Please cite the Published Version

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines)

Unveiling Social Relations: Leveraging Interpersonal Similarity Learning for Social Relation Recognition

Wang Tang, Linbo Qing, *Member, IEEE*, Haosong Gou, Li Guo , and Yonghong Peng

Abstract—**Identifying social relationships from images is a challenging yet promising research area with great potential for improving human health and enhancing our understanding of social networks. However, present endeavors in this field tend to concentrate on leveraging visual features for the exploration of social relationships, while disregarding certain concealed information that lies beneath these features, such as interpersonal similarity. These methodologies may result in inadequate visual data encoding, thereby imposing limitations on the accuracy of social relationship recognition. In light of this, we propose a novel framework that utilizes interpersonal similarities within images to provide more information for identifying social relationships, thereby mitigating the issue of insufficient feature exploration. Furthermore, our proposed framework incorporates an innovative CF-Loss function that effectively incentivizes the identification of accurate social relationships while penalizing incorrect identifications, ultimately bolstering the model's capacity to discriminate between distinct social relationships. Our experimental findings demonstrate the superiority of our proposed framework over state-of-the-art methods on public datasets, confirming its effectiveness and accuracy in identifying social relationships.**

Index Terms—**Social relation recognition, interpersonal similarity learning, confusion loss function.**

I. INTRODUCTION

SOCIAL relationships are the cornerstone for individuals' engagement in social activities and are established through interactions involving two or more people. The accurate and efficient identification of these relationships is crucial, not only for improving and maintaining human health, such as mental health and health-related behaviours [1], [2], but also for studying and understanding social networks and daily human life [3], [4], [5], [6], [7]. Additionally, social relationship recognition (SRR) offers valuable insights for other related tasks, such as evaluating the vitality of urban spaces [8], measuring human perception

Wang Tang and Linbo Qing are with the School of Electronic Information, Sichuan University, Chengdu 610065, China (e-mail: tangwang@stu.scu .edu.cn; qing_lb@scu.edu.cn).

Haosong Gou is with the China Mobile Communications Group, Sichuan Co., Ltd., Chengdu 611335, China (e-mail: gouhaosong@139.com).

Li Guo and Yonghong Peng are with the Department of Computing and Mathematics, Manchester Metropolitan University, M1 5GD Manchester, U.K. (e-mail: L.Guo@mmu.ac.uk; y.peng@mmu.ac.uk).
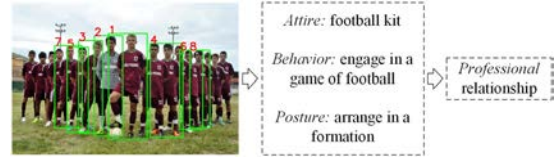
Fig. 1. Discernment of social relationships based on similarities in interpersonal traits.

of the urban environment [9] and user mobility modeling and check-in prediction [10].

Current research on image-based Social Relationship Recognition (SRR) can be broadly categorized into two main approaches: (a) direct classification, where visual features of people, objects, and scenes are extracted using convolutional neural networks [11], [12], [13], [14]; and (b) inferring social relationships through a graph neural network (GNN) using the extracted visual features [15], [16], [17], [18], [19], [20]. Notably, recent investigations on SRR have shown promising results by employing a visual transformer (ViT) as an encoder for precise person visual feature extraction [21]. This study provides compelling evidence for the effectiveness of transformers as feature extractors for SRR tasks. However, it is noteworthy that these studies primarily focus on fusing visual features using techniques such as concatenation, addition, and multiplication, without considering the implicit information embedded within these features prior to fusion. This oversight limits the ability of these approaches to fully comprehend social relationships.

