新しい映像・テキストの対応付け方法による映像検索の問題点の解決

研究代表者　　植　木　一　也　　明星大学　情報学部　情報学科　准教授

1 概要

現在，日々インターネット上にアップロードされる映像や公共の場で設置されている防犯カメラなどの大規模映像から，詳細な検索クエリを使用して特定の情報やイベントを抽出する高度な技術への需要が高まっている．検索クエリ文の表現が非常に多彩であることから，クエリ文に完全に合致する映像を収集して学習することは非常に困難である．近年では，画像とテキストがペアになっている大規模なキャプション付きの画像データセットを学習する CLIP[1]などのマルチモーダル視覚・言語モデルの学習技術の発展により，映像検索精度が大幅に向上している．これらの画像・テキストの埋め込みモデルでは，画像から得られた特徴ベクトルとテキストから得られた特徴ベクトルが 1 対 1 で対応するように学習している．しかしながら，現実世界において 1 枚の画像を説明するような文章は多数あり，1 つの文を表現する画像も無数存在する．そのため，条件を満たす映像が見つけられない，条件の一部しか満たしていない映像が検索されてしまうなどの問題が起きてしまうことがわかった．そのため本研究では，

1. 画像と文を 1 対多で対応させる手法
2. 文と画像を 1 対多で対応させる手法

の 2 つの方向について検討し，大規模映像データセットを用いた実験により評価を実施し，その有効性を検証することを目的とする．

2 先行研究

映像検索技術は,長年にわたり研究され,さまざまな手法が提案されている.多くの研究では,MSR-VTT[2], VATEX[3], LSMDC[4], DiDeMo[5]など，数万から数十万の規模のキャプション付き映像データセットを用いて提案した手法を評価している．

一方，米国国立標準技術研究所（NIST）が実施している TRECVID ベンチマーク[6]において，複雑なクエリ文を用いて，100 万を超えるインターネットアーカイブの大規模な映像の中から適切な映像を検索する Ad-hoc Video Search（以下，AVS）タスクが 2016 年から毎年実施されている．映像が多種多様であることや，検索のクエリ文が人物，物体，シーン，動作，時間などの複数の複雑な条件を含んでいることから，より実用性の高いベンチマークとなっている．現時点では，LAION-5B[7]（50 億枚）などの大規模なキャプション付き画像データセットを事前学習したマルチモーダル埋め込みモデルを複数統合する手法[8][9]が最も検索精度が高い．

事前学習済みの画像・テキストの埋め込みモデルを用いる手法では，クエリ文をテキストエンコーダに，検索対象のフレーム画像を画像エンコーダにそれぞれ入力し，出力された特徴ベクトル同士を比較して類似度を計算することにより，映像検索を実現している．しかし，画像とテキストの特徴量の分布にモダリティギャップ[10]があることも示されており，視覚・言語の対応付けにはまだまだ課題も多いのが現状である．

3 提案手法

提案手法の概要を図 1 に示す．現在主流となっている画像・テキストの埋め込みモデルを用いる手法による結果に加え，テキスト生成ベースの手法と画像生成ベースの手法による結果を統合することにより実現している．

