Contrastive Masked Auto-Encoders Based Self-supervised Hashing for 2D Image and 3D Point Cloud Cross-modal Retrieval

Rukai Wei[¶1], Heng Cui[¶1], Yu Liu[*1], Yufeng Hou[1], Yanzhao Xie[2], Ke Zhou[1]

[1]Huazhong University of Science and Technology [2]Guangzhou University

¶ Contribute equally * Corresponding author

Motivation

- Retrieving 3D point-cloud with 2D images (or vice versa) quickly is crucial in many real-world applications, like autonomous driving, augmented reality, and robotics.
- Cross-modal hashing is a good solution since it can convert high-dimensional multi-modal data into binary hash codes to achieve fast retrieval speed and low storage overhead.

Challenges when applying hashing to 2D-3D cross-modal retrieval:

- The irregular and unordered data structure of point-cloud data makes it difficult to capture meaningful semantics effectively.
- The feature variation and semantic gap between 2D pixels and 3D coordinates prohibit the learning of accurate correspondence between two modalities.

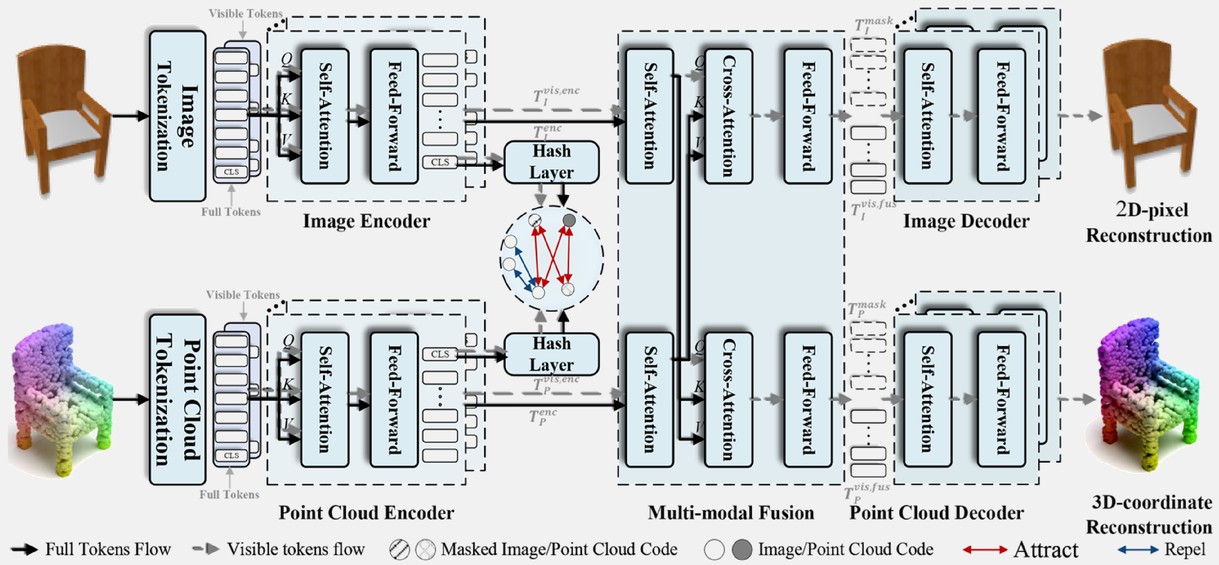
The Proposed CMAH

We propose CMAH, which is the first job that attempts to address the issue of self-supervised cross-modal hashing for images and point-cloud data. The key insight lies in harnessing the power of contrastive learning and masked auto-encoders to capture global data relationships while also acquiring fine-grained cross-modal associations.

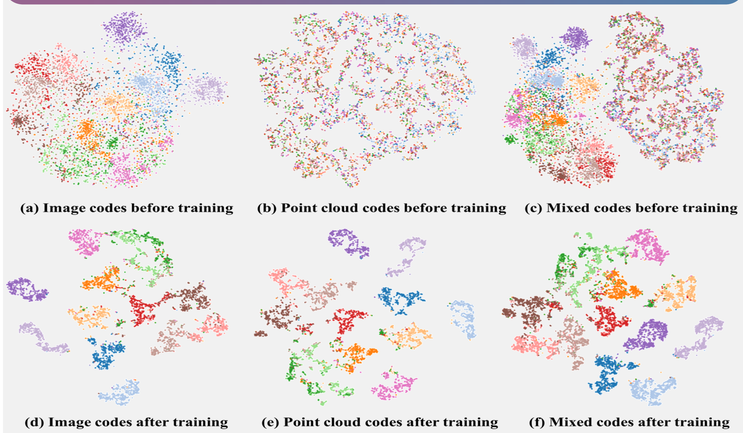
✓ In multi-modal contrastive learning, we contrast both full-full and masked-full pairs of the image and point-cloud data to achieve explicit modal alignment.

