

特徴タグ分析を用いたアニメーションに関する メタデータ作成の提案

Metadata Building Method for Anime Clip Scenes using Feature Tag Analysis

Shan Junjie^{1*} 石井 智也¹ 安尾 萌² 西原 陽子¹
Junjie Shan¹ Tomoya Ishii¹ Megumi Yasuo² Yoko Nishihara¹

¹ 立命館大学情報理工学部

¹ College of Information Science and Engineering, Ritsumeikan University

² 立命館グローバル・イノベーション研究機構

² Ritsumeikan Global Innovation Research Organization, Ritsumeikan University

Abstract: 近年、大規模言語モデル (LLM) の発展に伴い、画像や動画の内容に対し自然言語により概要を生成し、それを用いた検索を行う手法が流行している。しかし、自然言語による記述は、内容と形式の一貫性を保つことが困難であり、LLM によって生成された概要記述で画像や動画の統一的な分析・検索を行うのは困難である。本研究では、アニメーションに対する特徴タグ分析を行い、メタデータを構築する手法を提案する。分割されたアニメのカットシーンに対し、フレーム単位ごとに特徴タグを抽出することにより、統一的なタグ記述のあるアニメーションのデータセットを構築した。本論文では、アニメ動画の入力からカットシーンの特徴タグを生成するまでの処理の流れを説明し、生成された特徴タグを用いて異なるジャンルのアニメーションを分析した。

1 はじめに

近年、大規模言語モデルの発展により、ビデオの検索に注目が寄せられている。LLM を使用してビデオコンテンツの説明文 (Caption) を生成し、ユーザの検索クエリとのテキスト類似度を計算するのは一般的なアプローチである [1-3]。また、マルチモーダル LLM の発展により、直接特徴画像の入力からビデオのコンテンツから該当するシーンの特定も可能になった [4]。

しかし、LLM により生成された自然言語の説明概要は、自然言語自身の柔軟性により、用語、文法、形式などで記述的な統一性を保つことは困難である。例えば同じ入力ビデオやプロンプトであっても、モデル推論上の影響により、出力される記述文が変わる可能性も残っている。そのため、LLM が生成した記述文を用いて、ビデオコンテンツの統一的な分析や柔軟なインデックスを作成するためのメタデータの構築が難しくなる。

この問題に対し、本研究では、入力画像の内容的な特徴を分析し、タグとして出力するモデルを利用することで、ビデオコンテンツのシーン画面に対するメタデータの構築手法を提案する。提案手法では、世界中

で人気がある日本のアニメ動画をビデオコンテンツとして、アニメーションのメタデータの構築を実装した。

本研究では、主に以下の三つの内容を達成した：

1. 入力されたアニメ動画から、自動的に各カットシーンの画面内容に対する特徴タグのメタデータを作成するシステムを実装した。
2. 異なるジャンルのアニメ動画を用意し、10 万件以上のカットシーンを含む特徴タグのデータセットを作成した。
3. 作成されたアニメーションのタグデータセットを用いて、ジャンル別に内容の特徴の分析を行った。

2 関連研究

ビデオや動画に特徴タグを付けて、内容分析や検索、推薦を行うシステムに関する課題は様々存在する。中村らの研究では、ニュースの動画シーンに事前に手動でタグ付けが行われたデータを分析し、シーンを選択する手法を提案した [5]。上畑らの研究では、ニコニコ動画に付けられた動画のジャンル等のタグを利用し、タグ

*連絡先：立命館大学情報理工学部
〒567-8570 大阪府茨木市岩倉町 2 - 150
E-mail: shan@fc.ritsumei.ac.jp

数やタグの種類を用いて、動画の多様性を分析した [6]. 村上らの研究でも、ニコニコ動画に関するタグの研究を行っており、視聴者が付与したタグを階層化し、タグ階層化データから関連動画を提示する応用の可能性を示した [7]. 関連動画を提示することで検索への応用も期待される.

これらの研究により、特徴タグがビデオの内容の一つの要約として、幅広い場面での応用可能性が示された. しかし、既存する特徴タグは主に沢山のユーザや専門家から手動で付けられ、ビデオ内容の広いジャンルや全体的な説明までしか反映できないものである. 本研究では、入力画像の内容を表す特徴タグを推定するモデルを利用し、自動的にアニメビデオの入力から各カットシーンの特徴タグを、より細部な画面内容を反映できるタグデータを作成するシステムを提案した.

