Self-Supervised Video Copy Localization with Regional Token Representation: Supplementary Material

Minlong Lu, Yichen Lu, Siwei Nie, Xudong Yang, and Xiaobo Zhang

Ant Group, Hangzhou, China
{luminlong.lml,luyichen.lyc,jiegang.yxd,ayou.zxb}@antgroup.com

1 Comparison with Supervised Models by Categories

In Section *Comparison with Supervised Training* of the main text, we compare our self-supervised model (Ours-ssl) with its supervised counterpart (Ours-sup). In this section, we provide the comparison of segment-level F-scores for each topic categories on the VCSL dataset, as shown in Table 1.

Table 1: Performance comparison by categories on the VCSL dataset using our RTR feature.

	show	game	music	news	sport	life	ad	anim-ation	film	tv	all regu-lar	kichi-ku	all
Ours-sup	87.91	95.85	87.74	92.30	63.22	67.70	74.83	96.14	74.08	92.97	73.03	62.97	69.71
Ours-ssl	93.03	95.57	86.93	92.97	63.30	66.27	75.56	95.19	73.86	94.43	73.11	57.24	68.83
Ours-ft	86.88	95.90	89.25	91.50	66.09	71.83	76.05	96.73	75.51	95.27	75.02	63.92	71.56

The performance of Ours-ssl is comparable to the Ours-sup model across regular topic categories, but Ours-ssl model underperforms in the "kichiku" category. This can be attributed to the nature of "kichiku" videos, which remix or mash up content from various sources and often contain repetitive and rapid temporal edits, which are challenging for an algorithm to mimic effectively.

2 Comparison with Competition Winners: More Details

In Section *Comparison with Competition Winners* of the main text, we compare our model with recent competition winners. The strategies used to cope with picture-in-picture scenarios in the winning solutions [3, 4, 7] from recent image and video copy detection competitions [2, 5] are adapted and evaluated on the VCSL dataset. In this section, we provide more details, including how these methods are used for video copy localization and their time efficiency.

The approach in [7] employs a multiple crops strategy for image copy detection, which generates multiple crops from the input and extracts features for each crop, resulting in a feature set for each image for similarity measurement.

Table 2: Comparison with recent competition winning solutions on the VCSL dataset.

Method	Segment-level			Video-level			Test Time
	Recall	Precision	**F-score↑**	Recall	Precision	**F-score↑**	
MultiCrop [7]	74.38	69.93	<u>72.09</u>	93.10	99.79	<u>96.33</u>	5.2
ImConcat [3]	57.54	67.19	62.00	91.60	97.83	94.61	326
SAM [4]	65.87	67.51	66.68	91.60	96.40	93.94	5.7
Ours-ft	75.76	67.81	<u>71.56</u>	93.93	99.14	<u>96.46</u>	0.5

This method requires N forward passes to extract all the features, where N is the number of crops. As a result, there is an increase in computation by a factor of N when compared to a single feature extraction for the input. N varies in [7] as crops are human-defined regions and are also generated via region proposals and detection-based techniques. For the sake of efficiency, we implement a simplified 13-crop version of [7], which is still 13 times slower. The same detector backbone as our model is used as the temporal localization model.

Two images are concatenated as input for a Vision Transformer (ViT) in [3], which outputs a binary prediction indicating whether copied content exists in the input image pair. This method implicitly handles copies in local regions through the patch tokens and the self-attention mechanism. To apply this method for videos, each pixel of the similarity map is computed by feeding the corresponding frame pair to the model. The amount of computation is proportional to the number of pixels, which is extremely time-consuming. It takes about two weeks to generate the similarity maps for the test set, and we use TN [6] as the localization model to avoid additional cost for generating the training data.

