一般発表 | Track 5: 高度なデータ利活用・ドメイン応用（医療情報，教育，地理情報等）

2026年3月2日(月) 13:00 ~ 15:10 | I会場

[8I] マルチメディア

座長:北山 大輔(工学院大学)　コメンテータ:鎌原 淳三(大阪成蹊大学)　ジュニアコメンテータ:小島 優希也(東京都立大学)

13:00 ~ 13:20

[8I-01] Predicting YouTube Video Popularity Using Multi-Aspect Content and Engagement Features

*沈 秋桐[1]、木村 昌臣[1,2] (1. 芝浦工業大学、2. マレーシア・クランタン大学)

13:20 ~ 13:40

[8I-02] マルチモーダルLLMを用いた美術実技能力の向上支援

*三好 香利奈[1,2]、牛尼 剛聡[3] (1. 福岡県立春日高等学校、2. 九州大学未来創成科学者育成プロジェクト　QFC-SP、3. 九州大学大学院芸術工学研究院)

13:40 ~ 14:05

[8I-03] Diffusionモデル内にキャラクター情報の制約を追加したアニメ画像生成モデルの構築

*木畑 光貴[1]、落合 桂一[1]、戸田 浩之[1] (1. 横浜市立大学)

14:05 ~ 14:30

[8I-04] 高精度自車位置推定のための車載カメラ画像と三次元点群地図間の局所特徴照合に関する検討

*高津 悠生[1]、道満 恵介[2]、川西 康友[3]、久徳 遙矢[1] (1. 愛知工科大学、2. 中京大学、3. 理化学研究所)

14:30 ~ 14:55

[8I-05] ユニバーサル基板を使用した電子回路の不良原因推定に向けた深層学習による電子部品とその端子検出

*小泉 天翔[1]、林 哲矢[1]、田中 剛[1]、遠藤 雅樹[1]、寺田 憲司[1]、大野 成義[1] (1. 職業能力開発総合大学校)

Predicting YouTube Video Popularity Using Multi-Aspect Content and Engagement Features

Qiutong SHEN † and Masaomi KIMURA ‡

† Data Science/Engineering Lab, Shibaura Institute of Technology, Tokyo, Japan
†† University Malaysia Kelantan, Malaysia
E-mail: † ma24006@sic.shibaura-it.ac.jp, ‡ masaomi@sic.shibaura-it.ac.jp

Abstract Understanding which factors influence video popularity is important for online video platforms. This study analyzes YouTube video view using a dataset of trending videos collected from ten countries. After basic data cleaning and consolidation, multi aspect features are constructed to describe video content and user engagement, including features extracted from titles and tags, together with temporal and video level attributes. Video popularity is examined from different perspectives, and view prediction is treated as a regression task to study how content related, and time related factors are associated with viewer attraction. Due to the long-tailed distribution of view, the data is divided into low view and high view subsets, which are modeled using different prediction models. Videos with low view are modeled using a multilayer perceptron based on structured features. Videos with higher view are modeled using a collaborative tag aware graph neural network, which captures relationships between videos through shared tags and additional contextual information such as category and country. Experimental results show that the proposed approach improves prediction accuracy compared to baseline models and helps identify factors associated with video popularity on YouTube.

Keyword Social Media, Video, View Prediction, YouTube

1. Introduction

Video popularity is an important topic for online video platforms, as views reflect user attention and content visibility. On platforms such as YouTube, a large number of videos compete for limited attention, while only a small portion of videos achieve very high view. As a result, predicting video popularity and understanding which factors influence views remain challenging problems. Video popularity is affected by multiple factors related to content, timing, and user interaction, and the influence of individual features on views can vary across features.

Existing studies have explored video popularity prediction using different features and models, but prediction accuracy remains limited, and video information is often not deeply analyzed from multiple perspectives. In particular, textual information and timing related factors are not always fully considered, and the uneven distribution of video popularity is often not explicitly addressed. In this study, video popularity is analyzed using multi aspect features, and a distribution aware modeling strategy is adopted to better handle videos with different popularity levels.

2. Related Work

Video popularity prediction has been widely studied as a supervised learning problem using content, metadata, and engagement features. Early YouTube-focused studies relied on structured features such as views, comments, ratings and categories, and applied regression or classification models to predict popularity or popularity levels [1, 2, 3]. While these works demonstrated the usefulness of engagement and temporal features, they typically employed a single global model and did not explicitly address the highly long-tailed distribution of video views, leading to biased performance dominated by low-view samples.

To handle heterogeneity in popularity, several studies proposed two-stage or level-based prediction frameworks. Ouyang et al. first predicted future popularity levels and then applied specialized regressors conditioned on level transitions, showing improved accuracy over single model baselines [4]. Related classification approaches also grouped videos into popularity categories before prediction [2]. These methods share our motivation to account for popularity heterogeneity, but their regime definitions are typically heuristic, and routing decisions are optimized for level prediction rather than end-to-end view regression. In contrast, our work defines popularity regimes using data-driven knee-point detection and selects routing thresholds by directly optimizing regression performance.

Another line of research emphasizes temporal dynamics of popularity. SMTPD introduced a large-scale benchmark for temporal popularity prediction and showed

that early popularity trajectories are highly predictive of future views[5]. Knowledge-graph-based temporal models further combined sequential modelling with relational reasoning to predict popularity evolution [6]. While effective in settings with rich temporal signals, these approaches rely on historical trajectories. Our work instead focuses on a static or early-stage setting, where popularity must be inferred from content, engagement, and relational structure without long observation windows.

Recent studies explored richer semantic representations through deep and multimodal models. Some approaches incorporated visual cues from video covers together with textual features, demonstrating that visual–textual fusion can improve popularity prediction [7]. Others employed heavy multimodal pipelines or retrieval augmented models to capture cross-video semantic similarity [8,9]. Very recent work further leveraged large language models to predict popularity and provide explanations, often using newly collected large-scale datasets [10]. While these methods enhance semantic expressiveness, they typically apply a single complex model to all videos. Our approach instead adopts lightweight textual semantics and emphasizes regime-dependent modeling.

Graph-based methods introduce relational structure into popularity prediction by modeling inter-video dependencies. GraphInf used graph convolutional networks to capture influence among videos in short-video networks [11], while collaborative tag-aware graph neural networks showed that tag-mediated message passing effectively captures implicit similarity in long-tail recommendation settings [12]. These works motivate our CTGNN design, but existing graph popularity models generally apply graph inference uniformly to all items. In contrast, we restrict graph modeling to the high-view regime, where relational neighbourhoods are sufficiently dense, and fuse multiple graph-derived and direct feature representations through a gating mechanism.

Interpretability-oriented studies further analysed feature contributions in popularity prediction. Xie et al. highlighted the dominant role of engagement-related signals in deep learning models for YouTube viewership prediction [13]. Our permutation feature importance analysis for low-view videos is consistent with these findings and provides additional insight by linking feature dominance to regime-specific data characteristics. More broadly, practical lessons from large-scale click prediction systems emphasize robust modeling under extreme imbalance and the importance of strong feature baselines, while advances in weakly supervised consistency learning provide methodological background for structure aware modeling under limited supervision [14].

Overall, prior work has explored feature-based, temporal, multimodal, and graph-based approaches to video popularity prediction. Our contribution differs by explicitly modeling the long-tailed distribution through a distribution aware routed framework that aligns model choice with data structure: a MLP for low-view videos and a collaborative tag aware graph neural network for high-view videos, with data-driven regime definition and prediction-oriented routing.

3. Methodology

3.1 Overview

We propose a routed prediction framework to address the long-tailed nature of video view. Given an input video, a routing classifier estimates the probability that the sample belongs to a high-view regime. The sample is then routed to one of two specialized regressors: (i) an MLP predictor for lower-view videos using structured features, and (ii) a collaborative tag aware graph neural network (CTGNN) for higher-view videos that leverages relational signals induced by shared tags and contextual nodes. The full model is trained and evaluated under a consistent high/low definition and routing policy, with selection of thresholds and hyperparameters reported in the experimental section.

3.2 Feature Representation

Each video is represented by multi-aspect features capturing engagement, video metadata, textual semantics, and external attention.

Structured features include engagement statistics (e.g., likes, dislikes, comment count), video metadata (e.g., duration, title length, tag count), and time related attributes derived from publishing and trending timestamps (e.g., interval from publish to trending). A binary weekend indicator is incorporated only in the MLP branch to provide a lightweight temporal cue for low-view prediction, while the graph model focuses on relational/content signals and contextual nodes; in practice, such coarse temporal indicators tend to be weakly discriminative for the high-view regime and can be partially absorbed by category/country context.

To capture semantic information beyond scalar features, we include (i) dense textual embeddings derived from title content (e.g., word2vec-based embeddings) and (ii) category representations. Country information is used

explicitly for the routing and MLP predictors (e.g., via one-hot encoding) and is also modelled in CTGNN through a dedicated country node type.

To quantify external media attention related to video content, we construct a keyword set from video titles and use these keywords to retrieve the number of related news mentions within a fixed time window around the publish date. The maximum mention count among the retained keywords is used as a compact proxy feature for external attention. The concrete window length and data source used for mention retrieval are described in the experimental setup.

Titles may appear in multiple languages. For keyword extraction, we prioritize the original title when its language is supported; otherwise, we fall back to an English translated title to ensure consistency for downstream processing.

We apply YAKE to extract candidate key phrases from each title. Stop words are constructed as the union of language-specific stopword lists across supported languages. Extracted phrases are normalized by lowercasing and punctuation removal, and phrases that are empty, too short, or stopwords are discarded. YAKE is configured to extract up to trigrams, reflecting the observation that many salient title concepts are multi-word expressions rather than isolated tokens.

In addition to per-title extraction, we construct a corpus TF-IDF vocabulary over all unique titles. TF-IDF is computed using an n-gram range of 1–3 to capture both single word keywords and multi-word entities (e.g., person names, events, products). This 1–3 gram design is intentional: unigram-only vocabularies often fragment important phrases, whereas trigrams allow the model to preserve semantically coherent expressions that better align with how titles convey topics.

To improve robustness against noisy or overly generic terms, the final keyword set for each title is defined as the intersection between (i) the filtered YAKE keywords extracted for that title and (ii) a global set of high-importance TF-IDF n-grams. This design keeps keywords that are both salient locally (YAKE) and informative globally (TF-IDF), reducing sensitivity to either method's failure modes.

3.3 Proposed Model

Our proposed model consists of three components: a routing classifier, a low view regressor based on a multilayer perceptron, and a collaborative tag aware graph neural network for high view prediction, as illustrated in Figure 3.3. The routing classifier outputs a probability p_{high} indicating whether a video belongs to the high view regime. During inference, p_{high} is compared with a routing threshold to determine which prediction branch is activated. The boundary between high and low view regimes is determined in a data-driven manner via knee-point detection on the training view distribution, while the routing threshold is selected based on validation performance.

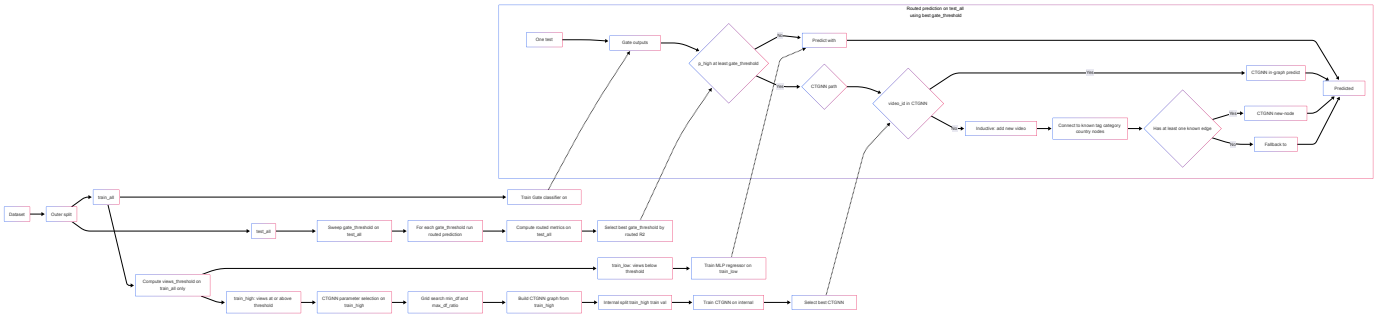


Table 3.3 Overall Workflow of the Gate-Based Routing Framework

3.3.1 Low-view Regressor (MLP)

For lower-view videos, we use an MLP regressor trained on standardized structured features. Figure 3.3.1 summarizes the architecture: repeated Linear → BatchNorm → LeakyReLU → Dropout blocks followed by a final linear output layer. Both inputs and targets are standardized on the training low split, and predictions are inverse transformed to the original view scale for reporting.

Figure3.3.1 Flowchart of the MLP Regressor

3.3.2 High-view Regressor (CTGNN)

CTGNN operates on a heterogeneous graph with four node types:

- video nodes: one node per unique video,
- tag nodes: representing tags associated with videos,
- category nodes: representing the video category,
- country nodes: representing the market/context where the video is observed.

Edges are created from each video to its associated tag, category, and country nodes, together with reverse edges to enable bidirectional message passing.

To mitigate noise from extremely rare tags and reduce the dominance of overly frequent tags, we filter tags using document frequency (DF) constraints computed over the set of videos used to build the graph. Let N be the number of videos and $DF(t)$ the number of videos containing tag t. A tag is retained if:

$$\min\_df \leq DF(t) \leq [\max\_df\_ratio \cdot N]$$

The lower bound min_df removes extremely rare tags that are unlikely to yield reliable collaborative signals, while the upper bound max_df_ratio suppresses overly frequent tags that act as hubs and introduce indiscriminate connectivity.

CTGNN derives multiple source-wise representations for each video through heterogeneous message passing over tag, category, and country nodes, inspired by collaborative tag aware graph learning that leverages tags to capture implicit similarity among items [12].

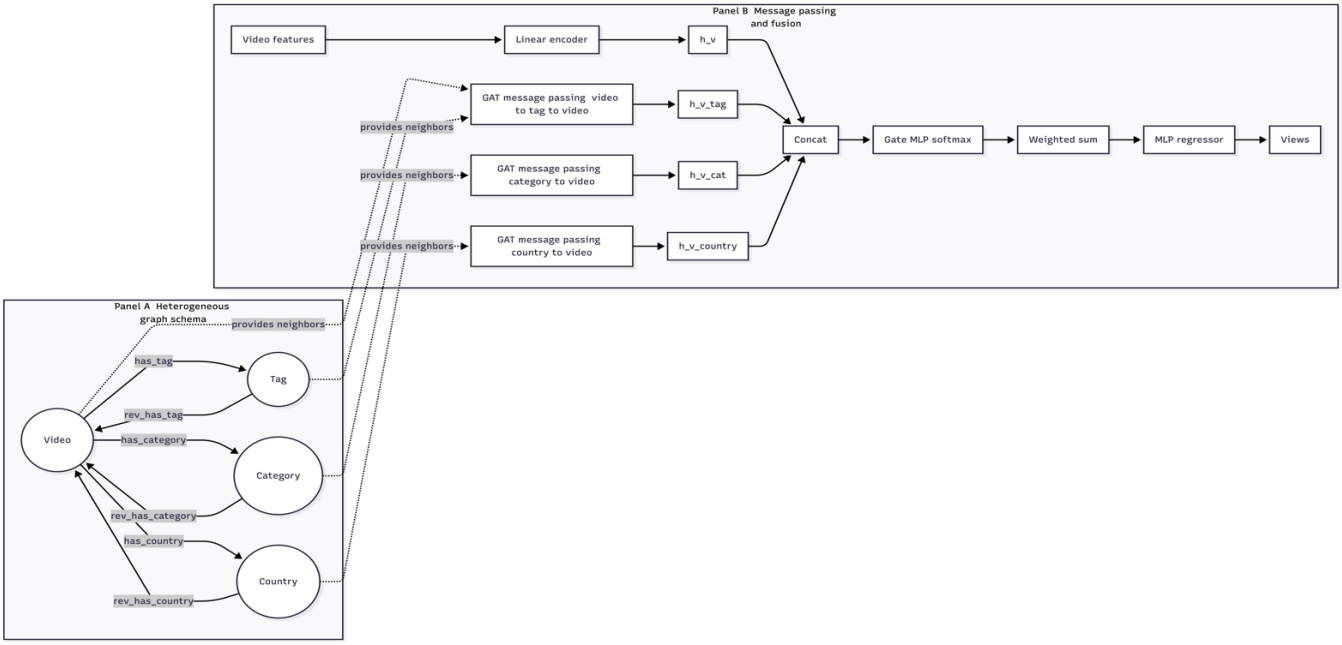


Figure 3.3.2 CTGNN Architecture and Message Passing Flow

As illustrated in Figure 3.3.2, collaborative information is propagated by performing message passing from video nodes to tag nodes and subsequently from tag nodes back to video nodes. This mechanism enables videos sharing common tags to exchange information indirectly. Category and country nodes supply contextual features, which are fused through a gating network for prediction.

CTGNN produces four video representations from different information sources: the tag collaborative representation h_t, the category contextual representation h_c, the country contextual representation h_{co}, and the direct feature representation h_v obtained from the video encoder. Since the reliability of these sources can vary across videos (e.g., some videos have sparse tag connections while others benefit strongly from tag-mediated neighbors), we adopt a learnable gating mechanism to adaptively weight them on

a per-video basis.

Concretely, we first concatenate the four representations and feed them to a small gating network $g(\cdot)$ that outputs a 4-dimensional weight vector:

$$\alpha = softmax(g([h_t; h_c; h_{co}; h_v])), \alpha \in \mathbb{R}^4$$

The function $g(\cdot)$ acts as a source-wise gating mechanism that adaptively determines the relative importance of different information sources for each video. By taking the concatenated representations as input, it produces fusion weights that allow the model to emphasize reliable collaborative or contextual signals while attenuating noisy or weak ones. Unlike node attention used during message passing, this gating mechanism operates at the representation level and controls how multiple source-specific embeddings are combined for final prediction.

Here, α_i indicates the relative importance assigned to the source for the current video. The softmax operation ensures $\alpha_i \geq 0$ and $\sum_{i=1}^{4} \alpha_i = 1$, making the fusion a convex combination and keeping the weights interpretable. The fused representation is then computed as:

$$h_{fuse} = \alpha_1 h_t + \alpha_2 h_c + \alpha_3 h_{co} + \alpha_4 h_v$$

Finally, a lightweight MLP is adopted as a regression head to map the fused representation h_{fuse} to the predicted view count. As CTGNN focuses on representation learning over the heterogeneous graph, the MLP acts as a standard readout module that converts the graph aware embedding into a scalar output. Combined with the gating mechanism, this design allows the model to emphasize relational and contextual signals when they are informative, while naturally falling back to the direct feature pathway when graph evidence is weak or unavailable.

4. Experiment

4.1 Dataset and Preprocessing

The dataset is based on the Trending YouTube Video Statistics collection from Kaggle, which provides daily statistics for trending YouTube videos. Trending videos collected between November 14, 2017 and June 14, 2018 are used in this study. Data from ten countries are included, namely Canada, Germany, France, the United Kingdom, India, Japan, South Korea, Mexico, Russia, and the United States. Each record contains basic video metadata such as a unique video identifier (video_id), title, publish time, trending date, category, and engagement statistics.

Since the same video has appear multiple times across different trending dates in the dataset, the raw dataset contains duplicate records for a single video. In this study, a collaborative tag aware graph neural network is constructed using video_id as the unique node identifier. Without proper deduplication, records from different trending dates would be merged into the same node, causing features, labels, and edges from different time points to be incorrectly combined. To ensure semantic consistency in graph construction, duplicate videos with the same video_id are removed by retaining only the record with the most recent trending date. Videos with missing or incomplete information are also excluded.

After data cleaning, a set of multi-aspect features is constructed to describe different properties of each video. These features include information related to video duration and the market where the video trended, as well as temporal characteristics derived from publish time and trending date, such as the number of days from publishing to trending, whether the video was published on a weekend, and the day of the week of publication. Text based features are extracted from video titles and descriptions, including sentiment scores, title length measured by word count, and the number of tags. For videos whose titles are not in English, titles are translated into English to ensure consistency in text analysis.

We extract compact keywords from each video title using YAKE, and further filter them using global TF-IDF statistics.

YAKE extraction. For each row, we select the original title if its language is supported; otherwise, we use the English translated title. We extract up to 3 keywords per video using YAKE with n=3 (allowing up to trigrams), returning the top candidates after stopword and punctuation filtering.

Global TF-IDF vocabulary. To obtain a global set of salient phrases, we compute TF-IDF on de-duplicated titles using $ngram\_range =$ (1,3). We then rank terms by their average TF-IDF across documents and select the top-N terms. The final keyword set for each video is the intersection between its YAKE keywords and this global top-N vocabulary, plus any global top N terms that appear in the title.

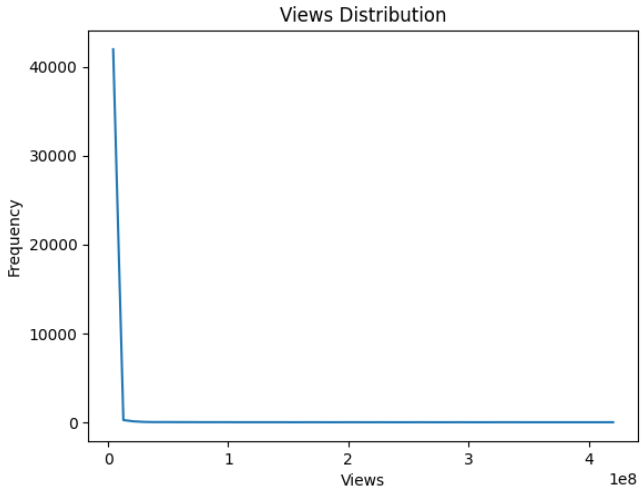
We sweep N and measure how much the global vocabulary covers the extracted YAKE keywords. We select N at the point where the marginal increase in coverage becomes small (coverage gain < 0.01 between consecutive candidates), and use a fixed N (e.g., 30k) for stable downstream processing.

