Partial video copy detection: A large-scale dataset & Performance characterization of 2D CNN features

PhD student:
Van-Hao Le

Supervisors:
Hubert Cardot
Mathieu Delalandre

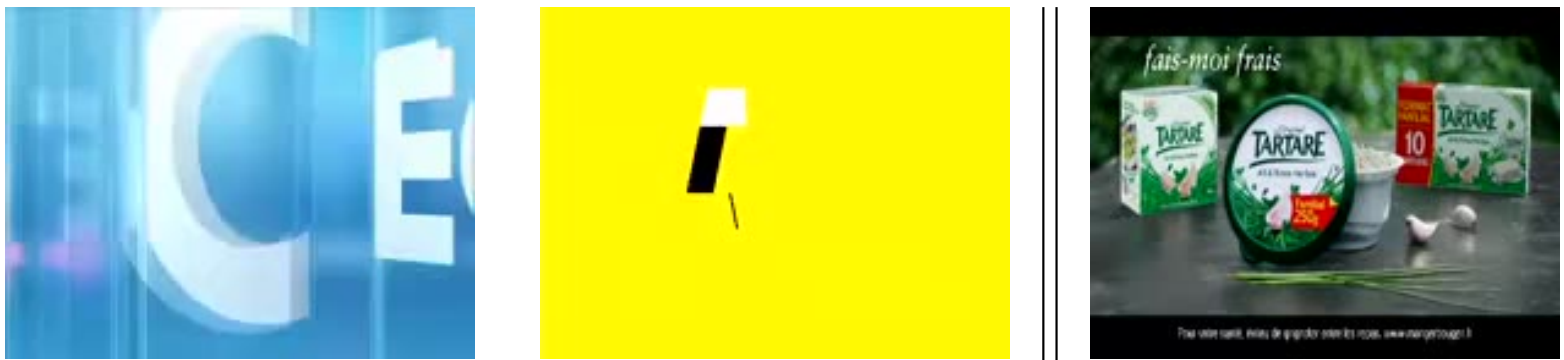
LIFAT Laboratory, University of Tours, France

May 23^{th}, 2024

Outline

Introduction (1/3)

- Partial video copy detection (PVCD) aims at finding short segment(s) which have transformed into long video(s).



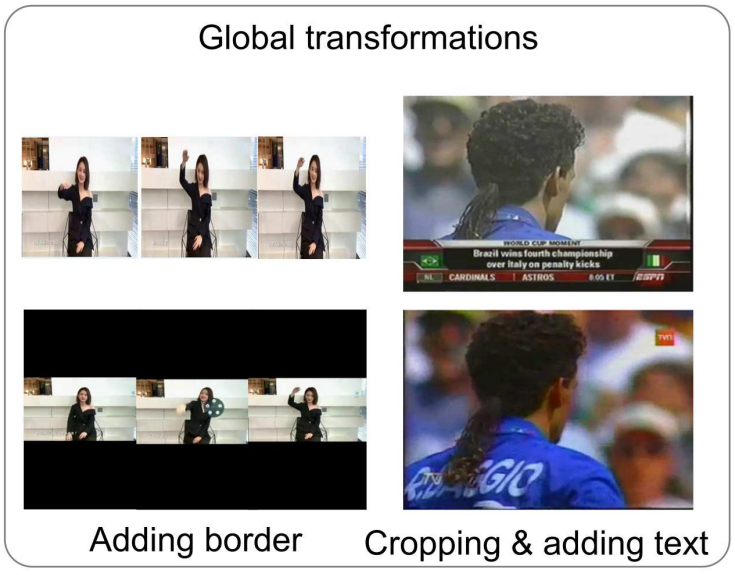
short segments / reference videos | a long video

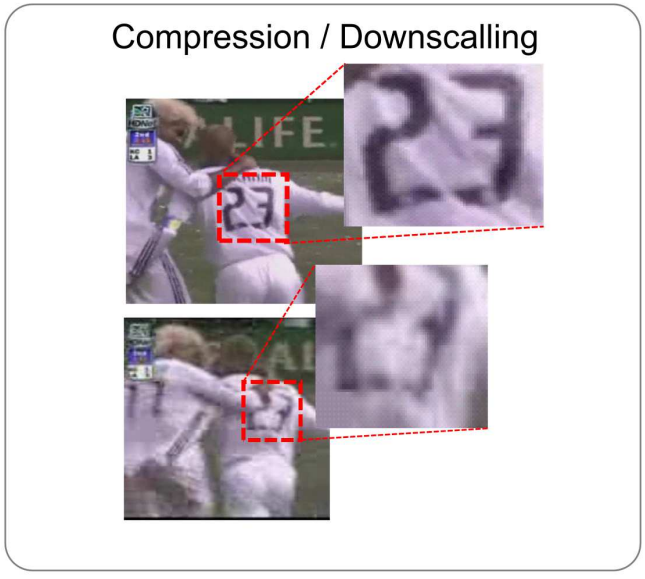
- PVCD includes several application domains (copyright protection, video retrieval, recommendation, etc.) [Han, 2021; Jiang, 2021; Tan, 2022].
- It is a key topic in the Computer Vision field including research [Jiang, 2019; Han, 2021; Tan, 2022] & practical[1] aspects.

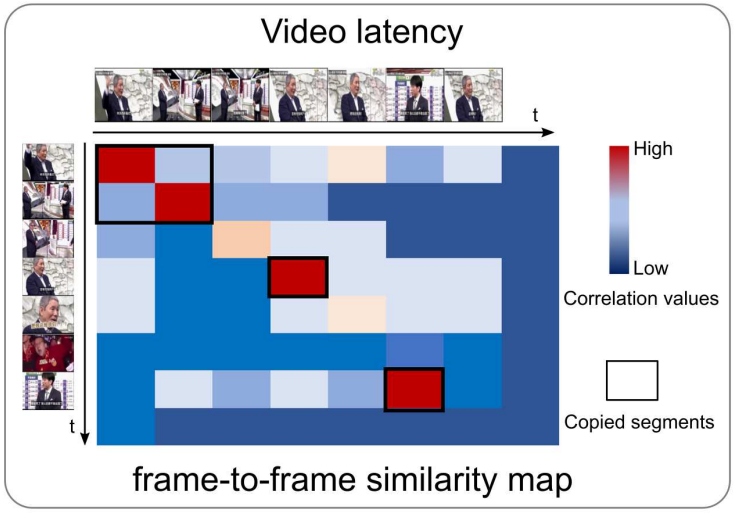
[1] Meta AI Video Similarity Challenge 2023 (100,000 USD): https://www.drivendata.org

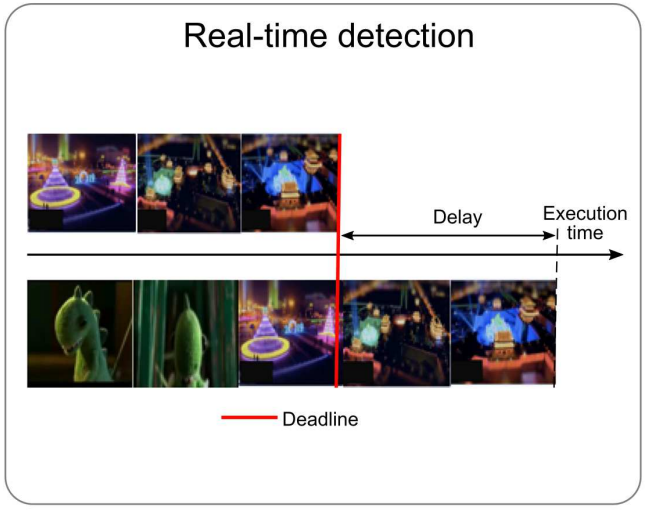
Introduction (2/3)

- PVCD addresses different detection problems [Kordopatis-Zilos, 2017; Tan, 2022; Zhao, 2022].









Introduction (3/3)

- Needs of public datasets for performance evaluation
  - 4 datasets have been published with protocols.
  - Protocols process mainly from Web videos with real degradations and manual annotation.