この章では，どのような問題に焦点を当てたかについて述べ，その後，テキスト生成ベースの手法と画像生成ベースの手法の詳細について述べる．

公益財団法人電気通信普及財団
研究調査助成報告書　第39号　2024年度

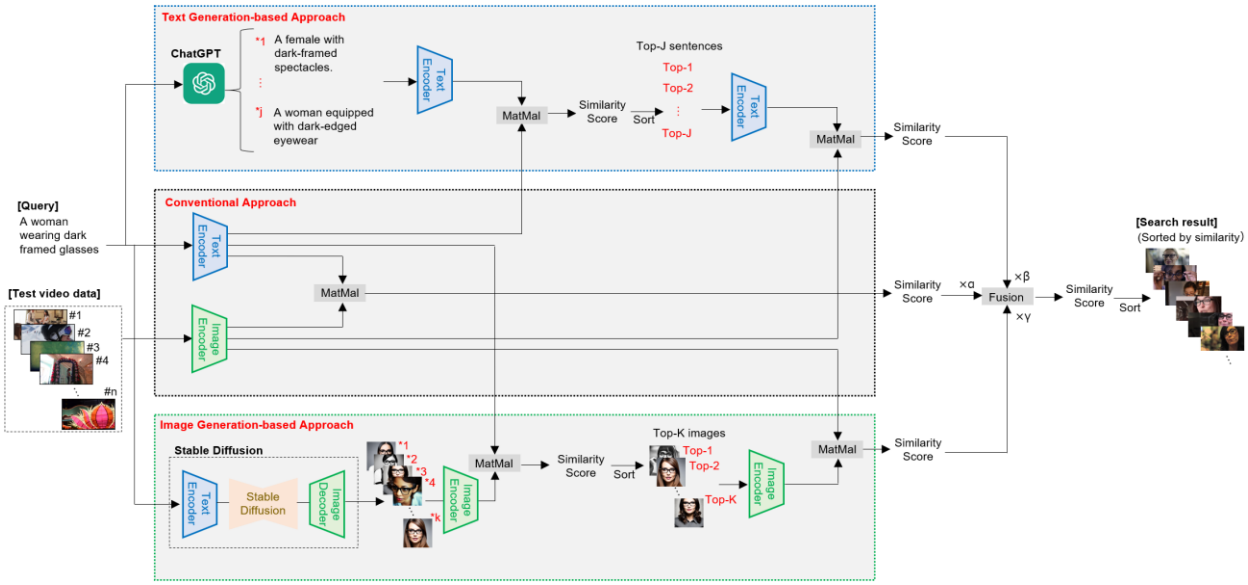

図 1. テキスト生成と画像生成を組み込んだ映像検索手法の概要図．画像・テキストの埋め込みモデルを使ったベースライン手法の結果と，上部に示すテキスト生成，下部に示す画像生成を使った手法の結果を統合することにより最終結果を出力する．

3-1 解決したい課題

CLIP に代表される画像・テキストの埋め込みモデルは，画像とその対応する説明文を用い，画像の特徴とテキストの特徴の類似度が高くなるように学習する．つまり，画像とその説明文がそれぞれ一対一に対応付けられるよう学習を行う．しかし，実際には画像とテキストは一対多または多対多で表現すべき特徴ではないだろうか．例えば，TRECVID 2023 の AVS タスクで実際にクエリ文として出題された「Two or more persons are seen in front of a chain link fence」を考えてみる．2 人以上（= Two or more people）という条件は，画像中に 2 人のみが映っている場合だけでなく，大勢が映っている場合にも満たされている．映っている人は，子供であっても学生であっても大人であっても構わない．金網（= chain link fence)が写っている場所についても，サッカー場やテニスコートなどの運動場周辺や，建設作業中の場所周辺など，さまざまな環境が考えられる．つまり，このクエリ文を表現する画像は一種類だけでなく無数存在するため，テキストと画像は一対多の関係といえる．

図 2. この画像を説明するテキストは一種類だけだろうか？

逆に，画像とテキストとの対応を考えてみる．図 2 に示す画像の説明文はどのようなものが考えられるだろうか．単純に「犬が一匹いる」というのも正しい表現であり，「栗色の可愛いミニチュアダックスフンドが広い公園の芝生の上で散歩している」という詳細な説明も正しい表現である．後方に写っている木に着目し，「公園にある大きな木」のような表現も可能である．一枚の画像に対しても複数の説明文が考えられること

から，画像とテキストの関連性も一対多と考えることができる．

一方，ChatGPT に代表されるテキストの生成技術や，Stable Diffusion[11]等のテキストから画像を生成する技術の発展が著しい．これらの技術を用いることで，検索クエリに関する多彩なテキスト・画像表現を獲得でき，映像検索のベースライン精度を向上できるのではないかと考えた．

以上のことから,本研究におけるリサーチクエスチョンは,「映像検索のベースライン精度が改善のために,テキスト生成技術による検索クエリの拡張や，画像生成モデルを使用したクエリ文に対応する画像を複数生成することは有効か」を検証することである．