4.2 Training Protocol and Motivation

In preliminary experiments, we observed that using a single MLP regressor to predict video views often leads to unsatisfactory performance.

Due to the long-tailed nature of the view distribution as shown in Figure 4.2, the model is dominated by low-view samples, which results in biased predictions and systematic underestimation for highly popular videos.

Figure 4.2 Views Distribution



These observations motivate a regime-based modeling strategy, where videos with different popularity levels are handled by specialized models.

The gate model is trained on the full training set using standardized features, with a fixed internal train–validation split used for validation and performance reporting.

MLP is trained exclusively on the training low subset, allowing the model to specialize in predicting view for low popularity videos.

CTGNN is trained on the training high subset. Within this subset, 80% of the video nodes are used for parameter learning, while the remaining 20% are held out for validation.

4.3 High/Low Definition: views_threshold from Knee Point

To define high-view vs low-view regimes in a data-driven manner, we compute a threshold on the training view distribution using a knee-point method in log scale. The knee index is obtained by the maximum distance to the line connecting the endpoints of the normalized curve. Videos with $views \geq threshold$ are treated as high-view.

4.4 CTGNN Tag Filtering Parameter Selection (min_df, max_df_ratio)

We select min_df and max_df_ratio through a grid search. For each setting, we evaluate:

- validation R^2 / RMSE on training high internal validation nodes,
- graph structure health: (i) vocabulary size, (ii) fraction of videos with zero kept tags, and (iii) average kept tags per video.

We first filter settings that violate structure constraints (e.g., too many zero tag videos or too small tag vocabulary). Among feasible settings, we select a near best R^2 configuration (within a slack) and break ties using a composite score that rewards both predictive performance and healthy graph connectivity.

	min_df	max_df_ratio	val_r2	...	vocab_size	mean_kept_tags	score
4	20	0.05	0.934093	...	214	3.112689	2.592880
2	10	0.05	0.917212	...	539	5.115437	3.025861
5	20	0.10	0.838886	...	217	3.331196	2.522289
3	10	0.10	0.825588	...	542	5.333944	2.946466
0	5	0.05	0.819763	...	1453	7.774622	3.351136
1	5	0.10	-0.168826	...	1456	7.993129	2.368985

4.5 Routing Threshold Selection (gate_threshold)

We sweep the routing threshold $threshold_g$ on the outer test set over a predefined range and select the value that maximizes routed regression R^2. This choice reflects the goal of optimizing end-to-end prediction quality rather than pure classification accuracy.

gate_threshold	gate_acc	...	routed_r2	TN_FP_FN_TP
0.93	0.958643	...	0.930040	[7759, 146, 205, 377]
0.95	0.959350	...	0.929946	[7777, 128, 217, 365]
0.94	0.958996	...	0.929924	[7770, 135, 213, 369]
0.92	0.958053	...	0.929762	[7748, 157, 199, 383]
0.91	0.958525	...	0.929717	[7746, 159, 193, 389]
0.90	0.957936	...	0.929673	[7737, 168, 189, 393]
0.87	0.957111	...	0.929381	[7717, 188, 176, 406]
0.85	0.955815	...	0.929210	[7697, 208, 167, 415]
0.83	0.955108	...	0.928965	[7681, 224, 157, 425]
0.80	0.954519	...	0.928951	[7666, 239, 147, 435]
0.70	0.948627	...	0.928245	[7591, 314, 122, 460]

4.6 Evaluation

Metrics include RMSE, MAE, and R^2, reported overall and separately for true low and true high subsets. We additionally provide gate performance (accuracy, AUC, confusion matrix) and route counts to analyze model behavior.

On the outer test set, the MLP only model achieves strong performance on the True LOW subset with an RMSE of 1.95×10^5, an MAE of1.23×10^5, and an R^2 of 0.729.

For the True HIGH subset, the CTGNN only model achieves an R^2 of 0.972 with an RMSE of 5.99×10^6 and an MAE of 2.96×10^6, with full inductive coverage.

Using the routed model with the best gate threshold = 0.93, the gate achieves an accuracy of 0.959 and an AUC of 0.972. The overall routed regression attains an RMSE of 1.92×10^6, an MAE of 3.62×10^5 , and an R^2 of 0.930.

When evaluated by true popularity regimes, the routed model yields an RMSE of 4.16×10^5 ($R^2 = -0.227$) on True LOW videos and an RMSE of 6.00×10^6 ($R^2 = 0.926$) on True HIGH videos. In total, 7,964 samples are routed to

the MLP and 523 samples are handled by CTGNN via inductive inference.

4.7 Feature Importance

To analyze the contribution of individual input features to low view prediction, we compute single feature permutation importance for the MLP model on the outer test True LOW subset, with detailed results summarized in Table 4.7. A baseline root mean squared error is obtained using the original test features. For each feature independently, its values are randomly permuted across samples while all other feature columns are kept unchanged. Permutation based feature importance is adopted because it provides a model agnostic measure of feature contribution by quantifying the degradation in predictive performance when the information carried by a feature is disrupted, while keeping its marginal distribution unchanged. This permutation breaks the association between the selected feature and the target variable. The model is evaluated on the permuted data, and the increase in prediction error relative to the baseline is recorded. This procedure is repeated multiple times with different random permutations, and the mean and standard deviation of the resulting RMSE increase are reported as the feature's importance score.

Table 4.7 Feature Permutation Importance of the MLP on the True LOW Subset

feature	rmse_increase_mean	rmse_increase_std	baseline_rmse
likes	1.44E+05	8.26E+02	1.95E+05
dislikes	8.99E+04	7.40E+02	1.95E+05
tag_count	1.64E+04	1.18E+03	1.95E+05
Duration	1.47E+04	7.56E+02	1.95E+05
title_length	1.45E+04	1.06E+03	1.95E+05
comment_count	1.33E+04	2.00E+02	1.95E+05
sentiment_score_description	9.88E+03	2.19E+02	1.95E+05
sentiment_score	5.22E+03	2.64E+02	1.95E+05
interval	4.48E+03	3.33E+02	1.95E+05
is_weekend	2.06E+03	3.18E+02	1.95E+05
max_keyword_mentions	1.84E+03	4.10E+02	1.95E+05

Engagement related features, including likes, dislikes, and comment count, lead to the largest increases in RMSE when permuted, indicating that the MLP relies primarily on observable signals of audience interest for low-view prediction.

In this context, both positive and negative feedback are informative, as they reflect that viewers were sufficiently interested in the video to watch it and express an opinion. In contrast, content semantic features, such as word embeddings and sentiment scores, exhibit relatively lower importance, suggesting that early audience interest provides more reliable predictive cues than intrinsic content semantics in the low-view regime.

5. Conclusion

This paper investigates YouTube video popularity prediction using multi-aspect content and engagement features under a highly skewed view distribution. To address the heterogeneous characteristics of videos at different popularity levels, we propose a routed framework that dynamically selects between a feature-based regressor and a graph model.

Extensive experiments show that a feature-based MLP is sufficient for low-view videos, where engagement statistics and basic content attributes dominate prediction and relational structure is sparse or noisy. In contrast, high-view videos exhibit stronger and more consistent relational patterns, such as shared tags, categories, and countries, forming dense neighborhoods that cannot be effectively captured by independent feature modeling. For this regime, the collaborative tag aware graph neural network (CTGNN) becomes necessary to exploit cross-video dependencies through message passing on a heterogeneous graph.

By routing videos to different predictors according to their estimated popularity regime, the proposed model effectively matches model capacity with data structure. Further analysis confirms that engagement and attributes primarily drive low-popularity prediction, while relational and contextual information is crucial for modeling high-popularity videos. Overall, this work demonstrates that adaptive integration of content, engagement, and relational signals provides a principled and practical solution for large-scale popularity prediction.

References

[1] Mekouar, S., Zrira, N. & Bouyakhf, E.-H. Popularity prediction of videos in YouTube as case study: a regression analysis study. In: Proceedings of the 2nd International Conference on Big Data, Cloud and Applications. https://doi.org/10.1145/3090354.3090406 (2017).

[2] Li, Yuping, Kent X. Eng and Liqian Zhang. "YouTube Videos Prediction: Will this video be popular?" (2019).

[3] He, X., et al.: Practical lessons from predicting clicks on ads at Facebook. Proceedings of the Eighth International Workshop on Data Mining for Online Advertising pp. 1–9 (2014)

[4] S. Ouyang, C. Li, and X. Li, "A Peek Into the Future: Predicting the Popularity of Online Videos," IEEE Access, vol. 4, pp. 3026–3033, Jun. 2016, doi: 10.1109/ACCESS.2016.2580911.

[5] Y. Xu et al., "SMTPD: A New Benchmark for Temporal Prediction of Social Media Popularity," 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2025, pp. 18847-18857, doi: 10.1109/CVPR52734.2025.01756.

[6] Liu, P., Yu, Z., Sun, Y., Xi, M. (2023). Video Popularity Prediction Based on Knowledge Graph and LSTM Network. In: Yu, Z., et al. Data Science. ICPCSEE 2023. Communications in Computer and Information Science, vol 1879. Springer, Singapore. https://doi.org/10.1007/978-981-99-5968-6_32

[7] Y. Tian and X. Wang, "Predicting video popularity based on video covers and titles using a multimodal large-scale model and pipeline parallelism," Applied and Computational Engineering, vol. 41, no. 1, pp. 182–189, Feb. 2024, doi: 10.54254/2755-2721/41/20230741.

[8] Haixu Liu, Wenning Wang, Haoxiang Zheng, Penghao Jiang, Qirui Wang, Ruiqing Yan, and Qiuzhuang Sun. 2025. Multi-Modal Video Feature Extraction for Popularity Prediction. arXiv:2501.01422 [cs] doi:10.48550/arXiv.2501.01422

[9] Zhong, T., Lang, J., Zhang, Y., Cheng, Z., Zhang, K., Zhou, F.: Predicting micro-video popularity via multi-modal retrieval augmentation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2579–2583 (2024)

[10] Pratik Kayal, Pascal Mettes, Nima Dehmamy, and Minsu Park. 2025. Large Language Models Are Natural Video Popularity Predictors. In Findings of the Association for Computational Linguistics: ACL 2025, pages 11432–11464, Vienna, Austria. Association for Computational Linguistics.

[11] Zhang, Yuchao, Pengmiao Li, Zhili Zhang, Chaorui Zhang, Wendong Wang, Yishuang Ning and Bo Lian. "GraphInf: A GCN-based Popularity Prediction System for Short Video Networks." International Conference on Web Services (2020).

[12] Z. Zhang, Y. Zhang, M. Dong, K. Ota, Y. Zhang, and Y. Ren, "Collaborative tag-aware graph neural network for long-tail service recommendation," IEEE Trans. Serv. Comput., vol. 17, no. 5, pp. 2124–2137, Sep./Oct. 2024, doi: 10.1109/TSC.2024.3349853.

[13] Xie, Jiaheng and Xinyu Liu. "Unbox the Black-Box: Predict and Interpret YouTube Viewership Using Deep Learning." Journal of Management Information Systems 40 (2020): 541 - 579.

[14] M. Xie, J. Xiao, and S.. Huang, "Label-aware global consistency for multi-label learning with single positive labels," in Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS*)*, New Orleans, LA, USA, 2022.

マルチモーダル LLM を用いた美術実技能力の向上支援

三好　香利奈[†§]　牛尼　剛聡[‡§]

†福岡県立　春日高等学校　〒816-0811　福岡県春日市春日公園 5 丁目 17 番
‡九州大学大学院芸術工学研究院　〒819-8540 福岡市南区塩原 4-9-1
E-mail:　†k6ja11@gmail.com,　‡ushiama@design.kyushu-u.ac.jp

あらまし　生成 AI は文章生成に加え，画像理解や対話にも利用可能となり，芸術教育への応用が進んでいる。しかし，多人数指導が前提となりやすい美術教育では，学習者の制作途中作品に対して段階的に助言を与える個別指導が十分に行えない場合がある。そこで本研究では，マルチモーダル LLM を用いて，制作途中の作品画像に対して段階（ラフ，下描き，清書，完成）に応じた助言を提示する学習支援アプリケーションを構築した。被験者実験により，システムのユーザビリティと内発的動機づけへの影響を調査した。結果として，ユーザビリティには改善余地が残る一方で，制作活動の楽しさや価値認知など一部の動機づけ指標において肯定的傾向が観察された。本手法は，生成 AI を「コンテンツ生成」だけでなく，学習者の制作プロセス支援へ拡張する可能性を示す。

キーワード　美術教育，マルチモーダル LLM，段階別フィードバック，メタプロンプト，Dify

1. はじめに

近年，ディープラーニングの発展により，大規模言語モデル（LLM）を中心とした生成 AI が普及し，対話や画像理解を含む多様な支援が可能となった[1]。美術教育においても画像生成の授業導入などの応用が報告されているが，制作途中の作品に対する個別のフィードバックを継続的に提供することは，授業運営上の制約から容易ではない。そこで本研究は，マルチモーダル LLM を用いて制作途中作品に助言を与えることで，学習者の制作プロセスと動機づけにどのような影響が生じるかを検証する。

近年、教員不足や業務負担の増加により，授業内で学習者一人ひとりへ十分な個別支援を提供することが難しい状況が報告されている。美術の制作指導では，学習者の制作過程を観察し，段階に応じた助言を反復的に与えることが重要であるが，多人数指導ではその実施が制約されやすい。この課題は，制作のつまずきや苦手意識の形成にも繋がりうる。

個別の助言が不足すると，学習者は自身の作品の課題を把握しにくく，改善の方向性を誤ったまま制作を進める可能性がある。その結果，意図した表現に到達できない経験が増え，制作活動への自信や内発的動機づけが低下することがある。したがって，制作途中における適切なフィードバックを，学習者の特性に合わせて提供する仕組みが求められる。

近年，LLM の高度化と，学習規模の拡大により，生成 AI は文脈を踏まえた対話，推論，指示追従といった能力を急速に向上させてきた[2]。さらに，画像入力を扱えるマルチモーダル LLM の普及により，文章だけでなく視覚情報を含む状況理解にもとづいて助言を生成できるようになり，教育支援の適用範囲が広がっている[3]。これらを活用することで，教員が限られた時間で行ってきた観察，助言，振り返りの一部を AI が補助し，学習者の状態に応じた「その場の支援」を提供することが期待できる。とりわけ美術の制作学習では，完成物のみならず制作過程の意思決定と試行錯誤が学習成果に直結するため，制作段階に応じた形成的フィードバックを継続的に提供する仕組みが重要である。一方で，生成 AI の出力には誤りや不確実性が含まれうるため，学習者の主体性を損なわず，教員の指導意図と整合する形で活用する設計と検証が求められる。

以上を踏まえ，本研究では，美術教育における制作途中作品を入力として受け取り，制作段階に応じた助言を継続的に提示する仕組みを，マルチモーダル LLM により実現し，その教育的効果と課題を検討することを目的とする。具体的には，(1) 制作段階（ラフ，下描き，清書，完成）を前提とした段階別フィードバックの設計，(2) ユーザ属性や画材などの条件を反映した助言生成の枠組みとプロトタイプ実装，(3) 被験者実験に基づくユーザビリティと動機づけの探索的評価，を主な貢献とする。

本論文の構成は次のとおりである。第 2 章で関連研究を整理し，第 3 章で提案手法と実装を述べ，第 4 章で評価実験と結果を示し，第 5 章で考察と今後の課題を述べる。

2. 関連研究

これまでにも，AI による画像生成を授業へ導入し，創造性や鑑賞活動への影響を検討した報告がある[3]。また，群馬県では，県内の複数の小中学校において対話型生成 AI を活用し，児童生徒と生成 AI の対話を通じて個性や関心を引き出すことを目的とした取組が報告されている[4]。一方で，学習者の制作途中作品を入力として受け取り，制作段階に応じた具体的な改善助言を反復提示する「段階別フィードバック」を，マル

チモーダル LLM により個別最適化して提供する枠組みは，体系的に整理されていない。そこで本研究は，途中作品への段階別助言に焦点を当て，プロトタイプ実装と被験者実験により有用性を探索的に評価する。

3. 提案手法

3.1 システムの概要

本研究では，マルチモーダル LLM を用いて描画学習を支援する手法を提案する。ユーザはモチーフ画像を入力し，制作途中の作品画像を制作段階（ラフ，下描き，清書，完成）ごとにアプリケーションへアップロードする。システムは，ユーザ属性（年齢，経験），画材，制作段階，および画像入力に基づき，次に行うべき作業や改善点を対話形式で提示する。

3.2 プロンプト設計

個別指導を実現するため，指導方針を規定するメタプロンプト（固定）と，ユーザ属性に応じて生成されるユーザ適応プロンプト（可変）を用いる。メタプロンプトは，（1）指導者としての役割，（2）段階別の評価観点（例：構図，形の正確さ，明暗，質感），（3）助言の粒度と語調，（4）出力フォーマット，を定義する。ユーザは年齢と描画経験を入力し，システムはそれに基づきユーザ適応プロンプトを生成する。以降，モチーフ画像，画材，制作段階，途中作品画像を入力として，各段階に適した助言を提示する。

3.3 実行フロー

提案手法のプロトタイプを，ノーコード AI アプリケーション開発ツール Dify を用いて実装した。以下に実行フローを示す。

① ユーザは年齢と描画経験を入力する（図 1）。
② 入力情報に基づき，ユーザ適応プロンプトを生成する（図 2）。
③ ユーザはモチーフ画像をアップロードし，使用画材を入力する（図 3）。
④ システムは制作段階に応じた助言を提示し，ユーザは途中作品画像を送信して助言を更新する（図 4）

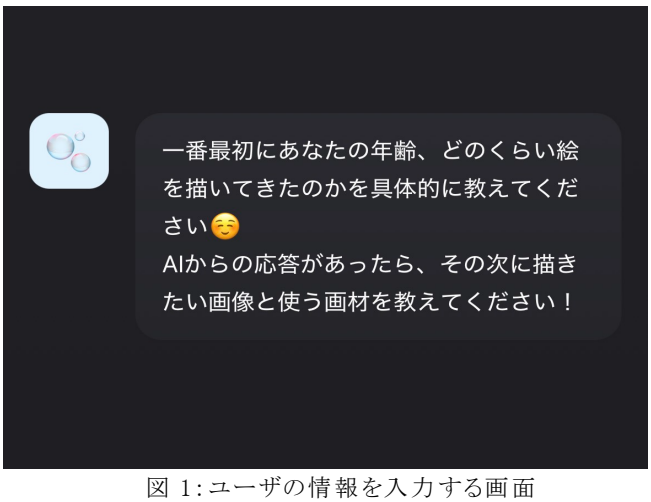


図 1：ユーザの情報を入力する画面

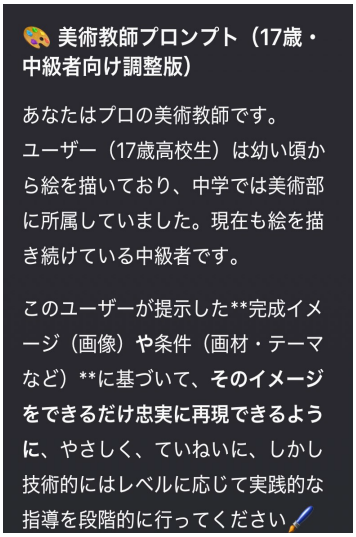


図 2:メタプロンプトによって生成されたユーザに適合したプロンプト

図 3：モチーフの画像をアップロードした画面

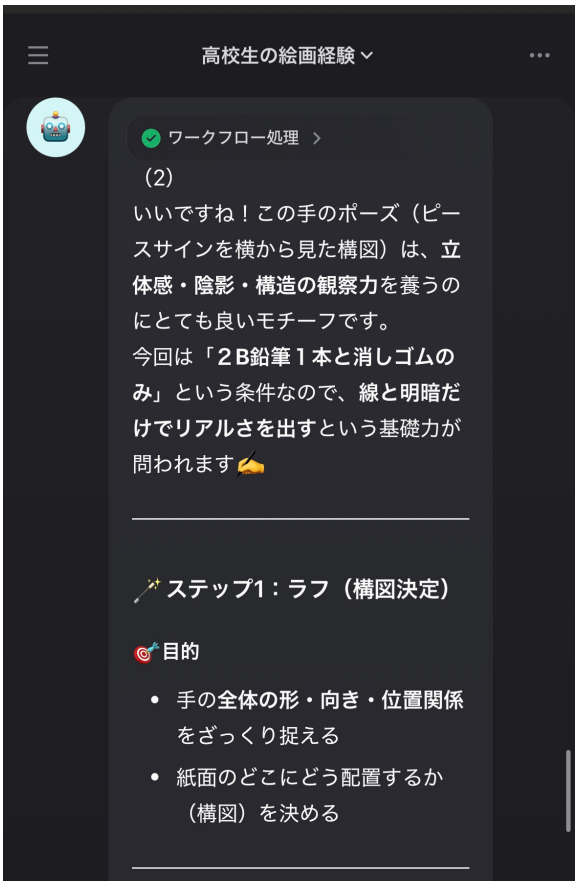


図 4：ユーザの作品に対して与えられたコメントの例

3.4 段階別フィードバック

指導の順序は以下の通りである。

① ラフ: 構図と大まかな形状を決定する段階である。
② 下描き: 線を描き込み，形や比率，バランスを調整する段階である。
③ 清書: 線画や色塗りを進め，明暗や質感表現を整える段階である。
④ 完成: 仕上げを行い，作品の最終的な評価と次回に向けた改善点を提示する段階である。

