らが使用した會下らのデータセット [2] の種類を参考に、14 種類の料理について、インターネット上のレシピサイトから、レシピと画像を収集した。表 1 に収集した料理名、レシピ数、画像数を示す。

會下らのデータセットでは、総計 4,877 枚であったところ、今回収集した画像は、総計 37,035 枚であり、約 7.6 倍となっている。カテゴリごとにレシピ数や画像枚数にばらつきはあるものの、各カテゴリ 100 枚以上、1000 枚超のカテゴリも存在する。レシピが多いほど、カロリー量の偏りも少なくなるため、カロリー量認識器の学習に適するようになる。

表 2 カロリー認識の結果。太文字は、全データで最もよい精度を誇っているもの。赤文字は、それぞれのデータセットで優位であるもの。

		オリジナルモデル		再学習モデル	
		會下らのデータ	収集データ	會下らのデータ	収集データ
カロリー	絶対誤差 [kcal] ↓	87.5	161.0	166.4	**80.0**
	相対誤差 [%] ↓	27.8	68.1	61.7	**25.5**
	誤差 20%以内割合 [%] ↑	53.6	30.1	24.0	**62.3**
カテゴリ	Top-1[%] ↑	**89.2**	27.9	46.8	73.0

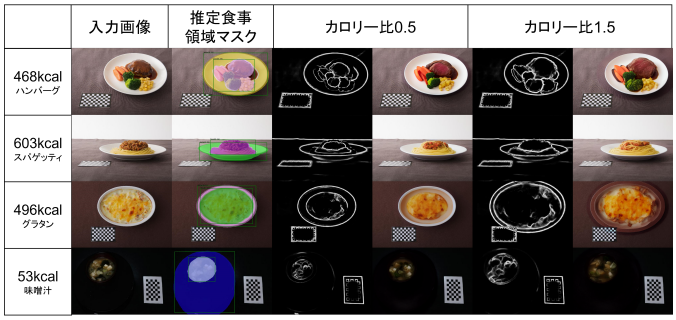

図 4 生成画像の結果。入力画像の左側にはカロリー量推定器で推定したカロリー量とカテゴリを記述している。

これらの今回収集した画像のみを使って、前田らのモデルを再学習し、それぞれのデータセットで評価を行った。表 2 に、前田らのモデルと再学習したモデルのカロリー認識の結果を示す。また、評価の際に用いられている画像は、各データの 1/5 であり、會下らのデータでは、約 1,000 枚、今回収集したデータでは、約 7,400 枚である。

結果から、それぞれ学習したデータセットのカロリー量認識率が高いことがわかる。しかし、学習枚数の多さと画像の解像度が高さ、評価枚数を考えると、本研究の再学習モデルが有効であると考えられる。カテゴリ推定については、オリジナルモデルの方が有効に見える。しかし、會下らのデータセット 7 つのレシピサイトの整備されたデータを用いているため、レシピや写真の撮影方法が似ているためとも考えられる。本研究では、学習枚数も評価枚数も多い収集データを用いた、再学習モデルを用いて、カロリー量を調節した画像生成を行うことにした。

4.2 生成画像

生成画像の結果を図 4 に示す。どの画像でも、カテゴリ推定があっており、カロリー量の値も妥当なものになっている。また、segmentation map が取れており、SoftEdge 画像の編集がなされ、生成画像の食事領域が変化していることがわかる。一番上の段のハンバーグ (hamburger steak) のように、“steak” など、別の食品の単語があると、元画像とは外観が離れてしまう場合も存在することがわかる。し