3-2 テキスト生成ベースの手法の導入

入力する検索クエリ文のバリエーションを豊富にするため，自然言語処理や対話生成の高度な機能を備えた ChatGPT 等のテキスト生成器を活用し，新たな文を生成する．例えば「Give me 10 sentences that mean exactly the same as “元のクエリ文” with slight changes.」というプロンプトを入力することで，意味は変更せずに表現のみを変化させた文を生成することができる．

ChatGPT を用いて生成した文は，概ね元のクエリと同様の意味をもつ異なる表現の文が得られていたが,中には“Building with columns during daytime”という元のクエリ文に対して“Creating with columns in the day hours”といったように意図しない意味の文が生成される事例も存在する．この問題に対処するため，生成された文の中から適切なものを自動的に選別する手法を採用する．具体的には，元のクエリ文をテキスト生成器に入力し，新たに j 個の文を生成したのち，画像・テキストの埋め込みモデルにおけるテキストエンコーダを用いて，生成されたテキストの特徴ベクトルと元のクエリ文の特徴ベクトルの類似度を計算する．その後，類似度が高い上位の J 個の文を選択する．テキストの選別の際に学習済みの画像・テキストの埋め込みモデルを使用する理由は，このモデルが画像とテキストを同時に埋め込むことを考慮しているため，画像を表現する文として適切と考えたからである．

3-3 画像生成ベースの手法の導入

マルチモーダル埋め込みモデルにおける画像とテキスト間の特徴量の分布にはモダリティギャップ[10]があることから，画像生成技術を用いてそのギャップを克服する．画像生成技術を用いることで，単一テキストに対して複数の関連画像を生成することも可能なため，これによりテキストと画像の一対多の関係を効果的に表現できる．生成された画像群は，テキストの内容をより幅広くカバーし，映像検索における精度と多様性の向上に寄与することが期待される．画像生成には Stability AI 社が公開している学習済みの Stable Diffusion モデルなどを用いることができる．

以下，具体的な方法について説明する．検索クエリ文を画像生成モデルに入力し，複数の画像（k 枚）生成する．次に，生成された画像が元のテキストと完全に一致するとは限らないため，画像・テキストの埋め込みモデルを使用して，生成画像の特徴ベクトルと元のクエリ文の特徴ベクトルの類似度を計算する．この計算に基づき，類似度が高い上位 K 枚の画像を選択する．この選択された K 枚の画像は，映像検索のクエリとして使用される．具体的には，検索対象のデータベースの画像数を n とした場合，n × K の類似度計算を行い，検索対象のデータそれぞれに対して K 個の類似度の平均を算出する．

3-4 テキスト生成・画像生成ベースの手法との統合

次に，画像・テキストの埋め込みモデルを用いるベースライン手法と，テキスト生成・画像生成ベースの手法との統合する．ベースライン手法，テキスト生成ベースの手法，画像生成ベースの手法それぞれから，類似度スコアを計算することができる.それぞれの手法で得られる類似度スコアのスケールを合わせるため，まずは各手法から得られる類似度スコアを，全テストデータで最高が 1.0，最低が 0.0 となる min-max 正規化する．次に，ベースライン手法，テキスト生成ベースの手法，画像生成ベースの手法の統合比率を，それぞれ $\alpha:\beta:\gamma$ の比率の重みで統合する．最終的な類似度スコアを求めた後，類似度スコア順にソートすることにより映像の検索結果を得ることができる．

複数のモデルの結果を統合する際にも,類似度スコアの重み付き平均で簡単に統合することが可能である．この単純な統合方法を用いた場合でも，映像検索の精度向上に効果があることが実験によって明らかになっている．

4 検証実験

4-1 使用する映像データセットと評価方法

提案手法の有効性を確認するため，2022 年の TRECVID ベンチマークで使用された V3C2 映像データセットを使用した．V3C2 データセットは，インターネットアーカイブで公開されている映像で，全体で 1,300 時間(1.6 TB)，1,425,454 本の短いセグメントから構成されている．本実験では計算量削減のため，映像のすべてのフレームを用いるのではなく 10 フレームおきに画像を抽出することとした．検索クエリ文は，TRECVID 映像ベンチマークで 2022 年の出題された 30 種類（クエリ ID 701～730，以後 tv22）を用いた．