なお，各段階の終了時にユーザは途中作品画像を送信し，システムはその時点の状態に基づく助言を提示する。これにより，ユーザは作品の課題を認識し，改善の方向性を得た上で制作を継続できることが期待される。

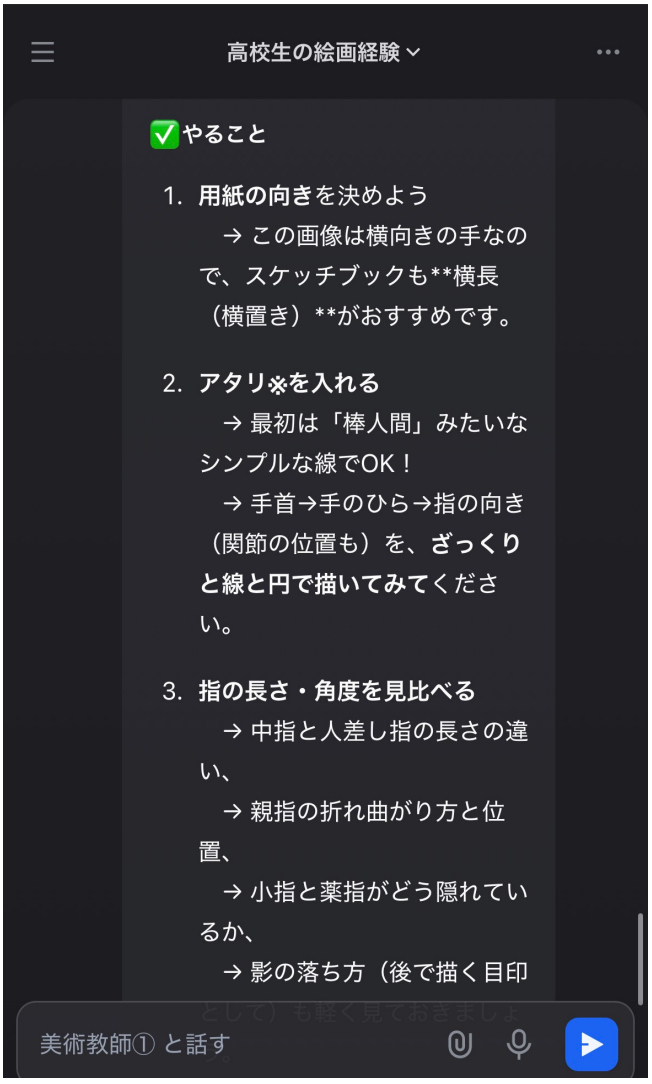


図 5: 指導の例

4. 評価

4.1 実験方法

提案手法を評価するため，被験者実験を実施した。被験者は 2 課題（立方体，球体）についてデッサンを行い，一方の課題ではシステムを使用する条件（AI あり），他方ではシステムを使用しない条件（AI なし）で描画した。順序効果を抑制するため，課題順（立方体→球体／球体→立方体）と条件順（AI あり→AI なし／AI なし→AI あり）を組み合わせた 4 群を設定し，カウンターバランスを行った（表 1）。被験者は主に大学生であり，各自のスマートフォンを用いてシステムを利用した。なお，実験中のトラブルにより群間の人数は均等ではない。

表 1: 被験者実験の構成

群	1 回目	2 回目	人数
①	立方体・AI あり	球体・AI なし	3
②	立方体・AI なし	球体・AI あり	4
③	球体・AI あり	立方体・AI なし	2
④	球体・AI なし	立方体・AI あり	3

4.2 評価指標

評価は(1) ユーザビリティ，(2) 動機づけ，(3) 作品品質の 3 観点から行った。ユーザビリティは System Usability Scale（SUS）により測定し，AI あり条件の描画後に回答させた。動機づけは Intrinsic Motivation Inventory（IMI）[7,8] の一部尺度を用いて測定し，各課題の描画後に回答させた。IMI は各項目の回答を被験者ごとに平均し，AI あり／AI なし条件で比較した。アンケートは Google フォームで実施した。作品品質は専門家（美術教育を専門とする大学教員）により評価し，評価項目は構図，形の正確さ，明暗，質感とタッチ，表現力の 5 項目とした。

図 6:立方体のモチーフ

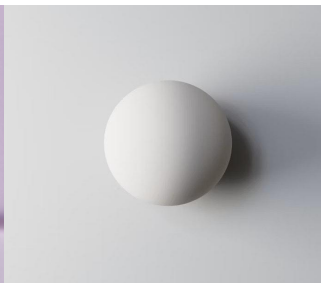


図 7:球体のモチーフ

4.3 結果

SUS の平均は 47.5 点であった。これはユーザビリティに改善余地があることを示唆する。実験中には，(i) 送信した作品画像が参照されない，(ii) モチーフや作品内容の解釈が不適切になる，といった不具合が確認され，操作負荷や期待外れの応答が評価に影響した可能性がある。

IMI の結果を図 8 に示す。多くの項目で AI あり条件が AI なし条件を上回る傾向が見られた。一方で，「この活動をうまくやれていると思った」は AI なし条件が高かった。これは，システムの不安定さや操作負荷により助言を活用しきれなかった可能性，あるいは助言が学習者の自己判断を一時的に阻害した可能性などが考えられる。

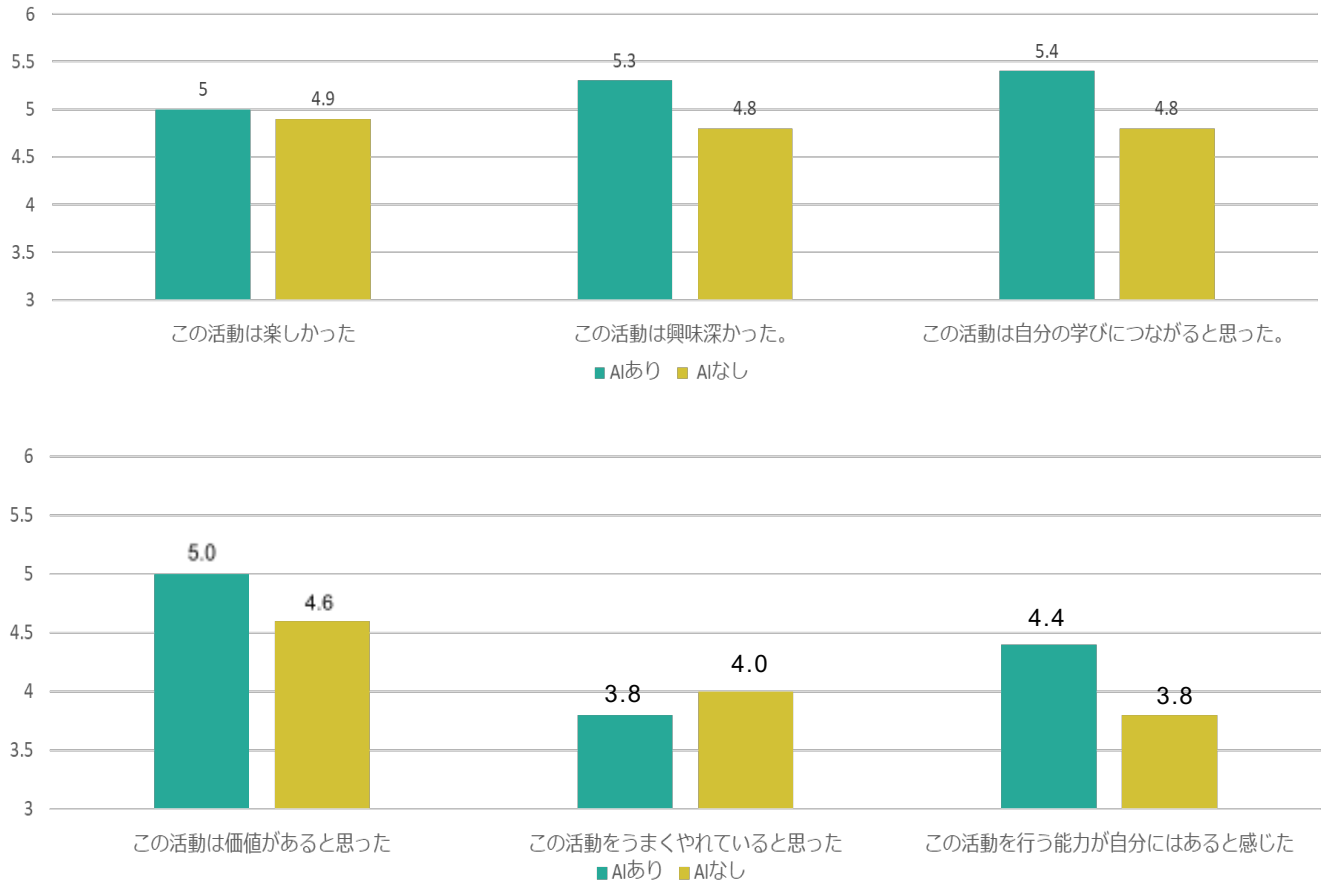


図 8:内発的動機づけ尺度アンケートの結果（※左上から順に、「この活動は楽しかった」,「この活動は興味深かった」,「この活動は自分の学びにつながると思った」,「この活動は価値があると思った」,「この活動をうまくやれていると思った」,「この活動を行う能力が自分にはあると感じた」）

専門家による作品評価（図 9）では，条件間の差は大きくは見られなかった。被験者数が限られること，および評価者が単独であることから，本結果は探索的な知見として位置づけ，今後は評価者数の増加やルーブリックの明確化により再検証する。

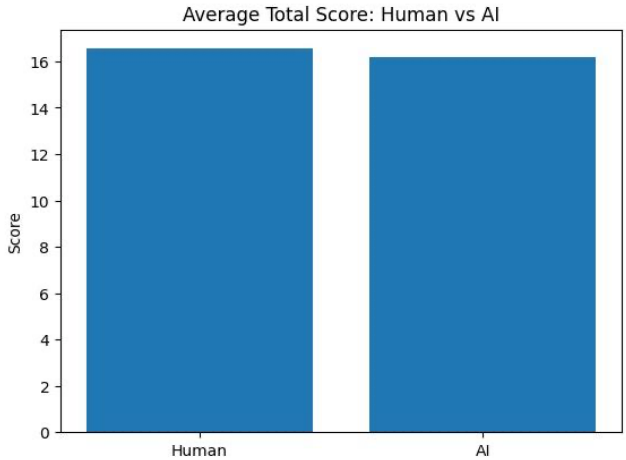


図 9: 作品評価に関する専門家による評価の平均

5. おわりに

本論文では，マルチモーダル LLM を用いて制作途中の作品画像に対し，制作段階（ラフ，下描き，清書，完成）に応じた助言を提示する美術実技学習支援手法を提案した。Dify を用いてプロトタイプを実装し，被験者実験により，ユーザビリティ，内発的動機づけ，および作品品質の観点から探索的に評価した。

評価の結果，SUS の平均は 47.5 点であり，ユーザビリティには改善余地があることが示唆された。実験中には，作品画像の参照が不安定になる等の不具合が確認され，操作負荷や応答品質のばらつきが評価に影響した可能性がある。一方で，IMI では多くの項目において AI あり条件が AI なし条件を上回る傾向が観察され，制作活動に対する楽しさや価値認知など，動機づけに関して肯定的影響を与える可能性が示された。作品品質に関しては，専門家評価において条件間の差は大きくは見られなかった。

本研究には，（１）被験者数が限られること，（２）実験中のトラブルにより群間人数が均等ではないこ

と，（３）作品品質評価の評価者が限られること，といった制約が存在する。したがって，本稿の結果は探索的知見として位置づけ，今後は条件統制を強化した実験設計と，評価手法の精緻化により再検証する必要がある。

今後の課題として，まず入力フローの簡素化と動作安定化によりユーザビリティを改善する。次に，助言の提示量と粒度を制御し，学習者の主体性を損なわない対話設計を検討する。さらに，段階別フィードバックの観点をルーブリックとして明確化し，複数評価者による作品評価や制作過程指標(修正回数,制作時間,自己説明の質など）を導入することで，教育的効果を多面的に評価する。以上により，生成 AI をコンテンツ生成に留めず，制作プロセス支援へ活用するための設計指針と実証的知見の蓄積を目指す。

参　考　文　献

[1] 総務省　令和６年版　情報通信白書の概要　特集②　進化するデジタルテクノロジーとの共生, https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r06/html/nb000000.html

[2] 坂口慶祐，大規模言語モデルの現状と今後の展望（1）大規模言語モデルのインパクトと直面する課題，電子情報通信学会誌，107 巻 6 号，pp.529-533，2024.

[3] 宮井淳行，JMMMU: A Japanese Massive Multi-discipline Multimodal Understanding Benchmark の研究過程，自然言語処理，32 巻 3 号，2025

[4] 田中博之, 画像生成 AI を用いた図画工作科教育の授業開発に関する試行的研究 -創造的鑑賞のための小単元における Adobe Firefly の活用を通して-, 学術研究 : 人文科学・社会科学編, Vol. 75, pp. 169-189, 2024.

[5] 指導難しい図工と美術に生成ＡＩ活用、子どもと対話重ね個性や関心引き出す…群馬県教委が導入へ，読売新聞オンライン 2024/06/03, https://www.yomiuri.co.jp/kyoiku/kyoiku/news/20240530-OYT1T50003/

[6] Adi Bhat　システムユーザビリティスケールとは？ |QuestionPro

[7] 最上多実子、中込和幸、亀島信也, 内発的動機づけ尺度（Intrinsic Motivation Inventory）日本語作成, 日心第 73 回大会（2009)

[8] 尾無徹、畠山陽介、川原恭一、澤田彩華、猪浦智史　内発的動機づけ尺度（Intrinsic Motivation Inventory）の日本語作成と新任期保健師への適応可能性の検討, 日本公衆衛生看護学会誌（2025)

Diffusion モデル内にキャラクター情報の制約を追加したアニメ画像生成モデルの構築

木畑 光貴† 落合 桂一† 戸田 浩之†

†横浜市立大学データサイエンス学部 〒236-0027 神奈川県横浜市金沢区瀬戸 22-2

E-mail: †{d224026f,ochiai.kei.dk, toda.hir.xg}@yokohama-cu.ac.jp

あらまし 拡散モデルによる画像生成技術は進展したが，アニメ画像の生成，特に複数キャラクター生成時におけるキャラクターの身体的特徴の取り違え(属性混同)や描画崩壊が課題となっている．既存研究の多くは写実的画像を対象としており，アニメ画像の線画，平坦な塗りによる深度情報，特有の表現への適応が不十分である．そこで本研究では，LLM を用いてプロンプトから各キャラクターの外見的特徴や配置情報を抽出・構造化し，これらを制約条件として組み込む生成モデルを提案する．本研究では，既存の個別生成プロセスをアニメ画像に適するように調整した上で基盤とし，そこに上述の構造化された属性制約を統合する手法を提案する． 実験では既存手法と比較評価を行い，アニメ調の複数キャラクター生成における本モデルの有効性を示す.

キーワード 画像生成，拡散モデル，漫画・コミック

1. はじめに

近年，Stable Diffusion [1]などを通じて，誰もが高品質な画像を容易に生成できる環境が普及している．これらは入力された自然言語(プロンプト)に基づいて画像を出力する，テキスト条件付き拡散モデルというアプローチを採用している．テキスト条件付き拡散モデルの核となる拡散モデルは，データに微細なガウスノイズ(平均 0・分散一定の正規分布に従う確率ノイズ)を段階的に加えていく「拡散過程」と，そのノイズを推定して取り除き画像を段階的に復元する「逆拡散過程」という 2 つのプロセスから構成される．拡散モデルの学習の安定性や段階的なノイズ除去による精緻な画像再構成という特徴により，従来の敵対的生成ネットワーク(Generative Adversarial Networks，GAN)ベースの手法[2]を大幅に上回る自然言語から高度な画像を生成する表現力や，多様なドメインへの適応を可能にする高い汎用性といった利点を持つ．

しかしながら，テキスト条件付き拡散モデルによるアプローチでは，複数キャラクターを含む画像を生成する際に，画像生成時において個体間の身体的特徴が混同されたり，空間的な整合性が失われたりしやすいという課題が指摘されている[3].

第一に，同一の属性を持つキャラクター間での属性混同の問題である．例えば「黒髪の A」と「黒髪の B」のように類似した属性を持つキャラクターを同時に生成しようとした場合，本来対応すべき画像上の位置関係だけでなく，他方のキャラクター(B)の位置にも誤って影響を与えてしまう現象が確認されている．その結果，意図しない属性の混同が発生し，キャラクターごとの属性が正しく分離されないという問題が生じる[3，4]．図 1 にこの問題が生じた生成画像例を示す．この画像は，“Two girls with black hair standing side by side. The girl on the left has a long ponytail with black hair. The girl on the right has a short bob cut with black hair.”という入力プロンプトから生成されたものであるが，右側のキャラクターには，本来左側のキャラクターに指定したポニーテールの特徴と，ボブカットの特徴の両方が同時に現れていることが確認できる．

第二に，構図の破綻および作画崩壊の問題である．複数のキャラクターが近接する構図において，個体間の境界が不明瞭となり，身体の一部が融合するなどの構造的な崩壊が発生しやすい[3, 4]. これらの課題は，入力文中の単語やキャラクター属性を強調させる従来のプロンプトエンジニアリングや矩形領域による指定だけではモデルの内部計算を直接制御するのが難しいため解決が困難であり，モデル内部の生成プロセスに対するより直接的な介入が求められている[5]. 図 2 にこの問題が顕在化した生成画像の例を示す．この画像は，“Two anime girls hugging each other tightly.They are cheek to cheek.The girl on the left has long blue hair，and the girl on the right has short pink hair. Detailed frilly dresses. ”という入力プロンプトから生成されたものであるが，2 人の上半身が融合し，腕の本数がおかしくなっていることが確認できる．

これらの問題は，アニメ画像が明確な輪郭線によって形状が表現されることや，現実の人体構造とは異なるデフォルメ表現が多用され，わずかな属性の混同や線の不整合であっても，視覚的な破綻として強く知覚されやすいことに起因しており，これらの問題は，アニメイラストとしての成立を根本的に妨げる要因となっている．これに伴い，現在の制作フローでは，意図した構図が得られるまで生成を繰り返す試行錯誤や，生成後の画像に対して部分修正や手動での加筆修正を行う事後処理に多大な時間が割かれている．このよう

図 1 第一の課題の画像例

図 2 第二の課題の画像例

な非創造的な作業負担の増大は，AI を活用した創作活動における大きな障壁となっている．

したがって本研究では，アニメ画像において複数キャラクターを含む画像を生成する際に生じる上述の課題を解決すべき課題領域と定め，大規模言語モデル(Large Language Model，LLM)によるプロンプトの構造化(意味解析)と，その結果を拡散モデル内部への明示的な制約として導入する手法を組み合わせた新たなテキスト条件付き拡散モデルを提案する． 具体的には，LLM を用いて入力プロンプトからキャラクターごとの属性およびキャラクターの関係性を構造化データとして抽出し，それに基づいて各個体の占有領域および深度を規定する属性制約を，U-Net[6]内部で動的に制御する.これにより，アニメ画像生成におけるスタイルと構造の一貫性を担保する．ここで U-Net とは拡散モデルにおいて各ステ ップで加えられたノイズを予測し，元の画像データ（または潜在表現）を復元するための中心的なニューラルネットワークである．

本研究の貢献は以下のとおりである．

- 属性分離と構造的整合性の両立 ： 本研究では，LLM によるプロンプトで指定される構造に関わる情報を U-Net 内部の制御に直接組み込む新枠組みを提案する．これにより，従来のテキスト条件付けのみでは困難であった複数キャラクター生成時の「属性混同」を原理的に抑制し，キャラクターが密接に重なり合う複雑なシーンにおいても，個別の属性（髪色・服装等）と身体構造を正確に保持することが可能となり，従来手法で頻発していたキャラクターの融合や作画崩壊を著しく低減させ，高密度かつ高品質な画像生成を実現する．
- クリエイターの制作ワークフローの効率化への寄与 ： 本手法は，複雑なマスク画像の手動作成や過度なプロンプトエンジニアリングを必要とせず，自然言語指示のみで意図通りの制御を実現する．これにより，生成 AI を用いた創作活動における試行錯誤の回数を大幅に削減し，クリエイターが構図や演出の検討に集中できる効率的な制作支援環境を提供する．

本研究の構成は以下の通りである．

第 2 章では，アニメ画像生成，LLM を用いた画像生成の制御，および拡散モデルの内部構造への介入に関する関連研究について述べ，本研究の立ち位置および新規性について整理する．第 3 章では，提案手法の詳細について述べる．具体的には，LLM を用いたプロンプトの意味解析手法，および U-Net 内部でキャラクターの属性や配置を制御するための深度マップと損失関数の設計について詳述する．第 4 章では，提案手法を用いた評価実験の方法，設定を示す． 第 5 章では既存手法との比較を通じて，アニメ調の複数キャラクター生成における属性混同の抑制と構造的一貫性の向上について定性評価，定量評価の結果から考察する．第 6 章では，本研究の結論をまとめ，現状の課題および今後の展望について述べる．