Datasets	CC_WEB [Wu2007]	SVD [Jiang2019]	**VCDB** [Jiang, 2014]	**VCSL** [**He2022**]
No longer in use[2]	✓			
Small size	✓	✓		
Unbalanced data			✓	
Costly annotation[3]		✓	✓	✓
None / Little control of degradations	✓	✓	✓	✓
None frame-level annotation	✓	✓	✓	✓

- PVCD uses 2D CNN features, its performance has been little discussed [Kordopatis-Zilos, 2017; Hu, 2019].

[2] The mAP score of **CC_WEB** reached at $\simeq$ 0.976 [Kordopatis, 2017]

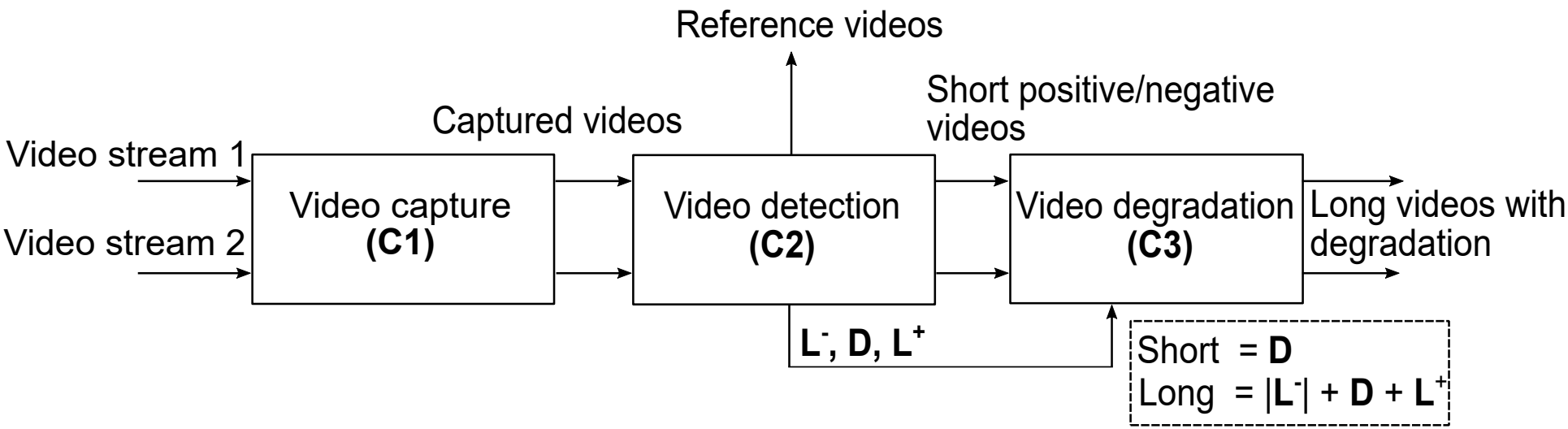
[3] About $\simeq$ 700 man-hours for labeling 528 positive & 100,000 negative videos in **VCDB**

Objectives of our work for PVCD

- **(1)** We propose a TV-based protocol to design a new dataset:
  - scalable with balanced data (positive / negative distribution),
  - applicable to several performance characterization tasks,
  - with a frame-level annotation for timestamping.
- **(2)** We characterize the performance of PVCD systems.
  - We report large-scale experiments to characterize 2D CNN features,
  - We propose a method for the characterization of key-frames.

(1) STVD dataset with TV-based protocols (1/8)

- We propose a new protocol able to:
  - separate captures similar to [Joly, 2007; Law-To, 2007],
  - avoid a full-search strategy with a metadata support,
  - extract real / true-life partial copies,
  - have a fine control of degradations.
- The system architecture includes 3 main components:
  - **(C1)** captures videos with a TV workstation,
  - **(C2)** annotates reference, positive and negative videos,
  - **(C3)** generates test sets with synthetic degradations.



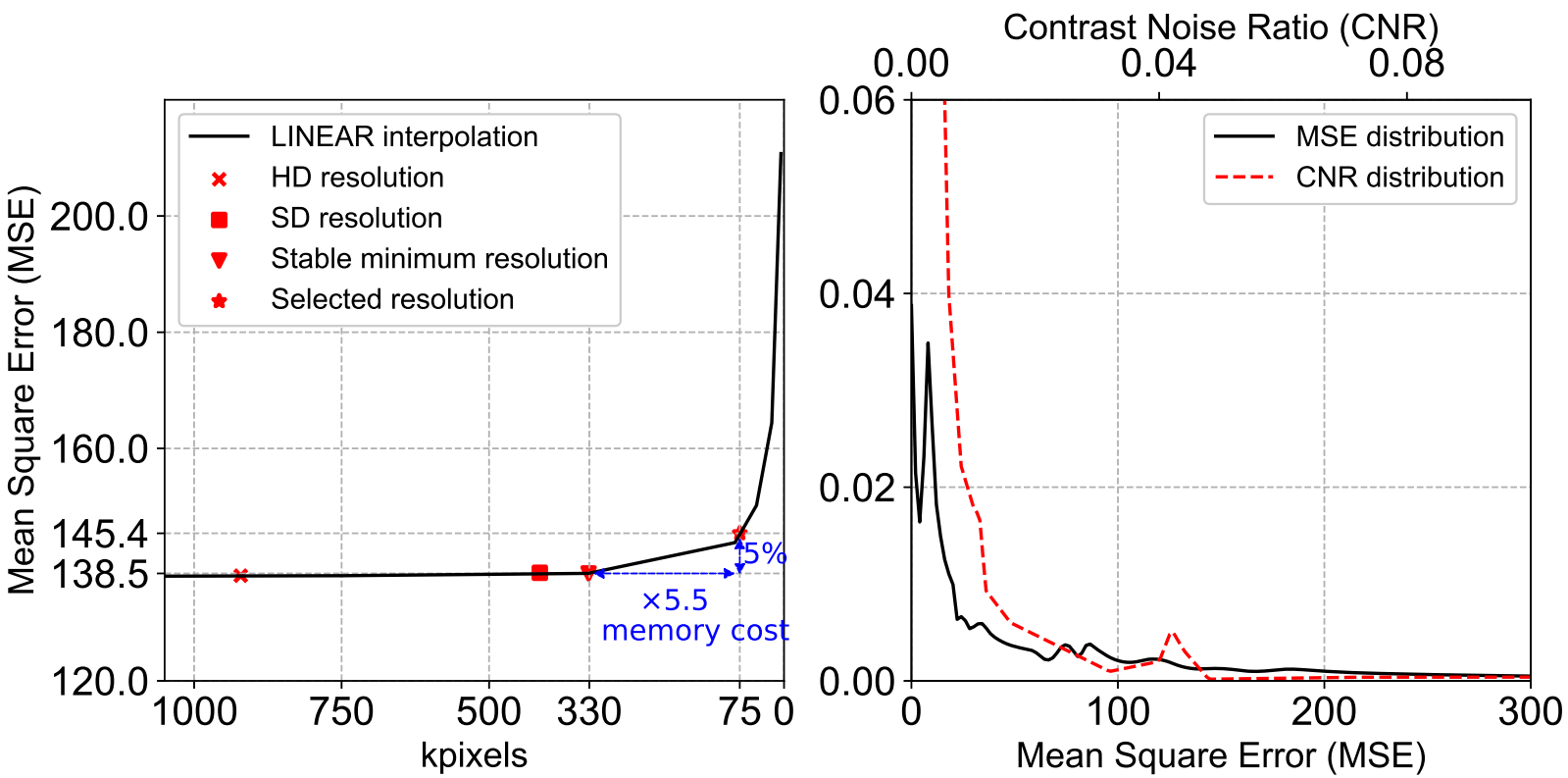
(1) STVD dataset with TV-based protocols (2/8)

Video capture **(C1)**

- Capture French DTT using a TV workstation setup:

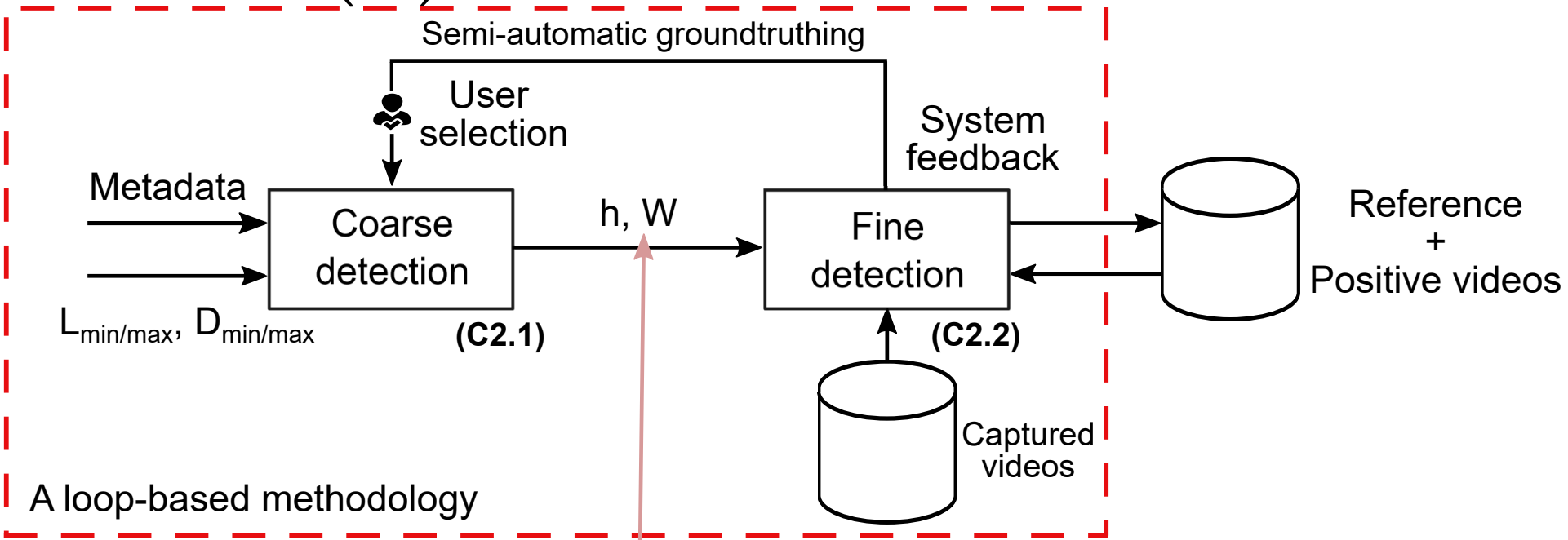
Channels per month	Daily file					Total size (TB)	
	Resolution	kbps	Aspect ratio	Length	FPS	Files	Size
8	320×240	560	4 : 3	20 h	30	720	3.46

- An optimized strategy of memory cost and a noise level



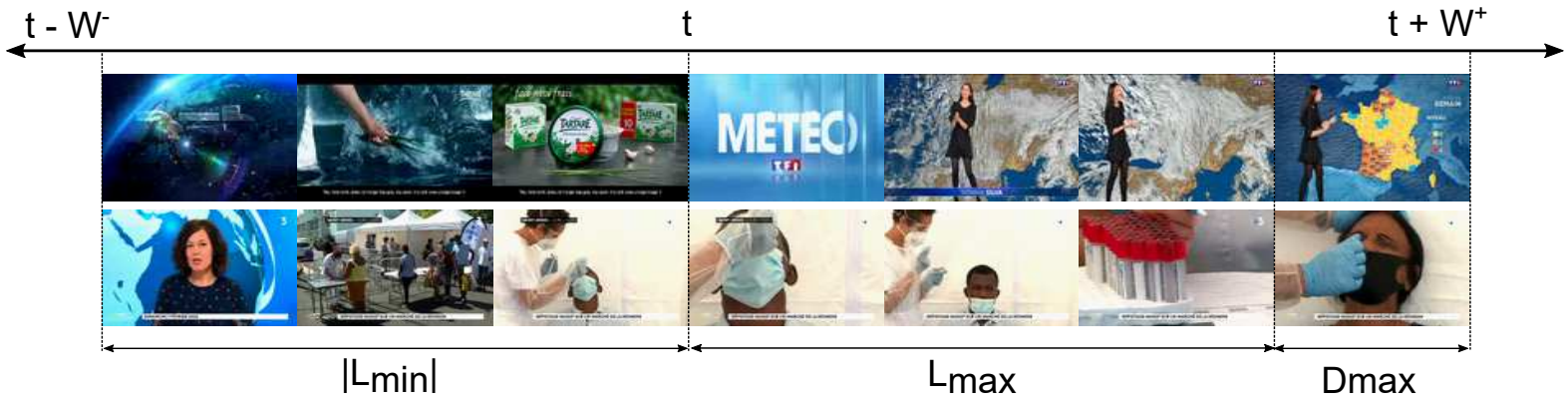
(1) STVD dataset with TV-based protocols (3/8)

Video detection **(C2)**



Hashcode: h = hashing(channel + normalized_title) = 2c76 ... 93fd

Capture window: $W = W^- + W^+$ where $W^- = |L_{min}|$, $W^+ = D_{max} + L_{max}$



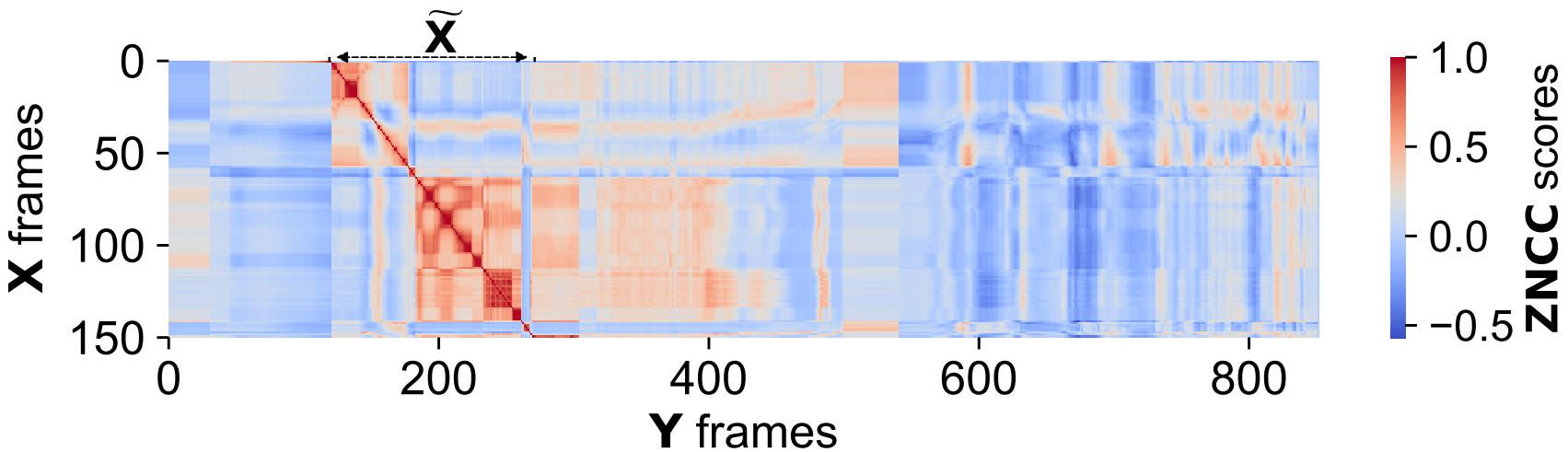
(1) STVD dataset with TV-based protocols (4/8)

Video detection **(C2)** - Zero-mean Normalized Cross Correlation

- **ZNCC** between 2 images $\mathbf{I}, \mathbf{I}^*$ with means $\bar{\mathbf{I}}, \bar{\mathbf{I}}^*$ and standard deviations $\sigma_{\mathbf{I}}, \sigma_{\mathbf{I}^*}$ are given as follows.

$$\mathbf{ZNCC}(\mathbf{I}, \mathbf{I}^*) = \frac{\sum_{\forall \mathbf{x}} (\mathbf{I}(\mathbf{x}) - \bar{\mathbf{I}})(\mathbf{I}^*(\mathbf{x}) - \bar{\mathbf{I}}^*)}{\sigma_{\mathbf{I}} \sigma_{\mathbf{I}^*}}$$

- A visual frame-matching between two videos, X, Y. Accelerate with GPU computing.