手法の精度比較は，ベンチマークのルールと同様，各クエリに合致している映像の 1 位から 1,000 位までのランキングを出力して映像検索精度を計測する．本評価では，ベンチマーク終了後に提供された正解データを利用して近似的に算出した推定平均適合率（inferred mean average precision：以下，mAP）を用いて，検索システムの性能を比較する．実験に使用する学習済み画像・テキストの埋め込みモデルについては，OpenCLIP で公開されている数多くのモデルのうち，比較的精度が高い 29 個のモデルと BLIP で公開されている 4 種類のモデルを評価した．

4-2 実験条件

テキスト生成ベースの手法については，ChatGPT に下記に示す 5 種類のプロンプトを入力し，新しい文を生成した：

1. Give me 10 sentences that mean exactly the same as “元のクエリ” with slight changes.
2. Give me 10 examples of “元のクエリ” that means exactly the same thing, but with a slight change in the sentence.
3. List 10 sentences that mean the same as “元のクエリ” with slight modifications.
4. List 10 examples that mean the same thing as “元のクエリ,” but with a slight change in the sentence.
5. Give me 10 sentences that mean the same as the following sentence with a slight change of wording: “元のクエリ”

tv22 の各クエリ文に対し，上記 5 種類のプロンプトを 2 回ずつ入力することにより，全部で 100 個の新しい文を獲得した．重複を除いたところ，最小 38 個，最高 100 個，平均で 85.2 個の新しい文を獲得することができた．生成された文の中から元のクエリ文と意味がかけ離れている文を除くため，画像・テキストの埋め込みモデルのテキストエンコーダを用い，テキストの類似度が 0.9 以下のものを排除した．

画像生成モデルについては，Stability AI 社が提供している Stable Diffusion v2.1 モデル（以下，SD 2.1）と，Stable Diffusion XL 1.0 モデル（以下，SDXL 1.0） を用いた．SDXL 1.0 は 2023 年 7 月にリリースされたばかりのモデルで，画像の品質も高く，プロンプトの理解力も向上している．今回の実験では，1 つの検索クエリに対して，SD 2.1 と SDXL 1.0 のそれぞれのモデルから 400 枚ずつ，合計で k=800 枚の画像を生成した．その後，生成された画像と入力クエリの類似度が高い画像を全体の 25%である K=200 枚を選択して映像検索に利用した．

4-3 検証結果

ベースライン法（B），テキスト生成ベースの手法（C），画像生成ベースの手法（S）と，複数の手法を統合した結果を表 1 に示す．ここで，複数手法を統合する際の類似度スコアの統合比 $\alpha:\beta:\gamma$ については，特に最適化の処理は行わず，単純にそれぞれ 1.0 と設定した．各手法の単体の精度を見てみると，ベースラインの手法を上回っているものはそれ程多くないが，ベースライン法とテキスト生成ベースの手法の統合(B+C)，ベースライン法と画像生成ベースの手法の統合（B+S），すべての統合（B+C+S）では，ベースライン精度に比べて 2 から 6 ポイント程度向上するなど，概ね大きな精度向上が確認できる．

画像生成ベースの手法との組み合わせでは，統合前はベースライン精度よりも大幅に低いにも関わらず，統合によって精度向上が見られるケースが多かった．これは，画像生成により，元のクエリと相補的な特徴が得られている，テキストと画像間のモダリティギャップが埋められているといった利点があるものと考えられる．一方，画像生成ベースの手法で精度が極端に低い原因を見てみると，検索クエリ「Two or more persons in a room with a fireplace」に対して，Stable Diffusion で生成したほとんどの画像に対し，「fireplace

（暖炉）」は画像の中にあるものの，「Two or more persons」が無視され，人が写っていないといった問題があった．

表 1. OpenCLIP や BLIP で提供されている事前学習済みのマルチモーダルモデルを利用し，ベースライン手法，テキスト生成ベースの手法，画像生成ベースの手法の映像検索精度（mAP）の比較．B:ベースライン手法，C：ChatGPT によるテキスト生成ベースの手法，S：Stable Diffusion による画像生成ベースの手法