2. 関連研究

2.1 アニメ画像生成の研究

アニメ画像の生成に関しては，GAN を用いた手法から発展し，拡散モデルを用いた手法が高い品質を達成している．

例えば，Cao らは AnimeDiffusion [3]において，線画情報と条件付き拡散モデルを組み合わせた，線画を構造的制約とするアニメ画像生成手法を提案した．彼らは，数百万枚規模の画像を含む独自の AnimeDiffusion Dataset を構築し，豊富なタグ情報を活用することで，

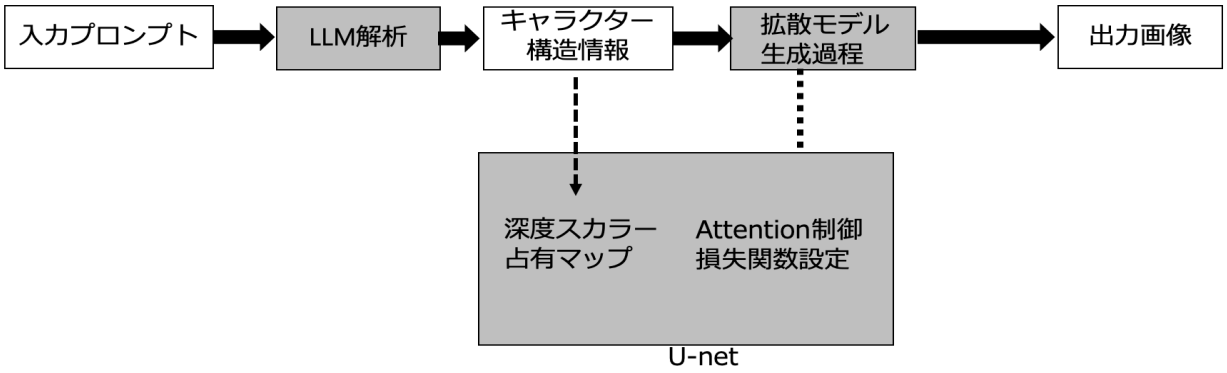


図 3 提案手法図

従来の GAN ベースの手法を大幅に上回る生成品質を実現している．また，拡散過程におけるノイズ予測モデルに対して新たな条件式を導入し，線画の構造を維持しつつ，参照画像やタグに基づいた高精細なイラストレーションの生成を可能にした．

しかしながら，AnimeDiffusion は主に「線画からの着色」や「単一のキャラクター」の生成に主眼を置いており，テキストプロンプトのみから複数のキャラクターが相互作用する複雑なシーンや，全身を含む動的な構図をゼロから生成するタスクに関しては，十分な検討がなされていない．

2.2 LLM を使用した拡散モデル型画像生成の研究

画像生成の制御性とプロンプト理解能力を高めるため，LLM を生成プロセスに統合する研究が活発に行われている．

Tang らは，LLM と拡散モデルを層レベルで統合する Deep Fusion[7]を提案し，複雑なテキストの意味内容を画像へ詳細に反映させることを試みている．また，Li らは Multi-Agent Collaboration-based Compositional Diffusion (MCCD)[8] において，LLM を用いてプロンプトを要素分解し，マルチエージェントシステムで各領域を協調的に生成する手法を提案している．

本研究においても，これらの先行研究と同様に，LLM をプロンプトの解釈や生成プロセスの補助として活用するアプローチを採用している．

2.3 拡散モデルにおける内部構造への介入

拡散モデルの生成プロセスに対して，明示的な条件式や制約を組み込むことで，モデルの挙動を制御しようとする研究が行われている．

Zampini らは，再学習を伴わずに拡散過程へ制約最適化を組み込み，物理法則や構造要件を強制的に満たすフレームワークを提示した[9]． また，Song らは，拡散モデルのサンプリング段階でノイズを制御するスコア関数を条件式として新たに導入し，生成段階での情報を条件付きのスコアで制御することで，既存のモデルを再学習せずに，クラス条件付き生成，画像修復，カラー化の 3 タスクを単一モデルで行っている[10]．

一方で，特定のキャラクターや複数のオブジェクトを一貫性を保って生成するための手法も提案されている．

Ruiz らは DreamBooth[11]において，数枚の画像から特定の主題を学習し，その識別子をプロンプトに紐付けることで，特定のキャラクターや物体のアイデンティティを維持したまま，多様な状況下での生成を可能にするファインチューニング手法を提案した． また，複数キャラクター生成における重なりの問題に対しては，Yang らが LaRender[12]を提案している．これは，潜在空間上でのレンダリングを通じて，マスク誘導によるレイヤーごとの生成と合成を行うことで，再学習なしにオブジェクト間の前後関係や重なりを制御するものである．

2.4 関連研究のまとめ

本研究で提案する手法は，Zampini や Song らが示した「推論時の制約導入」というアプローチを基盤としつつ，DreamBooth が目指した「個体のアイデンティティ保持」と，LaRender が扱った「複数物体の空間的整合性」を，LLM による構造化と条件制約によって同時に解決しようとするものである． 本研究は「2.2 LLM による高度な意味解析」を用いて「2.3 学習不要な内部介入技術」を，アニメ画像生成という特定のドメイン課題に最適化して統合した手法と言える．

3. 提案手法

3.1 概要

本研究では，拡散モデルを用いた画像生成において，

複数のキャラクターが登場する際に発生する「キャラクターの描画崩壊」および「属性の混同」という課題を解決するための手法を提案する．提案手法の核となるのは LLM によるプロンプトの構造的解析と，生成モデル（U-Net）内部への空間的・構造的な介入の組み合わせである． ここで U-Net とは．拡散モデルにおいて各ステップで加えられたノイズを予測し，元の画像データ（または潜在表現）を復元するための中心的なニューラルネットワークである．
具体的には，以下の 3 つのステップで構成される．

1. プロンプト解析: LLM を用いて入力プロンプトから各キャラクターの属性情報を抽出し，生成プロセスをキャラクターごとの「スロット」として管理する．
2. 空間的制御情報の予測: U-Net の中間特徴量から，各キャラクターが画像内で占める領域を示す「占有マップ」と，その前後関係を規定する「深度スカラー」を予測する軽量な追加ヘッドを導入する．
3. 特徴量の統合と干渉: 予測された深度情報に基づき，手前に位置するキャラクターの特徴を優先的に反映させる重み付け合成関数を適用する．同時に，アテンションマップに対して占有マップを用いた干渉を行うことで，各キャラクターの属性が正しい位置に配置されるよう制御を強める

3.2 LLM によるプロンプト解析

画像生成において， 複数のキャラクターが登場するプロンプトは，画像とテキストの対応関係を学習したモデルである CLIP[13]や ALIGN[14]等のテキストエンコーダ内で情報の絡み合いを引き起こしやすく，これが属性混同の主たる原因となる．そこで本手法では，拡散 モデルへの入力に先立ち，LLM を用いて自然言語プロンプトを構造化されたデータへと変換する前処理を行い，入力されたプロンプト(参考付録)から「誰が」「どこにいるか」という情報をスロットとしてキャラクターごとに抽出・格納する．

3.3 深度マップ推定

本節では，3.2 節で抽出・スロット化されたキャラクター情報を基に，画像内での空間的な前後関係を制御する手法について述べる． 本手法の核は，各スロットに対して「占有マップ」と「深度スカラー」という物理的な制約を割り当てる点にある．3.2 節の LLM 解析によって得られた各キャラクターのスロットiに対し，U-Net の中間特徴量Fから以下の 2 つの情報を予測する軽量な追加ヘッドを導入する．

- 占有マップ($S_i \in [0,1]^{\{H \times W\}}$) : キャラクターk が画像平面上で描画されるべき 2 次元的な領域を表す空間的な重みである．U-Net 内の Cross-Attention マップ，プロンプト内のトークンと画像領域の対応関係を強く保持している．
- 深度スカラー($z_i \in \mathbb{R}$) : 各キャラクターに対して単一の深度スカラー$z'' \in [0, 1]$を割り当てる．本手法では「キャラクターA はキャラクターB よりも手前である」という順序関係をロバストに扱うため，キャラクターごとに代表的な深度値を推定する．この深度スカラーは，プロンプト解析（3.1 節）で得られた「手前・奥」という言語的な制約条件と， 現在の生成画像から推定される深度情報の整合性が最適化される．

3.4 損失関数

本手法では，U-Net の中間特徴量Fから予測される各キャラスロットiの占有マップS_i，および深度スカラーz_iを最適化し，複数キャラクター間の描画崩壊や属性混同を抑制するために，以下の損失関数および微分可能な合成関数を導入する．

3.4.1 特徴量合成のための指数重み付け関数

キャラクター間の前後関係を考慮した合成を行うため，各スロットiの寄与率$W_i(p)$を，占有マップS_iと深度スカラーz_iを用いて以下の指数関数によって定義する．

$$W_i(p) = \frac{S_i(p) \times \exp(-\beta \times z_i)}{\sum_j S_i(p) \times \exp(-\beta \times z_i) + \epsilon}$$

ここで，βは深度の差に対する感度を制御するパラメータ，ϵは数値的安定性のための微小値である．この関数はz_iが小さいほど重みWが指数的に大きくなる性質を持ち NeRF [15]等で用いられる不透明度の概念に近い処理を行うことで，手前の物体を優先的に描画する制御を可能にする．また，z_iに対して勾配が流れる設計とすることで,「特定のキャラクターをより手前に描画すべき」という誤差逆伝播を可能にしている．

3.4.2 Attention Map

Cross-Attention 機構は，プロンプト内のトークンと画像領域の対応付けを担う重要なプロセスである[16]が，標準的な拡散モデルでは 2 次元的な相関のみに基づき計算されるため，奥行き方向の整合性は考慮されない．これが，手前のキャラクターの色が背景に混ざる，あるいは奥のキャラクターが手前のオブジェクトを透過して描画されるといった不自然な生成結果の主要因となる．そこで本手法では，推定された深度情報

を用いて Attention Map を動的に制限する Depth-Hard Attention Masking を導入する．これは,「あるピクセルの深度と，そこに描画されるべきトークンの想定深度が乖離している場合, その Attention を物理的に無効化する」という制約を数式化したものである．

$$A'_{i,\ p} = A_{i,\ p} \times exp\left(-\alpha \cdot |D_p - z_i|\right)^2$$

ここで，D_pは各ピクセルごとにおける深度マップの値であり，αは深度不整合に対するペナルティの強さを制御する係数である．

すなわち，ピクセルの深度D_pがトークンの設定深度z_iに近い場合，この項は 1 に近づき元の Attention が維持される．逆に，深度が大きく異なる場合（例：手前のキャラクターを描くべきトークンが，背景の深度を持つピクセルに反応している場合)， この項は 0 に近づき，Attention スコアは強制的に抑制される．この機構により，各トークンは自身が割り当てられた深度に対してのみ影響力を持つことが保証される．結果として，手前のキャラクターが奥のキャラクターに埋もれる現象や，属性情報が深度の異なる領域へ漏れ出す問題を，幾何学的な制約によって未然に防ぐことが可能となる．

4. 評価方法

4.1.1 評価データセット

複数キャラクターの描画において属性の混同や構造的破綻が発生しやすいシナリオを想定し，20 件の評価用プロンプトを作成した．各プロンプトは，「左に青い髪の少女，右に赤い髪の少年」や「手前にメイド服の女性，奥にスーツの男性」のように，2 名から 5 名程度のキャラクターに関する，異なる外見的特徴と空間的配置制約を含む記述により構成される．

4.1.2 比較手法

提案手法の有効性を検証するため，以下の 3 つのモデルと比較を行った．

- **Baseline:** 標準的な Stable Diffusion.
- **Regional Prompter[17]:** 画像平面を矩形分割し, 領域ごとにプロンプトを適用する手法.
- **Structure Diffusion[18]:** レイアウト情報に基づき, Cross-Attention マップに空間的変調を加える手法.
- **Proposed (提案手法):** LLM の意味解析と深度情報を統合した本研究の手法.

画像全体の統計的な距離を測る FID[19]等では「意味的な整合性」を十分に評価できないため，高度な画像理解能力を持つ大規模視覚言語モデル(Vision Language model，VLM)を用いた自動評価パイプラインを採用した．

4.2 実験設定

4.2.1 実験環境

本研究で提案する手法の有効性を検証するため，以下の環境および設定で実験を行った．

- 基盤モデル: 画像生成エンジンには Stable Diffusion v1.5 をベースとする．
- プロンプト解析（LLM）: プロンプトからの属性抽出および JSON 形式への構造化には，GPT-4o mini を使用する．
- 計算資源: 実験には RTX 6000 Ada を搭載したワークステーションを使用する．

4.2.2 実験設定

本実験では，全手法において公平な比較を行うため，基本的な画像生成パラメータを統一した．推論ステップ数は 50 ステップ，Classifier-Free Guidance (CFG) スケールは 7.5 に固定している．提案手法における固有のハイパーパラメータは，予備実験に基づき以下の 16 通り設定した．予備実験に関して，深度感度パラメータ β およびペナルティ係数 α の最適値を決定するため，予備実験を行った．具体的には，β を {2,5,10,20}，α を {5,10,20}の範囲で変化させ，さらに範囲を絞り生成画像を比較し，属性の分離精度と画像の自然さが最もバランス良く保たれる値を採用した．

3.3 節で定義した深度配慮型特徴量合成における深度感度パラメータ β は 6.0 とした．この値は，手前に位置するキャラクターの特徴を十分に強調しつつ，奥に配置されたキャラクターが完全に消失することなく自然な背景として描画されるためのバランスを考慮して決定された．

3.4 節で述べたペナルティ係数 α は 10.0 に設定した．これにより，キャラクター間の属性混同を防ぐための明確な分離性能を確保しつつ，過度に急峻なマスク適用によって生じる画像の不自然な切れ目や不自然な生成結果の発生を抑制している．

プロンプト解析を行う GPT-4o に対しては，Temperature を 0.0 に設定し，出力の決定論性を最大限に高めた状態で推論を行った．

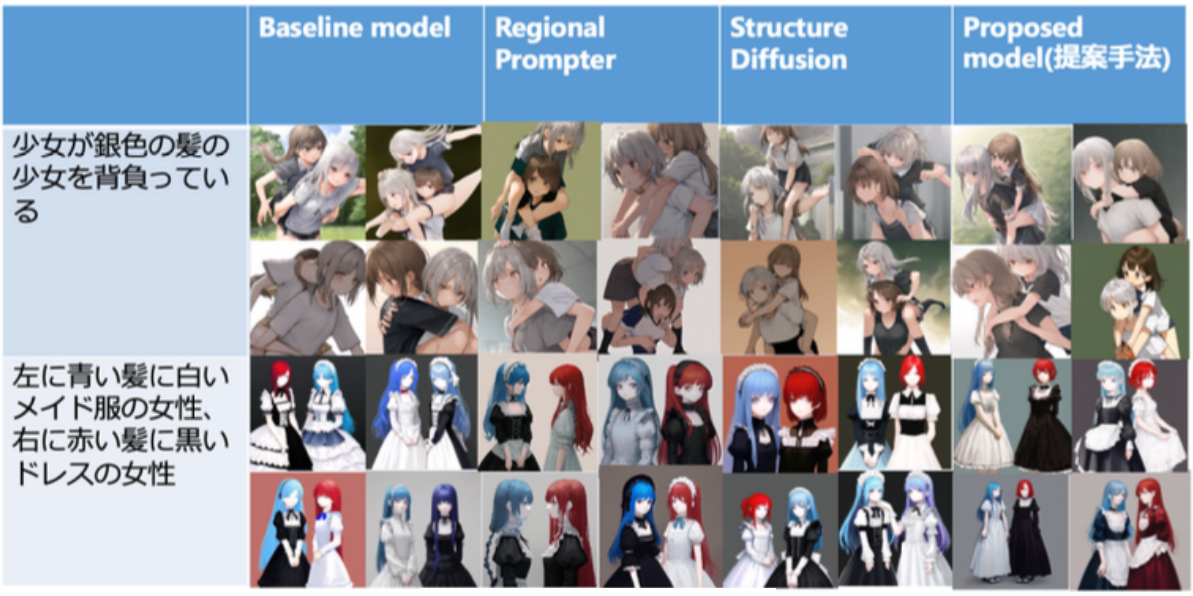


図 4 定性評価結果

5. 評価結果

5.1.1 定性評価結果

図 4 に，Baseline (Stable Diffusion)，Regional Prompter，Structure Diffusion，および提案手法 (Proposed model) による生成画像の比較を示す．ここでは，複雑な身的接触 を伴う構図と，属性の分離が求められる「左右への配置」 構図の 2 つのシナリオにおいての結果を示す．

- 構造的整合性と前後関係
  「少女が銀色の髪の少女を背負っている」というプロンプトにおいて，各手法の差異が現れた．Baseline および Structure Diffusion では，2 人の身体が融合する，腕の本数がおかしいなどの身体の構造的な破綻などが多く見られた．Regional Prompter は，領域を分割することで個体の分離には成功しているものの，キャラクター同士が複雑に絡み合う「おんぶ」のような密接な相互作用を自然に描画することに苦戦しており，他手法と比較すると優れているが，腕の本数が多いケースや 3 人に見えるケースが確認された．一方，提案手法は，手前のキャラクターが奥のキャラクターを支える構造を比較的正しく捉えている．特に，背負われる側の足や腕が手前のキャラクターに自然に重なる表現において，他手法よりも高い整合性を示していることがわかる．ただし，提案手法においても，課題は完全には排除できておらず，生成されるサンプルによっては依然として身体構造の崩れが見られる点は課題として残る．
- 属性分離の正確性
  「左に青い髪に白いメイド服の女性，右に赤い髪に黒いドレスの女性」というプロンプトにおいては，Baseline を除く 3 つの手法（Regional Prompter, Structure Diffusion, Proposed）がいずれも高い属性分離性能を示した．Baseline では「左右両方のキャラクターが青い髪になる」といった属性混同が発生しているのに対し，提案手法は左右のキャラクターの髪色と衣装を比較的分離できている．このタスクに関しては，空間を明確に分割する Regional Prompter も非常に良好な結果を示しており，提案手法との間に視覚的な大差は見られなかった．これは，被写体同士の重なりが少ない単純な左右配置においては，既存の領域分割アプローチでも十分な制御が可能であることを示唆している．「左に青い髪に白いメイド服の女性，右に赤い髪に黒いドレスの女性」というプロンプトにおいては，Baseline を除く 3 つの手法（Regional Prompter, Structure Diffusion, Proposed）がいずれも高い属性分離性能を示した．Baseline では「左右両方のキャラクターが青い髪になる」といった属性混同が発生しているのに対し，提案手法は左右のキャラクターの髪色と衣装を比較的分離できている．このタスクに関しては，空間を明確に分割する Regional Prompter も非常に良好な結果を示しており，提案手法との間に視覚的な大差は見られなかった．これは，被写体同士の重なりが少ない単純な左右配置においては，既存の領域分割アプローチでも十分な制御が可能であることを示唆している．

5.1.2 定量評価結果

表 1 定量評価結果

	Baseline model	Regional Prompter	Structure Diffusion	Proposed Model(提案手法)
平均順位	2.91	2.37	2.45	2.26

VLM を用いた 100 件のテストケースに対する順位付け評価の結果を表 1 に示す．平均順位は数値が小さいほど評価が高いことを意味する．

実験の結果，提案手法（Proposed）は平均順位 2.26 を記録し，比較した全手法の中で最も優れた性能を示した．これはベースラインである Stable Diffusion のスコアと比較して大幅な改善であり，提案手法が属性混同や構造崩壊の抑制に極めて有効であることを示している．比較手法については，Regional Prompter が 2.37，Structure Diffusion が 2.45 という結果となった．Regional Prompter はベースラインに次いで良好なスコアを記録しており，矩形領域によるハードな空間分割が，左右配置などの単純な分離タスクにおいて一定の効力を発揮していることが確認できる．これは，Structure Diffusion の Attention 変調が比較的緩やかであるため，アニメ調の画像において求められる明確な属性分離に対し，拘束力がやや不足していた可能性が考えられる．

提案手法は，これら既存手法と比較しても良い性能を示している．この優位性の要因は，Regional Prompter の明確な領域分割と，Structure Diffusion が目指す自然な馴染ませの両立にあると考えられる．提案手法は，前後関係が発生する境界付近でも破綻なく属性を分離できるため，VLM による実体の分離と空間的整合性の両観点で評価された結果と言える．

5.2 考察

評価実験の結果，提案手法は平均順位 2.26 を記録し，比較した全手法の中で最も優れた性能を示した．

次点の Regional Prompter は，単純な左右配置では提案手法と同等の分離性能を示したが，平均順位では下回った．この差は，複雑な構図への対応力にある．Regional Prompter が採用する矩形分割は，キャラクター同士が密接する「おんぶ」等のシーンにおいて，境界の破綻や不自然な切断を招きやすい．対して提案手法は，深度情報に基づく「Soft-Depth Compositing」を採用しているため，接触面や前後関係を自然に処理できる．この「分離性能」と「空間的な馴染み」の両立が，VLM による評価において Regional Prompter を上回る要因となったと考えられる．

一方，Structure Diffusion については，Attention 制御が比較的緩やかであり，アニメ調画像に求められる厳密な属性分離に対しては，拘束力が不足していたと推察される．以上より，複雑な複数キャラクター生成においては，単なる平面的な領域分割だけでなく深度情報を統合した幾何学的な制御が生成品質の向上に有効であると言える．

6. まとめ

本研究では，拡散モデルによる複数キャラクター生成時の課題である「属性の混同」と「構造崩壊」に対し，LLM の意味解析と深度情報を活用した新たな生成制御手法を提案した．

提案手法は，LLM が抽出した属性・空間情報を U-Net 内部の占有マップと深度推定に反映させ，空間的変調（Depth-Hard Attention Masking）と特徴量合成（Soft-Depth Compositing）を行うものである．実験の結果，既存手法では困難であった「キャラクター同士が密接に重なり合う複雑な構図」においても，3 次元的な整合性を保った自然な描画が可能であることを確認した．一方で，現状の制御対象は属性と深度に限定されており，LLM が持つ表情や相互作用といった豊かな文脈情報を完全には活用しきれていない．また，VLM による自動評価の客観性確保に向けた，人間による主観評価との相関検証も今後の課題である．

今後は，本手法を動画生成へ拡張し，フレーム間での時間的一貫性を確保する仕組みの構築を目指す．さらに，身体部位レベルでの密な深度制御の導入や，プロンプト解析の階層化による微細なディテールの再現性向上を図り，より実用的で高品質なアニメーション制作支援技術へと発展させていく．

参 考 文 献

[1] R. Rombach1, A. Blattmann1, D. Lorenz1, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, Proc. CVPR'22, 2022.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. “Generative Adversarial Networks, Proc. NIPS 2014, 2014.