The Similarity Alignment Model (SAM) proposed in [4] is also evaluated. It uses an HRNet model to process the similarity map, which outputs a clean and refined map. Postprocessing with connected component detection and RANSAC regression is then applied to localize the copied segment. The HRNet input resolution is 128×128; for larger similarity maps, non-overlapping patches are cropped for processing and the results are merged. The mean resolution of similarity maps in the VCSL testing set is 286×286 pixels, requiring an average of 9 crops for each video pair.

In summary, these methods contain excessive processing that requires computational costs at least an order of magnitude larger than our model, as shown in Table 2. The Test Time is measured in hours and represents the duration required to complete the evaluation of the entire testing set using one machine with 8 V100 GPU and 98 CPUs. Our model, while being much simpler and more time-efficient, achieves similar or better performance than these methods.

3 Visualization

Figure 1 presents additional visualizations of the attention maps of our model. The first row shows example frames that do not contain picture-in-picture editing. It is noted that the Regional Token has no attention focused on local regions

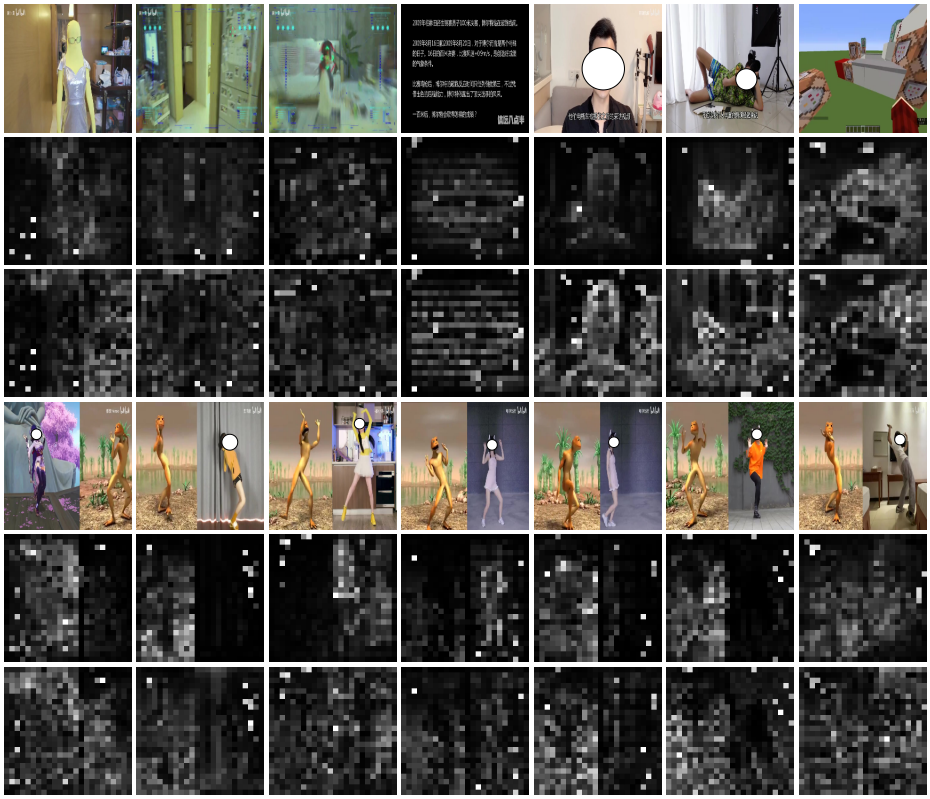

Fig. 1: Attention maps of our model. Row 1 and 4: the original frames; Row 2 and 5: the attention maps for the Regional Token; Row 3 and 6: the attention maps for the CLS token.

for these examples. Row 4 presents some ambiguous cases where two images are concatenated, and it is unclear which image could be the picture-in-picture region. In most cases, the attention of our Regional Token appears to be randomly focused on one of the regions, which is understandable given the ambiguity. In the final example of Figure 1, our Regional Token appears to falter in identifying local regions. Further investigation on this ambiguous type of data might provide insights into possible improvements.

