Yanzheng WANG, Yujun WANG, Fengshuo DAI, Xin YANG, Xiaoheng JIANG, Pei LV, Shizhe HU, Mingliang XU, 2026. Multi-stage contrastive multi-modal clustering with consistency retained. *ENGINEERING Information Technology & Electronic Engineering*, 27(8):250185. https://doi.org/10.1631/ENG.ITEE.2025.0185

Multi-stage contrastive multi-modal clustering with consistency retained

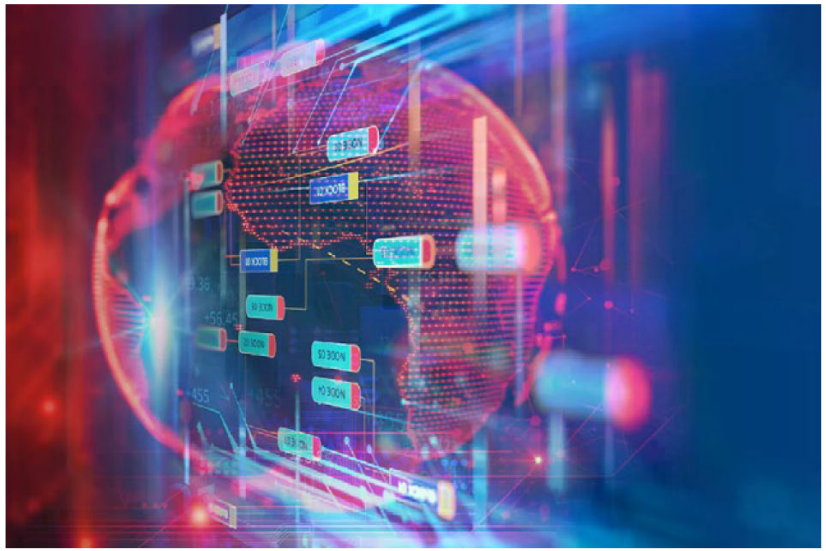
Key words: Multi-modal clustering; Contrastive learning; Consistency retention; Negative sample selection

Shizhe HU; Mingliang XU
E-mail: ieshizhehu@zzu.edu.cn
iexumingliang@zzu.edu.cn
ORCID: Shizhe HU, https://orcid.org/0000-0003-1301-2396
Mingliang XU, https://orcid.org/0000-0002-6885-3451

Methodology

Contrastive multi-modal clustering (CMC) combines contrastive learning (CL) with multi-modal clustering (MMC) to mine consistent and discriminative structures from multi-source data. CL enhances semantic alignment and discrimination between modalities by constructing positive and negative sample pairs, addressing the fundamental challenge of cross-modal representation learning.

- Semantic alignment: CMC integrates CL into MMC, enhancing semantic alignment and discrimination between modalities through positive/negative sample pair construction.
- Broad applications: MMC has broad applications in medical informatics, recommendation systems, Internet of Things (IoT), and machine vision measurement, benefiting from richer multi-source data representations.
- Stronger generalization: CL provides stronger feature representation and generalization capabilities, making CMC more effective in capturing cross-modal correlations.



Multi-modal data fusion—image, text, and audio representations

Motivation

1. Though CL boosts multi-modal alignment, existing clustering methods only adopt one/two-stage contrastive constraints with crude negative sampling, introducing noisy pairs that degrade fused feature quality.
2. Most contrastive MMC models ignore cross-modal consistency preservation, making representation learning vulnerable to low-quality modalities and hurting final clustering performance.
3. Current methods only enforce cluster-level cross-modal alignment without sample-level consistency regularization, failing to capture complete global data structures for complex multi-modal data.

Main idea

1. A three-stage contrastive MMC framework is built to perform contrastive constraints on single features, fused representations, and cluster assignments synchronously.
2. A dual-prior negative sample filtering strategy combining the original feature *K*-means and fusion pseudo-labels is proposed to eliminate misleading false negative pairs.
3. A dual consistency loss covering cluster and sample levels is introduced to mitigate interference from low-quality modalities and maintain cross-modal semantic uniformity.