一方、アニメ内容に関する特徴抽出やメタデータ作成の課題も多数存在する. 高山ら [8], 山田ら [9], 鈴木ら [10] はアニメのキャラクタを対象にして、色や位置などの視覚情報を利用する特徴抽出や類似検索の課題を行った. 高山らは肌の色やあごの形状などの特徴を用いて、アニメキャラクタの顔の特徴データセットを構築しており、山田らはグレースケール画像向きの顔の特徴抽出手法を提案した. 鈴木らはアニメキャラクタの顔パーツ検出手法に用いる学習データセットの作成案を提案した. アニメコンテンツに関するメタデータ付きのデータセットとしては、Manga109 [11] が挙げられる. 漫画作品をページ単位にして、「コマ」「キャラクタ」「テキスト」の三つの内容に対して人手で領域とラベルのアノテーションを行った. 品質の高いメタデータを構築した一方、大量なアノテーション作業やチェック管理にコストを要した. 本研究では、Web上に公開された画像をマルチラベルに分類するモデルを活用し、自動的にアニメ動画からカットシーン単位の特徴タグを付与する手法を提案する.

3 提案手法

3.1 システムの流れ

図1に提案したアニメ動画の入力からカットシーンの特徴タグを作成するまでの処理の流れを示す. 本研究では、アニメ動画のビデオデータからカットシーンを分割し、画像のマルチラベル分類モデルを用いてフレームごとに特徴タグを推定することにより、シーン画面の内容を表すタグデータを作成する.

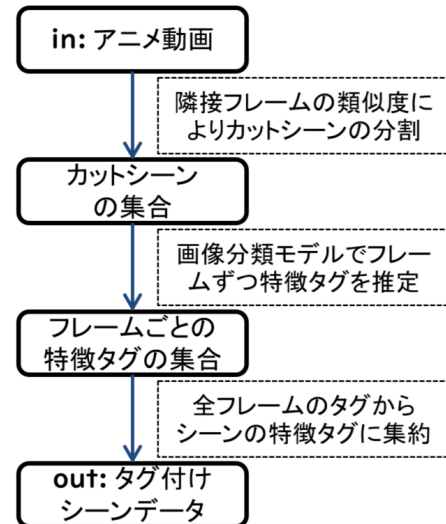


図1: アニメ動画からカットシーンの特徴タグを作成する流れ

3.2 カットシーンの分割

本研究では、動画構成の最小単位である「カット」を、特徴タグを付与するシーン対象として扱っている. 「カット」とは、一つの構図を表現し、同一オブジェクトが連続する画面であり、内容的には一貫性のある連続するフレームの集合である^{1 2}.

そのため、画面の内容的な連続性のあるカットシーンを分割する際に、隣接するフレームの類似度を測り、設定された閾値を超える（類似度が小さい）隣接フレームを前後に分けて、前のフレームまでの連続するフレーム集合を一つのカットシーンとして抽出する. 本研究では、アクション・スポーツ・日常系の三つのジャンルからそれぞれ10個のカットシーンに対する予備的な調査により、以下のシーン分割のルールを設定した.

1. 隣接フレームの類似度の測り方は、OpenCVのテンプレートマッチングにある「TM_CCOEFF_NORMED」方法³を利用する.
2. 類似度「TM_CCOEFF_NORMED」の閾値を0.6に設定し、類似度が0.6以下である隣接フレームを前後に分ける.
3. フレーム数が10枚以下のカットシーンが連続して出現した場合、再生時間の順番に合わせて、前からフレーム数10枚未満のカットシーンを30枚以上までに合併する.

以上のルールに従って分割されたカットシーンを、三つのジャンルからそれぞれランダムに1000件抽出し、

¹<https://kotatsu.info/2022-06-09-movie-terminology/>

²<https://orita-ani.net/scene-different-from-sequence/>

³https://docs.opencv.org/4.x/df/dfb/group_imgproc_object.html

表 1: 分析する特徴タグの情報とサンプル.