4 More Results and Analysis

In Section *Comparison with Supervised Training* of the main text, we show that when finetuned with only 1% of the manually-labeled data, our model achieves similar or slightly better results than the fully-supervised model (Our-sup), which is trained using the entire training set. In Table 3, we provide the performance of the supervised model trained with 1% of the VCSL training data

Table 3: Performance of the supervised model trained with 1% of the VCSL training data (indicated in italics), evaluated on the VCSL testing set using eff256d and our RTR features.

Feature	Method	Segment-level			Video-level		
		Recall	Precision	**F-score↑**	Recall	Precision	**F-score↑**
eff256d	Ours-sup-1%	55.78	60.30	*57.95*	70.01	99.35	*82.14*
	Ours-sup	71.85	66.77	69.22	88.25	99.11	93.37
	Ours-ft-1%	71.98	66.44	69.10	90.31	99.65	94.75
	Ours-ft	73.46	68.19	**70.73**	90.94	99.80	**95.17**
RTR	Ours-sup-1%	56.65	60.00	*58.28*	73.94	99.62	*84.88*
	Ours-sup	74.52	65.49	69.71	91.38	98.77	94.93
	Ours-ft-1%	70.21	69.58	69.90	90.51	99.89	94.97
	Ours-ft	75.76	67.81	**71.56**	93.93	99.14	**96.46**

Table 4: More performance comparison and analysis. All the methods are trained using the VCSL training set and evaluated on the VCSL testing set.

Method	Segment-level			Video-level		
	Recall	Precision	**F-score↑**	Recall	Precision	**F-score↑**
ViT (1cls)	73.81	64.35	68.76	89.53	97.28	93.24
ViT (1cls-512d)	73.87	64.49	68.86	89.71	98.24	93.78
DINO [1]	70.05	66.95	68.47	87.22	97.63	92.14
Ours-sup	74.52	65.49	69.71	91.38	98.77	94.93

as a reference. With this small amount of data, the supervised model does not perform well.

The CLS token and Regional Token features of our model have a dimensionality of 256d. For comparison, we test a vanilla ViT with a 512d CLS token feature. As shown in Table 4, this model's performance is comparable to that of the ViT model with a 256d CLS token feature (denoted as ViT (1cls)). This demonstrates that the performance gain of the Ours-sup model is not due to the increase in feature dimensionality. DINO, as introduced in [1], is applied to image copy detection, where the CLS token and GeM pooled patch tokens are concatenated as the feature for computing cosine similarity. We applied this method to the VCSL dataset in our experiments. As indicated in Table 4, our model surpasses the performance of the DINO model. These experiments further validate the effectiveness of our Regional Token for video copy localization.

References

1. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021)

2. Douze, M., Tolias, G., Pizzi, E., Papakipos, Z., Chanussot, L., Radenovic, F., Jenicek, T., Maximov, M., Leal-Taixé, L., Elezi, I., et al.: The 2021 image similarity dataset and challenge. arXiv preprint arXiv:2106.09672 (2021)
3. Jeon, S.: 2nd place solution to facebook ai image similarity challenge matching track. arXiv e-prints pp. arXiv–2111 (2021)
4. Liu, Z., Ma, F., Wang, T., Rao, F.: A similarity alignment model for video copy segment matching. arXiv preprint arXiv:2305.15679 (2023)
5. Pizzi, E., Kordopatis-Zilos, G., Patel, H., Postelnicu, G., Ravindra, S.N., Gupta, A., Papadopoulos, S., Tolias, G., Douze, M.: The 2023 video similarity dataset and challenge. arXiv preprint arXiv:2306.09489 (2023)
6. Tan, H.K., Ngo, C.W., Hong, R., Chua, T.S.: Scalable detection of partial near-duplicate videos by visual-temporal consistency. In: ACMMM. pp. 145–154 (2009)
7. Wang, W., Sun, Y., Zhang, W., Yang, Y.: D2lv: A data-driven and local-verification approach for image copy detection. arXiv preprint arXiv:2111.07090 (2021)