[3] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, Proc. NIPS 2017, 2017.

[4] Y. Cao , X. Meng , P. Y. Mok, X. Liu, T-Y. Lee, P. Li, Animediffusion: Anime face line drawing colorization via diffusion models, Proc. ICSCC'23, 2023.

[5] J. Zhang, S. Guo, P. Dong, J. Zhang, Z. Liu, Y. Yu, X. Wu, Easing Concept Bleeding in Diffusion via Entity Localization and Anchoring, Proc. ICML 2024, 2024.

[6] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, Proc. CVPR 2015, 2015.

[7] B. Tang, B. Zheng, X. Pan, S. Paul, S. Xie, Exploring the Deep Fusion of Large Language Models and Diffusion Transformers for Text-to-Image Synthesis, Proc. CVPR'25, 2025.

[8] M. Li, X. Hou, Z. Liu, D. Yang, Z. Qian, J. Chen, J. Wei, Y. jiang, Q. Xu, L. Zhang, MCCD: Multi-Agent Collaboration-based Compositional Diffusion for Complex Text-to-Image Generation, Proc. CVPR'25 2025.

[9] S. Zampini, J. Christopher, L. Oneto, D. Anguita, F. F ioretto, Training-free constrained generation with stable diffusion models, arXiv:2502. 05625 8 Feb 2025.

[10] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole, Score-based generative modeling through stochastic differential equations, Proc. ICLR'21, 2021.

[11] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman, DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, Proc. ICCV 2025, 2025.

[12] X. Zhan, D. Liu, LaRender: Training-Free Occlusion Control in Image Generation via Latent Rendering, Proc. ICCV 2025, 2025.

[13] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G . Krueger, I . Sutskever, Learning Transferable Visual Models From Natural Language Supervision," Proc. ICML, 2021, 2021.

[14] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. Le, Y. Sung, Z. Li, T. Duerig, Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *Proc. ICML*, 2021, 2021.

[15] B. Mildenhall, P. Srinivasan, M. Tancik, J. Barron, R. Ramamoorthi, R. Ng, NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, Proc. ECCV2020. 2020.

[16] W. Zhang, H. Liu, J. Xie, F. Faccio, M. Shou, J. Schmidhuber, Cross-Attention Makes Inference Cumbersome in Text-to-Image Diffusion Models, 2024.

[17] hako-mikan. 2023. Github. https://github. com/hako-mikan/sd-webui-regional-prompter

[18] W. Feng, X. He, T. Fu, V. Jampani, A. Akula, P. Narayana, S. Basu, X. Wang, W. Yang Wang, Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis, Proc. ICLR 2023, 2023.

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, Proc. NIPS. 2017, 2017.

参考付録

LLM 意味解析プロンプト

""" You are a precision parser for text-to-image prompts.
Task: extract character count
and inter-character relations
(overlap/front/behind/left/right). Return STRICT JSON
only, no prose.
Rules:
• Do NOT invent characters not grounded in the prompt.
• If prompt explicitly enumerates characters (e.g., 'Left character:', 'Right
character:'), treat them as distinct.
• If tags like '2girls' appear, interpret as 2 distinct characters unless explicitly
stated identical; when in doubt, choose the conservative smaller number.
• Extract overlaps and depth ordering only if clearly implied by words like 'in
front of', 'behind', 'overlapping', 'foreground', 'background', 'left',
'right', 'center'.
• Keep results compact. """

高精度自車位置推定のための車載カメラ画像と三次元点群地図間の局所特徴照合に関する検討

髙津　悠生†　道満　恵介††　川西　康友†††　久徳　遙矢†

† 愛知工科大学　〒 443-0047 愛知県蒲郡市西迫町馬乗 50-2
†† 中京大学　〒 470-0393 愛知県豊田市貝津町床立 101
††† 理化学研究所　〒 619-0237 京都府相楽郡精華町光台 2-2-2
E-mail: † takatsu.h@aut.kyutoku.jp, kyutoku-haruya@aut.ac.jp,
†† kdoman@sist.chukyo-u.ac.jp, ††† yasutomo.kawanishi@riken.jp

あらまし　近年，自動運転技術が大きな注目を集めており，なかでも自車位置推定はその中核を担う重要な要素技術である．その際データベースとして三次元点群地図を用い，入力センサとして安価で広く普及しているカメラを利用した高精度な位置推定が可能となれば，多様な車両への適用が期待できる．その実現に向け我々はこれまでに，おおよその自車位置が既知であることを前提とし，三次元点群地図から反射強度情報と位置情報に基づいて鳥瞰視点の反射強度画像を作成し，カメラ画像と二次元特徴点照合をして位置推定する手法を提案した．本稿ではこのおおよその位置を求めるための，三次元点群地図から直接抽出した特徴点とカメラ画像から抽出された特徴点間の照合による自車位置推定手法を提案する．提案手法では，それぞれから抽出された特徴点の位置に基づき対応関係の真値を作成し，異なるモダリティ間で距離を測る距離学習モデルを構築することで，異なるモダリティ間における特徴点の対応付けを行う．そして，得られた三次元点群地図上の点とカメラ画像上の点の対応から幾何的拘束を用いて自車位置を推定する．実験の結果，多くのシーンでは十分な精度の位置推定を行えなかったが，一部のシーンにおいては高精度な位置推定が可能であることが確認された．

キーワード　自車位置推定，三次元点群地図，距離学習，車載カメラ，LiDAR

1　はじめに

次世代モビリティ社会の実現へ向けた自動運転技術は，近年目覚ましい発展を続けている．このさらなる発展，および一般社会への普及は，ヒューマンエラーによる交通事故の削減や，交通安全性の向上に繋がる．さらに，高齢化社会に伴う問題の解決や，ドライバ不足問題の解決など，人々の暮らしやすさへの貢献に期待が寄せられている．そして，このような自動運転システムによる移動手段の提供が，世界中で試験的ながら実現されつつある．この次世代モビリティ社会の実現のために必要不可欠である自動運転技術は，様々な技術により成り立っている．その中でも，走行中の車両の現在位置を詳細に特定する自車位置推定は，適切な走行経路計画や制御などの安全な走行を実現するために必須となる基盤技術である．

自車位置推定技術の研究・開発は多岐に渡って行われており，様々なセンサを用いた手法が存在する．この代表的なものとして，GPS (Global Positioning System）に代表される衛生測位システムである GNSS (Global Navigation Satellite System）を用いた手法が広く普及している．GNSS は適切な条件下であれば，数 m から数 10 cm の範囲の誤差で位置推定が可能である．一方都市部では，信号が建物に反射して複数の経路から受信機に到達する多重経路伝搬により，正確な位置推定を行えない状況が発生する．さらにトンネル内においては，信号遮断により原理的に測位できなくなる．そこで，GNSS の測位情報に高精度な慣性計測装置 (IMU: Inertial Measurement Unit）を組み合わせた高精度化が試みられている [1]．しかし IMU は相対的な移動量を推定するものであり，誤差が蓄積していくため，高精度な GNSS による測位情報が定期的に必要となる．

そのため，自動運転システムの多くにおいて，車載外界センサによる街並みの観測情報および対応する位置情報を持つ地図をデータベースとしてあらかじめ用意し，これと走行中の車載外界センサ情報との照合で自車位置を推定するアプローチが広く採られている．例えば，レーザ光を用いて周囲の三次元環境を高精度に計測するセンサである LiDAR を用いて図 1 のような三次元点群地図を構築し，走行中の車両に積んだ LiDAR からの情報との照合により位置を推定する手法が広く用いられている [2]．LiDAR は，観測方向へ照射したレーザ光が観測対象によって反射され，戻って来るまでの時間（ToF: Time of Flight）を計測することで，観測対象までの距離とその反射強度を取得するセンサである．しかし LiDAR は一般的な車両に搭載するにはコストが高く，その特性上車両設計にも大きく制約が生じる．そこで自動運転システムのより幅広い車両における実用化へ向け，道路構造や道路構造物の詳細な情報から成る図 2 のような高精度三次元地図をデータベースとして用意し，これと車載カメラを用いた照合によるアプローチが推進されている [3]．この高精度三次元地図はセンチメートル未満の誤差の

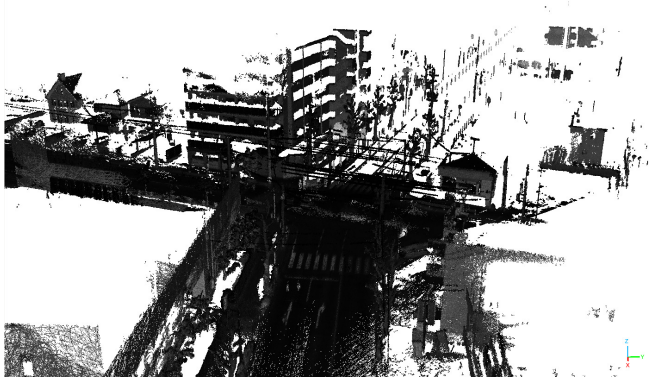


図 1　三次元点群地図の例

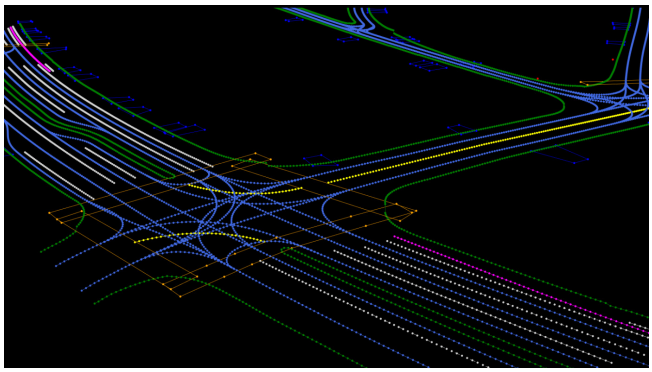


図 2　高精度三次元地図の例

道路構造や道路構造物の詳細な情報を持つベクタ形式の地図であり，専用車両により構築された三次元点群地図などを基に構築される．一方で，高精度三次元地図の作成・更新には多大な人手と労力を必要とするため，地図の対応エリアが限定されやすく，さらに更新頻度の低下などが問題となる可能性がある．

これらの課題を踏まえ本研究では，高精度三次元地図のベースとなる三次元点群地図をデータベースとして用い，安価で広く普及しているカメラを車載センサとして利用し，高精度に自車位置を推定する手法の実現を目的とする．この実現により，幅広い車両で，かつ広範囲な道路でのデータベースとの照合に基づく高精度な自車位置推定が可能となる．このとき，三次元点群地図と走行中の車載カメラ画像間で局所領域同士の対応が得られれば，幾何学的拘束から自車位置を算出できる．その実現に向けた初期検討として我々は，おおよその自車位置が既知であることを前提とし，その位置に対応する三次元点群地図を反射強度画像へ変換した上で，カメラ画像との局所特徴量照合から自車位置を推定する手法を提案した [4]．本稿では，この前提となるおおよその位置を求めるため，三次元点群地図とカメラ画像それぞれから直接抽出された局所特徴量を対応付け，得られた対応関係から自車位置を推定する手法を提案する．ここで，これらの特徴量は三次元点群情報およびカメラ画像それぞれから独立に抽出するため，全く異なる特徴量空間上に存在し，直接的な照合が困難である．そこで提案手法では，異なる空間に存在する特徴同士の対応関係を距離学習により求めることで，得られた対応関係から幾何的拘束を解き，自車位置を推定する．本研究の主な貢献は以下の通りである．

- 三次元点群地図とカメラ画像からそれぞれ抽出された特徴量間の対応関係を表現する距離学習モデルの提案
- 距離学習を実現するための 2D-3D 特徴の対応データセットの自動構築
- 構築したデータセットを用いた評価に基づく提案アプローチの実現可能性の検証

2　関 連 研 究

本研究における重要な要素技術である二次元画像や三次元点群からの局所特徴抽出手法，および三次元点群地図を二次元画像へ変換した上で照合する自車位置推定手法について紹介する．

2.1　二次元画像からの特徴抽出手法

画像からの局所特徴抽出手法についてはこれまでに多くの研究が行われており，SIFT [5] や SURF [6]，AKAZE [7] など，多数の手法が提案されている．これらの手法は，人手による設計に基づく局所特徴量であり，画像の回転や拡大・縮小，照明変化に対して一定の耐性を有することが知られている．

一方で，近年では深層学習を用いて特徴点検出および特徴記述を行う手法も提案されており，その代表例として SuperPoint [8] が挙げられる．SuperPoint は，自己教師あり学習により屋外シーンを含む実画像に適応した局所特徴抽出手法である．HPatches データセット [9] を用いた評価において，照明変化や視点変化を伴う屋外環境下でも従来の局所特徴抽出手法と同等，あるいはそれ以上に幾何的に整合した特徴点対応性能を示すことが報告されている．そのため提案手法では，二次元画像からの局所特徴抽出手法として SuperPoint を採用する．

2.2　三次元点群からの特徴抽出手法

三次元点群からの特徴抽出は，幾何構造を安定して表現することが重要な課題となっている．この課題に対し，これまでに ISS [10] や Harris3D [11] といった特徴点検出手法や，FPFH [12]，SHOT [13] などの局所特徴記述子が提案されてきた．これらは人手による設計に基づく手法であり，点群の回転やスケール変化に対する不変性を備える一方で，環境条件や点密度の変化に対しては性能が低下する場合がある．

近年ではこうした課題に対応するため，深層学習を用いて三次元点群から特徴点および特徴量を抽出する 3Dfeat-Net [14] や CED [15] が提案されている．これらの手法は，幾何構造を元に学習ベースで特徴抽出を行うものであり，従来の人手による設計手法と比較して，より幾何的に一貫した特徴抽出が可能であることが報告されている．

本研究では，カメラで取得した色情報を持つ二次元画像から抽出された特徴点と，三次元点群から抽出された特徴点間の対応付けを目的としている．しかし，三次元点群地図には色情報が存在しておらず，構造物の外観を直接表現することができない．そこで，構造的には平面であっても，材質の違いなどに起因して値が変動する反射強度情報に着目し，これを構造物の外観の情報として利用する．提案手法では，色情報付き点群からの特徴抽出を目的に設計された CED を採用し，反射強度情報

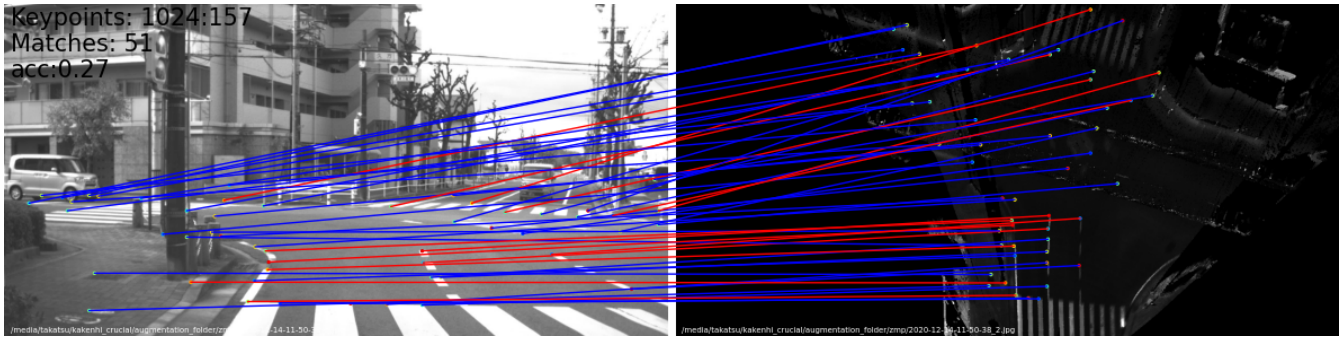


図 3　位置推定誤差が最も低い照合結果 赤線: 正対応

を擬似的なグレースケールの色情報として与えることで，幾何構造に加えて反射強度情報も利用した特徴抽出を行う．

2.3　反射強度情報と可視光情報間の照合に基づく自車位置推定

我々はこれまでに，異なる性質を持つ LiDAR 情報とカメラ情報間の照合が実現可能か確認するため，情報の性質のみに着目した調査を行った [16]．この調査では，カメラと同様の投影用パラメタを用いて三次元点群を二次元画像へと変換し，同一物体の情報が同一箇所に描画された画像を生成した．これにより，視点の差異などによる照合の失敗要因を排除し，各画素が持つ情報の性質の差異のみが現れるデータとした．以上のように生成したカメラ画像と LiDAR ベースの画像を局所画像特徴 [8] に基づいて照合 [17] し，得られた結果について分析を行った [16]．結果として，カメラ画像と視野が概ね一致した視点から生成された反射強度画像において，高精度に局所画像特徴を照合可能であることを確認した．

しかし実際の自車位置推定用途を想定した場合，自車の位置・姿勢を求める段階であるため，同一視点からの反射強度画像を生成することはできない．そこで我々は，異なる視点から生成した反射強度画像とカメラ画像間の照合に基づく自車位置推定手法を提案した [4]．まず，大まかな自車位置を得られている状況を前提とし，三次元点群地図を鳥瞰視点で投影した反射強度画像を生成する．そして，これとカメラ画像間で局所画像特徴量を対応付け，その幾何的拘束から自車位置を推定した．具体的には，画像からの局所画像特徴抽出は SuperPoint [8] を利用し，異なるモダリティ間の局所特徴量の対応付けは SuperGlue [17] を学習することで実現した．そして，鳥瞰視点の反射強度画像とカメラ画像間から得られた局所特徴量に基づく対応点から基本行列を推定することで，車載カメラ位置の推定を行った．その結果，平均 4.43m の精度で位置推定が可能であることを確認した．例として，最も位置推定誤差が低い結果における対応付け結果を図 3 に示す．

この手法は大まかな自車位置が既知であることを前提としており，本研究はこの大まかな自車位置を推定することを目的としている．

3　提 案 手 法

三次元点群地図とカメラ画像間の局所特徴照合に基づく位置推定手法を提案する．ここで，LiDAR と可視光カメラのように異なる原理を持つセンサでは，取得されるデータ表現が本質的に異なる．具体的には，それぞれから抽出される特徴量や特徴次元が異なるだけでなく，LiDAR では幾何構造や反射強度に関する情報が，カメラ画像では輝度やテクスチャに関する情報が反映されている．そのため，三次元点群地図とカメラ画像から抽出されたこれらの特徴量を直接対応付けることは容易ではない．この問題に対し提案手法では，異なる次元・モダリティから得られた特徴量間の対応関係に基づき，対応する特徴量同士が近接し，非対応の特徴量同士が分離されるような特徴間の距離関係を表現する距離学習モデルを構築する．これにより，各特徴量は共通の埋め込み空間へ射影され，距離に基づく比較が可能となる．

提案手法の処理の流れを図 4 に示す．まず三次元点群地図およびカメラ画像から，それぞれのデータに対応する局所特徴抽出器を用いて特徴点およびその特徴量を抽出する．そして，抽出された特徴量をあらかじめ構築した距離学習モデルに入力し，埋め込み空間上での特徴量表現を得る．最後に，特徴量空間内での距離を基に特徴点間の対応候補を求め，RANSAC を用いて PnP（Perspective-n-Point）問題を解くことで，外れ値を除去しつつ自車位置を推定する．

3.1　学習データ作成

距離学習モデルの学習には，対応する特徴量ペア（正例）と対応しない特徴量ペア（負例）から成る学習データが必要となる．しかし，カメラ画像から得られる画像特徴量と三次元点群地図から得られる点群特徴量は，特徴次元や表現の性質が異なるため，それらが対応しているか否か，すなわち実世界での同一位置から得られた点であるか否かを，特徴量の値のみから判断することはできない．そこで提案手法では，高精度な車両位置・姿勢情報およびセンサ間の位置関係と，幾何的制約を用いて画像特徴量と点群特徴量に対応関係を与え，距離学習に用いる学習データを構築する．学習データの作成の流れを図 5 に示す．