タグ種類	サンプルタグ	タグの説明
Character (キャラ関連 のタグ, 合計 2748 個)	blue_eyes	画面上に青い目 のキャラがいる
	blonde_hair	画面上に金髪の キャラがいる
	kimono	画面上に着物を着る キャラがいる
	hair_over_eyes	目が前髪に隠される キャラがいる
	shouting	画面に大声を出そうと しているキャラがいる
Background (背景関連の タグ, 合計 394 個)	cloud	雲が背景にある
	cliff	背景に崖がある
	classroom	背景に教室の風景 がある
	indoors	室内である背景
	in_water	キャラが水の中にいる
Composition (構図関連の タグ, 合計 122 個)	from_above	ハイアングルの構図 、俯瞰
	from_outside	室外からの視点
	pov	一人称視点
	eye_focus	目のクローズアップ
	vanishing _point	消失点のある構図

合計 3000 件のカットシーンに対して, 人手でカットシーンの分割精度を評価した結果, 平均的な分割精度は 93.83% となった.

3.3 フレーム内容の特徴タグの推定

本研究では, Web 上に公開された訓練済みのアニメ画像向きの内容特徴を推定するモデル⁴を利用し, 分割されたカットシーンのフレーム画面に対し一枚ずつ特徴タグの推定を行った. 分析可能な特徴タグ数は合計で 6891 個であり, 本研究では, その中に含まれる三種類の特徴タグを用いて, 画面内容を表すメタデータを構築した. 表 1 にフレーム画面の内容を表すタグとして抽出し, カットシーンに付与した三種類の特徴タグの情報とサンプルを示す. 本研究では, キャラクターデザイン関連のタグ (Character), 画面背景関連のタグ (Background) と構図関連のタグ (Composition) の三種類から, アニメのカットシーンに対する内容のメタデータを作成している. キャラクターデザイン関連のタグは 2748 個, 画面背景関連のタグは 394 個, 構図関連のタグは 122 個を用意した⁵.

タグ推定モデルによりフレーム画面を一枚ずつ推定し, 推定された内容タグの結果を合算し, 一つのカット

⁴<https://github.com/KichangKim/DeepDanbooru>

⁵上記三種類のタグ以外に, オブジェクトを示すタグ (ペットボトル, ラーメン, ゲーム機など) や画像の表現を示すタグ (グレースケール画像, 四コマ漫画など) も多数ある. 今回はこれらのタグをカットシーンのメタデータ構築に使用しない.

```
CaptainTsubasa_005_cut0281.mp4:
[44, [0, 0, 44, 0, 0, ..., 0, 5, 0, 0, 0]]
カットシーンのID
Jujutsu_004_cut0049.mp4:
[55, [0, 3, 11, 0, 0, ..., 0, 0, 0, 0, 0]]

UmaMusume_S1_005_cut0065.mp4:
[226, [0, 0, 0, 0, 0, ..., 0, 70, 0, 0, 0]]
カットシーンの合計フレーム数 当該タグの出現したフレーム数

HeroAcademia_010_cut0329.mp4:
[36, [0, 22, 0, 0, 0, ..., 0, 0, 0, 0, 0]]
長さ6891の内容タグの位置に対応する格納ベクトル

HibikeEuphonium_S1_004_cut0334.mp4:
[56, [0, 56, 56, 0, 0, ..., 0, 0, 0, 0, 0]]
```

図 2: カットシーンの記録サンプル

シーンに対応するベクトルに格納する. 図 2 にフレームずつ合算したカットシーンの記録サンプルを示す. 一つのカットシーンに対して, シーンの ID, シーンのフレーム数と, フレームずつ合算した内容タグの格納ベクトルの三つの情報を記録する.

3.4 カットシーンの特徴タグ

同じカットシーンにしても, キャラクターやオブジェクトの移動, そしてカメラワークの動きにより, 画面の内容も常に変わっている. 一枚前のフレーム画面に存在する内容タグは, 一枚後のフレーム画面から消える可能性もある. そのため本研究では, 各カットシーンに対して, フレームごとに推定して合算された内容タグの格納ベクトルから, カットシーンの合計フレーム数の半分 (50%) 以上の出現頻度のあるタグを抽出し, 当該カットシーンの特徴タグとしてシーンに付与する.

例えば, 図 2 に示す一番下の記録のように, 56 枚のフレームがあるカットシーンの場合, 出現頻度は 28 回以上のタグ (=56*50%, 同じタグは 28 枚以上のフレーム画面から検出された) を, 当該シーンの内容特徴を代表するタグと考え, カットシーンの特徴タグとして抽出する. 出現頻度が 28 回未満の内容タグは, カットシーン全体の内容特徴を反映できないとみなし, 特徴タグから削除する.