Regarding SRR, the similarity between individuals can provide potent and valuable latent features that aid in identifying social relationships. Fig. 1 illustrates how this similarity contributes to our understanding of social relationships. The individuals in the figure share a strong resemblance in terms of their attire (football kit), behaviour (participate in a football match.), and posture (queue formation). By leveraging both interpersonal similarity and location information, along with scene details, one can deduce that they share a professional relationship. The figure elucidates how analogous traits can provide valuable insights into the social relationships between individuals. Importantly, research studies rooted in psychology corroborate that exploring interpersonal similarity constitutes an effective approach to investigating individual interactions [22]. As such, this study will incorporate interpersonal similarity as a key feature into our framework prior to feature fusion, in order to enrich the model's comprehension of social relationships.
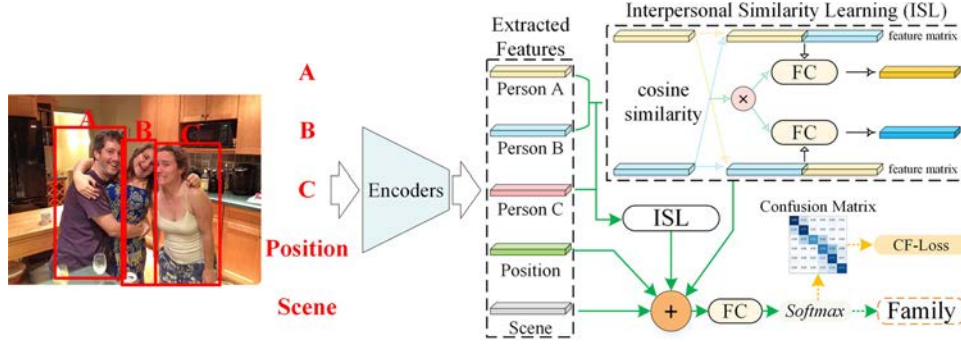
Fig. 2. Overall architecture of the ISL framework.

In addition, the issue of data imbalance can significantly impact the effectiveness of model training. In various research domains, scholars typically leverage the loss function to fine-tune the model, allowing it to allocate greater emphasis to the smaller samples [23], [24], [25]. Moreover, these loss functions incorporate conventional data augmentation techniques to further lessen the impact of data imbalance. For the issue of data imbalance in SRR task, Wang et al. [21] propose a new loss function. This loss function optimizes model training by increasing the inter-class distance while reducing the intra-class distance. While the model demonstrates satisfactory performance in overall metrics, it falls short when it comes to recall for individual classes. In practice, the randomness of input data means that the imbalance of data distribution for each epoch may not be absolute. Hence, the method proposed in [21] also bears the risk of failre. Addressing these limitations is crucial to enhancing the model's precision in identifying and discerning between distinct social relationships. Against this backdrop, our study proposes a novel confusion loss function (CF-Loss) to mitigate this challenge.

To implement our proposed methodology, we have developed a novel framework that leverages Interpersonal Similarity Learning (ISL) to identify social relationships. In order to provide a thorough comparison with the approach presented in [21], we still use ViT for encoding visual features of individuals depicted in images. We then propose an ISL model that focuses on exploring similarities between individual visual features. Additionally, we have introduced a new CF-Loss function which is designed to improve the model's ability to differentiate between different social relationships.

Our contributions can be summarized as follows:
1) This study presents a novel framework for SRR that utilizes latent interpersonal similarities within visual features to enhance the understanding of social relationshipsm, which effectively improves the model's ability of recognizing social relationships.
2) To improve the discriminative capacity of the model for different social relationships, a new CF-Loss function is proposed to optimize its training.
3) The proposed model has achieved superior performance compared to state-of-the-art results on both the PISC [12] and PIPA [13] datasets.

The letter is structured as follows: Section II presents the proposed approach, Section III showcases the experimental results demonstrating its effectiveness, and Section IV concludes the letter.

II. THE PROPOSED METHOD

A. Overview

The study's framework involves several key steps, as shown in Fig. 2. An encoder is used to encode individuals, their positions, and scene information. Visual features are extracted using ViT for individuals, FC for position information, and ResNet-50 for scene information [26]. These features are then input into the ISL model to compute interpersonal similarity. The similarity, location, and scene information are integrated for social relationship recognition. Model training is optimized using the CF-Loss function.

B. Feature Extraction

For fair comparison with [21], similar approaches were used for feature extraction. ViT was employed to extract visual features of individuals using bounding boxes and coordinates. All individuals were uniformly resized to 224×224 and normalized before inputting into ViT, resulting in a 512-matrix output. The scene image was also resized to 224×224, resulting in a 512-matrix output. Position coordinates were processed using the FC model to obtain a 512-matrix.