まず，カメラ画像からは SuperPoint [8] を，三次元点群地図からは CED [15] をそれぞれ用いて特徴点および特徴量を抽出

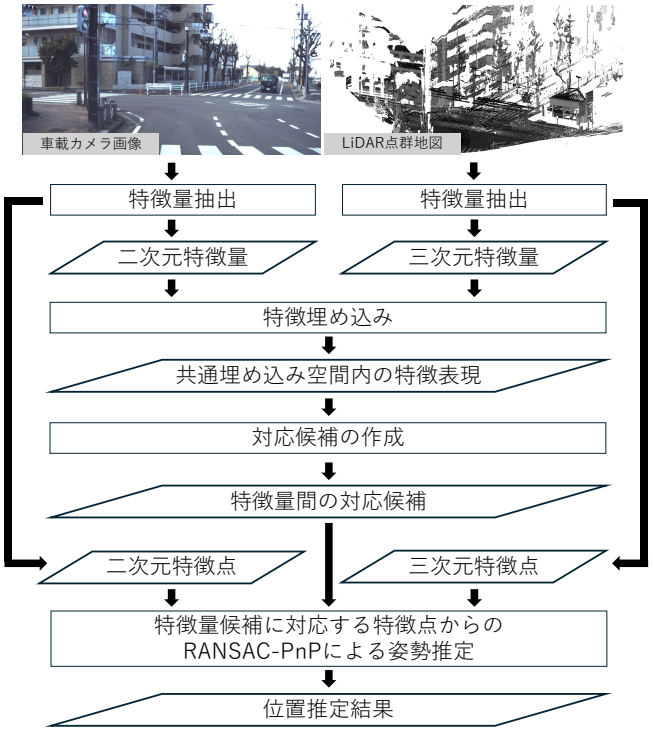


図 4　局所画像領域に基づいた照合と位置推定の流れ

する．次に，点群特徴点を車両の位置姿勢情報およびカメラパラメータに基づいた投影変換処理によりカメラ画像平面へ投影する．ここで，移動体や植生に由来する特徴点は，地図生成時刻と撮影時刻の差や形状変化により画像特徴量と点群特徴量の間で正確な対応関係が得られないため，対応付けにおいてノイズとなる．そのため，画像のセマンティックセグメンテーションを用いて植生，車両，歩行者の領域をマスクとして抽出し，そのマスク領域中の特徴点を動的構造物上の点であるとみなして除外する．このようなマスク処理により選別された特徴点に対し，投影後の位置が D px 以内に存在する特徴点同士の中から，最も近接するペアを正しい対応とみなし，対応付けの正対応候補として選定する．

ここで，カメラパラメタや各センサの取り付け位置を表す外部パラメタには，わずかな誤差が含まれる．また，カメラ画像と三次元点群地図とでは，それぞれ異なる局所特徴検出器を利用して特徴抽出を行っているため，同一箇所から特徴が得られるとは限らない．そのため，選定された正対応候補には誤対応が含まれる可能性がある．そこで，構築された真値ペアに対して RANSAC-PnP を適用し，再投影誤差に基づく幾何的整合性を満たすか否かを検証する．さらに，本処理で得られたインライア特徴点ペアは何らかの幾何的な一貫性を満たしていると判定されたに過ぎず，偶然幾何的拘束を満たした可能性もある．そのため，得られたインライア特徴点ペアを基に推定した自車位置が L m 以内であった場合のみ，これらを最終的な正対応ペアの真値として採用する．

3.2　距離学習モデル

提案手法では，三次元点群地図とカメラ画像という異なるモダリティかつ異なる局所特徴抽出器から抽出された特徴量を同一の埋め込み空間へ射影し，対応関係を学習する距離学習モデルを構築する．

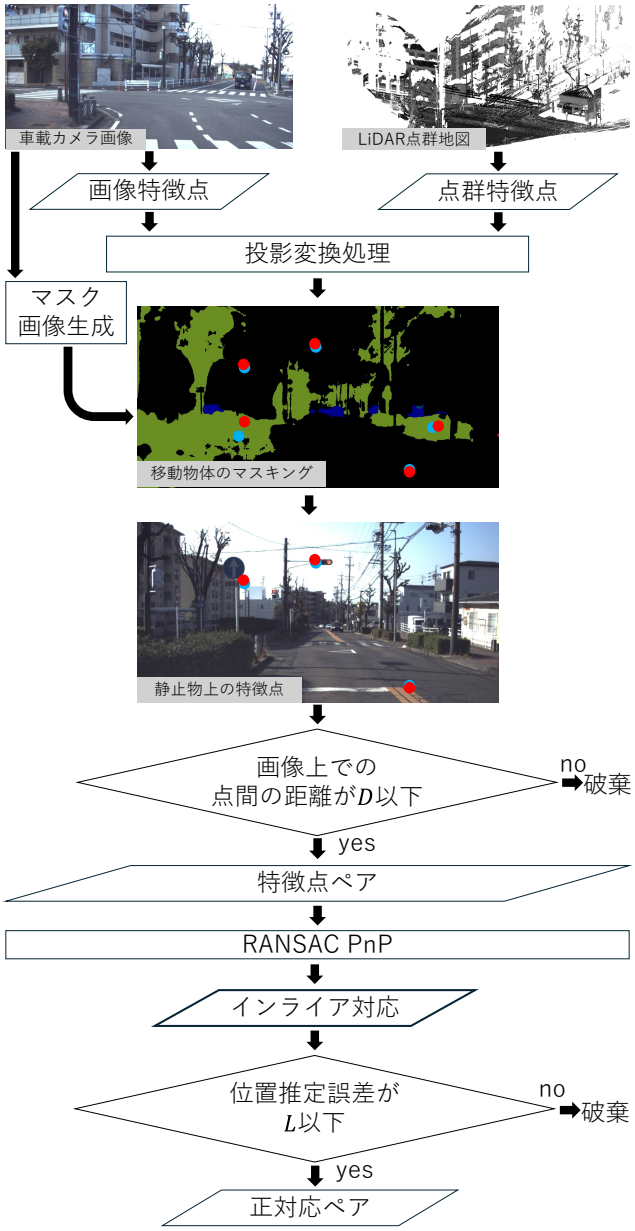


図 5　学習データ作成の流れ

距離学習モデルは三次元点群地図およびカメラ画像から抽出された特徴量をそれぞれ入力とする 2 つの射影ネットワークから構成され，各特徴量を共通の d_{emb} 次元の共通埋め込み空間に射影し，出力を L2 正規化することで距離表現を生成する．損失関数には，対応する特徴量同士が埋め込み空間内で近接し，非対応の特徴量同士が分離されるようなマージン付き距離損失を用いる．具体的には，モデルが出力する L2 正規化された画像特徴量および点群特徴量間の距離をコサイン距離で表現し，正例距離 d^+ と負例距離 d^- の距離差にマージン m を導入したヒンジ型損失により学習を行う．また，画像特徴量から点群特徴量への対応と，点群特徴量から画像特徴量への対応の両方向に対して同一のマージン付き距離制約 $L_{\mathrm{img}\to\mathrm{lidar}}$ および

$L_{\text{lidar}\to\text{img}}$ を適用する．

$$L_{\text{img}\to\text{lidar}} = \max(0, d^{+} - d^{-} + m) \quad (1)$$

$$L_{\text{lidar}\to\text{img}} = \max(0, d^{+} - d^{-} + m) \quad (2)$$

これら 2 つの損失を統合する際に，画像と点群のいずれを基準としても一貫した対応関係が得られるようにするため，最終的な損失関数をそれらの平均として次式のように定義する．

$$loss = \frac{1}{2}(L_{\text{img}\to\text{lidar}} + L_{\text{lidar}\to\text{img}}) \quad (3)$$

3.3 位置推定

学習済み距離学習モデルを用いて，三次元点群地図とカメラ画像間の特徴量対応に基づく位置推定を行う．まず，カメラ画像および三次元点群地図から抽出された特徴量を，学習済み距離学習モデルに入力し，共通の埋め込み空間に写像する．次に，得られた画像特徴量と点群特徴量の埋め込み間のユークリッド距離を算出し，相互最近傍となるペアを対応点候補として抽出する．そして，得られた対応ペアに含まれる画像上の二次元特徴点座標と，三次元点群地図上の対応する三次元点座標を用いて，RANSAC-PnP によりカメラの位置・姿勢を推定する．最後に，推定された三次元点群地図の座標系（ENU 座標系）で表現されるカメラ位置を，カメラの外部パラメータを用いて車両位置へ移動することで，最終的な自車位置推定結果とする．

4 評価実験

提案手法の有効性を確認するため，実際の走行データを用いて評価を行った．以降，評価に用いたデータや実験条件，実験結果について順に述べる．

4.1 実験データ

実験データには，高精度な位置情報を取得可能な Applanix POSLV 610 および LiDAR の Livox Mid-100，ステレオカメラの ZMP Robo Vision2s を搭載した図 6 に示す実験車にて，豊田工業大学周辺の図 7 に示すルートを周回して取得したデータを利用した．Applanix POSLV 610 は RTK-GNSS，高精度 IMU，高解像オドメトリセンサなどの複数のセンサから構成され，これらの情報を統合利用することで位置情報を算出する MMS（Mobile Mapping System）である．さらに，測位データ解析・後処理ソフトウェアである POSPac MMS を利用して最適化することで，高精度かつ安定した位置・姿勢情報を算出可能である．この情報を，以降の実験における各センサ情報取得時点の車両の正しい位置・姿勢として用いた．ZMP Robo Vision2s は左右それぞれ幅 1,280× 高さ 960px の画像を取得できるステレオカメラであり，本実験では左側のカメラから撮影した画像をカメラ画像として利用した．ただし，取得された画像には特徴抽出の際にノイズとなる車両のダッシュボードなどのフロントガラスへの反射の映り込みが含まれるため，幅 1,280× 高さ 640px に切り出したものを利用した．

4.2 三次元点群地図の構築とノイズ除去

学習や評価に用いる特徴点マップを生成する際に利用する三次元点群地図の構築は，LiDAR および MMS（Mobile Mapping System）を搭載した車両により取得された点群データを統合することで構築した．本地図の構築では，Applanix POSVL 610 で取得し，POSPac MMS により最適化された各時刻における高精度な位置姿勢データを用いることで，LiDAR から得られたセンサ座標系における三次元点群を世界座標系へ変換して結合することで，三次元点群地図を構築した．

ここで，単純に結合した三次元点群地図には，車両や歩行者などの移動体に由来するノイズが含まれてしまう．そこで本研究では，複数周回で一貫して観測される静的構造のみを残すことを目的として，ボクセル整合性に基づくノイズ除去を行った．対象とするデータは，同一ルートを 5 周走行して取得された点群であり，全周回の点群に対して共通の三次元ボクセルグリッドを定義した．具体的には，地図座標の原点を基準にボクセルサイズ 0.2 m の格子に空間を分割した．つぎに，各ボクセルについて周回ごとの点数を集計し，ある周回において当該ボクセル内の点数が 5 点以上である場合を「観測あり」と定義した．その上で，「観測あり」を満たす周回数が 3 周回以上であるボクセルのみを静的構造として保持し，それ以外のボクセルは移動体に由来するノイズとみなし除去した．そして，保持されたボクセルに属する点のみを各周回から抽出し結合することで，ノイズを大まかに除去した三次元点群地図を構築した．

図 6 実験車

図 7 走行ルート

4.3 実験条件

図 7 に示したルートで取得した周回データを，走行方向に沿って 20 m 間隔で分割し，各区間に対応する局所的な三次元点群地図情報とカメラ画像の組を 1 箇所として切り出すことで

データセットを構築した．そして，これらから図 5 におけるインライア対応を得られた地点のみを残し，学習データ 3,143 箇所，検証データ 448 箇所，および評価データ 897 箇所を抽出し，本実験に用いた．このうち学習データおよび検証データについて，図 5 の処理で絞り込み，それぞれ 1,106 箇所，151 箇所を学習に用いた．

三次元点群地図の特徴抽出には CED [15] を用いた．CED は，ステレオカメラにより取得された高密度な色付き三次元点群を対象として，点群の幾何構造および色情報に基づき，特徴点および特徴量を抽出することを目的に設計された手法である．一方で，本研究では実スケールの市街地環境を対象とした，広範囲かつ高密度でない三次元点群地図を扱い，点群が持つ反射強度情報をグレースケールの色情報として利用する．そして，点群密度のばらつきに対して安定した特徴点抽出を行うことを目的として，ボクセルサイズを 0.17 m に設定し，近傍探索半径および CED における重心距離のしきい値を拡大することで，幾何構造および反射強度差に基づく特徴点が抽出されるよう調整した．

3.1 節におけるノイズ除去のためのセマンティックセグメンテーション手法には, MMsegmentation が提供する, CityScapes データセット [18] で学習された DeepLabV3+ [19] を利用した．また，学習データ作成時における正対応判定のしきい値は，$D = 7\,\mathrm{px}$ および $L = 1\,\mathrm{m}$ とした．さらに，距離学習モデルが射影する埋め込み空間の次元は $d_{\mathrm{emb}} = 16$ とし，損失関数に用いるマージンは $m = 0.6$ に設定した．

推定位置誤差を求める際には，RANSAC-PnP にて推定された位置姿勢を地図座標系へ変換し，POSPAC MMS によって得られた真値の水平方向の位置および姿勢の差を求めることで位置推定誤差を算出した．

4.4 実験結果

各データに対して提案手法による位置推定を行い，PnP を解けたデータ数，およびそれらの位置推定誤差を表 1 に示す．表から，PnP 成功割合が非常に少ないこと，および位置推定精度が非常に低いことが確認された．

評価データにおいて 1 m 未満の自車位置推定が行えた際の対応付け結果を図 8 に示す．また，評価データにおいて自車位置推定誤差が最大であった際の対応付け結果を図 9 に示す．図 8 では各点に対して正しく一対一の対応付けが行えているが，図 9 では明らかな誤対応ばかりであり，かつ画像の視野角内で偏った位置の特徴点のみから幾何的拘束を解いたことがわかる．以上から，構築されたモデルでは特徴量間の正確な対応関係を十分に学習できておらず，かつ汎化性能が担保されていないと考えられる．

4.5 考察

幾何的拘束を解けるシーンが少なく，かつ位置推定精度が低かった要因として，三次元点群とカメラ画像という異なるモダリティから抽出された特徴量を対応付ける必要があるという困難な問題設定であった点が挙げられる．構築した距離学習モデルにおける，学習データにおける正対応ペアおよび誤対応ペアそれぞれの距離の分布を図 10 に示す．このように，正対応ペアに比べて僅かに誤対応ペアの方が距離が遠くなる傾向は学習できているものの，それらを十分に分離できなかったことが分かる．そのため，これらのモダリティを横断する対応関係を埋め込み空間上でより適切に表現化可能な距離学習モデルの構築が必要である．

さらに，学習データの品質の低さも要因となり得る．ここで，車両が静止している際のデータの場合，センサ間の時間同期のズレや車両の振動などの影響を無視できるため，高品質な学習データが得られると考えられる．そこで本実験で抽出したデータとは別途，図 7 のルートを走行したデータの中で，交差点等における静止シーンから学習データを用意し，距離学習モデルを構築した際の距離の分布を図 11 に示す．この距離分布では，概ね良好に正対応ペアおよび誤対応ペアの距離を測れていることが分かる．このことから，より正確な対応ペアを大量に用意

表 1　位置推定可能であった評価シーンの自車位置推定誤差

		シーン数および位置推定誤差
位置推定成功シーン数（成功数/全シーン数）		60/897
位置推定誤差	最小値 (m)	0.89
	最大値 (m)	227.70
	平均値 (m)	57.28
	中央値 (m)	53.39

図 8　位置推定誤差が 1 m 未満の対応付け結果
青点：カメラ画像の特徴点 赤点：三次元点群地図の特徴点
黄線：距離学習モデルによって算出された対応関係

図 9　位置推定誤差が最大時の対応付け結果
青点：カメラ画像の特徴点 赤点：三次元点群地図の特徴点
黄線：距離学習モデルによって算出された対応関係

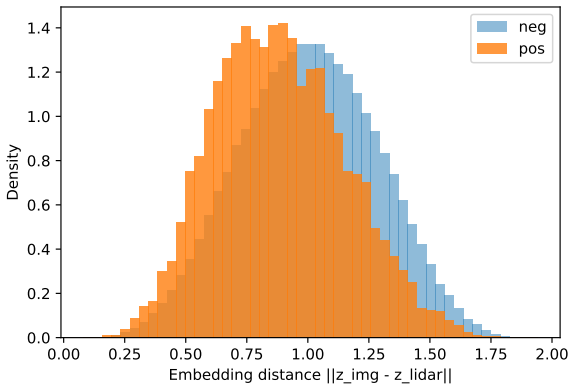


図 10　学習データにおける正対応および誤対応の距離の分布

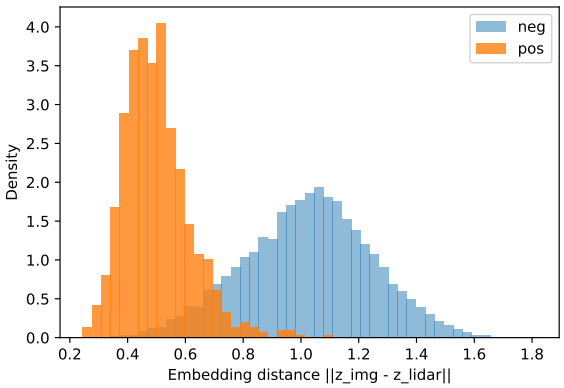


図 11　静止データにおける正対応および誤対応の距離の分布

して学習することで，より高精度な位置推定が可能となると考えられる．

さらに，それぞれ異なる局所特徴抽出モデルを用いているため，位置的に異なる箇所から抽出された特徴同士を正対応ペアと見做して対応付けることになっている可能性もある．これらの問題点を解決するためには，カメラ画像および三次元点群地図の双方に対応し，同一箇所から局所特徴を抽出可能なモデルの構築が求められる．

5　ま　と　め

本稿では，低コストかつ汎用性の高い自車位置推定手法の構築を目的として，三次元点群地図とカメラ画像間の照合による自車位置推定手法を提案した．提案手法では，異なる次元・モダリティのデータそれぞれから抽出された局所特徴量間の対応付けが困難であるという課題に対し，特徴量間の対応関係を学習する距離学習モデルを構築した．さらに，幾何的拘束条件などから制約をかけることによる，正しい対応関係を表す学習データの自動収集を試みた．実験の結果，一部のデータについては良好な位置推定が可能であったものの，汎化性能の低い結果となった．今後の課題としては以下が挙げられる．

学習データの高品質化

走行中のデータから構築したデータセットは，センサ間の時間同期のズレの影響を受ける．また，キャリブレーションパラメータの精度が十分でない場合，各種投影時にズレが生じる．これらの問題を解決し，より高品質なデータを用いた評価を行う必要がある．

各モダリティで共通した特徴点抽出

提案手法では，カメラ画像および三次元点群地図それぞれについて，画像用および点群用に設計された独立の局所特徴抽出手法を用いた．これらは各モダリティ内において特徴的である点を抽出するため，空間的に同一の点が検出されない可能性がある．そのため，両モダリティで共通して特徴的となる点を検出し，その特徴量を記述可能な局所特徴抽出手法を設計する必要がある．

謝　　辞

本研究の一部は阪神高速若手研究者助成基金および JSPS 科研費 JP22K17916 の支援による．また，本実験に使用した車載センサデータを提供頂いた豊田工業大学に感謝する．

文　　献

[1] Kana Nagai, Matthew Spenko, Ron Henderson, and Boris Pervan. Evaluating ins/gnss availability for self-driving cars in urban environments. *Proceedings of the 2021 International Technical Meeting of The Institute of Navigation*, pp. 243–253, 2021.

[2] Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *IEEE robotics & automation magazine*, Vol. 13, No. 2, pp. 99–110, 2006.

[3] Chengcheng Guo, Minjie Lin, Heyang Guo, Pengpeng Liang, and Erkang Cheng. Coarse-to-fine semantic localization with hd map for autonomous driving in structural scenes. *2021 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp. pp.1146–1153, 2021.

[4] 高津悠生, 道満恵介, 川西康友, 久徳遙矢. 三次元点群地図に基づく鳥瞰反射強度画像とカメラ画像間の局所特徴照合による自車位置推定の検討. 画像の認識・理解シンポジウム 2025, pp. IS2–143, 2025.

[5] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, pp. pp.91–110, 2004.

[6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. *CVPR*, pp. pp.346–359, 2008.

[7] Pablo F. Alcantarilla, Jesús Nuevo, and Adrien Bartoli. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *BMVC*, pp. pp.13.1–13.11, 2013.

[8] D. DeTone et al:. Superpoint: Self-supervised interest point detection and description. *CVPR2018 Workshop on Deep Learning for Visual SLAM*, pp. 337–349, 2018.

[9] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. *CVPR*, 2017.

[10] Y. Zhong. Intrinsic shape signatures: A shape descriptor for 3d object recognition. *in THE IEEE International Conference on Computer Vision Workshops*, 2009.

[11] I. Sipiran and B. Bustos. Harris 3d: a robust extension of the harris operator for interest point detection on 3d meshes. *The Visual Computer*, p. pp. 963 – 976, 2011.

[12] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. *IEEE*

international conference on robotics and automation, pp. 3212–3217, May 2009.

[13] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique signatures of histograms for local surface description. *Proc. ECCV*, Vol. 6313, pp. 356–369, September 2010.

[14] Zi Jian Yew and Gim Hee Lee. 3dfeat-net: Weakly supervised local 3d features for point cloud registration. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 607–623, September 2018.

[15] Hanzhe Teng, Dimitrios Chatziparaschis, Xinyue Kan, Amit K. Roy-Chowdhury, and Konstantinos Karydis. Centroid distance keypoint detector for colored point clouds. *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1196–1205, 2023.

