人とロボットのインタラクションにおけるマルチモーダル情報を用いた異常検出

Anomaly Detection in Human-Robot Interaction using Multimodal Information

望月翔太[1*]　山下紗苗[1]　湯浅令子[2]
窪田智徳[3]　小川浩平[3]　東中竜一郎[1]
Shota Mochizuki[1], Sanae Yamashita[1], Reiko Yuasa[2]
Tomonori Kubota[3], Kohei Ogawa[3], and Ryuichiro Higashinaka[1]

[1] 名古屋大学大学院情報学研究科
[1] Graduate School of Informatics, Nagoya University
[2] 名古屋大学情報学部
[2] School of Informatics, Nagoya University
[3] 名古屋大学大学院工学研究科
[3] Graduate School of Engineering, Nagoya University

Abstract: In this study, to detect anomalies in human-robot interactions, we created a dataset and constructed anomaly detection models. We created the dataset by collecting videos of human-robot interactions in a framework where humans intervene when a dialogue breakdown occurs, and labeled the scenes where humans intervened as anomalies. Additionally, we constructed anomaly detection models by fine-tuning existing video and audio classification models. Beyond the conventional classification approachs, we applied deep metric learning to fine-tuning methods and evaluated its effectiveness.

1　はじめに

対話システムの普及に伴い，音声対話が可能な対話ロボットが，雑談や実店舗での商品販促など，様々に活用されている [1, 2]．しかし，対話ロボットのインタラクション性能は未だ完全ではなく，しばしば，対話破綻 [3]（円滑な対話継続が困難な状態）を起こす．

そうした中で，人間と対話システムが協力することで，効率的に対話することを目指す取り組みとして，複数人同時対話の枠組みが提案されている [13, 17]．図 1 に，複数人同時対話の枠組みを表した模式図を示す．この枠組みでは，人間がオペレータとして複数の対話システムの対話を監視し，問題が生じた際にのみ対話に介入することで，同時に複数人と対話する．先行研究において我々は，大阪府に位置する施設「ニフレル」に，大規模言語モデルである GPT-3 [12] に基づいて案内対話を行う自律対話ロボットを配置し，複数人同時対話の枠組みを検証するフィールド実験を，2023 年 2 月 6 日から 3 月 3 日までの 26 日間実施した [4]．その結果，大規模言語モデルを用いることで高度な案内が可能である一方，対話者が話しかけていることの認識の失敗や音声認識誤りなどの要因により，対話者の発話に反応しない，適切な応答を返せないなどの問題が生じるという課題が明らかとなった．我々は，より効率的な複数人同時対話の実現のためには，このような問題のあるインタラクションを検出し，問題が発生していることを介入を行うオペレータに提示することで，オペレータの介入判断を補助することが有効だと考えている．

本研究では，フィールド実験でのインタラクション中に生じていた問題を検出することを目的として，フィールド実験で収集したインタラクション映像を用いた異常検出データセットの作成，および，マルチモーダル情報を用いた異常検出モデルの構築に取り組む．本論文では，まず，異常検出データセットの作成方法について述べる．続いて，作成したデータセットを用いて異常検出モデルを構築する手法と，構築したモデルの評価結果について報告する．最後に，評価結果を踏ま

*連絡先：名古屋大学大学院情報学研究科
〒 464-8601 名古屋市千種区不老町
E-mail: mochizuki.shota.k8@s.mail.nagoya-u.ac.jp

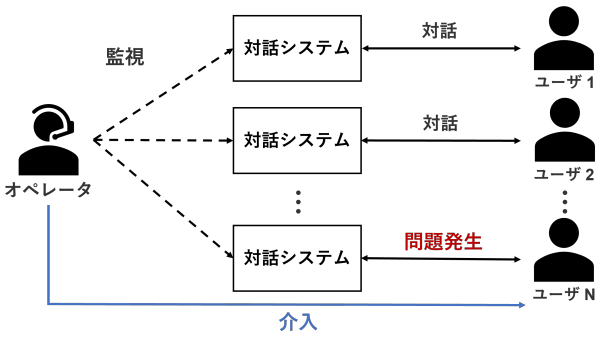


図 1: 複数人同時対話の枠組み．ユーザは対話システムと対話し，オペレータは対話の様子を監視する．対話に問題が発生した場合，オペレータが対話に介入し，対話を継続させる．