✓ In multi-modal fusion, we employ the co-attention mechanism to fuse fine-grained features from different modalities.

✓ In masked image/point cloud reconstruction, we reconstruct masked 2D patches and 3D coordinates, allowing the model to focus more on local semantics.



Visualization



(a) Image codes before training (b) Point cloud codes before training (c) Mixed codes before training

(d) Image codes after training (e) Point cloud codes after training (f) Mixed codes after training

✓Learned hash codes show compactness for the same category and dispersion for different ones.

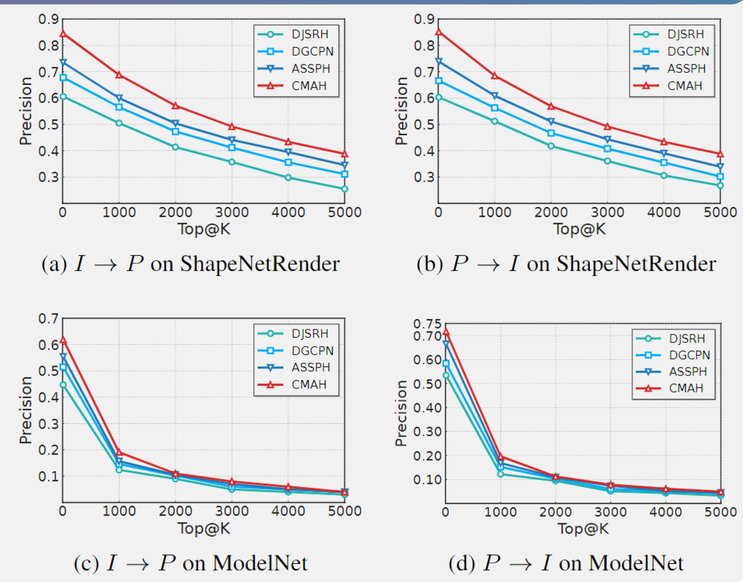
✓Hash codes of images and point clouds with similar semantics can be closely grouped together.

Comparisons with SOTAs

Task	Method	ShapeNetRender			ModelNet			ShapeNet-55		
		16-bit	32-bit	64-bit	16-bit	32-bit	64-bit	16-bit	32-bit	64-bit
I→P	DJSRH[5]	0.624	0.642	0.663	0.314	0.349	0.393	0.466	0.491	0.513
	DGCPN[1]	0.681	0.695	0.717	0.356	0.387	0.424	0.513	0.532	0.560
	ASSPH[9]	0.716	0.740	0.749	0.373	0.411	0.446	0.541	0.564	0.593
	CMAH	**0.760**	**0.775**	**0.791**	**0.409**	**0.466**	**0.501**	**0.577**	**0.619**	**0.637**
P→I	DJSRH[5]	0.610	0.647	0.661	0.357	0.384	0.429	0.474	0.506	0.531
	DGCPN[1]	0.674	0.697	0.711	0.386	0.422	0.465	0.522	0.549	0.576
	ASSPH[9]	0.709	0.736	0.744	0.409	0.448	0.487	0.557	0.581	0.614
	CMAH	**0.758**	**0.772**	**0.790**	**0.452**	**0.516**	**0.545**	**0.585**	**0.626**	**0.647**

✓ CMAH outperforms ASSPH by 5.6%, 12.3%, and 7.4% on the three benchmarks for the $I \rightarrow P$ task at 64 bits. Similarly, for the $P \rightarrow I$ task, CMAH achieves improvements of 6.2%, 11.9%, and 5.4%, respectively.

✓ CMAH excels at retrieving more semantically similar and relevant results compared to others, irrespective of changes in the search radius.



(a) $I \rightarrow P$ on ShapeNetRender (b) $P \rightarrow I$ on ShapeNetRender

(c) $I \rightarrow P$ on ModelNet (d) $P \rightarrow I$ on ModelNet