Method

1. To build a global multi-stage contrastive framework, we construct three CL branches covering early, middle, and late stages, which perform constraint optimization on single-modal raw features, adaptive fused representations, and cross-modal cluster assignments, respectively.

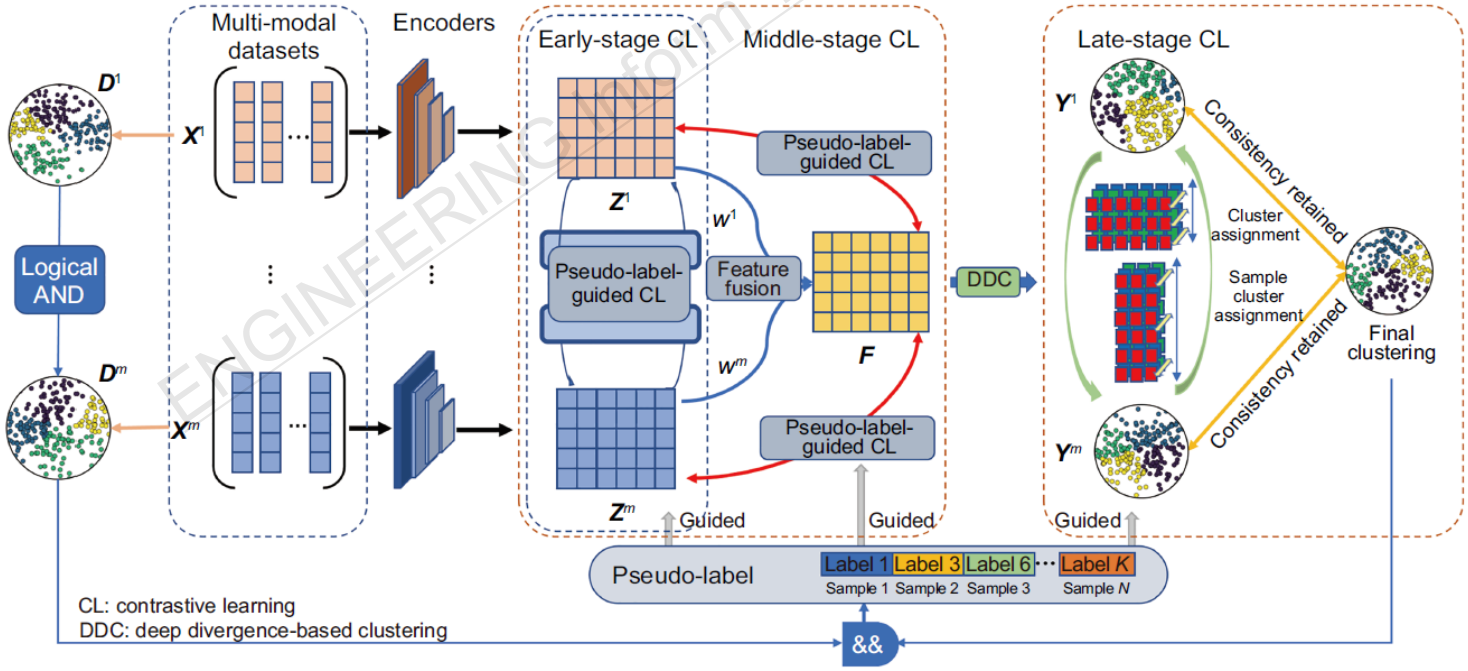


Fig. 1 Framework of the MCMC-CR method

Method (Cont'd)

2. To eliminate false negative sample pairs that mislead model training, we design a dual-prior screening strategy. We combine the original space *K*-means clustering results and fusion pseudo-labels to generate a mask matrix, which filters invalid negative samples at all three contrastive stages.

$$D_{i,j} = \begin{cases} 1, & \text{if } \bigwedge_{m=1}^{M} \left(D_{i,j}^{m} = 1 \vee D_{j,i}^{m} = 1 \right), \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

$$\mathcal{N}_i^{\text{false}} = \{ j \mid \mathcal{D}_{i,j} = 1,\ \text{pred}_i = \text{pred}_j,\ j \in [1, N] \}, \tag{3}$$

Method (Cont'd)