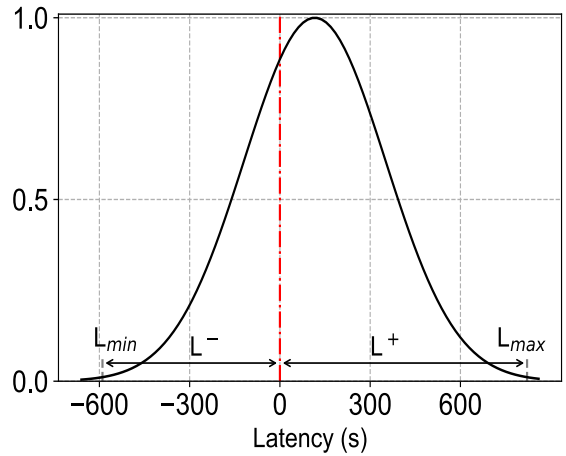


(1) STVD dataset with TV-based protocols (5/8)

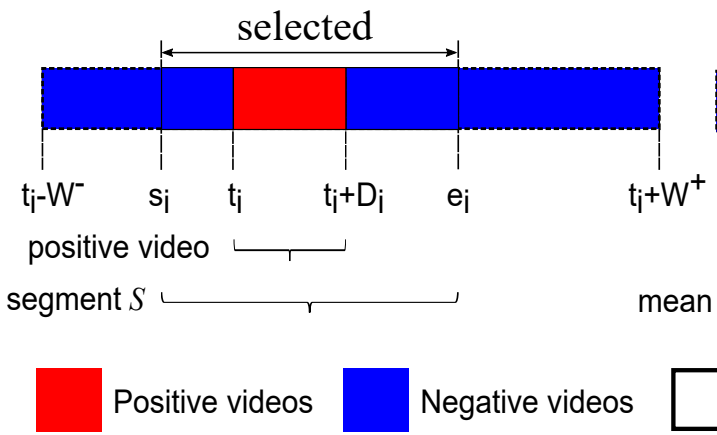
Video degradation **(C3)**

- **(C2)** extracts real-life partial video copies:
  - applying the latency model detected with **C2**,
  - detecting the overlapping cases for merging.

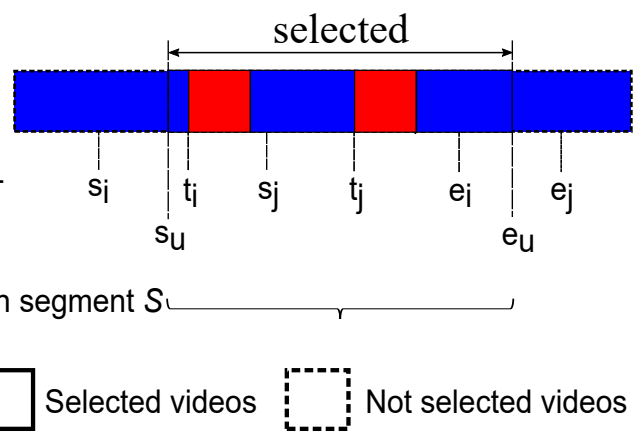
An approximately Gaussian distribution



A normal case of a partial copy



An overlapping case



(1) STVD dataset with TV-based protocols (6/8)

Video degradation **(C3)**

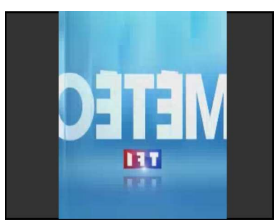
- The final dataset with degradation methods: 6 sets (A-F)

	Test set	Video cut	Downscaling	Compression	Flipping	Rotating	Black-border	Video speeding
		T_0	T_1	T_2	T_3	T_4	T_5	T_6
A	Root capture	✓						
B	'Hello World'	✓	✓	✓				
C	Pixel attack	✓	✓	✓				
D	Global transformations	✓	✓	✓	✓	✓	✓	
E	Video speeding	✓	✓	✓				✓
F	Combination	✓	✓	✓	✓	✓	✓	✓

Set A

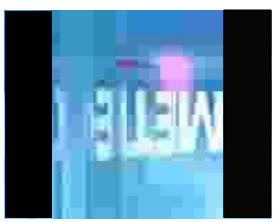
Set B

Set C



Set D


Set E


Set F

(1) STVD dataset with TV-based protocols (7/8)

- STVD statistics: included 6 test sets in **C3**

	(C1)		(C2)		(C3)	
	Channels	Duration	Videos	Duration	Videos	Duration
Positive videos	8	4,800 h	3,780	6 h	19,280	2,515 h
Negative videos	16	9,600 h	12,165	21 h	64,040	8,145 h

- The comparisons between STVD versus VCDB, VCSL

Datasets	**VCDB** [Jiang, 2014]	**VCSL** [He, 2022]	**STVD** Ours
Reference videos	28	122	243
Positive videos	528	9,207	19,280
Negative videos	100,000	N/A	64,040
Duration (h)	2,000	N/A	10,660
Positive pairs	9 K	281 K	1,688 K
Noise characterization	real noise	real noise	noise-free
Annotation cost (m-h)	700	20,000	105
Frame-level annotation	✗	✗	✓

(h): hours, (m-h): man-hours and N/A: not available

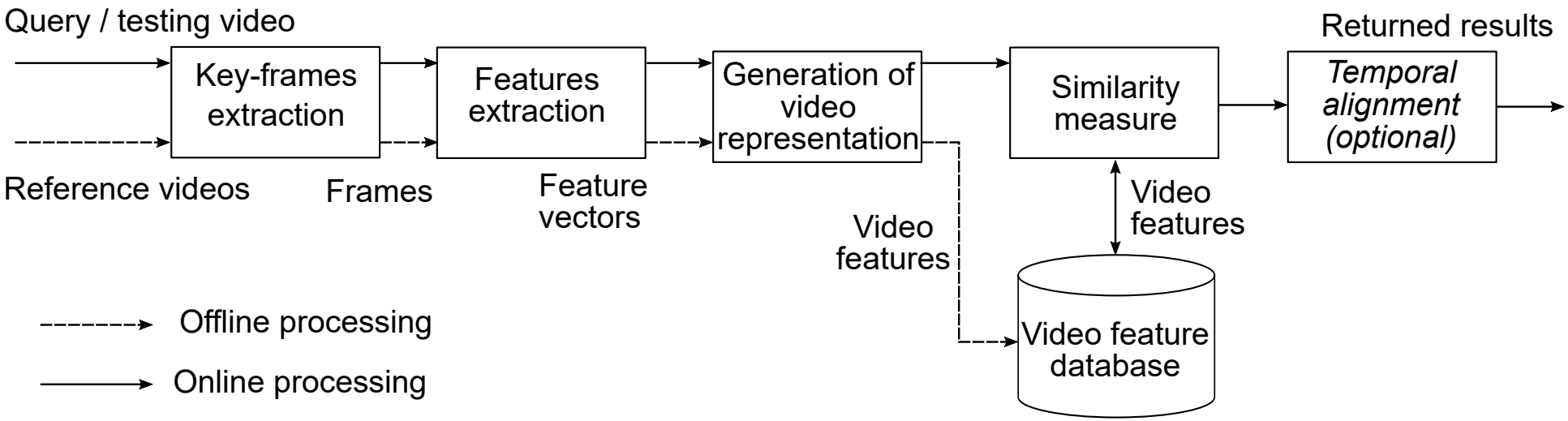
(1) STVD dataset with TV-based protocols (8/8)

- We proposed a protocol to design a dataset for PVCD:
  - ensuring the dataset scalability with balanced data,
  - offering a fine control of degradation,
  - able to annotate at a frame-level for groundtruthing.
- Our protocol can be adapted to other problems:
  - computer vision / fact checking,
  - operational research / parallel machine scheduling.
- We published a large-scale dataset, named STVD that
  - includes $\simeq$83K videos, 10K hours, 1,688 K positive pairs,
  - is publicly available at https://dataset-stvd.univ-tours.fr/(PVCD),
  - has been promoted in the CV/ML research community[4].

[4] cited papers: 3, CV/ML web platforms: 6, research mailings: 14, email contacts: 40.