[16] 高津悠生, 久徳遙矢, 谷本樹希, 道満恵介, 秋田時彦. Lidar 点群情報とカメラ画像間の局所画像特徴マッチングに関する初期検討. 令和五年度 電気・電子・情報関係学会東海支部連合大会, pp. D4–4, 2023.

[17] P.E. Sarlin et al:. Superglue: Learning feature matching with graph neural networks. *CVPR*, pp. 4938–4947, 2020.

[18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[19] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Computer Vision – ECCV 2018*, 2018.

ユニバーサル基板を使用した電子回路の不良原因推定に向けた深層学習による電子部品とその端子検出

小泉　天翔† 林　哲矢† 田中　剛† 遠藤　雅樹† 寺田　憲司†
大野　成義†

† 職業能力開発総合大学校　〒 187–0035 東京都小平市小川西町 2–32–1
E-mail: †{b22310,m243202,t-tanaka,endou,k-terada,ohno}@uitec.ac.jp

あらまし　ユニバーサル基板は, 電子回路製作において柔軟な回路設計や配線変更, 部品交換が可能であることから, 教育や製品試作の現場で用いられている．一方, 回路構成の自由度が高いため, 目視による不良箇所の特定には多大な時間と労力を要している問題がある．先行研究では, ユニバーサル基板の配線面画像から回路図を復元し, 正解回路図との比較による不良箇所の自動推定方法が提案されている．しかし, 回路図復元に必要となる電子部品および端子位置を画像から検出する方法については, 十分に検討されていない．そこで本研究では, 深層学習を用いてユニバーサル基板上の電子部品および端子を検出する方法を提案する．具体的には基板画像に画像処理を施して背景成分を抑制した後, 物体検出モデルにより電子部品および端子の位置を推定する．さらに, ユニバーサル基板ではスルーホールが等間隔に配置されている構造的特徴に着目し, 各電子部品が接続されるスルーホールの行および列を特定する．提案法により, 基板上の電子部品の種別および端子位置の情報を抽出可能であることを確認した．また, 本方法を先行研究に適用することで, 不良原因推定の自動化が可能となる見込みを得た．

キーワード　ユニバーサル基板,　不良原因推定,　電子部品検出,　深層学習,　物体検出

1　はじめに

ユニバーサル基板は, プリント基板とは異なりスルーホールが等間隔に配置された基板で, 柔軟な電子部品配置や配線変更が可能である．この特性により, 工学教育や製品試作の現場で広く活用されている．特に教育現場では, 学習者が回路設計から製作, 動作確認までの一連の流れを体系的に学ぶための重要な教材として位置づけられている [1], [2]．一方, ユニバーサル基板はその自由度の高さゆえに, 誤配線や電子部品配置ミスといった製作不良が発生しやすいという課題を持つ．現状, これら不良原因の特定は指導者や熟達者による目視確認に依存しており, 学習者ごとに異なる多様な回路に対してその特定を行うには多大な労力と時間を要する．このような確認作業の負担は, 学習者の待機時間の増大や教育現場における学習効率の低下を招く要因となっており, 問題視されている [3], [4].

この問題を解決するために, 教育現場における確認作業をデジタル化し, 効率化する取り組みが先行研究で進められている．林ら [5], [6] は, ユニバーサル基板の配線面画像を解析して配線パターンを抽出し回路図を復元する方法および復元回路図と正解回路図の比較に基づいて不良箇所を自動推定する方法を提案している．しかし, 回路図の復元には, 配線情報だけでなく基板上に実装された電子部品の種別とその端子位置を正確に把握する必要があるものの, 先行研究ではその具体的な検出方法が十分に提案されていない.

そこで本研究では, 深層学習に基づく物体検出により, 電子部品と端子を同時に検出し, スルーホールの行および列へ対応付けて, 回路図復元・不良原因推定に必要な情報を抽出する方法を提案する．深層学習を用いてユニバーサル基板上の電子部品面に実装された電子部品およびその端子を同時に検出可能な物体検出モデルを構築する．検出された電子部品と端子の位置関係に基づいて, 各端子が接続されているスルーホールの行列位置を特定し, 電子部品の種別と端子位置を行列形式で出力する方法を提案する．本方法により, ユニバーサル基板を使用した電子回路における不良原因推定に必要な電子部品および端子位置情報を抽出することを可能とする.

本論文の構成は以下のとおりである．はじめに, ユニバーサル基板における不良原因推定に関する先行研究を概観し, 本研究の位置づけを示す．次に, プリント基板上の電子部品検出に関する関連研究を紹介する．続いて, 提案方法の全体構成を示し, その詳細について説明する．さらに, 実験結果および考察を述べる．最後に, 結論と今後の課題について述べる.

2　先行研究

先行研究として, 工学教育の現場における確認作業の効率化を目的とした取り組みが報告されている．林ら [7] は, ユニバーサル基板における回路図比較を用いた不良原因自動推定方法を提案している．本方法は, 指導者が目視で行ってきた不良原因推定作業をデジタル化し, 電源を投入することなく回路の動作不良の有無およびその原因を推定することを目的としている．林らの方法では, ユニバーサル基板の電子部品面および配線面の画像を入力とし, 基板裏面における配線面画像から配線パターンを抽出することで回路図を復元する．復元された回路

図をあらかじめ用意された正解回路図と比較し, さらに回路シミュレーションを行うことで, 不良箇所および不良原因を推定する構成となっている．先行研究において, 基板裏面の配線面画像から配線パターンを抽出する方法については, 画像処理を用いた具体的な方法が提案されている．一方で, 回路図復元に不可欠な情報である基板表面の電子部品面画像から, 実装された電子部品の種別やその端子位置を抽出する具体的な方法については示されていない．

本研究により, 先行研究で提案されている回路図復元を介した不良原因推定において必要となる入力情報の提供が可能となり, 電子回路製作支援の高度化に寄与することが期待される．

3 関 連 研 究

本章では, 電子回路基板を対象とした既存の電子部品検出方法および回路製作支援システムに関する関連研究について述べる．

電子回路基板における自動光学検査 (AOI) の分野では, 画像解析や深層学習を用いたプリント基板上の電子部品検出方法が数多く報告されている．画像処理に基づく方法として, 茂木ら [8] は,HSV 形式の色情報や境界線追跡を用いてプリント基板上の電子部品を認識する方法を提案している．しかし, 同方法は照明条件や明暗の影響を受けやすく, 誤認識により検出精度が低下することが報告されている．一方, 深層学習を用いた方法として,Ong ら [9] は,Faster R-CNN や YOLOv3,SSD FPN などの物体検出モデルを用い, プリント基板上に表面実装された電子部品の識別および位置特定を行う方法を提案している．また,Kim ら [10] は,YOLOv5 モデルをソリッドステートドライブの基板に適用し, 微小な電子部品の検出を行う方法を示している．これらの研究はいずれも, あらかじめレイアウトされたプリント基板を対象としており, 電子部品の欠損や実装不良の検出を主目的としている．

しかし, 本研究が対象とするユニバーサル基板は, 電子部品配置や配線が自由であり, 回路構造が個々に異なるという特徴を持つ．そのため, プリント基板を前提とした既存の電子部品検出方法をそのまま適用した場合, 正確な電子部品検出や端子位置の検出が困難となり, 回路の接続関係を抽出することは容易ではない．ユニバーサル基板を対象とした関連研究として,Takemura ら [11] は,Web ベースの指示に従ってユニバーサル基板上の回路製作を支援する教育システムを提案している．このシステムでは, セグメンテーションやパターン認識を用いて基板上の回路構造を認識し,SPICE シミュレーション形式に変換することで回路動作の確認を行っている．同研究は, ユニバーサル基板上の回路構造認識を目的としている点で本研究と関連するが, ユニバーサル基板が持つスルーホールの行列形式に着目し, 電子部品およびその端子位置を体系的に特定する方法については十分に検討されていない．

以上より, 既存の研究では, プリント基板を対象とした電子部品検出方法や, ユニバーサル基板を対象とした回路製作支援システムが提案されているものの, ユニバーサル基板の構造的特徴を活用して電子部品と端子位置を整理し, 回路接続関係の特定に適した情報として出力する方法については十分に検討されていない．

4 研究のアプローチ

以下では, ユニバーサル基板を用いた電子回路における不良原因推定の自動化に向けて, 電子部品およびその端子位置を抽出することを目的とした本研究のアプローチについて述べる．まず, **4. 1** において, 研究対象であるユニバーサル基板が有する構造的特徴と, 先行研究による回路図復元において必要とされる情報抽出の形式を整理する．次に **4. 2** では, 基板の規則的な構造に着目することで本研究が採用した方法選択の根拠および基本的な考え方を示す．最後に **4. 3** において, 提案する処理の全体的な流れを概説する．

4. 1 研究対象の特性と情報抽出の目標

本研究が対象とするのは, 教育現場や製品試作の現場において広く用いられているユニバーサル基板上に実装された電子回路である．ユニバーサル基板は, 柔軟な配線および電子部品配置が可能であり, 回路設計の自由度が高いという特徴を持つ．構造的な特徴として, ユニバーサル基板には, 電子部品を挿入するための貫通孔であるスルーホールが等間隔に配置されている．これらのスルーホールは行列状に整列しており, 各ホールは基板上の行および列に変換することが可能である．この規則的な配置は, 基板上に実装された電子部品の位置を行列形式として表現するための基準となる．不良原因推定の自動化に向けては, 基板上の電子部品間の接続関係をデータとして取得する必要がある．そこで本研究では, 基板の行列形式を基準とし, 実装された電子部品の種別を示す電子部品ラベルと, 各電子部品の端子が接続されているスルーホールの行および列の座標値から構成される行列情報を抽出することを目標とする．

4. 2 基板における行列形式に基づく提案方法の選択

4. 1 で述べたように, ユニバーサル基板はスルーホールが等間隔に配置されているという明確な構造的特徴を有している．本研究では, この規則性を活用することで, 基板上の電子部品および端子位置を体系的に整理することを目指す．

電子部品および端子を画像処理のみによって識別する場合, 電子部品の種別ごとにしきい値や形状条件を個別に設定する必要があり, 撮影条件や外観のばらつきに応じた再調整が頻繁に必要となる．そのため, 本研究では, 外観のばらつきを学習によって吸収可能な深層学習に基づく物体検出モデルを用いて, 電子部品および端子の位置を推定する．

物体検出モデルは, 画像中に存在する複数の対象物を同時に検出し, それぞれの位置およびクラスラベルを推定できるため, 基板上に実装された電子部品と端子候補を効率的に抽出することが可能である．さらに, 検出された端子位置を, スルーホールの行列形式に基づいて行および列へ対応付けることで, 回路の接続関係を表現するために適した形式の情報を得ることができる．

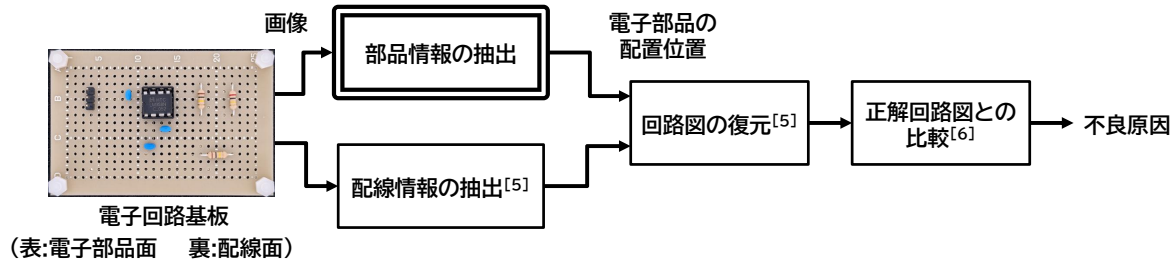


図 1　先行研究における全体フローと本研究の位置づけ
(二重枠線部分が本研究の提案範囲)

なお, 本研究における検出対象は, 電子部品の種別および端子位置の特定に限定し, 抵抗値や静電容量といった部品値による分類は対象外とする．これは, 抵抗器のカラーコードを用いた値の識別については既存研究において検討が進められていること [12], [13] ならびに, 基板を真上から撮影した画像のみでは部品値を正確に識別することが困難な場合があるためである．

4.3　提案方法の全体的な流れ

2 で述べた先行研究では, 電子部品面および配線面の情報を統合し, 回路図復元を介して正解回路図との比較を行うことで不良原因を推定する方法が提案されている．本研究は, その中でも十分に検討されていない電子部品面情報の抽出に着目し, 電子部品および端子位置を取得する方法を提案する．先行研究における全体的な処理フローと, 本研究の位置づけを図 1 に示す．図 1 において, 二重枠線で示された処理が本研究の提案範囲であり, それ以外の処理は先行研究の範囲に該当するものである．

本研究における処理の流れは, 基板画像から電子部品および端子を検出し, 端子位置をスルーホールの行および列へ変換して出力するものである．まず, 基板画像に対して背景抑制を目的とした画像処理を施し, 背景成分の影響を低減する．次に, スルーホールの配置に基づいてホール間距離を算出し, 後段の電子部品および端子の対応付けに用いる距離基準を設定する．

続いて, 深層学習に基づく物体検出モデルを適用し, 電子部品および端子の位置と, 電子部品のクラスラベルを推定する．その後, 検出された電子部品座標および端子座標に基づいて, 各端子を対応する電子部品へ割り当て, 端子がどの電子部品に接続されているかを特定する．最後に, 端子座標を行および列の情報として出力する．なお, 各処理の具体的な手順および設定条件については, **5** において詳述する．

5　提 案 方 法

本章では, ユニバーサル基板上に実装された電子部品およびその端子位置を検出し, 接続位置を行列形式で出力するまでの提案方法について述べる． **5.1** では基板画像の取得条件を示す． **5.2** では背景抑制および角度補正を含む前処理と, 画像回転によるデータ拡張方法について述べる． **5.3** では電子部品検出モデルおよび端子検出モデルの学習方法を示す． **5.4** では推論結果の統合方法について説明する． **5.5** では電子部品検出結果と端子検出結果の対応付け方法を示す． **5.6** では対応付けた端子位置をスルーホール配列に基づく行列形式へ変換し,CSV 形式で出力する手順を説明する．提案方法全体の処理フローを図 2 に示す．

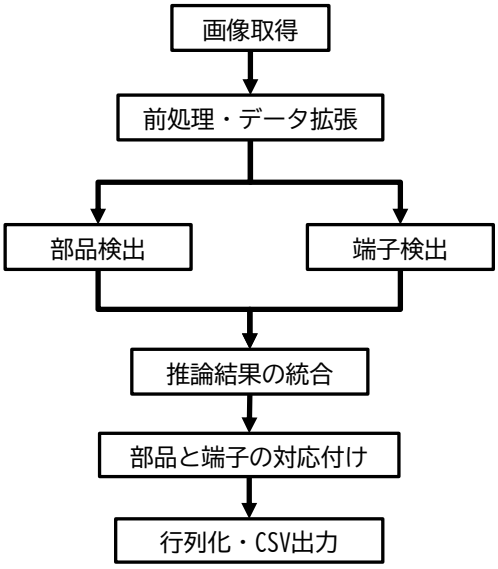


図 2　提案方法の全体フロー

5.1　基板画像の取得

基板画像の取得には, 入手性が高く, 教育現場や製品試作の現場においても追加設備を必要とせずに撮影環境を再現しやすいスマートフォンカメラを用いる．撮影は, 基板全体が画角内に収まるように真上から行い, 照明器具に可能な限り近づけた高さの距離に固定して行う．照明には白色 LED ドームライトを用いる．ドームライトは基板全体を均一に照射でき, 反射や影の影響を抑制しやすい特性を持つ．これにより, 撮影条件に起因する濃淡むらや局所的なハイライトを低減し, 後段の検出処理に対して安定した入力画像を取得する．使用した撮影環境および照明装置の詳細を表 1 に, 撮影環境の外観を図 3 に示す．

本研究では, 工学教育現場における電子回路実習において学生が製作した 82 枚のオペアンプ発振回路基板を撮影した画像を用いる．なお, 電子部品における配置のレイアウトは学生に委ねられており, 各基板において自由な配置となっている．

表 1　撮影環境および使用機器一覧

項目	内容
撮影装置	スマートフォンカメラ（Google Pixel 6a）
照明装置	白色 LED ドームライト（TD302 × 60 – 160W – 4）
撮影距離	168 mm
撮影角度	基板に対して垂直（真上）
対象基板	オペアンプ発振回路（82 枚）

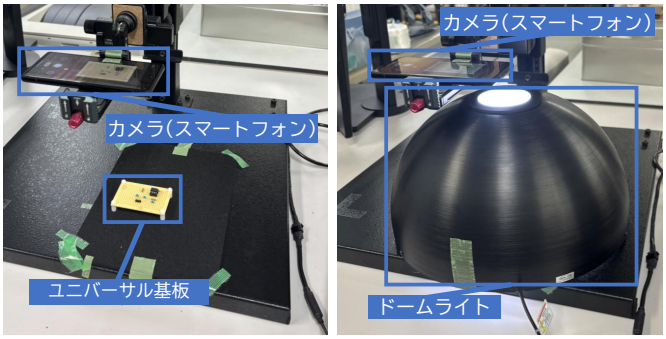


図 3　ユニバーサル基板の撮影環境

5.2　前処理とデータ拡張

本節では, 撮影した基板画像に対して背景成分の影響を抑える前処理を行い, 回転によるデータ拡張を施した学習用画像データを作成する．ユニバーサル基板の撮影画像には, 基板領域以外の背景が含まれるほか, 撮影時のわずかな傾きや位置ずれが生じる場合がある．これらはノイズとして後段の処理に悪影響を及ぼす可能性があるため, 本研究では, 基板外形の検出に基づく背景抑制および角度補正を前処理として実施し, 入力画像の条件を可能な範囲で統一する．

5.2.1　基板外形の検出に基づく切り抜きと角度補正

本節では, 基板外形の検出に基づく切り抜きおよび角度補正の手順について述べる．基板領域の抽出には, 林らの方法に基づき, 入力画像から基板外形を推定した後, 角度補正と切り抜きを行う．

まず, 入力画像に対してガウシアンフィルタを適用し, 微小なノイズを低減する．次に, 基板外形を検出して四隅のコーナ位置を推定し, それらの配置関係から基板の傾き角を算出する．推定した傾き角に基づいて画像を回転させ, 基板の長辺および短辺が画像座標系に対して水平および垂直になるように角度補正を行う．

角度補正後の画像に対しては, 再度基板外形の検出およびコーナ推定を行い, 得られたコーナ座標を用いて基板領域を切り抜く．切り抜いた画像は, 後段の物体検出処理に適した解像度を確保するため, リサイズ処理を施す．この一連の処理により, 撮影環境に依存する背景成分や基板の傾きが低減され, 電子部品および端子の特徴を安定して抽出できる入力画像が得られる．切り抜きおよび角度補正後の基板画像の例を図 4 に示す．

5.2.2　回転によるデータ拡張

本節では, 回転処理によるデータ拡張について述べる．深層学習モデルの検出性能を十分に発揮させるためには, 多様な学習データが必要となる．しかし, 実際に撮影可能な基板画像の

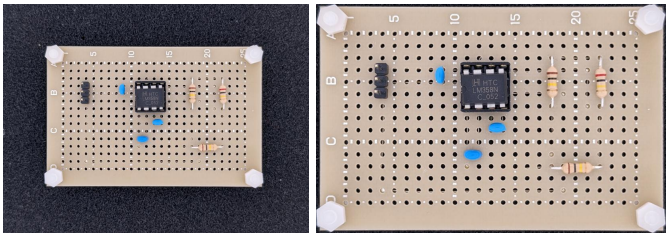
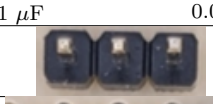


図 4　前処理による基板画像の切り抜きと角度補正例
(左: 元画像 右: 切り抜きおよび角度補正後の画像)

表 2　検出対象とする電子部品

対象とする電子部品	電子部品画像
抵抗器	 100 kΩ　200 kΩ
セラミックコンデンサ	 0.1 μF　0.01 μF
3 端子ピンヘッダ	
ジャンパ線	
8 ピン IC	 LM358N

枚数には限りがある．本研究では, 全 82 枚の基板画像のうち 12 枚を最終的な方法評価に用いるテストデータとして独立させ, 残りの 70 枚を学習データとして使用する．学習データに対しては, 限られた画像枚数から疑似的にデータの多様性を高めるため, 回転によるデータ拡張を行う．具体的には, 前節の前処理によって標準化した画像に対し,0° から 315° まで 45° 刻みで回転させる．この処理により, 撮影時に基板がどの方向を向いていても, 電子部品および端子の特徴を識別できるような学習サンプルを生成する．画像内の電子部品および端子位置を示す正解ラベルの座標情報についても, 回転後の位置に合わせて変換し, 学習用データセットを合計 560 枚に拡張する．

5.3　電子部品検出モデルおよび端子検出モデルの学習方法

本節では, ユニバーサル基板上に実装された電子部品および端子位置を特定するための学習方法について述べる．検出対象とした電子部品の一覧を表 2 に示す．なお, **4.2** で述べたように, 本研究では電子部品の種別および端子位置の特定に検出対象を限定し, 抵抗値や静電容量といった部品値による分類は対象外とする．

本研究では, 電子部品および端子の検出に深層学習に基づく物体検出モデルを用いる．具体的には,YOLOv11 [14] を採用する．YOLOv11 は, 画像全体を一度の処理で解析する構造を持ち, 複数の対象物を高速かつ高精度に検出できるという特徴を有する．また, 検出と同時に各物体の位置をピクセル座標とし

て出力できるため, 基板上の電子部品および端子位置の特定に適している.