えて，本研究で作成したデータセットの性質や今後の課題について考察する．

2　異常検出データセットの作成

異常検出データセットは，フィールド実験で収集された映像から正例・負例を抽出することで作成した．

2.1　収集されたインタラクション映像

ニフレルでのフィールド実験は，来館したユーザと対話する自律対話ロボットを施設内に 6 台配置し，対話に問題が発生した際に 2 人のオペレータが介入するという設定で実施し，実験を通して，ユーザによる発話が 7 万回以上行われ，オペレータによる介入は 2,160 回行われた（詳細は [4] を参照のこと）．

実験期間中，対話ロボットの背面に設置された広角のカメラとロボット前面に設置されたマイクにより，人とロボットのインタラクションの様子を記録した．図 2 に，収集された映像の例を示す．

6 台のロボットを合計して，インタラクションが生じていない箇所も含め約 1,300 時間分の映像が収集されている．本映像を用いて，異常検出データセットを作成する．

2.2　正例・負例の抽出

収集された映像から，問題が生じているインタラクションを収めた正例と，正常なインタラクションを収めた負例の抽出を行った．フィールド実験では，ユーザとロボットのインタラクションに問題が生じた際，オペレータによる介入が行われた．そのため，介入が行われる直前のインタラクションには問題が生じていたと考えられる．したがって，介入が行われる直前の映像を正例，ユーザが発話しているがオペレータによる介入が行われていない時間の映像を負例として抽出した．正例，負例の詳細な抽出方法は以下の通りである．

正例 オペレータによる介入が行われる直前の 10 秒間の映像．

負例 オペレータによる介入が行われておらず，かつ，ロボットに対してユーザが発話している時間から抽出した 10 秒間の映像．ユーザが発話していることを条件とした理由は，インタラクションが生じていない映像が負例に含まれることで，人や音声の有無により正例と負例の識別が容易になるのを防ぐためである．

介入に基づく抽出により，1,943 件の正例を獲得した．正例の抽出は，介入が生じた時刻に基づき自動的に行った．正例の件数が実験で生じた介入回数よりも少ない理由は，カメラやシステムの不具合により介入直前の様子を録画できていない場合や介入が行われた時刻を正確に断定できない場合など，自動的な抽出が難しい例が含まれていないためである．負例についてもランダムに同数抽出することで，インタラクションの様子を収めた映像 3,886 件からなる異常検出データセットを作成した．

図 2: 収集された映像の例．左から右に時間が流れている．ユーザがロボットと対話する様子が収められている．

3　異常検出モデルの構築

作成した異常検出データセットを用いて既存の分類モデルをファインチューニングすることで，異常検出モデルを構築した．本実験で扱うモデルの詳細，および，異常検出の結果について述べる．

3.1　モデル

本実験では，扱うモダリティの異なる 3 つのエンコーダを用いて，2 クラス分類モデル，および，深層距離学習 [14] モデルを構築した．

3.1.1　エンコーダ

動画エンコーダ 動画（複数のフレーム）を入力とするエンコーダとして，Transformer ベースの動画分類モデルである VideoMAE [5] を用いる．本実験では，動画分類用のデータセットである Kinetics-400 [6] で事前学習済みのモデル[1]を使用する．本モデルは，224 × 224 の 16 枚のフレームを入力とするため，10 秒間の映像から等間隔に 16 枚のフレームを抽出して入力する．

音声エンコーダ 音声を入力とするエンコーダとして，畳み込みニューラルネットワークベースの音声分類モデルである VGGish [7] を用いる．本実験では，音声分類用のデータセットである YouTube-8M [8] で事前学習済みのモデル[2]を使用する．

マルチモーダルエンコーダ 我々は，上述の動画エンコーダと音声エンコーダを組み合わせることで，動画と音声の両方を入力とするエンコーダを実装した．本エンコーダでは，異なるモダリティから得られた特徴に Cross Attention を取る手法である Cross-Modal Attention [11] を用いる．図 3 に本エンコーダのアーキテクチャを示す．VideoMAE が出力した動画特徴，および，VGGish が出力した音声特徴について，それぞれ一方の特徴を Query，他方の特徴を Key, Value とした Multi-Head Attention を適用する Cross-Modal Attention を取る．その後，Attention からの出力を結合することで動画と音声を組み合わせた特徴を獲得する．