モデル名	B	C	S	B+C	B+S	B+C+S
ViT-H-14 (laion2b_s32b_b79k)	15.38	14.44	14.44	15.79	20.06	20.49
ViT-g-14 (laion2b_s12b_b42k)	15.85	15.24	13.54	16.09	19.43	19.80
ViT-g-14 (laion2b_s34b_b88k)	16.10	15.71	14.95	16.85	21.06	21.82
ViT-bigG-14 (laion2b_s39b_b160k)	15.95	15.24	14.44	17.02	19.78	21.02
xlm-roberta-base-ViT-B-32 (laion5b_s13b_b90k)	14.15	13.93	10.62	14.68	15.94	16.71
xlm-roberta-large-ViT-H-14 (frozen_laion5b_s13b_b90k)	17.75	16.74	14.60	18.15	20.94	21.78
convnext_large_d (laion2b_s26b_b102k_augreg)	15.68	15.69	13.30	16.39	18.96	19.77
convnext_large_d_320 (laion2b_s29b_b131k_ft)	16.05	15.98	13.26	16.69	19.02	19.98
convnext_large_d_320 (laion2b_s29b_b131k_ft_soup)	16.67	16.68	13.53	17.29	19.72	20.68
convnext_xxlarge (laion2b_s34b_b82k_augreg)	15.63	15.75	14.48	16.57	19.64	20.17
convnext_xxlarge (laion2b_s34b_b82k_augreg_soup)	15.40	15.24	14.37	16.10	19.46	20.07
convnext_xxlarge (laion2b_s34b_b82k_augreg_rewind)	15.64	15.31	14.46	16.14	19.47	20.07
coca_ViT-L-14 (mscoco_finetuned_laion2b_s13b_b90k)	22.44	22.47	23.24	23.46	26.93	27.19
EVA01-g-14 (laion400m_s11b_b41k)	14.78	13.34	14.80	15.00	18.39	18.49
EVA01-g-14-plus (merged2b_s11b_b114k)	18.85	18.42	8.95	19.45	17.48	19.60
EVA02-B-16 (merged2b_s8b_b131k)	18.17	17.62	9.71	18.57	17.65	19.37
EVA02-L-14 (merged2b_s4b_b131k)	18.11	17.06	8.21	18.53	17.83	19.59
EVA02-L-14-336 (merged2b_s6b_b61k)	18.32	17.20	8.05	18.66	18.19	19.81
EVA02-E-14 (laion2b_s4b_b115k)	16.58	14.91	15.39	16.70	21.52	21.93
EVA02-E-14-plus (laion2b_s9b_b144k)	16.07	15.15	15.64	17.21	20.82	21.53
ViT-B-16-SigLIP (webli)	14.53	13.31	10.55	14.81	17.80	18.12
ViT-B-16-SigLIP-256 (webli)	14.71	13.57	10.82	14.89	18.06	18.34
ViT-B-16-SigLIP-i18n-256 (webli)	13.58	12.85	11.36	13.79	17.31	17.65
ViT-B-16-SigLIP-384 (webli)	15.10	14.04	10.32	15.33	18.40	19.01
ViT-B-16-SigLIP-512 (webli)	16.28	15.16	11.03	16.60	19.48	20.10
ViT-L-16-SigLIP-256 (webli)	15.97	15.78	8.52	16.67	17.55	18.81
ViT-L-16-SigLIP-384 (webli)	15.56	16.44	8.14	16.84	16.84	18.50
ViT-SO400M-14-SigLIP (webli)	15.60	15.51	7.21	16.30	16.85	18.27
ViT-SO400M-14-SigLIP-384 (webli)	16.66	16.75	7.64	17.36	17.74	19.10
blip-itm-base-coco	14.67	15.18	18.45	15.88	21.24	21.36
blip-itm-large-coco	18.43	17.58	20.25	19.09	22.72	22.64
blip-itm-base-flickr	13.29	11.82	14.60	13.10	18.05	17.20
blip-itm-large-flickr	16.96	16.69	17.21	17.68	20.28	20.42

