Research Article Image Duplication Detection Based on Computer Vision

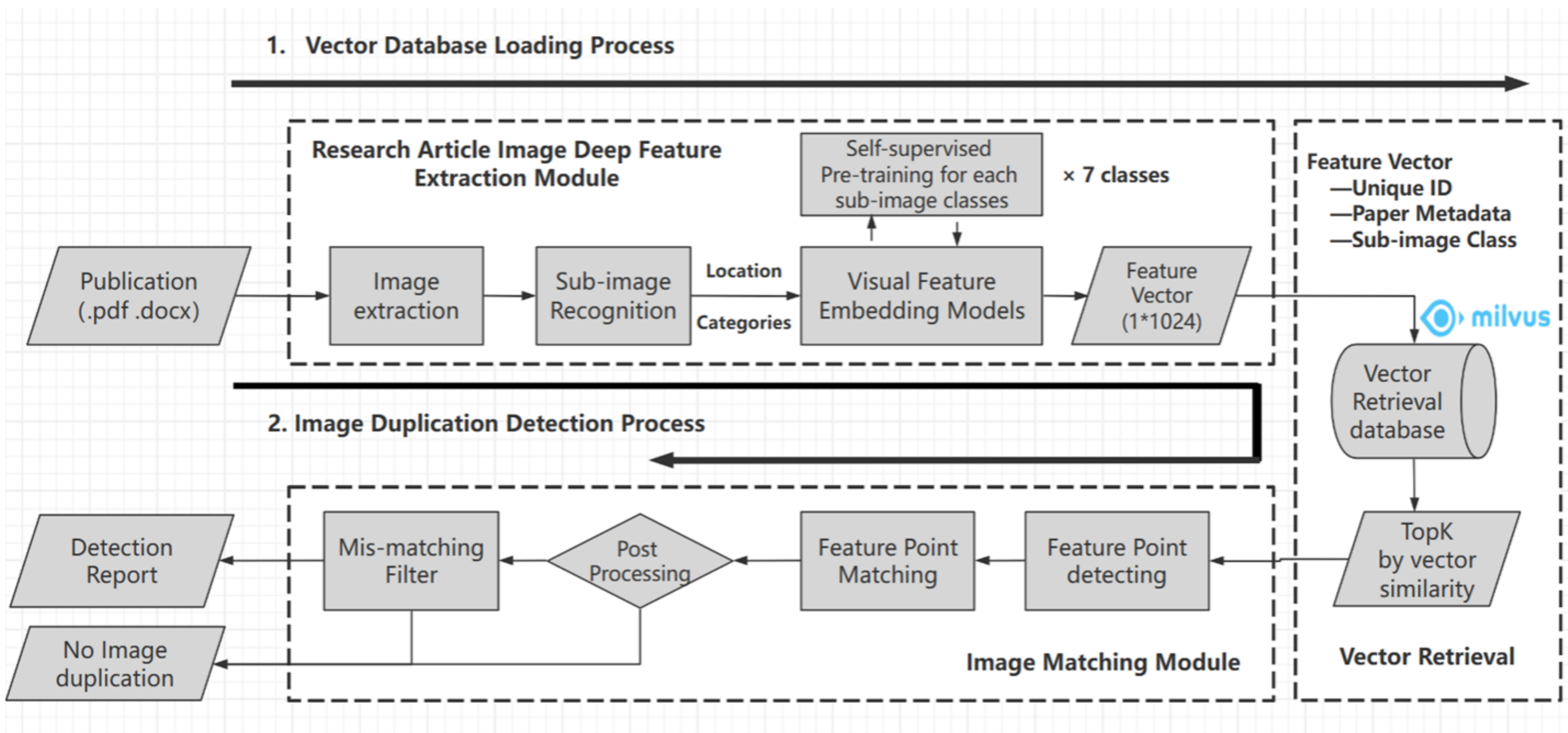
Ding Junpeng[1], Liu Jianhua [2], Hu Tianyi [1], E Haihong[1]*

[1] School of Computer Science, Beijing University of Posts and Telecommunications, China

[2] Beijing Wanfang Data Co., Ltd., China

Overview of our study

Research Article Image Duplication Detection :



- **Challenges:**

a) Images in research articles are mostly composed of multiple sub-images, and plagiarism most likely occurs in sub-images rather than entire image;

b) Current image duplication detection methods based on Siamese Network necessitate the specification of input image pairs for detection, leading them ill-suited for large-scale image screening task.

c) Improper duplication of image content in papers often involves image tampering(local manipulations), such as scaling and rotation, which significantly diminish the accuracy of detection.