学習時には，Vision and Language Model である LLaVA [15] の学習手法を参考に，2 段階のプロセスで学習を行う．LLaVA では，まず，Vision Encoder と LLM のパラメータをフリーズして，中間の Projector のパラメータのみを更新することで，事前学習済みの LLM への入力として適した特徴への変換を学習させる．その後，End-to-End なファインチューニングを実施している．本研究では，まず，VideoMAE および VGGish をユニモーダルで学習したパラメータでフリーズし，Multi-Head Attention のパラメータを更新することで，2 つのエンコーダの出力を組み合わせてクラス分類用に変換することを学習させる．その後，VideoMAE, VGGish のパラメータも含めてモデル全体のパラメータを更新する．

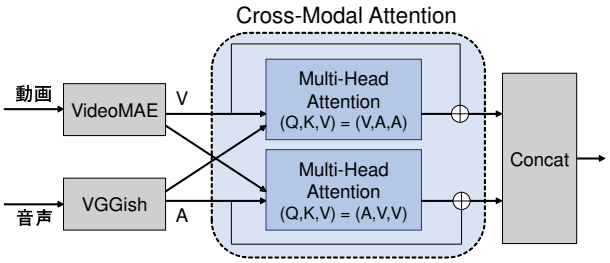


図 3: マルチモーダルエンコーダのアーキテクチャ．Cross-Modal Attention の実装は，Cheng らの研究 [11] を参考にした．

3.1.2　モデルの構築

異常検出データセットを用いて分類モデルや深層距離学習モデルを学習することで，異常検出モデルを構築した．データセットは，訓練データ 70%, 検証データ 15%, テストデータ 15%にランダムに分割し，モデルの学習および評価に用いた．分割時には，各データで 6 台のロボットに関する動画の割合に偏りが生じないようにした．

本実験では，各エンコーダごとに，以下の 2 つのモデルを構築した．

クラス分類モデル 最終層のユニット数を 2 としたニューラルネットワークを各エンコーダの末尾に接続し，softmax 関数の出力に基づいて 2 クラス分類を行う．損失関数として Cross-entropy Loss を用い，モデルのパラメータを更新する．

深層距離学習モデル 深層距離学習とは，異なるクラスに属するデータから得られた埋め込み間の距離は遠く，同一クラスに属するデータから得られた埋め込み間の距離は近くなるように学習する手法である [14]．本実験では，最終層のユニット数を 512 としたニューラルネットワークを各エンコーダの末尾に接続することで，入力データに対し

[1] https://huggingface.co/MCG-NJU/videomae-base-finetuned-kinetics

[2] https://github.com/tensorflow/models/tree/master/research/audioset/vggish

表 1: テストデータに対する評価結果．深層距離学習を用いたモデルは，k 近傍法によるクラス分類における k の値を括弧内に示している．各尺度ごとに最も高いスコアを太字で，2 番目に高いスコアを下線で表す．* は，マクネマー検定において，同一のモダリティを入力とするクラス分類モデルと比較して $p < 0.05$ で有意差が認められたことを示す．

	モデル	Accuracy	Precision	Recall	F1-score
クラス分類	動画	0.636	0.657	0.610	0.626
	音声	0.783	0.765	**0.829**	0.793
	マルチモーダル	0.789	0.777	<u>0.819</u>	0.796
深層距離学習	動画 ($k = 1$)	0.630	0.641	0.607	0.623
	動画 ($k = 5$)	0.646	0.657	0.628	0.641
	動画 ($k = 10$)	0.655	0.655	0.676	0.663
	動画 ($k = 30$)	0.662	0.669	0.664	0.664
	動画 ($k = 50$)	0.664	0.674	0.655	0.662
	音声 ($k = 1$)	0.753	0.765	0.737	0.750
	音声 ($k = 5$)	0.806	0.840	0.763	0.799
	音声 ($k = 10$)	0.817	0.835	0.797	0.815
	音声 ($k = 30$)	0.823	0.862	0.777	<u>0.816</u>
	音声 ($k = 50$)	**0.830***	**0.878**	0.773	**0.821**
	マルチモーダル ($k = 1$)	0.784	0.783	0.791	0.787
	マルチモーダル ($k = 5$)	0.810	0.824	0.794	0.808
	マルチモーダル ($k = 10$)	0.819	0.839	0.795	<u>0.816</u>
	マルチモーダル ($k = 30$)	0.822	0.857	0.778	0.815
	マルチモーダル ($k = 50$)	<u>0.824</u>*	<u>0.863</u>	0.775	<u>0.816</u>