(2) Performance characterization of 2D CNN features (1/8)

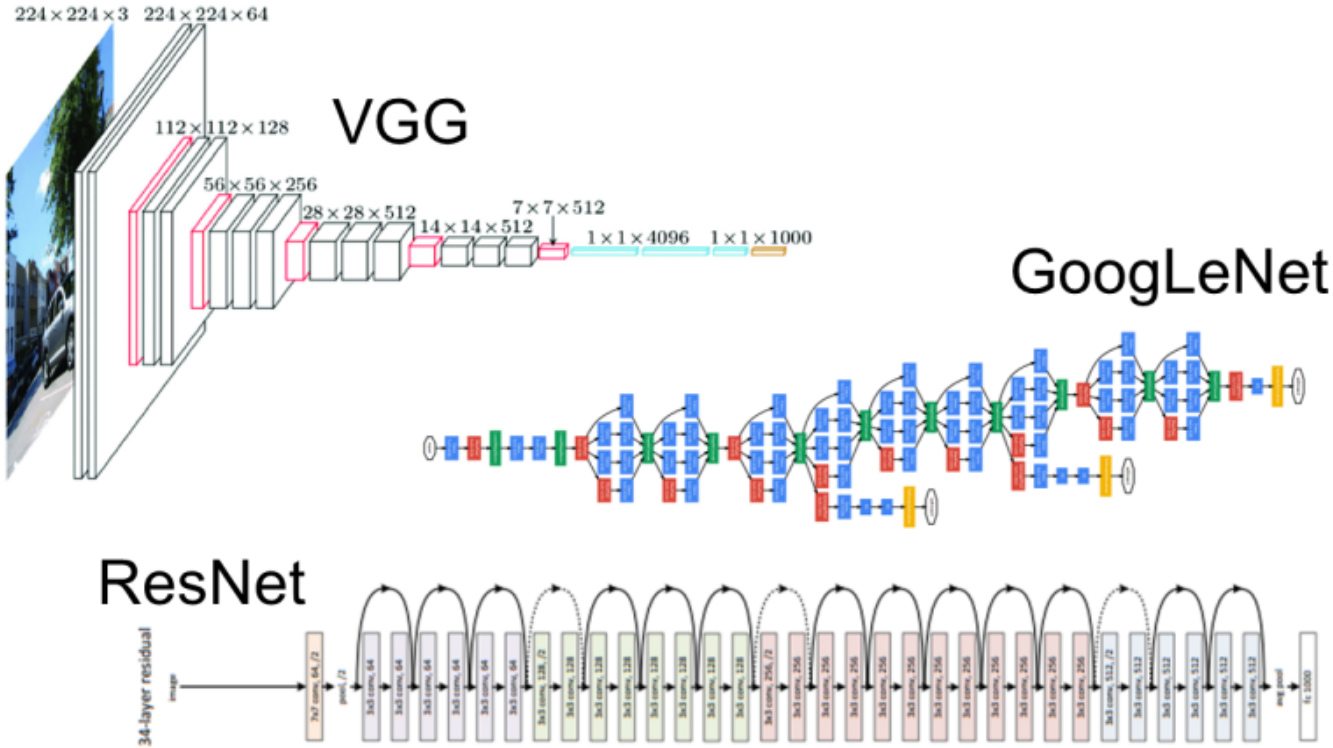
- A general framework for the PVCD system.



- No single approach seems to be optimal for all PVCD [Shen, 2020].
  - key-frames extraction: fixed by FPS, learning methods,
  - feature extraction: selected CNN (VGG, ResNet), methods (FC, MAC),
  - video representation: video-level (single vector), frame-level (multiple),
  - similarity measures: Euclidean distance, Cosine similarity,
  - temporal alignment: dynamic programing, temporal network.

(2) Performance characterization of 2D CNN features (2/8)

- Various pre-trained CNN models: AlexNet, VGG, InceptionNet, ResNet.



- Various versions among models: VGG-13/16/19, ResNet-50/100/152.
- Selected models: VGG-16, Inception-v1, ResNet50.

(2) Performance characterization of 2D CNN features (3/8)

- Various extraction methods: Last FC, MAC, R-MAC[5].

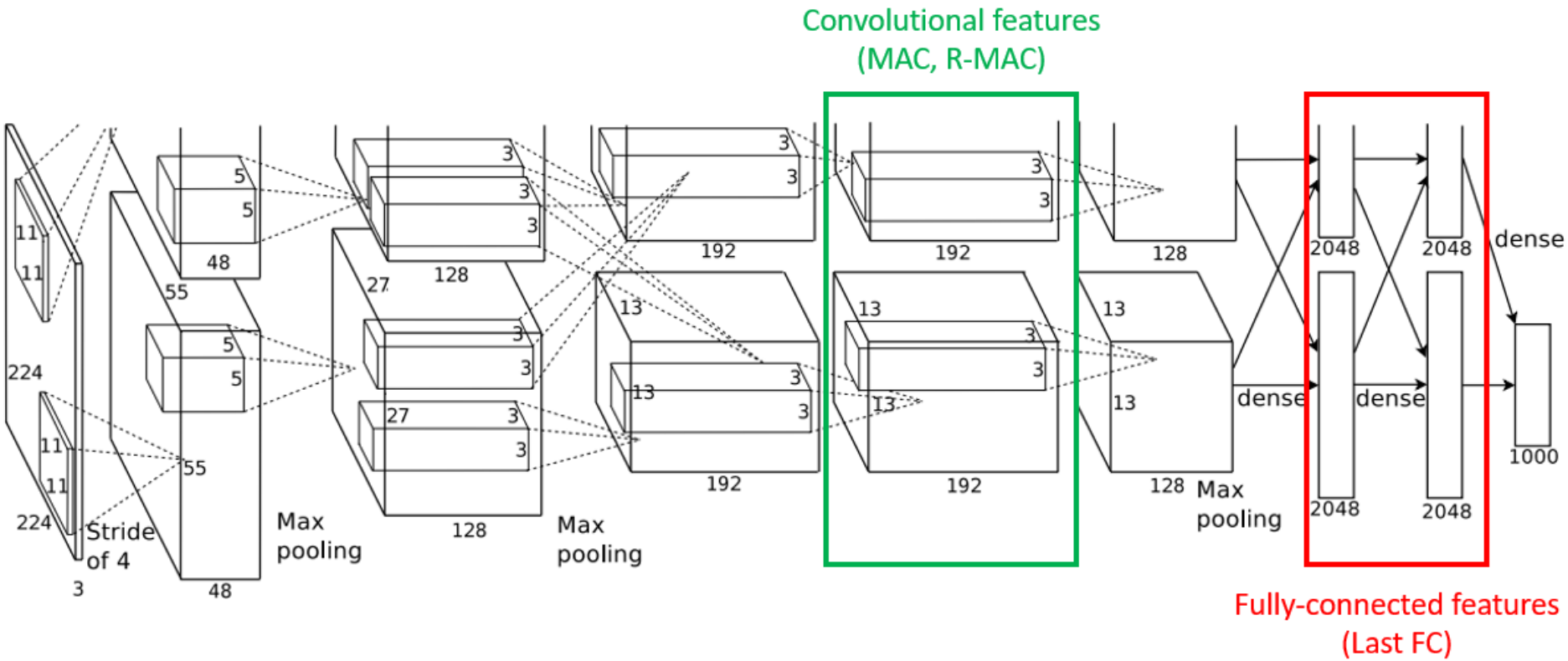
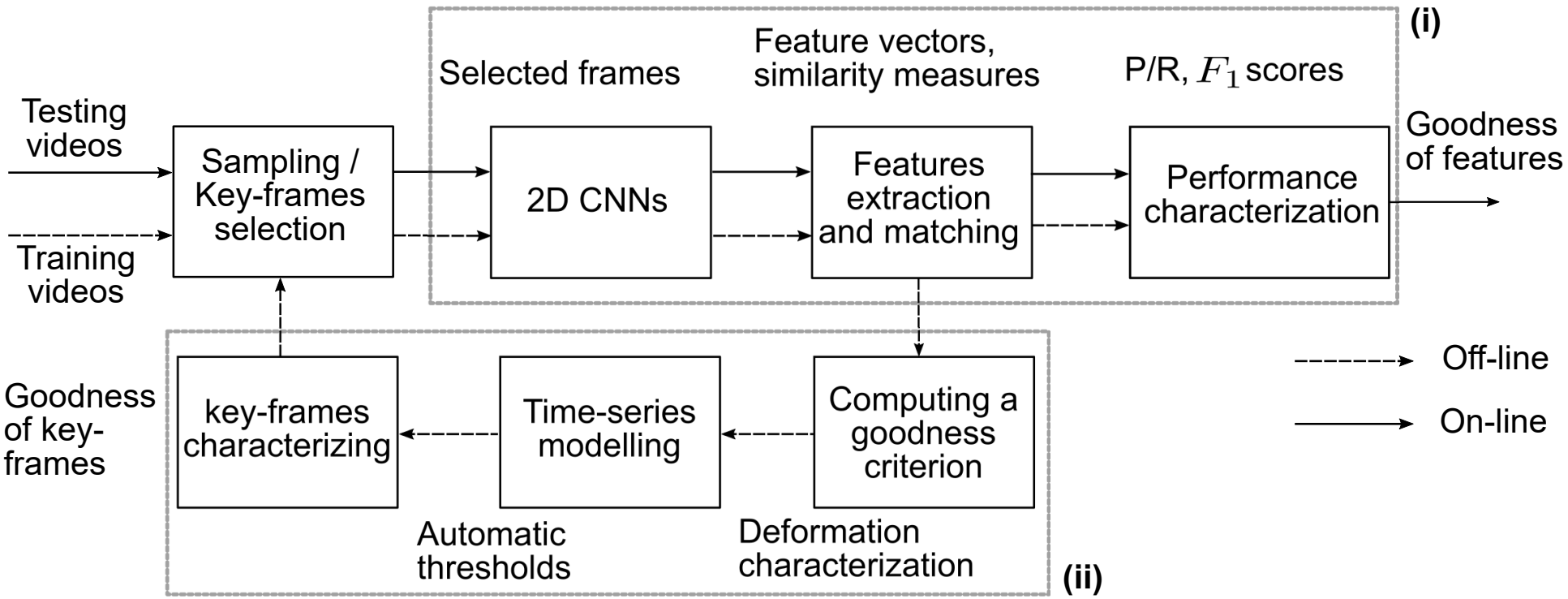