C. ISL

With the individual visual features extracted, a thorough exploration of their interpersonal similarities can be conducted. The initial stage involves computing the similarity between their visual features utilizing cosine similarity. The formula utilized for this computation is as follows:

$$r_{i,j} = \frac{x_i^T x_j}{|x_i| \, |x_j|} \qquad (1)$$

where x_i and x_j represent the visual features of person i and j, respectively, while $r_{i,j}$ denotes the cosine similarity result. The acquisition of interpersonal similarity learning can be achieved through multiplying the cosine similarity coefficient with the feature matrix. It is worth noting that the feature matrix is obtained by concatenating the visual features of two distinct individuals in different orders, as shown in Fig. 2. The formula employed for this computation is expressed as follows:

$$ISL_i = r_{i,j} \cdot FC\left(x_i, x_j\right) \qquad (2)$$

$$ISL_j = r_{i,j} \cdot FC\left(x_j, x_i\right) \qquad (3)$$

	PISC-C				PISC-F							PIPA
	Int.	Non.	No.	mAP	Fir.	Fam.	Cou.	Pro.	Com.	No.	mAP	Acc.
Dual-glance [12]	73.1	84.2	59.6	79.7	35.4	68.1	76.3	70.3	57.6	60.9	63.2	59.6
GRM [16]	81.7	73.4	65.5	82.8	59.6	64.4	58.6	76.6	39.5	67.7	68.7	62.3
MGR [17]	-	-	-	-	64.6	67.8	60.5	76.8	34.7	70.4	70.0	64.4
SRG-GN [15]	-	-	-	-	25.2	80	100.0	78.4	83.3	62.5	71.6	53.6
GR2N [20]	81.6	74.3	70.8	83.1	60.8	65.9	84.8	73.0	51.7	70.4	72.7	64.3
SRR-LGR [18]	89.6	84.6	78.5	84.8	83.9	52.4	35.9	64.0	54.0	63.6	73.0	66.1
HF-SRGR [19]	89.1	87.0	75.5	84.6	82.2	39.4	33.2	60.0	47.7	71.8	73.3	65.9
MT-SRR [21]	91.8	91.8	75.2	86.8	71.6	69.7	62.5	88.0	34.2	72.7	74.6	72.5
ISL (ours)	92.8	91.6	75.7	**87.0**	78.2	72.6	70.0	88.6	42.9	81.1	**75.6**	**73.0**

We use per-class recall (in %) and mAP (in %) for PISC, and ACC (in %) for PIPA, with the best results highlighted. Note that: Int.: *'intimate'*, Non.:*'non-intimate'*, No.: *'no relation'*, FRI.: *'friend'*, FAM.: *'family'*, COU.: *'couple'*, PRO.: *'professional'*, COM.: *'commercial'*.

D. CF-Loss

The CF-Loss function proposed in this study is derived from the confusion matrix generated at every epoch of the model training process. During the process of model training, it can provide rewards for accurately classified relationships, while imposing penalties for relationships of confusion. The mathematical formula for its computation can be expressed as follows:

$$CF_{Loss} = \varphi_{s,t} \cdot log\left(1 + e^{-\omega_p}\right) + (1 - \varphi_{s,t}) \cdot log\left(1 + e^{\omega_n}\right) \tag{4}$$

In the positive set, $\varphi_{s,t}$ takes on a value of 1 to signify that both actual and predicted values are true. Conversely, in the negative set, $\varphi_{s,t}$ remains at 0, indicating a true actual value but a false prediction. Additionally, weight ω is derived from the confusion matrix and divided into two components: ω_p and ω_n, which respectively correspond to positive and negative weights. The formulas for computing ω_p and ω_n are as follows:

$$\omega_p = 1 - \frac{\xi_{s,t}}{n^2}, s = t \in N^* \tag{5}$$

$$\omega_n = \frac{\xi_{s,t}}{n^2}, s \neq t \in N^* \tag{6}$$

where $\xi_{s,t}$ represents the original data in the confusion matrix, and n represents the number of relationship categories.