5 まとめ

本研究では，映像検索のタスクにおいて，豊富な文を生成可能な ChatGPT を使用して言語側の表現を充実させることと，Stable Diffusion を用いて生成した画像を直接的に用いることにより，検索精度の向上が可能であることを示した．ただしその際は，生成されたテキストや画像をそのまま使うのではなく，学習済みの画像・テキストの埋め込みモデルを利用して，テキストや画像を選別することも重要である．本手法は新しい画像・テキストの埋め込みモデル，テキスト生成モデル，画像生成モデルを利用した際に，ベース精度を向上させる効果があるため，高い汎用性がある．

今後は，テキスト生成技術を活用して映像に関連する豊富なメタデータやコンテキストを生成し，これを画像生成技術によって視覚的な要素と組み合わせるなど，テキスト・画像生成技術の相互作用を探究することで，より意味豊かで精緻な検索結果が得られるかどうかを検証していく予定である．また，テキスト・画像の認識・生成技術を，画像認識や自然言語処理の幅広いタスクに拡張し，革新的な技術の進展を促進する

ことで，異なるモダリティ間での情報の統合や相互理解を深めることを目指す.

【参考文献】

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” arXiv:2103.00020, 2021.

[2] J. Xu, T. Mei, T. Yao, and Y. Rui, “MSR-VTT: A Large Video Description Dataset for Bridging Video and Language,” In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[3] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, “VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research,” In Proc. of IEEE International Conference on Computer Vision (ICCV), 2019.

[4] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, “Movie Description,” International Journal of Computer Vision, vol.123, no.1, pp.94-120, 2017.

[5] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, “Localizing Moments in Video with Natural Language,” In Proc. of the IEEE International Conference on Computer Vision (ICCV), pp.5803-5812, 2017.

[6] G. Awad, K. Curtis, A. A. Butt, J. Fiscus, A. Godil, Y. Lee, A. Delgado, E. Godard, B. Chocot, L. Diduch, Y. Graham, and G. Quénot, “TRECVID 2023 - A series of evaluation tracks in video understanding,” In Proc. of TRECVID 2023, 2023.

[7] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, “LAION-5B: An open large-scale dataset for training next generation image-text models,” In 36th Conference on Neural Information Processing Systems (NeurIPS), 2022.

[8] J. He, R. Li, J. Guo, H. Zhang, M. Li, Z. Wu, Z. Wang, B. Du, and C. Liang, “WHU-NERCMS at TRECVID 2023: Ad-hoc Video Search (AVS) and Deep Video Understanding (DVU) Tasks,” In Proc. of TRECVID 2023, 2023.

[9] K. Ueki, Y. Suzuki, H. Takushima, H. Sato, T. Takada, H. Okamoto, H. Tanoue, T. Hori, and A. M. Kumar, “Waseda_Meisei_SoftBank at TRECVID 2023,” In Proc. of TRECVID 2023, 2023.

[10] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, “Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning,” in Proc. of the Conference on Neural Information Processing Systems (NeurIPS), 2022.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” In Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.10684--10695, 2022.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
Query Extension for Video Retrieval Using ChatGPT	The 12th International Conference on Information Communication and Applications (ICICA2023)	2023 年 5 月
ChatGPT を利用したクエリ拡張による映像検索精度の向上	第 26 回 画像の認識・理解シンポジウム (MIRU2023)	2023 年 7 月
Waseda_Meisei_SoftBank at TRECVID 2023	Notebook paper of the TRECVID 2023 Workshop	2023 年 11 月
アドホック映像検索における事前学習済みモデルの有効な活用法	精密工学会誌，vol.89，no.12	2023 年 12 月

Improving Video Retrieval Performance with Query Expansion Using ChatGPT	The 7th International Conference on Image and Graphics Processing (ICIGP2024)	2024年1月
Exploring the Evolution of Video Retrieval: From Early Systems to Latest Techniques	The 5th International Conference on Computer Vision and Computational Intelligence (CVCI 2024)	2024年1月
High-Precision Video Retrieval through Image Generation	The 8th IIEEJ International Conference on Image Electronics and Visual Computing (IEVC 2024)	2024年3月