抽出されたカットシーンの特徴タグを, 3.3 節に示す内容タグの格納ベクトルと同じ順に, 長さ 6891 の One-Hot ベクトルの形で保存する (図 3).

```
HeroAcademia_005_cut0073.mp4 [0, 0, 0, 0, 0, ..., 1, 0, 0, 0, 0]
MIX_S1_001_cut0093.mp4 [0, 1, 1, 0, 0, ..., 0, 1, 0, 0, 0]
DragonBall_107_cut0275.mp4 [0, 1, 1, 0, 0, ..., 1, 0, 0, 0, 0]
```

図 3: カットシーンの One-Hot 特徴タグベクトル

表 2: ジャンル別のカットシーンの統計.

ジャンル	カットシーン数	平均フレーム数
アクション	34673	127.97
スポーツ	31864	119.91
日常系	32601	120.54
合計	99138	122.94

表 3: 内容タグのカットシーンでの出現割合.

出現頻度	タグ数割合
10%未満	0.3934
10% ~ 20%未満	0.1116
20% ~ 30%未満	0.0723
30% ~ 40%未満	0.0552
40% ~ 50%未満	0.0446
50% ~ 60%未満	0.0401
60% ~ 70%未満	0.0352
70% ~ 80%未満	0.0320
80% ~ 90%未満	0.0337
90% ~ 100%未満	0.0556
=100%	0.1262

表 4: 「アクション」アニメでの種類別頻出タグ Top10.

タグ種類	タグ内容	出現シーン数 (割合)
Character	1girl	11228 (32.38%)
	1boy	8722 (25.15%)
	black_hair	8405 (24.24%)
	short_hair	5353 (15.44%)
	open_mouth	5344 (15.41%)
	long_hair	5191 (14.97%)
	smile	5134 (14.81%)
	looking_at_viewer	4759 (13.72%)
	brown_hair	4286 (12.36%)
	shirt	4119 (11.88%)
Background	sky	8205 (23.66%)
	outdoors	7813 (22.53%)
	cloud	6824 (19.68%)
	blue_sky	4264 (12.30%)
	cloudy_sky	4187 (12.07%)
	tree	3566 (10.28%)
	scenery	3360 (9.69%)
	building	2996 (8.64%)
	mountain	2066 (5.96%)
	city	1937 (5.59%)
Composition	male_focus	9034 (26.05%)
	close-up	3897 (11.24%)
	upper_body	3533 (10.19%)
	depth_of_field	2276 (6.56%)
	from_behind	567 (1.63%)
	lens_flare	567 (1.63%)
	from_side	567 (1.63%)
	eye_focus	479 (1.38%)
	gradient	469 (1.35%)
	monochrome	441 (1.27%)

4 タグ付けシーンデータの分析

本研究では、アクション・スポーツ・日常系の三つのアニメジャンルからそれぞれ 10 タイトル、各タイトルに対して 10 話の合計 300 話 (=3*10*10) のアニメ動画をデータセットとして、提案システムによりタグ付けされたアニメのシーンデータを作成した。表 2 に分割されたジャンル別のアニメシーンの統計情報を示す。提案手法で分割されたカットシーン数はアクション、スポーツ、日常系でそれぞれ 34673 本、31864 本と 32601 本となり、平均フレーム数も三ジャンルで同じ 120 枚程度であった。

4.1 内容タグの分布の統計

フレームごとに記録した各カットシーンの内容タグ (3.3 節) の結果を基にして、各内容タグのアニメのカットシーンでの出現分布を調査した。表 3 に全ての内容タグの、カットシーン内で出現したフレーム数の割合の統計結果を示す。

統計結果により、四割弱 (39.34%) の内容タグがカットシーン中で 10%未満のフレームにしか検出されなかった (フレーム数は 100 枚であるカットシーンの場合、当該タグは、10 枚未満のフレームにしか存在しなかった)。この結果から、カットシーンの画面に出た主な内容タグは、オブジェクトやカメラの動きなどにより、わずかなフレームからしか検出できなかったと考えられ、シー

ン全体を代表する特徴タグとして抽出する際に、タグの出現頻度によりフィルタリングする必要があると判断する。一方、12.62%の内容タグが、カットシーンの全フレーム画面に検出された (出現頻度=100%)。この結果から、分割されたカットシーンの画面に集中している内容もあり、タグ推定によりアニメシーンの特徴タグを作成するのは可能と考えられる。