Specificlly, in the process of iterative learning, the model adapts its attentional focus to different relationship categories in response to errors made during the previous iteration. This mechanism contributes to improved recognition accuracy for each category, ultimately leading to an overall enhancement of the model's accuracy.

III. EXPERIMENTS AND RESULTS

A. Datasets

PISC [12]: The PISC dataset has 22,670 images with 76,568 relation samples and a hierarchical task structure, i.e., PISC-C and PISC-F. The evaluation metric used for this dataset is the mean average precision (mAP), borrowed from previous SRR methods.

TABLE II
MODEL TRAINING SETTINGS

Items	Parameters	Items	Parameters
Environment	Pytorch 1.13	Batch size	64
GPU	Nvidia GeForce RTX 3090	Learning rate	0.0001
Optimizer	Adam	Epochs	200

PIPA [13]: The PIPA dataset has 26,915 pairs of people categorized into five social domains and 16 specific social relations. Our study focuses on SRR, with evaluation based on the dataset's 16 social relations using top-1 accuracy (Acc), a metric adopted from previous SRR methods.

B. Implementation Details

Data Augmentation: We used data augmentation techniques to address the data imbalance in the PISC-F dataset. Specifically, we manually modified pairs by reversing person order and horizontally flipping minority labelled samples.

Training Setting: Similar to [18], [19], [21], we set the hyperparameters for this experiment as shown in Table II. In addition, for visual feature extraction, we employed a large ViT pre-trained model with an input size of 224 and a patch size of 16.

C. Comparison With State-of-The-Art Methods

To assess the effectiveness of our proposed ISL model, we conducted a comparative analysis against state-of-the-art methods, as presented in Table I. Our IFIL model outperformed existing methodologies, achieving superior overall recognition accuracy. Its impressive results on PISC-C, PISC-F, and PIPA were particularly noteworthy, with accuracy rates of 87.0%, 75.6%, and 73.0%, respectively. These represent an improvement of 0.2%, 1.0%, and 0.5% over MT-SRR [21]. Such outcomes demonstrate the efficacy of interpersonal similarity learning.

In addition, in the context of the PISC-F experiment, our proposed method demonstrated superior recall rates compared to MT-SRR. Specifically, the improvements are 6.6%, 2.9%, 7.5%, 0.6%, 8.7%, and 8.9% respectively. This results provide

TABLE III
RESULTS OF THE ABLATION EXPERIMENT ON OVERALL ACCURACY

Ablation Method	PISC-C (mAP)	PISC-F (mAP)	PIPA (Acc.)
(i) IF	79.6	69.3	69.5
(ii) IF+ISL	86.5	75.0	72.8
(iii) IF+ISL+CF-Loss	**87.0**	**75.6**	**73.0**

TABLE IV
RUNTIME COMPARISONS USING DIFFERENT METHODS

Method	PISC-C	PISC-F	PIPA
SRR-LGR	7'22"	20'42"	3'43"
HF-SRGR	8'03"	21'15"	4'01"
MT-SRR	14'38"	53'38"	10'23"
ISL	13'45"	48'32"	9'12"

Note: $X'Y'$ represents X minutes and Y second.

evidence for the effectiveness and superiority of our proposed CF-Loss method.

D. Ablation Study

(a) ISL study: To investigate the efficacy of our proposed ISL approach in accurately understanding social relationships, we employed t-distributed stochastic neighbor embedding (T-SNE) to reduce the dimensionality of the encoded features and conducted visualization analyses. As shown in Fig. 3, the visual features of the depicted couple are starkly divergent in the presence or absence of the ISL method. Specifically, their visual features exhibit a significant degree of dispersion in the absence of ISL. When attempting to merge the visual features of two distinct individuals without considering this aspect, the resulting model may lead to misleading or ambiguous inferences, such as incorrectly identifying the individuals as friends, siblings, or colleagues. However, through the utilization of ISL to encode shared features, it becomes apparent that a robust linkage is formed between their pivotal features and even overlap with one another, thus facilitating a deeper understanding of the intimate social interconnections among them by the model.