モデルの学習および評価には,5-fold の交差検証を用いる．これは, 学習用画像データを 5 つのグループに分割し, 各グループを 1 回ずつ検証データとして使用し, 残りの 4 グループでモデルを学習する方法である．この方法により, 学習データの偏りを抑えつつ, モデルの汎化性能を評価する.

本研究では, 電子部品検出と端子検出をそれぞれ独立したモデルとして学習させる．電子部品と端子は画像中で占める面積や外観的特徴が大きく異なる．そのため, 単一のモデルで両者を同時に検出する場合, いずれかの検出精度が低下する可能性がある．そこで, 本研究では検出対象ごとに特化したモデルを用いることで, 検出精度の向上を図る．このように, 電子部品検出モデルと端子検出モデルを独立して学習させ, それぞれのパラメータを調整することで, 基板上の電子部品および端子位置の高精度な検出を目指す.

5.3.1 電子部品検出モデルの学習

基板上に実装された電子部品の識別と位置推定を行うモデルの学習について述べる．本モデルの目的は電子部品の種別を正しく特定することにあり, 抵抗器のカラーコードから得られる詳細な抵抗値などといった電子部品の数値情報の特定までは行っていない.

具体的には, 前節で作成した学習用画像データ 560 枚を用いて, 学習用データセットを構成する．データの分割においては, 回転後の画像が学習用と検証用にまたがって混在してしまうことを防ぎ, 評価による検証用データには回転前の元画像から生成されたデータのみを含めるように注意する．これにより, モデルが学習時に見たことのない基板画像での性能を正確に評価できる．主要なハイパーパラメータとしては, バッチサイズ 32, エポック数 100, 入力画像サイズ 1024 × 1024 ピクセルと設定する．さらに, 撮影時のわずかな照明環境の変化に対するロバスト性を高めるため, 学習時のデータ拡張として, 彩度および明度のランダム変化を適用し, モデルが多様な外観条件に対応できるようにする.

5.3.2 端子検出モデルの学習

電子部品の接続位置を特定するために必要な, 基板上の電子部品に接続された端子位置を特定するモデルの学習について述べる．端子検出モデルでは **5.3.1** で説明した電子部品検出モデルと同様に, 前節で作成した学習用画像データ 560 枚を用いて学習を行う．データの分割においても, 電子部品検出モデルと同様に, 回転後の画像が学習用と検証用にまたがって混在しないよう注意する．主要なハイパーパラメータは, バッチサイズ 32, エポック数 100, 入力画像サイズ 1024 × 1024 ピクセルと設定する．さらに, 学習時のデータ拡張として明度をランダム変化させ, ロバスト性を向上させる.

5.4 推論結果の統合

本節では, **5.3** で学習した電子部品検出モデルおよび端子検出モデルを用いて推論を行い,5-fold 交差検証により得られた複数モデルの推論結果を統合する方法について述べる.

物体検出モデルの推論結果は, 各検出対象の位置を示すバウンディングボックス (以下,bbox), クラスラベル, および信頼度から構成される．交差検証により学習された各モデルは, 学習データの違いに起因して, 同一画像に対して bbox の位置や信頼度が異なる場合がある．これらのばらつきは, 後段で行う電子部品と端子の対応付けや端子位置の行列化に直接影響する.

そこで本研究では,fold 間の推論結果を統合することで検出結果の安定化を図る統合方法として, 複数モデルの検出結果を重なり度に基づいて融合する Weighted Boxes Fusion(WBF) [15] を用いる．本方法は, 信頼度を考慮しつつ bbox を統合できるため, 検出位置のばらつきを抑制できる．なお, 本処理は電子部品検出結果および端子検出結果に対して個別に適用する.

5.5 電子部品検出結果と端子検出結果の対応付け

本節では, **5.4** で統合した電子部品の検出結果 (部品 bbox) と端子の検出結果 (端子 bbox) を用いて, 各端子がどの電子部品に属するかを対応付ける方法について述べる.

単純に部品 bbox と端子 bbox の距離に基づいて対応付けを行う場合, 隣接部品の端子が誤って割り当てられる可能性や, 端子検出の過検出・見逃しに起因する対応付け誤りが生じる可能性がある．そこで本研究では, 部品種別ごとに想定される端子数を基準とし, 基本的には部品 bbox の内部に存在する端子に対して対応付けを行う.

以下では, 対応付けの基本方針を **5.5.1** で述べ, 対応付けに用いるホール間距離の推定方法を **5.5.2** で説明する．続いて,2 端子部品およびそれ以外の電子部品に対する具体的な対応付け手順を **5.5.3** および **5.5.4** で示す.

5.5.1 対応付けの基本方針

電子部品と端子の対応付けには, 部品 bbox および端子 bbox の中心座標を用いる．各電子部品に対して想定される端子数 (期待端子数) に基づき, 部品 bbox 内から端子候補を収集する.

端子候補数と期待端子数を比較し, 補正処理として, 候補が過剰な場合は信頼度に基づく選別を行い, 不足する場合は探索範囲の拡張を行う. それでも, 端子が見つからない場合は, 補完処理として端子位置の推測を行う．また, 同一端子が複数部品に割り当てられることを防ぐため, 一度割り当てた端子は他部品の候補から除外する.

探索範囲の拡張には, ユニバーサル基板のスルーホールが等間隔に配置される点に着目し, ホール間距離を基準として用いる．これにより, 画像ごとのスケール差や回転角の違いに依存しない一貫した基準で対応付けを行う．図 5 に対応付けの全体フローを示す.

5.5.2 ホール間距離の推定

本項では, 端子探索範囲の基準として用いるホール間距離の推定方法について述べる．ユニバーサル基板では, スルーホールが等間隔に配置されており, 画像上にも周期構造が現れる.

本研究では, 林らの方法に基づき, 画像に対して自己相関解析を行うことで周期構造を推定する．具体的には, 入力画像をグレースケール化し, モルフォロジー勾配処理によってスルーホール周辺のエッジ成分を強調する．続いて,2 次元フーリエ変換を

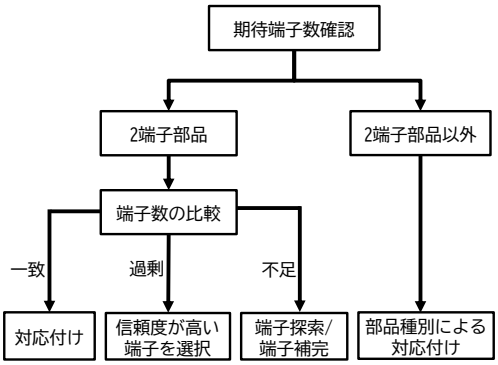


図 5　対応付けの全体フロー

用いた自己相関解析を行い, 得られた自己相関画像を水平方向および垂直方向に平均化する．1 次元波形上のピーク間隔から周期を推定し, 各方向の平均値をホール間距離として採用する．

5.5.3　2 端子部品の対応付け

抵抗器, セラミックコンデンサ, ジャンパ線は 2 端子部品であり, 本項ではこれらに対する対応付け手順を述べる．2 端子部品は, 部品 bbox が過大に周囲を包含しやすく, 隣接部品の端子が混入する可能性がある．

そこで, 部品 bbox のアスペクト比に基づいて処理順を制御し, アスペクト比が 1 から離れた部品 bbox から順に対応付けを行う．これにより, 過大包含の影響を受けやすい電子部品の誤対応付けを抑制する．

端子候補数が期待端子数と一致する場合は, 候補をそのまま採用する．一致しない場合においては, **5.5.1** で述べた補正, 補完処理を適用する．

5.5.4　2 端子部品以外の対応付け

3 端子ピンヘッダおよび 8 端子 IC については, 端子が隠れず, 高精度に検出できると考えられるため, 補完処理は行わない．

ピンヘッダについては, 部品 bbox 内の端子 bbox を候補として収集し, 電子部品の向きに基づいて順序対応付けを行う．IC については, 部品 bbox 内および必要に応じて近傍から端子 bbox を収集する．端子順序の特定には,IC に設けられた切り欠きの位置を用いる．IC 中心から切り欠き方向へのベクトルを基準角とし, 各端子候補へのベクトルの偏角を算出・正規化することで, 端子番号順に対応付けを行う．図 6 に IC 中心から切り欠きを基準とした端子順序特定の例を示す．

5.6　端子位置の行列化と出力

本節では, **5.5** で対応付けた電子部品の端子位置を, ユニバーサル基板の物理構造に基づく行列形式へ変換し,CSV 形式で出力する方法について述べる．

まず, スルーホールの配列に対応する行列の領域を定義する．**5.5.2** と同様に自己相関解析を用いてホール間隔および中心位置を推定し, 隣接ホール間の中点を境界として矩形領域を定義することで, 行列を構築する．

次に, 各端子座標が含まれる矩形領域を特定し, その行および列を端子の行列位置として割り当てる．最後に割り当てた行列

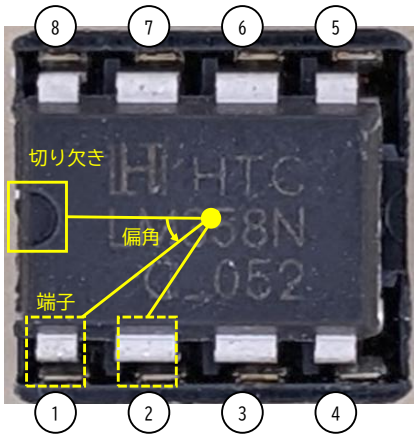


図 6　IC の端子順序特定例

位置に基づき, 各セルに電子部品ラベルおよび端子番号を記載した CSV ファイルとして出力する．これにより, 基板上の電子部品および端子位置を行列形式で表現し, 不良原因推定などの後続処理に利用可能なデータを得る．CSV ファイルの例を図 7 に示す．

```
部品名, 端子番号, 行, 列
抵抗 R1,1,3,18
抵抗 R1,2,7,18
コンデンサ C1,1,4,8
. . . .
. . . .
. . . .
```

図 7　電子部品配置を表す CSV ファイルの例

6　結果および考察

本章では, 深層学習モデルを用いた電子部品および端子の検出精度と, それらの結果を対応付けて行列形式へ変換した最終的な出力結果について述べ, 本研究で提案した方法の有効性を評価する．

6.1　電子部品および端子の出力結果

まず,YOLOv11 を用いた電子部品および端子の物体検出結果について評価する．評価指標として, 物体検出において一般的に用いられる mAP50 および mAP50-95 を採用する．mean Average Precision(mAP) は, 各クラスにおける適合率と再現率から算出される Average Precision(AP) の平均値であり, 検出精度を総合的に評価する指標である．mAP50 は,bbox と正解との Intersection over Union(IoU) が 50%以上の場合に算出される指標であり, 対象物を捉える基本的な検出能力を示す．一方,mAP50-95 は,IoU しきい値を 50%から 95%まで 5%刻みで変化させた際の平均精度であり, 位置特定の厳密さを含めた検出性能を評価する指標である．

電子部品検出の結果,mAP50 は 1.00,mAP50-95 は 0.855 と

なった．電子部品検出の混同行列を表 3 に示す．混同行列のしきい値は 0.5 とした．全ての電子部品が正しく検出されており，誤検出は確認されなかった．

端子検出においては,mAP50 は 0.980,mAP50-95 は 0.618 となった．端子検出の混同行列を表 4 に示す．一部の端子において誤検出が確認された．しかし，本研究の目的は物体検出精度そのものを最大化することではなく，検出結果を基に正確な接続情報を行列形式として出力することである．検出段階における誤検出や未検出については，後段の電子部品と端子の対応付け処理によって補正可能であり，その影響を含めた最終的な出力結果を次節で評価する．

表 3　電子部品検出の混同行列

		正解						
		セラミックコンデンサ	IC	IC の切り欠き	ジャンパ線	ピンヘッダ	抵抗器	背景
予測	セラミックコンデンサ	**36**	0	0	0	0	0	0
	IC	0	**12**	0	0	0	0	0
	IC の切り欠き	0	0	**12**	0	0	0	0
	ジャンパ線	0	0	0	**14**	0	0	0
	ピンヘッダ	0	0	0	0	**12**	0	0
	抵抗器	0	0	0	0	0	**36**	0
	背景	0	0	0	0	0	0	0

表 4　端子検出の混同行列

		正解				
		隠れ端子	IC の端子	端子	ピンヘッダの端子	背景
予測	隠れ端子	**34**	0	0	0	**2**
	IC の端子	0	**96**	0	0	0
	端子	0	0	**134**	0	**6**
	ピンヘッダの端子	0	0	0	**35**	**1**
	背景	**3**	0	**1**	**1**	0

6.2　電子部品と端子の対応付け後の行列出力結果

本節では，電子部品および端子の検出結果を基に対応付けを行い，最終的に得られた行列形式の出力結果について評価する．評価には，テスト用基板 12 枚 (1 枚あたり 15 行× 25 列，計 375 個のスルーホール) を用い，合計 4,500 個のスルーホールに対して，その実装状態の判定が正解と一致しているかを検証した．評価指標として，スルーホールの総数に対する正解判定数を示した．結果を表 5 に，出力例を図 8 に示す．

その結果，全てのテスト基板において，実装された電子部品の種別および接続端子位置を誤りなく行列形式へ変換できた．検出段階で端子位置にわずかな誤差が含まれていたものの，電子部品と端子の対応付けにおける補正，補完処理によって正しい接続情報を抽出できることが確認された．

表 5　評 価 結 果

端子種別	評価対象数	正解判定数
端子無し	4,196	4,196
抵抗器	72	72
コンデンサ	72	72
IC	96	96
ピンヘッダ	36	36
ジャンパ線	28	28
合計	4,500	4,500

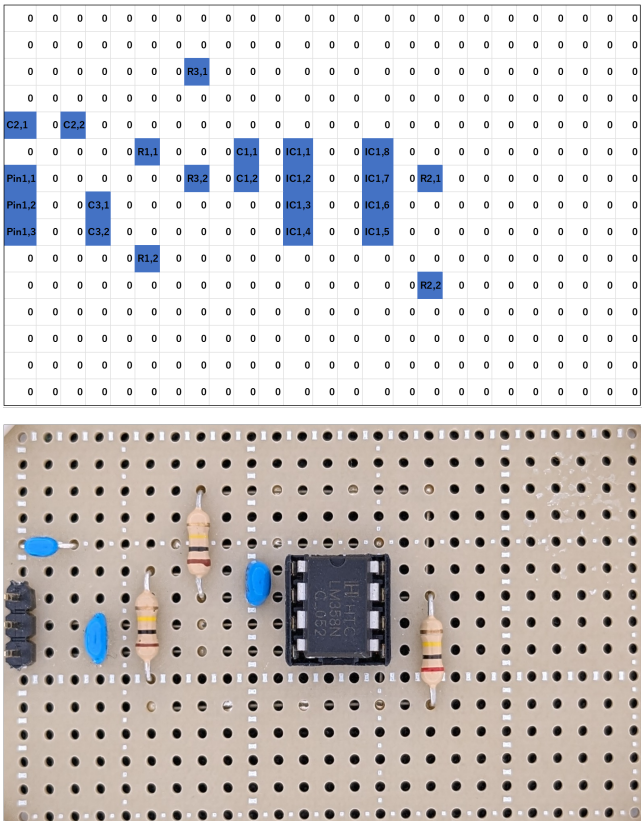


図 8　行列形式での出力結果例 (上: 行列出力結果, 下: 元画像)

6.3　考　　察

本研究では，深層学習に基づく物体検出と，ユニバーサル基板の行列形式を活用した後処理を組み合わせることで，電子部品および端子位置を高精度に検出し，接続関係を行列形式で出力する方法を実現した．これにより，回路図復元や不良原因推定に必要な基礎情報を取得できることを示した．

一方で，検討すべき課題も残されている．まず，物体検出モデル自体の精度向上に関する課題である．本研究では彩度や明度の変化に対するロバスト性を検証した．しかし，検出精度を最適化するためのパラメータ探索を十分に行ったとは言えない．今後は，前処理方法や学習条件を含め，より多彩な設定を検討する必要がある．また，本研究の対象以外の電子部品に対する端子検出についても，拡張の余地がある．

次に，本方法の限界について述べる．本研究では行列出力において誤りのない出力を確認した．しかし，これは前段の物体

検出において高い精度が得られたことに依存している．物体検出の段階で電子部品の検出漏れや誤検出が発生した場合, 対応付けにおける補正, 補完処理が困難となり, 行列出力の誤りにつながる可能性がある．そのため, 検出モデルのさらなるロバスト化が重要な課題である．隠れた端子の検出では, 電子部品の直下に存在している端子のみを検出可能としている．また, 補正処理において, 候補端子数が期待端子数を上回る場合の処理は, 正解の端子は誤って検出した端子よりも高い信頼度を持つと仮定して選別していることが制約条件として挙げられる．

さらに, 本研究では電子部品の種別と端子位置の特定に注力しており, 抵抗値や静電容量といった電子部品の具体的な値の判別は対象外としている．これは, 基板を真上から撮影した画像のみでは, 電子部品側面に記載された情報を正確に取得することが困難であるためである．今後の発展として, あらかじめ用意した部品表と検出結果を照合することで, 部品値の特定や実装漏れ・過剰実装の検出を行う仕組みを組み込むことが考えられる．このような機能が実現すれば, 教育現場における回路製作支援のさらなる高度化が期待される．

7 おわりに

本研究では, 電子回路製作の教育現場における確認作業の負担軽減を目的として, 先行研究において未検討であった電子部品面の情報取得を自動化し, 不良原因推定を支援するための方法を提案した．深層学習による物体検出とユニバーサル基板の構造的特徴を活用した座標補正を組み合わせることで, 基板上の電子部品の種別および各端子の接続先を行列形式として出力できることを確認した．

これにより, 先行研究の方法と統合することで, 指導者の目視確認に依存せずに, ユニバーサル基板上の電子回路における不良原因推定の自動化が可能となる見込みを得た．今後は, 本研究の対象外であった部品値の識別方法の導入に向け, あらかじめ用意された部品表と基板上の電子部品情報を照合する機能の拡張や, 対象とする基板枚数の増加, ならびに多様な部品種別への対応を進める．

謝　辞

本研究は科研費 JP24K15243 の助成を受けたものです．ここに記して深謝いたします．本研究を進めるにあたり, 北陸職業能力開発大学校より研究対象となる電子回路試作基板をご提供いただきましたことを厚く御礼申し上げます．

文　献

[1] Takemura, A. E-Learning System for Electronic Circuit Construction Using Handwriting Recognition and Mixed Reality Techniques. International Association for Development of the Information Society, 2018.
[2] 薮哲郎,「オンラインによる電気実験実習の実施」次世代教員養成センター研究紀要 = Bulletin of Teacher Education Center for the Future Generation, 7, 79-90, 2021.
[3] 斎藤正義,「回路基板の不良箇所を特定する技能を向上するための教材開発とその効果」, PTU フォーラム 2023, 第 31 回職業能力開発研究発表講演会論文集, 24-A-3, 5-6, 2023.
[4] Drew, D., Newcomb, J. L., McGrath, W., Maksimovic, F., Mellis, D., & Hartmann, B. The toastboard: Ubiquitous instrumentation and automated checking of breadboarded circuits. Proceedings of the 29th Annual Symposium on User Interface Software and Technology, pp. 677-686, 2016.
[5] 林哲矢, 田中剛, 遠藤雅樹, 寺田憲司, 大野成義.「ユニバーサル基板の配線画像解析による回路図復元」IEICE Conferences Archives, 2025.
[6] Hayashi, T., Tanaka, T., Endo, M., Terada, K., & Ohno, S. Initial Methodology for Estimating the Cause of Operating Defects by Schematic Restoration from Electronic Circuit Board Images. ICIM2025: 2025 International Conference on Information Management, 2025.
[7] 林哲矢, 田中剛, 遠藤雅樹, 寺田憲司, 大野成義,「電子回路試作基板の自動回路構成比較による動作不良原因の推定」電子情報通信学会研究会発表講演論文, R2025-52, vol.125, no.268, pp.19-24, 2025.
[8] 茂木友哉, 滑川光裕, 植田佳典.「電子基板の部品挿入もれ判定における基礎研究」第 75 回全国大会講演論文集, 597-598, 2013.
[9] Chiun, O. Y., & Ruhaiyem, N. I. R. Object detection based automated optical inspection of printed circuit board assembly using deep learning. International Conference on Soft Computing in Data Science, pp. 246-258, Springer Nature Singapore, 2023.
[10] Kim, P., Huang, X., & Fang, Z. SSD PCB Component Detection Using YOLOv5 Model. Journal of Information & Communication Convergence Engineering, 21(1), 2023.
[11] Takemura, A. Education System for Electronic Circuit Construction Involving Soldering on a Circuit Board. International Association for Development of the Information Society, 2019.
[12] 三谷芳弘, 杉村佑貴, 浜本義彦.「抵抗器読み取りのための色特徴に関する一検討」IEICE Conferences Archives, 2007.
[13] 鹿間信介, 杉原弘記.「画像認識による抵抗器読み取り手法に関する研究―カラーコード画像に重畳する鏡面ハイライトの回避の試み」電気学会論文誌 C（電子・情報・システム部門誌）, 136(11), 1532-1540, 2016.
[14] Khanam, R., & Hussain, M. Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725, 2024.
[15] Solovyev, R., Wang, W., & Gabruseva, T. Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing, 107, 104117, 2021.