- **Our Contributions:**

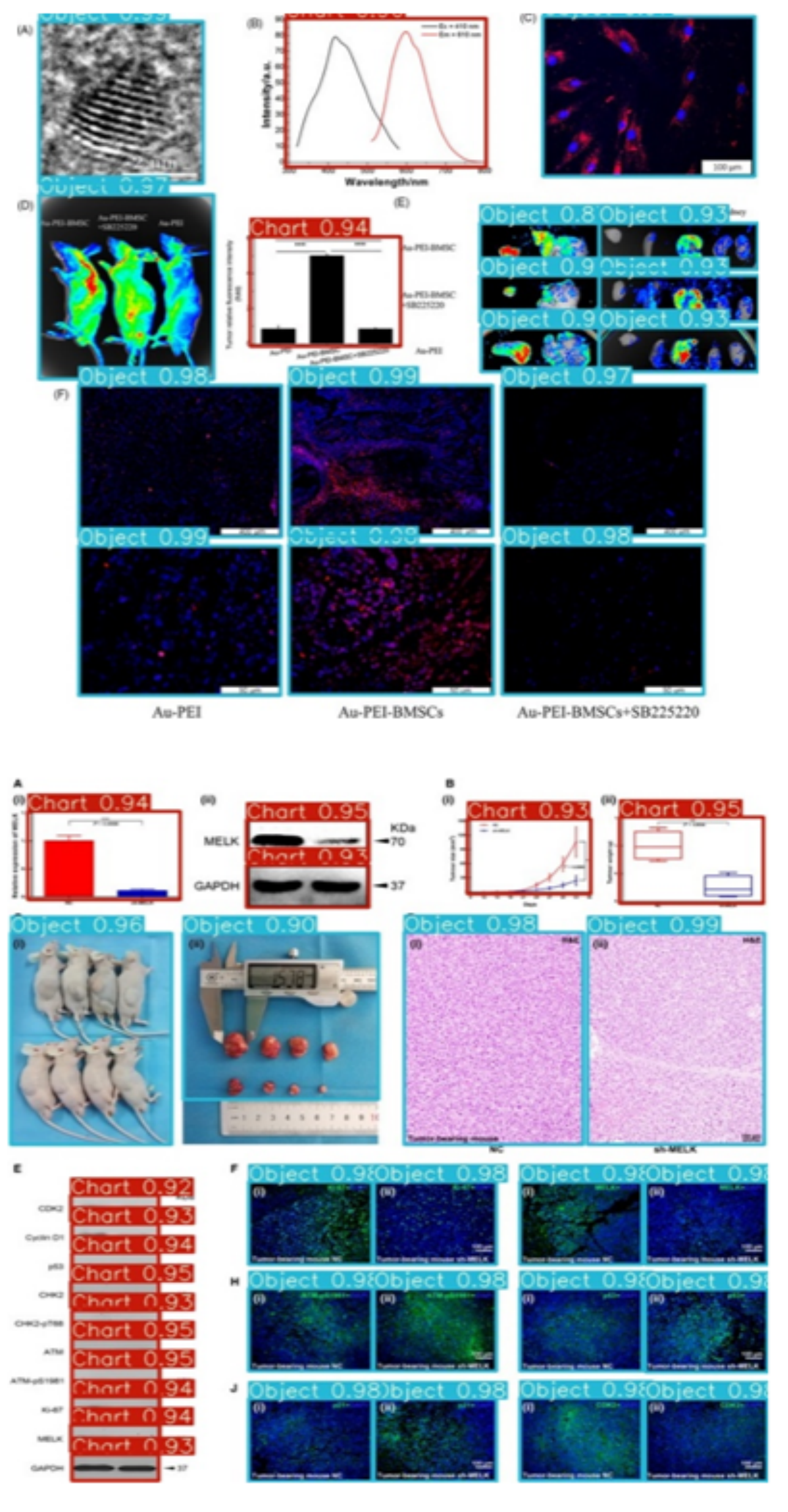
1. We trained a sub-image recognition model based on the object detection neural network.
2. We trained different visual feature embedding models for each sub-image category, realizing high-dimensional feature space representation for those sub-images.
3. We continually collect image from medical journals to expand our vector database, and have already inserted over 28 million vector into our Milvus database.