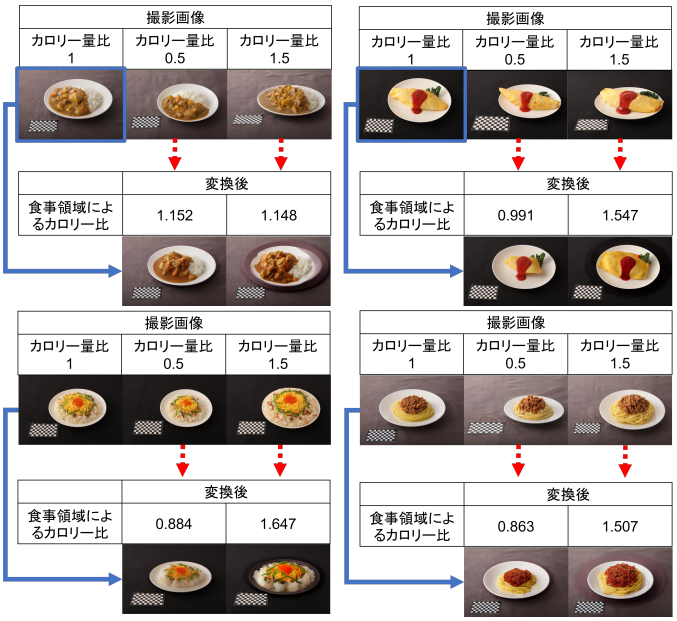

図 5 撮影画像と生成画像との比較。各カテゴリの左上の画像を入力として、同じ列のカロリー量 (基準カロリー量の 0.5 倍、1.5 倍) を入力したときの出力画像を下段に示している。撮影画像は、食事のサイズにより皿の大きさや画角が異なる。なお，0.5 倍，1.5 倍の撮影画像は比較のための参考画像であり，処理には使用していない.

かし、付け合わせなどは比較的形や色を保っており、食事量も変わっている。また、一番下段の味噌汁の画像については、かなりの領域が濃い黒色で染まっており、SoftEdge 画像で、皿の輪郭が取れない場合も存在することがわかる。こういった場合でも、食事量の変化を見ることができる。

4.3 実画像との比較

ここでは、独自に撮影したカロリー量付き画像と比較を行って、どの程度現実の食事量と差異があるかを示す。図 5 に、撮影画像・生成画像を示す。なお、撮影画像を同士を見比べると、食事のサイズにより皿の大きさや画角が異なる。なお、入力画像と変換後の画像の Grounded-SAM で検出した食事領域の比を面積比と捉えたとき、その体積比をカロリー比とみなすこととし、「(変換後の画像の画面全体に対する食事領域の割合/入力画像の画面全体に対する食事領域の割合)$^{(\frac{3}{2})}$」と計算した値を食事領域によるカロリー比とした。

撮影画像と生成画像を比較すると、形が異なるものも存在するものの、見た目の食事量は似ていると言える。撮影画像は、それぞれの皿のサイズも異なるため、単純な比較ができない。しかし、生成画像は入力画像と比較することで、量の大小を認識することができる。また、食事領域によるカロリー比についてみてみると、カロリー比 0.5 で編集した場合は指定したカロリー量比と遠い値となったが、カロリー比 1.5 で編集したものでは、指定したカロリー量比と近い値となっている。

表 3 各カテゴリ、各カロリー比に対して生成画像 100 枚に対する食事領域によるカロリー比．数値は，平均値±標準偏差 を示す.

食事カテゴリ	推定カロリー比 (0.5 倍時)	推定カロリー比 (1.5 倍時)
肉じゃが	0.522 ± 0.0417	2.485 ± 1.3268
炒飯	0.507 ± 0.0378	1.905 ± 1.0061
ちらし寿司	0.428 ± 0.1687	1.257 ± 0.5095
カレー	0.559 ± 0.0714	1.558 ± 0.6483
焼きそば	0.511 ± 0.0112	1.530 ± 0.2707
グラタン	0.498 ± 0.0544	1.397 ± 0.2116
ハンバーグ	1.184 ± 0.1941	2.683 ± 1.4070
味噌汁	1.321 ± 1.6865	3.576 ± 1.5958
炊込みご飯	1.505 ± 0.9103	1.716 ± 0.2736
オムライス	0.818 ± 0.1661	1.932 ± 0.9982
ピラフ	0.511 ± 0.0104	1.561 ± 0.5165
ポテトサラダ	0.730 ± 0.2682	1.523 ± 0.3880
スパゲティ	0.374 ± 0.1403	1.260 ± 0.6171
シチュー	0.572 ± 0.1448	1.381 ± 0.4155
全カテゴリ平均	0.717 ± 0.2790	1.840 ± 0.7275