4.2 ジャンル別の頻出特徴タグ

抽出されたカットシーンの特徴タグを 3.3 節に定義した三つのタグ種類に分けて、カットシーンでの頻出タグをアニメのジャンル別により調査した。表 4、表 5 と表 6 にそれぞれ「アクション」「スポーツ」「日常系」のアニメのカットシーンにおいて、「キャラクタ」「背景」「構図」の三種類から統計した頻出する特徴タグの上位 10 個の結果を示す。

「アクション」と「スポーツ」のアニメシーンにある「キャラクタ」に関する頻出特徴タグの上位 10 個の内容は全て一致し、「日常系」アニメの上位 10 個において唯一異なるのは「school_uniform」(学校制服)である (「1boy(一人の男がいる)」と入れ替える)。この結果から、アニメ動画はジャンルに関係なく、キャラクタのデザインに近い選択や構成があることが分かった。

表 5: 「スポーツ系」アニメでの種類別頻出タグ Top10.

タグ種類	タグ内容	出現シーン数 (割合)
Character	black_hair	12875 (40.41%)
	shirt	10695 (33.56%)
	lgirl	9281 (29.13%)
	lboy	9159 (28.74%)
	brown_hair	7626 (23.93%)
	open_mouth	7091 (22.25%)
	short_hair	6988 (21.93%)
	smile	5692 (17.86%)
	looking_at_viewer	5268 (16.53%)
Background	long_hair	4727 (14.83%)
	outdoors	8640 (27.12%)
	sky*	5917 (18.57%)
	blurry_background	4450 (13.97%)
	tree	4426 (13.89%)
	cloud	4074 (12.79%)
	indoors	3659 (11.48%)
	blue_sky*	3425 (10.75%)
	building	2997 (9.41%)
	scenery	2148 (6.74%)
Composition	cloudy_sky*	2037 (6.39%)
	male_focus	11258 (35.33%)
	upper_body	6032 (18.93%)
	depth_of_field	5205 (16.34%)
	close-up	2729 (8.56%)
	from_side	871 (2.73%)
	gradient	766 (2.40%)
	from_behind	747 (2.34%)
	lens_flare	496 (1.56%)
	halftone	420 (1.32%)
	letterboxed	416 (1.31%)

表 6: 「日常系」アニメでの種類別頻出タグ Top10.

タグ種類	タグ内容	出現シーン数 (割合)
Character	lgirl	13881 (42.58%)
	long_hair	13117 (40.23%)
	brown_hair	10490 (32.18%)
	short_hair	9327 (28.61%)
	open_mouth	7418 (22.75%)
	black_hair	7325 (22.47%)
	school_uniform	6800 (20.86%)
	smile	6071 (18.62%)
	shirt	5854 (17.96%)
Background	brown_eyes	4763 (14.61%)
	outdoors	6590 (20.21%)
	sky	5069 (15.55%)
	blurry_background	4308 (13.21%)
	indoors	4271 (13.10%)
	tree	4117 (12.63%)
	cloud	3584 (10.99%)
	blue_sky	3151 (9.67%)
	building	2799 (8.59%)
	scenery	2660 (8.16%)
Composition	window	2413 (7.40%)
	depth_of_field	5515 (16.92%)
	upper_body	3293 (10.10%)
	close-up	2440 (7.48%)
	male_focus	1681 (5.16%)
	gradient	674 (2.07%)
	from_side	661 (2.03%)
	from_behind	597 (1.83%)
	pov	513 (1.57%)
	film_grain	498 (1.53%)
	portrait	418 (1.28%)

「背景 (Background)」に関する特徴タグからも同じ傾向が見られる。収集されたアニメ動画はジャンルに関係なく、背景の部分に「outdoors」「scenery」「sky」「tree」など室外の景色に関する内容が多かった。この結果から、アニメシーンの背景や場所にも類似性があることが分かった。「アクション」アニメと違い、「スポーツ」と「日常系」のアニメシーンで頻出する背景に関する上位 10 個の特徴タグに「indoors (室内の風景)」の内容があり、「アクション」のジャンルにより、「スポーツ」と「日常系」のアニメには室内であるシーンが多いことが分かった。