Sub-image Recognition

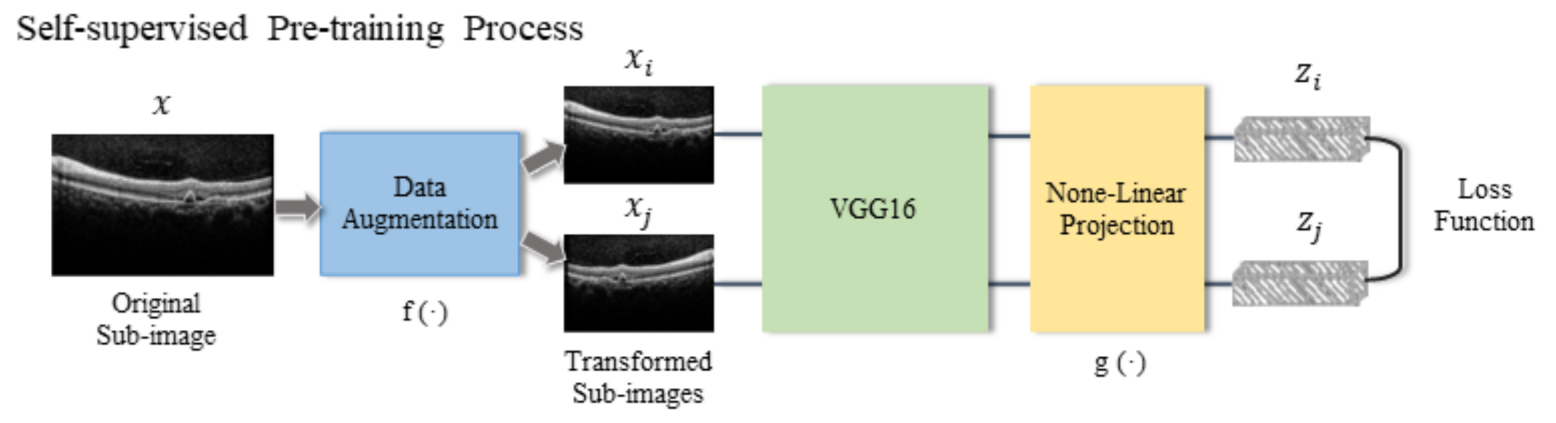
- We have built a dataset of over 200,000 images from journals in the fields of the medical and materials, and annotated the coordinates and categories of sub-images. We further trained a sub-image recognition model based on YOLO v7.
- The experimental results mentioned in tables below show that our model achieves an accuracy of 84.80% and a recall rate of 86.50%. ↓ →

Class	Test cases	Precision	Recall	mAP@.5	mAP@.5:.95
Statistical chart	754	94.40%	97.90%	97.00%	88.20%
Western blotting	557	95.70%	96.30%	96.70%	76.90%
Fluorescent staining	658	89.90%	93.80%	94.70%	86.70%
Diagrammatic sketch	54	73.00%	66.70%	73.90%	63.00%
Radiography & Angiograph	185	90.40%	87.60%	88.00%	78.60%
Physical image	20	98.90%	100.00%	99.50%	90.40%
Others	71	51.10%	63.40%	56.50%	36.30%
Average	2299	**84.80%**	**86.50%**	**86.60%**	**74.30%**

Class	Test cases	Precision	Recall	mAP@.5	mAP@.5:.95
Chart	1311	96.20%	97.50%	98.00%	84.60%
Object	917	92.60%	91.30%	93.40%	85.40%
Other	71	63.20%	53.50%	58.10%	37.50%
Average	2299	**83.70%**	**80.80%**	**83.20%**	**69.10%**



Deep Feature Embedding



- By using self-supervised pre-training, the image depth feature extraction models (each class has its own model) enhanced its ability to capture and distinguish subtle differences between sub-images in same class.
- The model embeds the input images into a latent space and outputs a 1024-dimensional vector representation. The vectorized image features are then stored into vector retrieval database (open-source Milvus vector database) for subsequent vector search and retrieval.