3. To address performance degradation stemming from imbalanced quality among multiple modalities, we introduce a consistency loss built on Kullback–Leibler (KL) divergence. It forces the clustering distribution of each individual modality to approach the unified clustering distribution derived from fused multi-modal features at the cluster level, preventing low-quality modalities from dominating the final clustering outcome and facilitating balanced utilization of valid information from all modalities.

$$\mathcal{L}_{\text{Consistency}} = \sum_{m=1}^{M} \mathrm{KL}\left(\boldsymbol{C}_{\mathrm{f}} \| \boldsymbol{Y}^{m}\right) = \sum_{i=1}^{N} \sum_{j=1}^{K} \sum_{m=1}^{M} c_{ij} \ln \frac{c_{ij}}{y_{ij}^{m}}, \tag{16}$$

Method (Cont'd)

4. The proposed method is optimized by minimizing the following loss function:

$$\mathcal{L}_{\text{total}}=\alpha(\mathcal{L}_{\text{Early}}+\mathcal{L}_{\text{Middle}}+\mathcal{L}_{\text{Late}})+\beta\mathcal{L}_{\text{Consistency}}+\mathcal{L}_{\text{DDC}},$$

where $\alpha(\mathcal{L}_{\text{Early}}+\mathcal{L}_{\text{Middle}}+\mathcal{L}_{\text{Late}})$ denotes the multi-stage CL loss, which integrates the CL losses from the early, middle, and late stages to ensure global alignment of consistency information between modalities. $\beta\mathcal{L}_{\text{Consistency}}$ is the consistency learning loss, which is used to align the clustering assignments of each modality with the clustering assignments of the fusion features. $\mathcal{L}_{\text{DDC}}$ is the deep divergence-based cluster loss.

Datasets & comparison methods

Our evaluation spans five diverse multi-modal benchmark datasets ranging from 1440 to 20 000 samples with 2–4 modalities, compared against 19 baselines covering single-modal, traditional, and deep MMC paradigms for comprehensive validation.

Table 1 Multi-modal dataset overview

Dataset	Number of samples	Number of categories	Number of modalities	Feature dimensions
Caltech-3V	1440	7	3	40\254\928
Caltech-4V	1440	7	4	40\254\928\512
NUS-Wide	20 000	8	4	100\100\100\100
Flickr	12 154	6	3	100\100\100
IAPR	7855	6	2	100\100

\ is used to distinguish between the features of different modalities within a dataset

Clustering results

Multi-stage contrastive multi-modal clustering with consistency retained (MCMC-CR) achieves state-of-the-art performance across all five benchmark datasets with an average ACC improvement of 4.1 percentage points over the second-best method, demonstrating robust multi-modal integration and generalization capabilities.