Fig. 1 An example of feature extraction on AlexNet [Krizhevsky, 2012].

[5] Maximum Activations of Convolutions, Regional Maximum Activations of Convolutions

(2) Performance characterization of 2D CNN features (4/8)

- **(i)** We report large-scale experiments to compare 2D CNN features:
  - comparison of 9 types of features with standard P/R and F_1 scores,
  - our conclusions and results are consistent with the CV state-of-the-art.
- **(ii)** We propose a method to characterize the goodness of key-frames:
  - a goodness criterion, time-series modelling & key-frames characterization,
  - highlights the difficulties of 2D CNN features for specific degradations.



(2) Performance characterization of 2D CNN features (5/8)

- **Video datasets**: VCDB [Jiang, 2016], SVD [Jiang, 2019], VCSL [He, 2022], and STVD [Le, 2022] which was selected[6].
- Divide randomly negative and positive videos into 2 sets as follows.

Tab. 1 Pre-processing of the STVD dataset for our experiments.

Videos	**No**	**Duration (h)**	**FPS**	**60% training**	**40% testing**	**Total**
Negative videos	12,165	1,545	0.08	259,050	172,700	431,750
Positive videos	3,869	415				
Copied segments	4,436	7.5	1	16,200	10,800	27,000
	Total: **(i)** 458,750 f $\simeq$ 7 GiB of 4,096-F dimension, **(ii)** 745,050 f					

- **Protocols**: standard P/R, F_1 scores using the Cosine similarity.
- Copied segment: **(i)** FPS=1 (16,200 f), **(ii)** FPS = 30 (486,000 f)

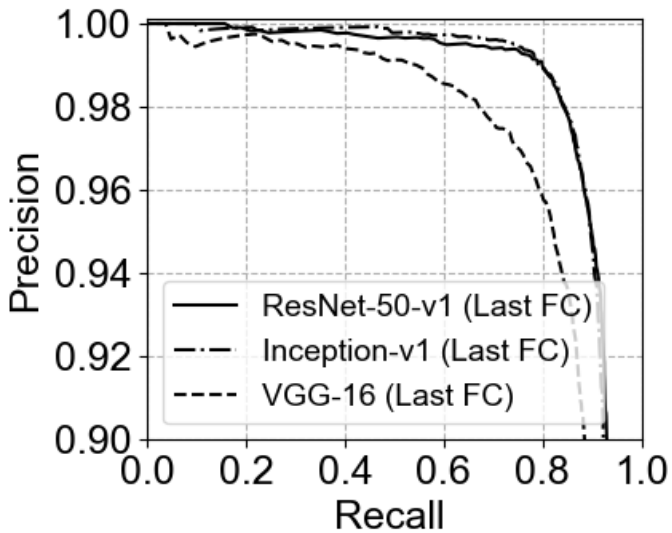
[6] fine control of degradations, large-scale, balanced positive / negative distribution, accurate timestamping

(2) Performance characterization of 2D CNN features (6/8)

- Large-scale experiments to characterize these 2D CNN features:
  - 9 CNN features (3 models $\times$ 3 methods[7]),
  - 4.4 **M** vectors, 445 **B** matchings.
- Comparison of 2D CNN features results.

Tab. 2 Top F_1 scores

	Last FC	**MAC**	**R-MAC**
ResNet50-v1	**0.926**	0.828	0.823
Inception-v1	**0.923**	0.738	0.782
VGG-16	0.894	**0.922**	0.918



- Our results highlight and be consistent with the state-of-the-art:
  - the separability of features is not achieved (even if $F_1 \simeq 0.93$),
  - recent 2D CNN (ResNet-50) outperform [He, 2016],
  - correlation between 2D CNN & methods (VGG & MAC) [Cools, 2022].

[7] Last Fully Connected (Last FC), Maximum Activations of Convolutions (MAC) and Regional-MAC (R-MAC)

(2) Performance characterization of 2D CNN features (7/8)

- **Goodness criterion**: $\phi(X) \geq 0$ using Cosine similarity (SC) is given

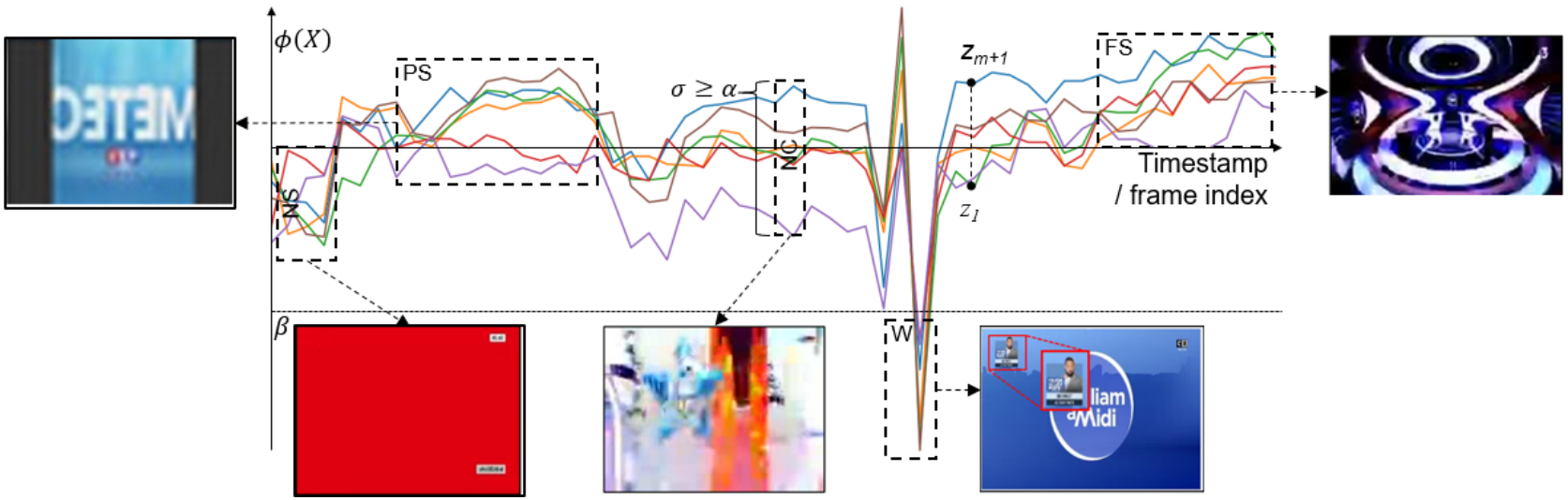
X is a feature vector

$$\phi(X) = SC_{\min}(X, \{\tilde{X}_1, \ldots, \tilde{X}_m\}) - SC_{\max}(X, \{Y_1, \ldots, Y_{n_1}\}, \{X_1^*, \ldots, X_{n_2}^*\})$$

$\tilde{X}$ is near-duplicate of X

Y is negative,
$X^* \neq X$ has a different reference

- **Time-series modelling**:



- **Key-frame categorization**: in 5 categories based on automatic thresholds, Not Consistent (NC), Worst (W), Not Separable (NS), Partially Separable (PS), Fully Separable (FS).