Class	Vector Amount
Western blotting	1,878,001
Fluorescent staining	10,497,469
Diagrammatic sketch	7,011,320
Radiography & Angiograph	3,427,590
Physical image	3,724,634
Others	1,476,649
Sum	28,015,663

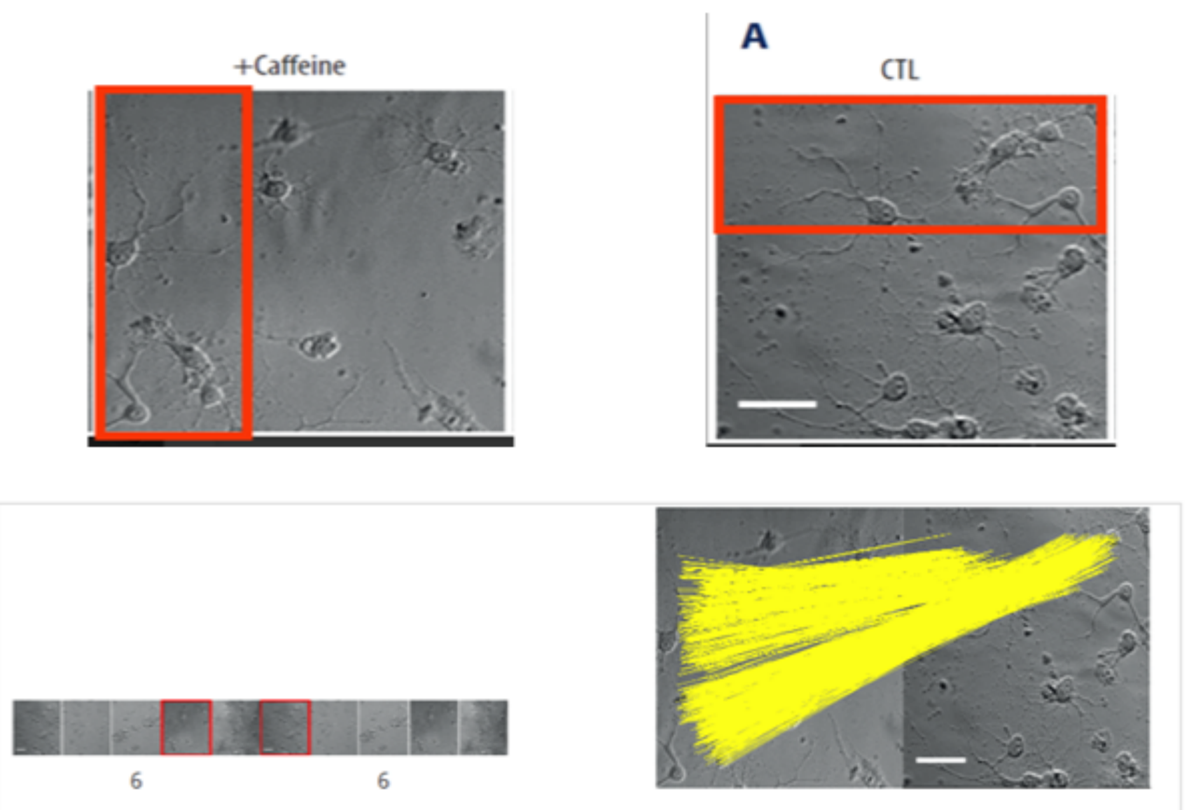
←

We have completed more than 1 million medical article automatic analysis and insert more than 28 million sub-image into vector database.

Vector Retrieval & Image Matching

- Feature point detection algorithm: ORB (Oriented FAST and Rotated BRIEF) based on the classical algorithm BRIEF (Binary Robust Independent Elementary Features)
- Feature point matching algorithm: GMS (Grid-based Motion Statistics).
- Our proposed method has been applied in the product of Wanfang Data Co., Ltd.
  - Time consumption of vector retrieval : 0.2s-0.5s per times
  - Time consumption of feature point matching: approximately 0.03s per times.
  - We constructed the test sets based on PubMed, and the results show that the overall precision rate is over 90%.

- Real Case in Retracted Article:



* Images come from retracted paper: *Caffeine Treatment Promotes Differentiation and Maturation of Hypoxic Oligodendrocytes via Counterbalancing Adenosine 1 Adenosine Receptor-Induced Calcium Overload*

- Detect Precision in Test Set:

		Others	Western blotting	Fluorescent staining	Fluorescent staining	Radiography & Angiograph	Physical image	overall
Test set 1	Number of sub-image	142	11	734	707	348	327	2269
	Number of detected for duplication	84	9	590	568	289	228	1768
	Number of accurate detection	66	7	563	560	283	213	1692
Test set 2	Number of sub-image	225	31	1043	815	487	360	2961
	Number of detected for duplication	173	18	934	638	347	240	2350
	Number of accurate detection	143	17	904	521	307	236	2128
	Precision	0.8061	0.8611	0.9608	0.9015	0.9319	0.9577	0.9032