Table 2 Clustering results (%) in terms of ACC and NMI on multi-modal datasets

Method	Caltech-3V		Caltech-4V		NUS-Wide		Flickr		IAPR	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
KM	46.3	31.3	54.6	46.7	26.8	16.7	40.9	22.5	38.9	17.2
Ncuts (TPAMI'00)	42.6	25.4	67.8	47.6	31.6	14.1	48.4	26.1	41.9	18.9
AmKM	46.9	31.5	44.9	30.6	26.8	15.2	41.0	21.6	40.4	17.0
AmNcuts (TPAMI'00)	43.7	25.5	41.8	24.9	30.4	16.1	48.2	26.2	42.2	18.9
CoregMVSC (NeurIPS'11)	54.4	45.3	64.9	54.5	26.3	16.8	41.0	26.8	35.1	18.4
RMKMC (IJCAI'13)	59.5	49.4	65.5	60.3	30.5	14.5	42.3	23.4	36.4	15.9
SwMC (IJCAI'17)	30.2	23.1	43.7	44.2	12.5	15.0	34.3	34.5	30.2	23.1
ONMSC (AAAI'20)	58.2	56.8	62.3	66.1	16.9	15.3	30.6	16.4	21.6	11.1
SMCMB (TBD'24)	67.2	54.5	74.4	67.0	32.4	21.8	52.8	34.1	34.8	16.4
EAMC (CVPR'20)	38.9	21.4	29.6	16.5	24.6	9.7	30.5	9.1	37.1	16.4
DEMC (InfoSci'21)	53.4	41.4	49.3	45.0	21.3	11.1	44.8	25.2	30.1	13.8
SiMVC (CVPR'21)	56.9	50.4	61.9	53.6	25.7	10.2	45.6	26.3	42.7	18.5
CoMVC (CVPR'21)	54.1	50.4	56.8	56.8	43.9	31.2	49.3	30.6	46.7	21.5
MFLVC (CVPR'22)	63.1	56.6	73.3	65.2	45.1	33.1	53.8	32.8	47.3	22.6
SEM (NeurIPS'23)	69.2	59.2	82.6	<u>75.3</u>	39.6	25.2	53.1	30.9	42.2	18.9
DIVIDE (AAAI'24)	60.9	53.8	64.3	57.9	36.3	26.0	52.3	33.5	45.6	23.0
SCMVC (TMM'24)	<u>75.9</u>	<u>66.3</u>	<u>84.4</u>	72.9	<u>48.8</u>	<u>34.2</u>	54.2	32.3	46.5	24.1
SSLNMVC (TMM'25)	64.4	58.3	82.1	72.8	38.9	28.3	51.2	33.0	46.4	24.0
MSDIB (AAAI'25)	74.6	66.2	66.4	55.2	47.6	33.4	<u>55.5</u>	<u>35.5</u>	<u>50.7</u>	<u>25.7</u>
Ours (MCMC-CR)	**80.6**	**69.9**	**87.5**	**78.3**	**53.9**	**36.4**	**59.2**	**38.2**	**54.6**	**30.4**

The bold and underlined values are the best and second best results, respectively