4.4 定量評価

このタスクには、決まった定量評価は存在しない。このカロリー量を考慮した画像生成においては、どれくらい実際に希望のカロリー量に即した食事領域の増減がなされているかが重要である。よって、4.3 節と同様に、食事領域によるカロリー比を計測することで、その評価を行った。

表 3 に各カテゴリ、各カロリー比に対して生成画像 100 枚に対する食事領域によるカロリー比を示す。

おおよそ指定したカロリー比に近い値が、食事領域によるカロリー比となっている。しかし、カテゴリによって差異があり、カレーや焼きそば、ピラフといった適切な値になるものや、ハンバーグや味噌汁、オムライスといった過剰にカロリー比が大きくなるものが存在する。平均値を見てみると、やや大きいカロリー比を取ることが多いとわかる。また、分散についてみてみると、やはりハンバーグや味噌汁は、値が散らばっていることがわかる。最後に標準偏差を見てみると、平均では 0.3 ほど変化していることがわかるが、カロリー比が 0.01 と小さい値しか変化しない組み合わせや、1 以上大きくなるカテゴリが混在していることがわかる。

5. おわりに

今回は、カロリー量を考慮した食事編集のためモデルの構築を行った。本研究の機構は簡素であり、入力画像のカロリー量を直接入力することで、カテゴリ関係なく食事量の増減ができる。また、カテゴリ推定やカロリー量推定さえできれば、編集可能な食事の種類を簡単に増やすことができる。今後の課題には、カテゴリ推定・カロリー推定できる食事品目を増やすことや、食事領域拡大時の皿のアーティファクトを軽減すること、食事の材料や盛り付け方、奥行きの考慮できるモデルを構築することがあげられる。

参考文献

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al.: An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929* (2020).

[2] Ege, T. and Yanai, K.: Image-based food calorie estimation using knowledge on food categories, ingredients and cooking directions, *Proc.of ACM International Conference on Multimedia*, pp. 367–375 (2017).

[3] Ege, T. and Yanai, K.: Simultaneous estimation of food categories and calories with multi-task CNN, *Proc.of IAPR International Conference on Machine Vision Applications (MVA)*, pp. 198–201 (2017).

[4] Epstein, D., Jabri, A., Poole, B., Efros, A. A. and Holynski, A.: Diffusion Self-Guidance for Controllable Image Generation, *Proc.of Neural Information Processing Systems* (2023).

[5] 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).

[6] Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I. and Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding, *Proc.of IEEE International Conference on Computer Vision*, pp. 1780–1790 (2021).

[7] Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N. et al.: Grounded language-image pre-training, *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 10965–10975 (2022).

[8] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows, *Proc.of IEEE International Conference on Computer Vision*, pp. 10012–10022 (2021).

[9] Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J. and Sun, J.: Objects365: A Large-Scale, High-Quality Dataset for Object Detection, *Proc.of IEEE International Conference on Computer Vision*, pp. 8429–8438 (2019).

[10] Smith, S. P., Adam, M. T. P., Manning, G., Burrows, T., Collins, C. and Rollo, M. E.: Food Volume Estimation by Integrating 3D Image Projection and Manual Wire Mesh Transformations, Vol. 10, pp. 48367–48378 (2022).

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need, *Proc.of Neural Information Processing Systems*, Vol. 30 (2017).

[12] Xie, S. and Tu, Z.: Holistically-nested edge detection, *Proc.of IEEE International Conference on Computer Vision*, pp. 1395–1403 (2015).

[13] Zhang, L., Rao, A. and Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models, *Proc.of IEEE International Conference on Computer Vision* (2023).

[14] 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).