て 512 次元の埋め込みを獲得する．そして，深層距離学習で用いられる損失関数の 1 つである ArcFace [9] に基づきパラメータを更新することで，獲得される埋め込みのクラス間距離を離すように学習する．

深層距離学習モデルで分類を行う際には，訓練データの埋め込みを用いた k 近傍法によりクラスを決定する．すなわち，入力されたデータから獲得した埋め込みと訓練データの埋め込みとのユークリッド距離を計算し，距離の近い訓練データ k 件のラベルの多数決により正例か負例か判定する．本実験では，$k = 1, 5, 10, 30, 50$ の 5 つの値で分類を行った．

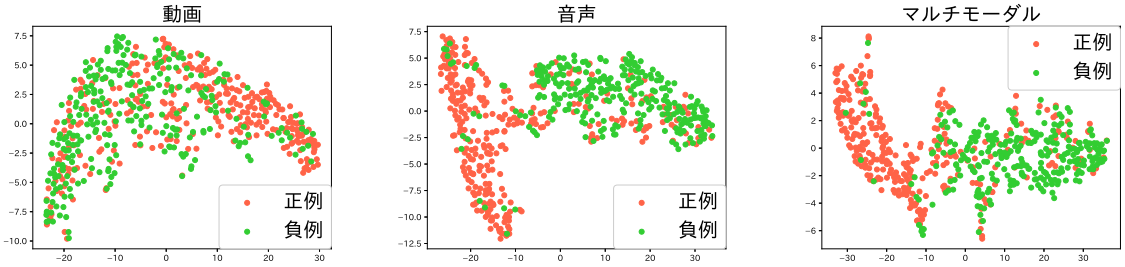
3.2 評価結果

表 1 に，各モデルの評価結果を示す．評価には，クラス分類モデルの標準的な評価尺度である Accuracy, Precision, Recall, F1-score の 4 つを用いた．いずれのモデルにおいても，Accuracy の値がチャンスレートの 0.5 を上回っており，動画・音声分類モデルをファインチューニングすることで，用いることでインタラクション中に問題が生じているか否かを識別可能であることが確かめられた．

扱うモダリティの比較では，動画モデルに比べて，音声モデルおよびマルチモーダルモデルのスコアが高く，本データセットにおける分類には音声情報が有用であることが明らかになった．これは，正例の映像には，ロボットがユーザの発話に応答できていない事例が多く含まれるためであると考える．図 4 に示すような，ロボットが応答できていない事例では，ロボットの音声が映像に含まれないのに対して，正常なインタラクションを収めた負例の映像の多くは，ロボットの音声が含まれているため，音声情報に正例と負例の差分が生じやすかったと考えられる．

クラス分類モデルと深層距離学習による埋め込みモデルとの比較では，どのモダリティを用いた場合も，深層距離学習モデルが，クラス分類モデルよりも高い Accuracy, F1-score を達成した．深層距離学習モデル ($k = 50$) とクラス分類モデルの分類結果に対してマクネマー検定を実施したところ，音声モデルとマルチモーダルモデルでは，有意差が認められた ($p < 0.05$)．図 5 に，深層距離学習を適用したモデルで獲得されたテストデータの埋め込みについて，次元削減アルゴリズムにである t-SNE [10] により 2 次元に削減してプロットした結果を示す．スコアの高い音声モデルおよびマルチモーダルモデルでは，正例と負例をよく分離できており，動画モデルにおいても，正例と負例が概ね離

図 4: 正例に多く含まれる，ロボットが応答できていない状況の例．



(a) 動画モデル　(b) 音声モデル　(c) マルチモーダルモデル

図 5: 深層距離学習を適用した各埋め込みモデルにおけるテストデータのプロット．赤色の点が正例（問題が生じているインタラクション），緑色の点が負例（正常なインタラクション）に対応している．

れるように分布している．このことから，深層距離学習により，クラス間の距離を離す埋め込みの学習が可能であることが確かめられた．k の値を変化させた場合の比較では，いずれのモデルにおいても，$k = 1$ の場合のスコアが最も低く，k を増加させるにつれてスコアが向上する傾向が見られた．

3.3 考察

動画と音声の両方を入力とするマルチモーダルモデルは，3 つのモデルの中で利用できる情報が最も多いことから，最も高いスコアを示すと思われたが，音声モデルと同程度の結果となった．これは，先に述べたように，正例の映像にはロボットが応答を返していない状況が多く含まれることから，本データセットにおける異常検出には音声情報が有用であったためであると考える．現状の対話ロボットでは，音声のインタラクションが基本になっているため，そのチャネルが途切れることがインタラクション中に生じた問題を検出する上で最もクリティカルな情報であったと考えられる．