Visualization clustering results

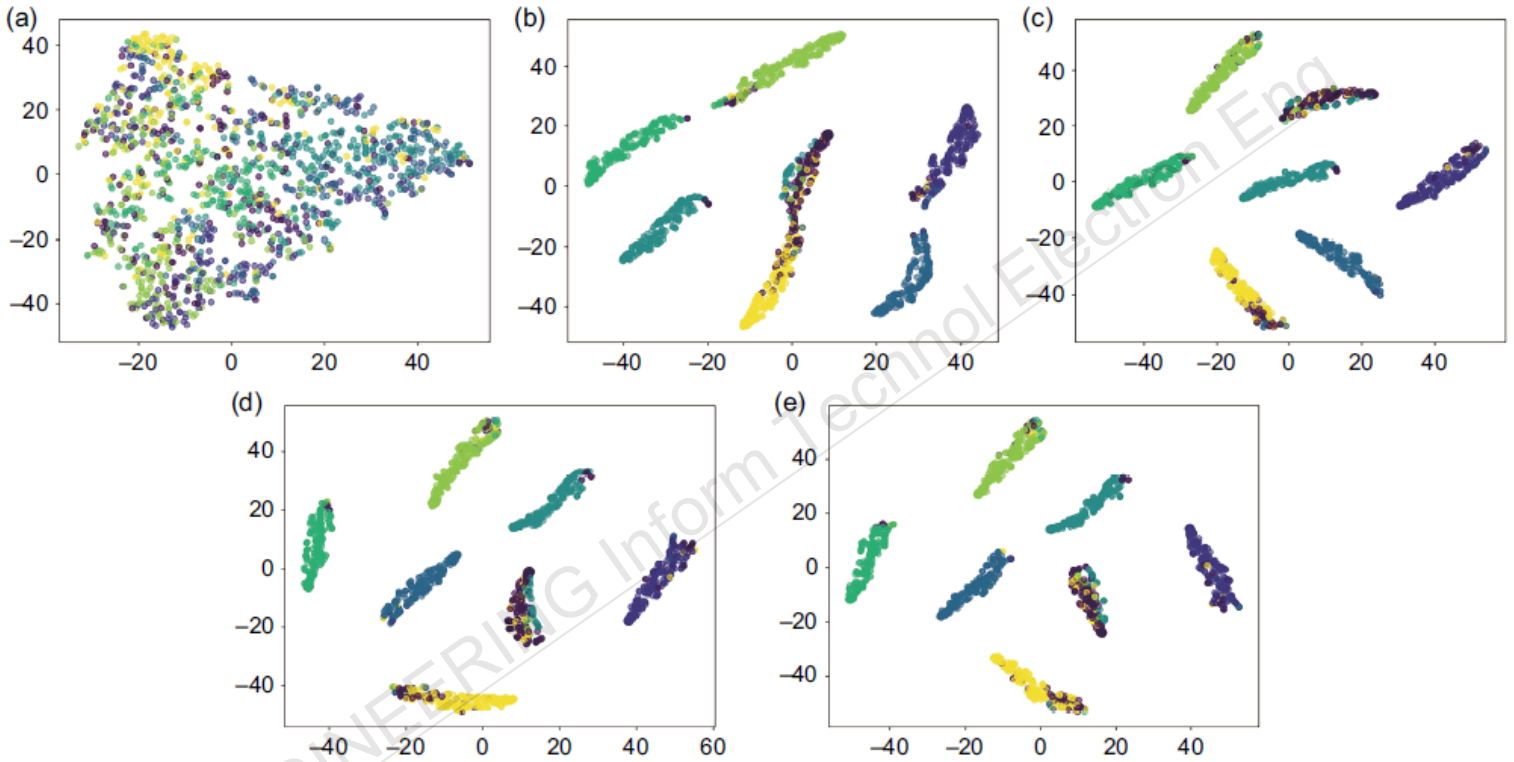


Fig. 4 Evolution of cluster assignments as the training goes on the Caltech-4V dataset: (a) epoch 0; (b) epoch 25; (c) epoch 50; (d) epoch 75; (e) epoch 100

As training progresses, data points evolve from a mixed distribution into well-separated clusters. This clear spatial partitioning visually verifies the model's effectiveness in learning discriminative features that distinguish different classes.