(b) Whole framework study: The ablation experiment was conducted stepwise following this detailed process: Firstly, the SRR task was performed using individual features (IF) without any integration, as shown in Table III. The PISC-C metric achieved 79.6% at this stage, while PISC-F and PIPA attained 69.3% and 69.5%, respectively. Secondly, visual features were incorporated into the ISL model, increasing 6.9%, 5.7%, and 3.3% in the metrics above. This underscores the exceptional efficacy of the ISL model in the experiment. Finally, CF-Loss optimization model training was integrated, further improving 0.5%, 0.6%, and 0.2% across all three metrics. The remarkable enhancement observed across all three stages of the model clearly indicates their positive contribution towards improving SRR.

(c) Runtime study: We compared the training duration of our proposed method with three recent studies, as shown in Table IV. The first two methods use convolution-based models for faster training but lower accuracy. The third method uses ViT and multiple transformers for accurate results but longer training time. Our method incorporates ViT and streamlines feature fusion, reducing training time while effectively leveraging interpersonal similarity for social relationship recognition.
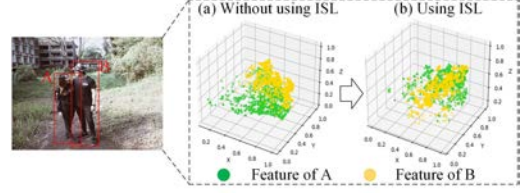

Fig. 3. Dimensionality reduction visualizations of the feature matrix. (a) Without using ISL and (b) using ISL.
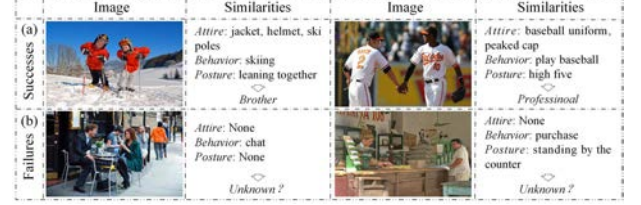

Fig. 4. (a) Successes and (b) failures in social relationship recognition using the ISL framework.

E. Qualitative Evaluation

We conducted a detailed investigation focusing on the example in Fig. 4 to understand the factors that contribute to the success and failure of our proposed methodology. In the correctly identified example (Fig. 4(a)), there are significant similarities in clothing, actions, and postures that lead to correct relationship identification. In the misidentified example (Fig. 4(b)), the lack of similarity between individuals leads to failure. In addition, considering only interpersonal similarity and overlooking contextual elements and logical constraints may also be another reason for failure. Therefore, future research could integrate additional visual cues or explore logical constraints to address this issue. Nevertheless, our proposed method outperforms previous recognition results.

IV. CONCLUSION AND DISCUSSION

In this letter, we propose a novel framework that leverages interpersonal similarities within images to enhance visual features and improve the accuracy of social relationship recognition. We introduce a novel CF-Loss function that enhances the model's ability to differentiate between various social relationships. Our experimental results demonstrate the superiority of our framework over state-of-the-art methods on the PISC and PIPA datasets, affirming the effectiveness and accuracy of our approach.

However, there are still opportunities for further research, e.g. further research should focus on developing methods for accurately recognizing social relationships in real-world application scenarios, especially in the presence of noisy or blurred images. Second, the dynamics of character interactions in public spaces should be explored, and the scalability of the models to handle larger datasets and different scenarios should be considered. It is also important to carefully consider the ethical and social implications associated with the field to ensure responsible application of social relationship recognition techniques.

REFERENCES

[1] D. B. Bugental, "Acquisition of the algorithms of social life: A domain-based approach.," *Psychol. Bull.*, vol. 126, no. 2, 2000, Art. no. 187.

[2] S. M. Platek et al., "Reactions to children's faces: Males are more affected by resemblance than females are, and so are their brains," *Evol. Hum. Behav.*, vol. 25, no. 6, pp. 394–405, 2004.

[3] Y. Teng, C. Song, and B. Wu, "Learning social relationship from videos via pre-trained multimodal transformer," *IEEE Signal Process. Lett.*, vol. 29, pp. 1377–1381, 2022.

[4] Y. Teng, C. Song, and B. Wu, "Recognizing social relationships in long videos via multimodal character interaction," *IEEE Signal Process. Lett.*, vol. 30, pp. 573–577, 2023.