「構図 (Composition)」では、「upper-body (上半身)」や「close-up」など、近距離から描写するシーンがジャンルに関係なく多数あり、「アクション」と「スポーツ」のアニメのほうが、男性のキャラクタを中心に描写する (「male.focus」) シーンは、「日常系」のアニメよりはるかに多くあった (26.05%, 35.33% VS 5.16%)。

5 おわりに

本研究では、画像内容の特徴タグを推定するモデルを利用し、アニメ動画のカットシーン向きのメタデータの構築手法を提案した。アクション・スポーツ・日常系の三つのアニメジャンルから 300 話のアニメ動画

を収集し、提案システムにより特徴タグが付与されたアニメのシーンデータの作成を試みた。作成したアニメシーンのタグデータを分析したことで、提案手法によりアニメシーンのメタデータの作成は可能であることが分かった。

最終的なメタデータの作成までに、まだ二つの課題が残っている。一つ目は、現在分類した三つのタグ種類の中に関連性のあるタグが存在する。例えば、表 5 の「Background」にある「sky」と「blue_sky」「cloudy_sky」のような上位・下位関係のあるタグが複数存在する。今後はタグの種類をより細かく分類し、タグ間の関係を付けた上で、内容タグのオントロジーにより、アニメシーンの特徴タグを付与する。二つ目は、各特徴タグと画面上の内容に対応する部分 (タグに関連する画面内容の位置情報) はまだ不明である。例えばシーンの背景に「sky」のタグがあるとしても、当該タグに関する画面内容の広さ、そしてタグ内容のシーンに対する重要度の判断はまだ達成されていない。今後は新たなタグ推定モデルを導入し、Attention の仕組みなどを用いることで、各タグの推定に関連するピクセルの抽出及び、関連する画面内容の位置情報もシーンデータに付与する。

参考文献

- [1] D. Han, E. Park, G. Lee, A. Lee, and N. Kwak: MERLIN: Multimodal Embedding Refinement via LLM-based Iterative Navigation for Text-Video Retrieval-Rerank Pipeline, *arXiv preprint*, arXiv:2407.12508 (2024).
- [2] Lei N, Cai J, Qian Y, Zheng Z, Han C, Liu Z, and Huang Q: A two-stage Chinese medical video retrieval framework with LLM, *CCF International Conference on Natural Language Processing and Chinese Computing*, Cham: Springer Nature Switzerland, (2023).
- [3] Xu J, Lan C, Xie W, Chen X, and Lu Y: Retrieval-based video language model for efficient long video question answering, *arXiv preprint*, arXiv:2312.04931 (2023).
- [4] Team Gemini: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, *arXiv preprint* arXiv:2403.05530 (2024).
- [5] 中村浩之, 小川祐樹, 諏訪博彦: 動画のタグ情報を利用したニュースシーンの選択手法の提案, 日本社会情報学会学会誌, Vol. 22, No. 2, pp. 27–38, (2011).
- [6] 上畑恭平, 伊東栄典: タグの類似度に着目した利用者投稿サイト動画の多様性分析, 電子情報通信学会技術研究報告: 信学技報, Vol. 115, No. 381, pp. 83–88, (2015).
- [7] 村上直至, 伊東栄典: 動画投稿サイトで付与された動画タグの階層化, 研究報告数理モデル化と問題解決 (MPS), Vol. 2010, No. 17, pp. 1–6, (2010).
- [8] 高山耕平, Henry Johan, 西田友是: 特徴抽出によるアニメキャラクターの顔認識, 情報処理学会第 74 回全国大会講演論文集, Vol. 2012, No. 1, pp. 673–674, (2012).
- [9] 山田太雅, 棟方渚, 小野哲雄: アニメキャラクターの顔における特徴量抽出手法の提案と実装, 情報処理学会研究報告エンタテインメントコンピューティング (EC), Vol. 2017, No. 6, pp. 1–6, (2014).
- [10] 鈴木絵理, 山内拓人, 鄭顕志: 色及び位置情報を用いたアニメキャラクターの目と口の抽出, 第 36 回人工知能学会全国大会論文集 (JSAI2022), Vol. 2022, p. 204-GS-7-03, 2022.
- [11] 藤本東, 小川徹, 山本和慶, 松井勇佑, 山崎俊彦, 相澤清晴: 学術用アノテーション付き漫画画像データセットの構築と解析, 映像情報メディア学会技術報告, Vol. 41, No. 5, pp. 35–40, (2017).