(2) Performance characterization of 2D CNN features (8/8)

- A proposed method to characterize key-frames using 2D CNN features:
  - a goodness criterion, time-series modelling & key-frames categorization,
  - $\simeq 0.8$ **M** feature vectors, $\simeq 244$ **B** matchings.
- Results of key-frames categorization.

(NC-Not consistent, W-Worst, NS-Not separable, PS-Partially Separable, FS-Fully Separable.)

Total	NC	W	NS	PS	FS
100 %	13.7 %	8.2 %	65 %	9.6 %	3.5 %

- Our results highlight:
  - an 'easy' categorization of key-frames,
  - a quantitative analysis of the goodness of key-frames,
  - only a small amount of 'good' key-frames ($\simeq 13\%$ in PS, FS),
  - difficulties to detect 'bad' key-frames ($\simeq 22\%$ in NC, W).

'good' key-frames

foreground / background


symmetrical

'bad' key-frames


blurred


near-constant

almost-duplicate

Conclusions & Perspectives

- Our contributions for performance characterization of 2D CNN features
  - We report large-scale experiments to characterize 2D CNN features:
    - 9 CNN features, 4.4 M vectors, 445 B matchings,
    - ResNet-50 outperforms, correlation CNN & methods.
  - We propose a method for the characterization of key-frames:
    - goodness criterion, time-series, categorization,
    - 0.8 M vectors, 244 B matchings,
    - categorization and analysis, performance limits of features.
- Our perspectives to further improve the PVCD performance:
  - protocol of automatic labeling for scalable frame classification,
  - robust key-frame selection and learning of 2D CNN features.

Thank you for your attention!

Appendix - Test sets

- The degradation methods used to generate the STVD.

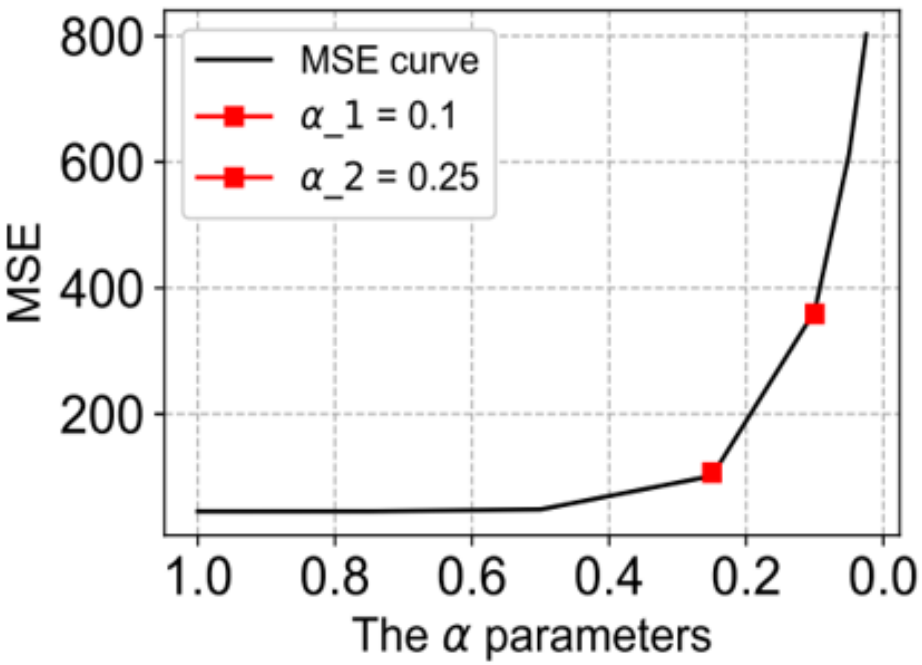
Label	Method	Parameters
T_0	video cut	introduced latency $\lvert\mathbf{L}^-\rvert, \mathbf{L}^+$
T_1	down-scaling	$\alpha \in [0.1, 0.9]$
T_2	compression	video bitrate $\frac{1}{\beta}$ with $\beta \in [1, 80]$ kbps
T_3	flipping	applies randomly (yes/no)
T_4	rotating	rotation $\in \{0, \frac{\Pi}{2}, \Pi, \frac{3}{2}\Pi\}$
T_5	black border & stretching	aspect ratio $\frac{w}{h} \in \{0.46, 1.78, 2.17, etc.\}$
T_6	video speeding	FPS $\in [15, 25]$

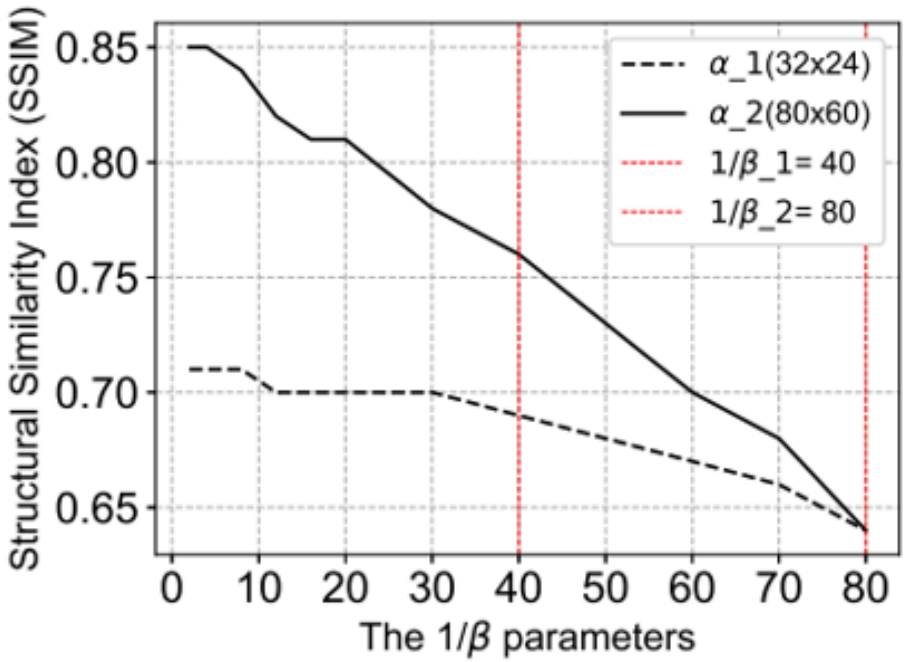
- 6 test sets are created considering different aspects.