また，動画モデルが音声モデルを下回った原因として，動画モデルに用いた VideoMAE が扱うフレームのサイズが 224×224 と小さく，インタラクションに問題が生じた際に対話者が示す動きや反応を捉えることが困難であったことが考えられる．本研究では，学習に

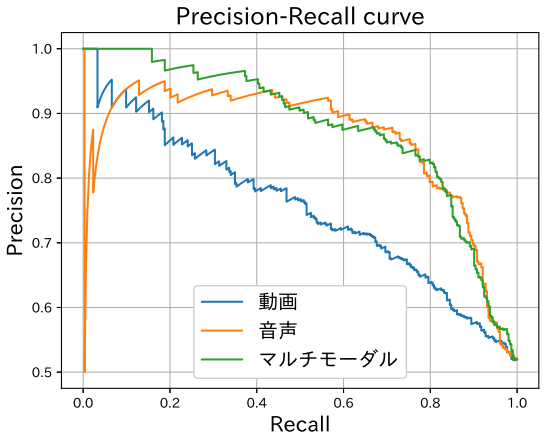


図 6: クラス分類モデルの Precision-Recall 曲線．

要する時間の観点から本モデルを採用したが，より大きなサイズのフレームを扱うことができる動画分類モデルが提案されているため，それらを利用することで性能の向上が期待できる．また，Xu ら [16] が，姿勢推定モデルの構築において，画像中から，人物が写っている領域を切り取ってモデルへと入力する手法を提案しているように，対話者の顔の周辺など，正例と負例とで動画特徴に差分が大きく現れると考えられる箇所を切り取って入力するアプローチも有効だと考える．

図 7: マルチモーダルモデルが高い確率で正例と予測した例.

図 6 に，クラス分類モデルの Precision-Recall 曲線を示す．評価結果では，マルチモーダルモデルは音声モデルと同程度のスコアであったが，Precision の高い箇所では，マルチモーダルモデルの方が良い性能を示している．図 7 に，マルチモーダルモデルが高い確率で正例と予測した例を示す．ユーザがロボットの陰に隠れていることで，ユーザの姿の認識ができておらず，そのためユーザの発話に応答できていないという，本データセットの正例に多く含まれる事例の 1 つである．この例では，ユーザがロボットの陰に隠れているという動画情報とロボットが応答していないという音声情報の両方が，異常検出に有用であるため，両方の情報を活用できるマルチモーダルモデルは 0.969 という高い確率で正例と予測できている．この例のように，動画と音声の両方の情報が活用でき，問題が発生していると高い確率で予測できる例については，マルチモーダルモデルは高い精度での検出が可能であると考えられる．

また，本研究で構築した異常検出モデルを複数人同時対話の枠組みに導入し，オペレータの介入判断を補助をする技術として活用するためには，インタラクション中に生じた問題を漏れなく発見できる Recall の高さがより重要になると考える．本研究で高いスコアを達成した音声モデルおよびマルチモーダルモデルにおいても，Recall が 90%ほどになるまでしきい値を下げると，Precision は 65%程度まで低下する．オペレータが並列して複数の対話を監視する状況では，モデルが問題発生と判断した際に，実際に介入すべきかをオペレータが判別する余裕が無い場合も想定される．したがって，並列する対話数が多い場合，Precision の高さも重要となるため，実用にはさらなる性能向上が必要である．

4　おわりに

本研究では，我々が実施したフィールド実験において，人とロボットのインタラクション中に生じていた問題をマルチモーダル情報を用いて検出することを目的として，異常検出データセットの作成および異常検出モデルの構築を行った．人とロボットのインタラクションの様子を収めた 10 秒間の映像 3,886 件からなるデータセットを構築し，本データセットを用いて既存の動画・音声分類モデルをファインチューニングすることで，80%の Accuracy でインタラクション中に生じた問題を検出できることを確認した．また，フィールド実験でのインタラクション中に生じていた問題を検出するためには，音声情報が有用であることが示された．さらに，深層距離学習により埋め込みを獲得し，k 近傍法により分類するアプローチにより，softmax によるクラス分類よりも有意に高い Accuracy を達成できることが実証された．