Convergence analysis

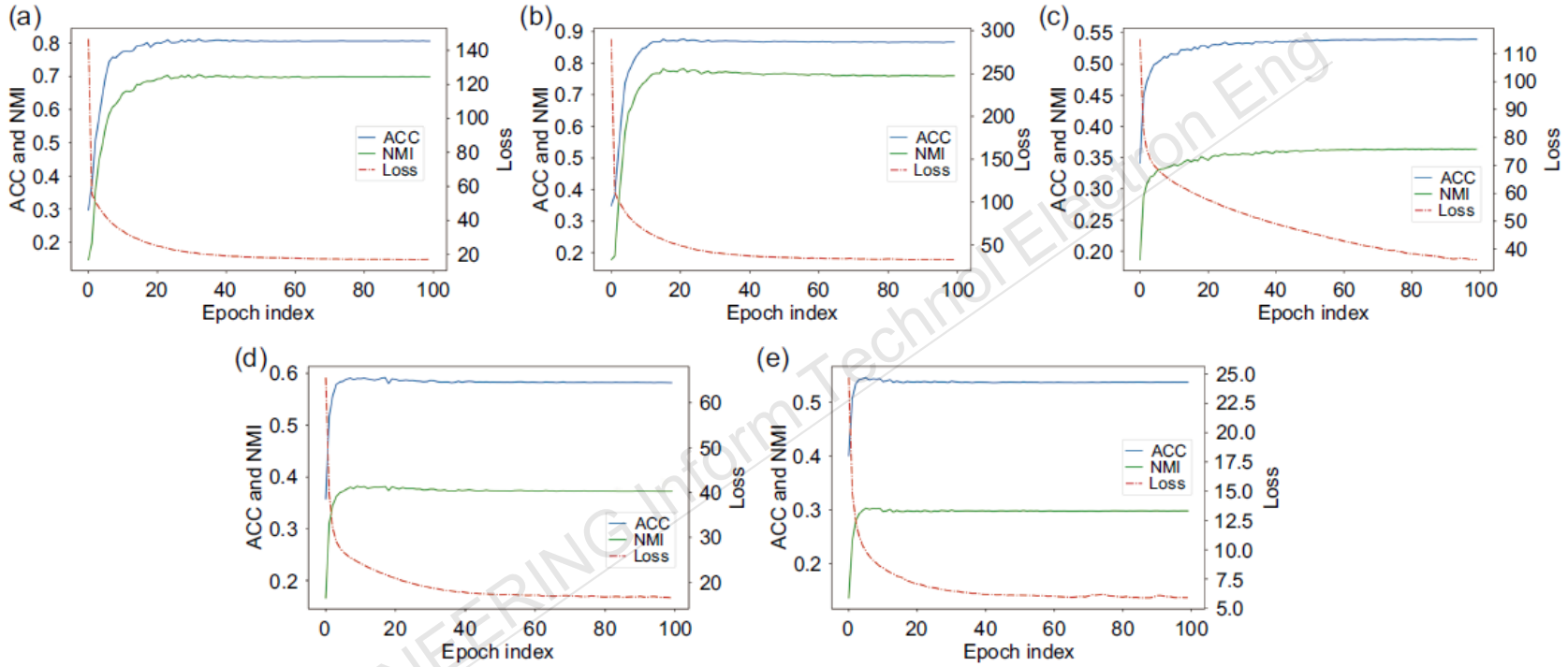


Fig. 5 Convergence analysis of MCMC-CR across Caltech-3V (a), Caltech-4V (b), NUS-Wide (c), Flickr (d), and IAPR (e) datasets

The loss function plummets and stabilizes within ~20 epochs, while ACC and NMI metrics steadily rise and converge to optimal values. This indicates that the model has a fast convergence speed and maintains excellent stability throughout training.

Conclusions & future work

Key contributions:

1. **Novel multi-stage CL framework:** achieves global alignment of cross-modal consistency information via early–middle–late CL, strengthening semantic integrity across different data modalities.
2. **Precise negative sample selection:** optimizes the quality of contrastive pairs, significantly enhancing the robustness of clustering and the discriminative power of learned feature representations.
3. **Sample-level consistency constraint:** enforces consistency on sample clustering assignments, enabling the model to better capture the holistic structural characteristics of multi-modal datasets.

Future directions:

We plan to extend this framework to tackle more complex real-world challenges, such as scenarios with **missing modalities** or **unaligned multi-modal data**, further validating its adaptability and effectiveness in practical applications.

Author Bio

Yanzheng WANG received the B.E. degree in software engineering from Henan University, China, in 2022. He is currently a master's candidate at the School of Computer Science and Artificial Intelligence, Zhengzhou University. His primary research focuses on multi-modal clustering algorithms and the theoretical advancements of the information bottleneck.

Yujun WANG received the B.E. degree in software engineering from Zhengzhou University, China, in 2024. He is currently pursuing a master's degree at the School of Computer Science and Artificial Intelligence, Zhengzhou University. His research interests include multi-modal clustering and information bottleneck theory.

Author Bio

Shizhe HU is an associate research fellow at the School of Computer Science and Artificial Intelligence, Zhengzhou University. He received his B.E. and Ph.D. degrees at the School of Computer Science and Artificial Intelligence, Zhengzhou University, China. His main research interests include information bottleneck theory and trusted multi-modal clustering. He has published several papers in peer-reviewed prestigious journals and conferences, such as *IEEE TPAMI*, *TIP*, *TKDE*, *TNNLS*, *TCYB*, *NeurIPS*, *ICML*, *CVPR*, *AAAI*, *IJCAI*, and *ACM MM*. He serves as an associate editor of *IEEE Transactions on Image Processing* and *Pattern Recognition* and as an editorial board member of *Information Processing and Management*. He also serves as the area chair of *ICLR* and *ICML* and as a senior program committee member of *IJCAI-ECAI*. More details can be found at https://shizhuhu.github.io.

Author Bio

Mingliang XU received the Ph.D. degree in computer science and technology from the State Key Laboratory of Computer-Aided Design and Computer Graphics, Zhejiang University, Hangzhou, China. He is currently a professor with the School of Computer Science and Artificial Intelligence of Zhengzhou University. His research interests include computer graphics, multimedia, and artificial intelligence. He has authored more than 100 journal and conference papers, including *ACM Transactions on Graphics*, *ACM Transactions on Intelligent Systems and Technology*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Cybernetics*, *IEEE Transactions on Circuits and Systems for Video Technology*, *ACM SIGGRAPH (Asia)*, *ACM MM*, and *ICCV*.