Test set	T_0	T_{1-2}	$\alpha \in$	$\beta \in$	T_{3-5}	T_6	Description
Set A	$\surd$						Root capture for tuning
Set B	$\surd$	$\surd$	$[0.25, 0.9[$	$[1, 40[$			"Hello world" test set
Set C	$\surd$	$\surd$	$[0.1, 0.25]$	$[40, 80]$			Pixel attack with scalability
Set D	$\surd$	$\surd$	0.6	20	$\surd$		Global transformations with scalability
Set E	$\surd$	$\surd$	0.6	20		$\surd$	Video speeding with scalability
Set F	$\surd$	$\surd$	$[0.1, 0.25]$	$[40, 80]$	$\surd$	$\surd$	Combination of sets C, D and E

Appendix - the alpha / beta parameters

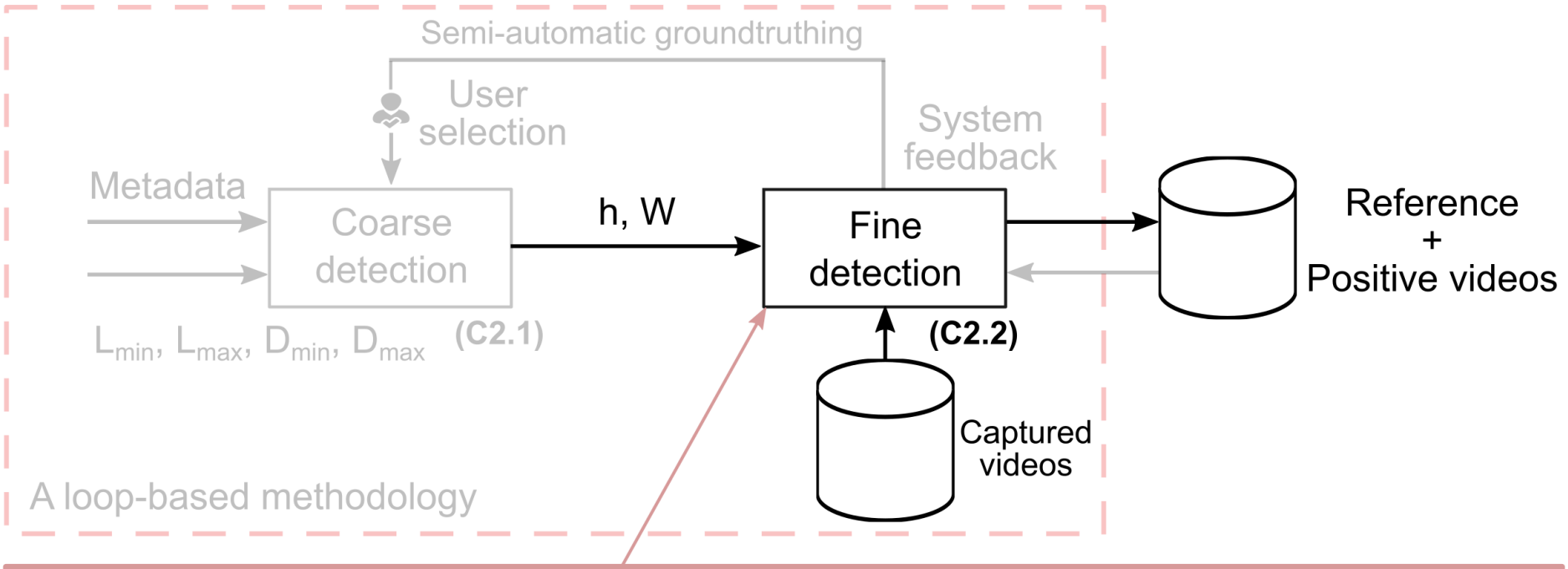
- The α and β parameters are selected based on the results.



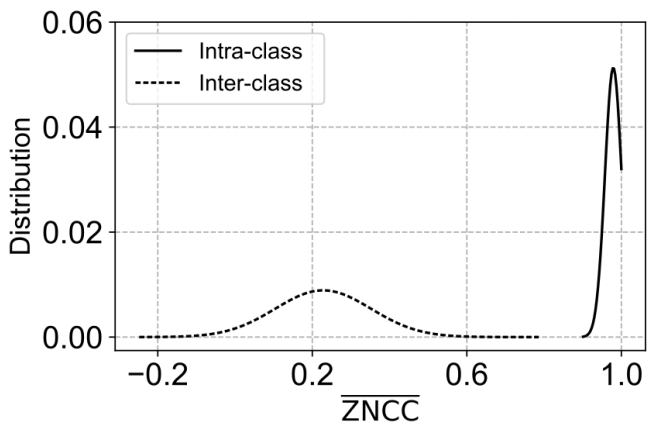


Appendix - ZNCC

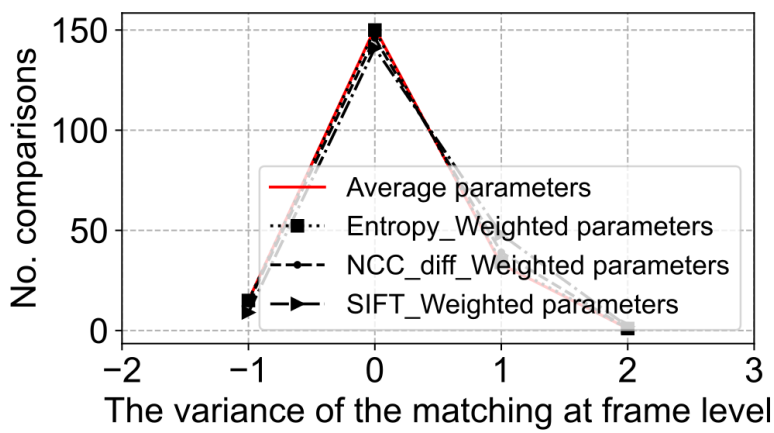
Video detection **(C2)**



The separation between the inter-class and intra-class distribution



The accuracy of weighting methods for timestamping



References

Cools, A., M.A. Belarbi, and S.A. Mahmoudi (2022). "A Comparative Study of Reduction Methods Applied on a Convolutional Neural Network". In: *Electronics* 11, p. 1422.

Han, Z. (2021). "Video similarity and alignment learning on partial video copy detection". In: *ACM International Conference on Multimedia (MM)*, pp. 4165–4173.

He, K. (2016). "Deep residual learning for image recognition". In: *Conference on computer vision and pattern recognition (CVPR)*, pp. 770–778.

He, S. (2022). "A Large-scale Comprehensive Dataset and Copy-overlap Aware Evaluation Protocol for Segment-level Video Copy Detection". In: *Computer Vision and Pattern Recognition (CVPR)*, pp. 21086–21095.

Hu, Y. et al (2019). "STRNN: End-to-end deep learning framework for video partial copy detection". In: *Journal of Physics: Conference Series*.

Jiang, C. (2021). "Learning segment similarity and alignment in large-scale content based video retrieval". In: *ACM International Conference on Multimedia (MM)*, pp. 1618–1626.

Jiang, Q.Y. (2019). "SVD: A large-scale short video dataset for near-duplicate video retrieval". In: *International Conference on Computer Vision (ICCV)*, pp. 5281–5289.

Jiang, Y.G. and J. Wang (2016). "Partial copy detection in videos: A benchmark and an evaluation of popular methods". In: *IEEE Transactions on Big Data* 2.1, pp. 32–42.

Jiang, Y.G. et al (2014). "VCDB: a large-scale database for partial copy detection in videos". In: *European Conference on Computer Vision (ECCV)*.

References

Joly, A. et al (2007). "Content-based copy retrieval using distortion-based probabilistic similarity search". In: *Transactions on Multimedia*.

Kordopatis-Zilos, G. (2017). "Near-duplicate video retrieval with deep metric learning". In: *International Conference on Computer Vision Workshops (ICCV)*, pp. 347–356.

Krizhevsky, A., I. Sutskever, and G.E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Conference and Workshop on Neural Information Processing Systems (NeurIPS)*. Vol. 25.

Law-To, J. et al (2007). "Video copy detection: a comparative study". In: *International Conference on Image and Video Retrieval (CIVR)*.

Le, V.H., M. Delalandre, and D. Conte (2022). "A large-Scale TV Dataset for partial video copy detection". In: *International Conference on Image Analysis and Processing (ICIAP)*. Vol. 13233. Lecture Notes in Computer Science (LNCS), pp. 388–399.

Shen, L. et al (2020). "Advance on large scale near-duplicate video retrieval". In: *Frontiers of Computer Science (Front. Comput. Sci.)* 14.5, pp. 1–24.

Tan, W., H. Guo, and R. Liu (2022). "A fast partial video copy detection using KNN and global feature database". In: *Winter Conference on Applications of Computer Vision (WACV)*, pp. 2191–2199.

Zhao, G. (2022). "STAR-GNN: spatial-temporal video representation for content-based retrieval". In: *International Conference on Multimedia and Expo (ICME)*, pp. 01–06.