[5] D. Shullani, D. Baracchi, M. Iuliani, and A. Piva, "Social network identification of laundered videos based on DCT coefficient analysis," *IEEE Signal Process. Lett.*, vol. 29, pp. 1112–1116, 2022.

[6] N. R. de Oliveira, D. S. Medeiros, and D. M. Mattos, "A sensitive stylistic approach to identify fake news on social networking," *IEEE Signal Process. Lett.*, vol. 27, pp. 1250–1254, 2020.

[7] H. Wang et al., "Learning social spatio-temporal relation graph in the wild and a video benchmark," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 6, pp. 2951–2964, Jun. 2023.

[8] T. Niu et al., "Small public space vitality analysis and evaluation based on human trajectory modeling using video data," *Building Environ.*, vol. 225, 2022, Art. no. 109563.

[9] J. Huang, L. Qing, L. Han, J. Liao, L. Guo, and Y. Peng, "A collaborative perception method of human-urban environment based on machine learning and its application to the case area," *Eng. Appl. Artif. Intell.*, vol. 119, 2023, Art. no. 105746.

[10] W. Liang and W. Zhang, "Learning social relations and spatiotemporal trajectories for next check-in inference," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 4, pp. 1789–1799, Apr. 2023.

[11] V. Ramanathan, B. Yao, and L. Fei-Fei, "Social role discovery in human events," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2475–2482.

[12] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Dual-glance model for deciphering social relationships," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2650–2659.

[13] Q. Sun, B. Schiele, and M. Fritz, "A domain based approach to social relation recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3481–3490.

[14] G.-N. Dong, C.-M. Pun, and Z. Zhang, "Kinship verification based on cross-generation feature interaction learning," *IEEE Trans. Image Process.*, vol. 30, pp. 7391–7403, 2021.

[15] A. Goel, K. T. Ma, and C. Tan, "An end-to-end network for generating social relationship graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11186–11195.

[16] Z. Wang, T. Chen, J. S. J. Ren, W. Yu, H. Cheng, and L. Lin, "Deep reasoning with knowledge graph for social relationship understanding," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 1021–1028.

[17] M. Zhang, X. Liu, W. Liu, A. Zhou, H. Ma, and T. Mei, "Multi-granularity reasoning for social relation recognition from images," in *Proc. IEEE Int. Conf. Multimedia*, 2019, pp. 1618–1623.

[18] L. Qing, L. Li, Y. Wang, Y. Cheng, and Y. Peng, "SRR-LGR: Local–global information-reasoned social relation recognition for human-oriented observation," *Remote Sens.*, vol. 13, no. 11, 2021, Art. no. 2038.

[19] L. Li, L. Qing, Y. Wang, J. Su, Y. Cheng, and Y. Peng, "HF-SRGR: A new hybrid feature-driven social relation graph reasoning model," *Vis. Comput.*, vol. 38, pp. 3979–3992, 2021.

[20] W. Li, Y. Duan, J. Lu, J. Feng, and J. Zhou, "Graph-based social relation reasoning," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 18–34.

[21] Y. Wang, L. Qing, Z. Wang, Y. Cheng, and Y. Peng, "Multi-level transformer-based social relation recognition," *Sensors*, vol. 22, no. 15, 2022, Art. no. 5749.

[22] T. E. Malloy, "Interpersonal attraction in dyads and groups: Effects of the hearts of the beholder and the beheld," *Eur. J. Social Psychol.*, vol. 48, no. 3, pp. 285–302, 2018.

[23] Z. Yuan, G. Li, Z. Wang, J. Sun, and R. Cheng, "RL-CSL: A combinatorial optimization method using reinforcement learning and contrastive self-supervised learning," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 4, pp. 1010–1024, Aug. 2023.

[24] Y. Zhang, Y. Liu, P. Zhu, and W. Kang, "Joint reinforcement and contrastive learning for unsupervised video summarization," *IEEE Signal Process. Lett.*, vol. 29, pp. 2587–2591, 2022.

[25] L. Yang, Z. Wu, J. Hong, and J. Long, "MCL: A contrastive learning method for multimodal data fusion in violence detection," *IEEE Signal Process. Lett.*, vol. 30, pp. 408–412, 2023.

[26] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.