今後は，より高性能なエンコーダの活用や，入力するデータの工夫により，異常検出性能の向上を試みたい．また，本研究で構築したモデルは入力されたインタラクションにおいて問題が生じているかを検出するものであるが，岡留ら [18] のように，生成モデルを用いて未来のインタラクションを生成することで，問題が生じそうかを予測することも検討している．さらに，複数人同時対話の枠組みにおいて異常検出モデルの検出結果をオペレータに提示するためのインタフェースの実装にも取り組んでいきたい．

謝辞

本研究は，JST ムーンショット型研究開発事業，JPMJMS2011 の支援を受けたものです．また，フィールド実験の実施にご協力いただいたニフレルのスタッフの皆様に感謝の意を表します．

参考文献

[1] Tatsuya Kawahara.: Spoken Dialogue System for a Human-like Conversational Robot ERICA, *In Proceedings of the 9th International Workshop on Spoken Dialogue Systems Technology*, pp. 65–75 (2018)

[2] Takuya Iwamoto, Jun Baba, Kotaro Nishi, Taishi Unokuchi, Daisuke Endo, Junya Nakanishi, Yuichiro Yoshikawa, Hiroshi Ishiguro.: The Effectiveness of Self-Recommending Agents in Advancing Purchase Behavior Steps in Retail Marketing, *In Proceedings of the 9th International Conference on Human-Agent Interaction*, pp. 209–217 (2021)

[3] Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, Masahiro Mizukami.: Integrated taxonomy of errors in chat-oriented dialogue systems, *In Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 89–98 (2021)

[4] Shota Mochizuki, Sanae Yamashita, Kazuyoshi Kawasaki, Yuasa Reiko, Tomonori Kubota, Kohei Ogawa, Jun Baba, Ryuichiro Higashinaka.: Investigating the Intervention in Parallel Conversations, *In Proceedings of the 11th International Conference on Human-Agent Interaction*, pp. 30–38 (2023)

[5] Zhan Tong, Yibing Song, Jue Wang, Limin Wang.: VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training, *In Proceedings of the Advances in Neural Information Processing Systems*, Vol. 35, pp. 10078–10093 (2022)

[6] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, Andrew Zisserman.: The Kinetics Human Action Video Dataset, *arXiv preprint arXiv:1705.06950*, (2017)

[7] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, Kevin Wilson.: CNN Architectures for Large-Scale Audio Classification, *In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 131–135 (2017)

[8] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, Sudheendra Vijayanarasimhan.: YouTube-8M: A Large-Scale Video Classification Benchmark, *arXiv preprint arXiv:1609.08675*, (2016)

[9] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, Stefanos Zafeiriou.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition, *In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4685–4694 (2019)

[10] Laurens van der Maaten and Geoffrey E. Hinton.: Visualizing data using t-SNE, *Journal of Machine Learning Research*, Vol. 9, No. 11, pp. 2579–2605 (2008)

[11] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, Yuejie Zhang.: Look, Listen, and Attend: Co-Attention Network for Self-Supervised Audio-Visual Representation Learning, *In Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3884–3892 (2020)

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei.: Language Models are Few-Shot Learners, *In Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 1877–1901 (2020)

[13] Dylan F Glas, Takayuki Kanda, Hiroshi Ishiguro, Norihiro Hagita.: Teleoperation of Multiple Social Robots, *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, Vol. 42, No. 3, pp. 530–644 (2012)

[14] Mahmut Kaya, Hasan S. Bilge.: Deep Metric Learning: A Survey, *Symmetry*, Vol. 11, pp. 1066 (2019)

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee.: Visual Instruction Tuning, *arXiv preprint arXiv:2304.08485*, (2023)

[16] Yufei Xu, Jing Zhang, Qiming Zhang, Dacheng Tao.: ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation, *In Proceedings of the Advances in Neural Information Processing Systems*, Vol. 35, pp. 38571–38584 (2022)

[17] Tatsuya Kawahara, Naoyuki Muramatsu, Kenta Yamamoto, Divesh Lala, Koji Inoue.: Semi-autonomous avatar enabling unconstrained parallel conversations –seamless hybrid of WOZ and autonomous dialogue systems–, *Advanced Robotics*, Vol. 35, No. 11, pp. 657–663 (2021)

[18] 岡留有哉, 中村泰.: MASK を用いた拡散確率モデルに基づく二者対話における振る舞い生成, 情報処理学会研究報告, Vol. 2023-MPS-143, No. 64, pp. 1–8 (2023)