JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

北京邮電大學
Beijing University of Posts and Telecommunications

UNIVERSITY OF SURREY

Zero-Shot Everything Sketch-Based Image Retrieval, and in Explainable Style

Highlight

Poster Session THU-PM.   Poster Location: West Building Exhibit Halls ABC 262

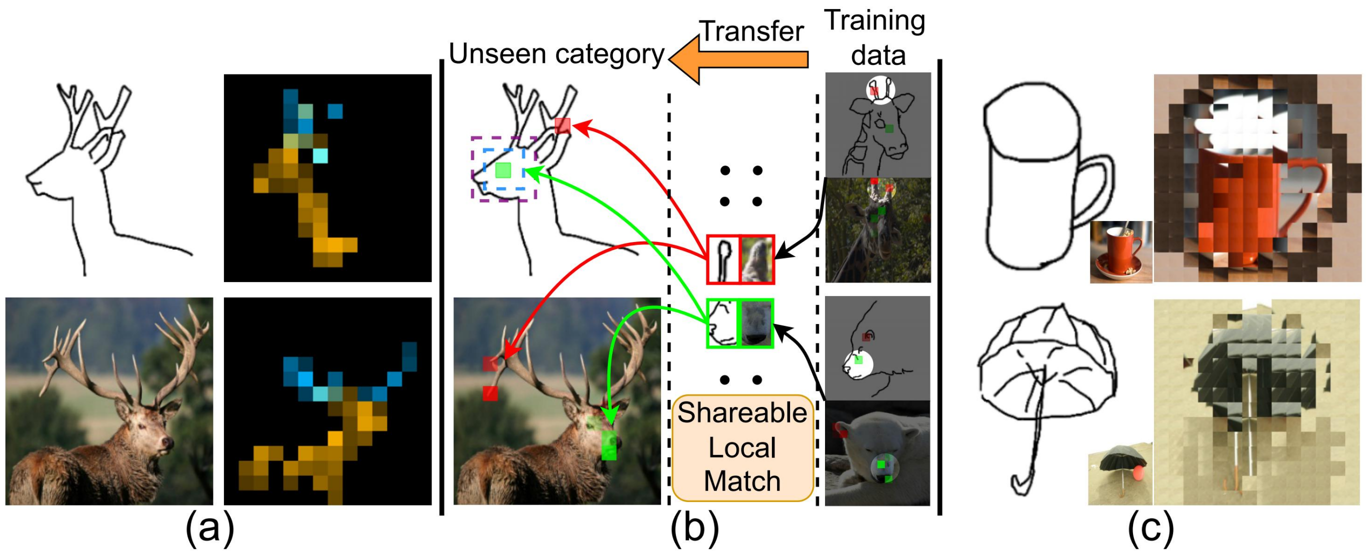
Fengyin Lin[1*], Mingkang Li[1*], Da Li[2†], Timothy Hospedales[2,3], Yi-Zhe Song[4], Yonggang Qi[1]

[1]Beijing University of Posts and Telecommunications   [2]Samsung AI Centre, Cambridge
[3]University of Edinburgh   [4]SketchX, CVSSP, University of Surrey

ZSE-SBIR Overview

- "E: Everything": We tackle three variants (inter-category, intra-category, and cross datasets) of ZS-SBIR with just one network.
- "E: Explainable": to understand how sketch-photo matching operates.
- A transformer-based cross-modal network was proposed with three specific designs:
  - ✓ Self-attention module with a learnable tokenizer
  - ✓ Cross-attention module
  - ✓ A kernel-based relation network

1. Introduction

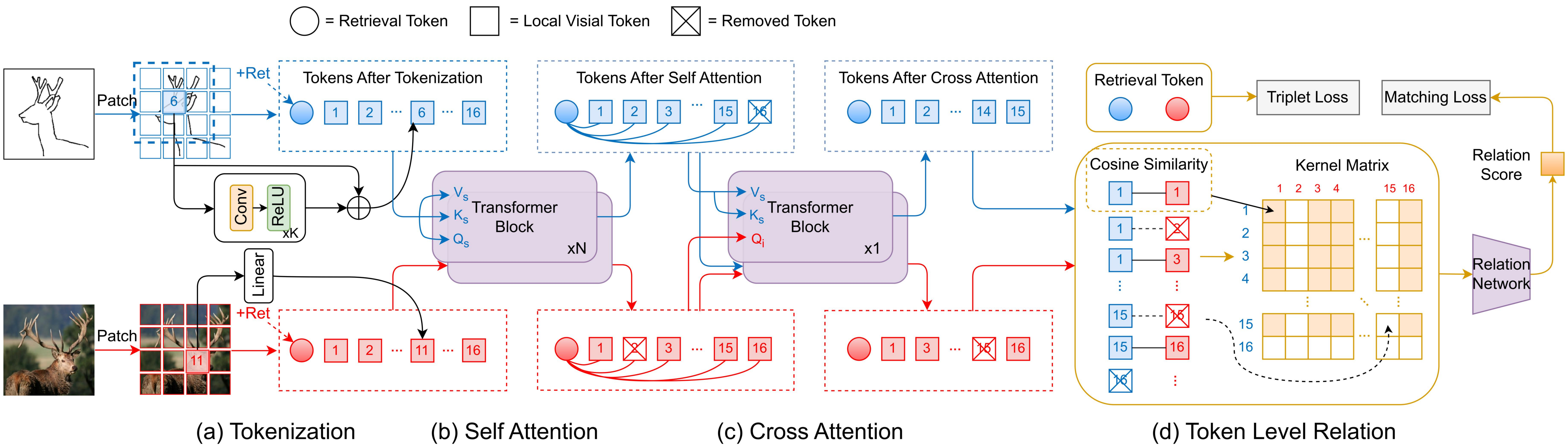


(a) The proposed retrieval token [Ret] can attend to informative regions.

(b) Cross-attention offers explainability by explicitly constructing local visual correspondence.

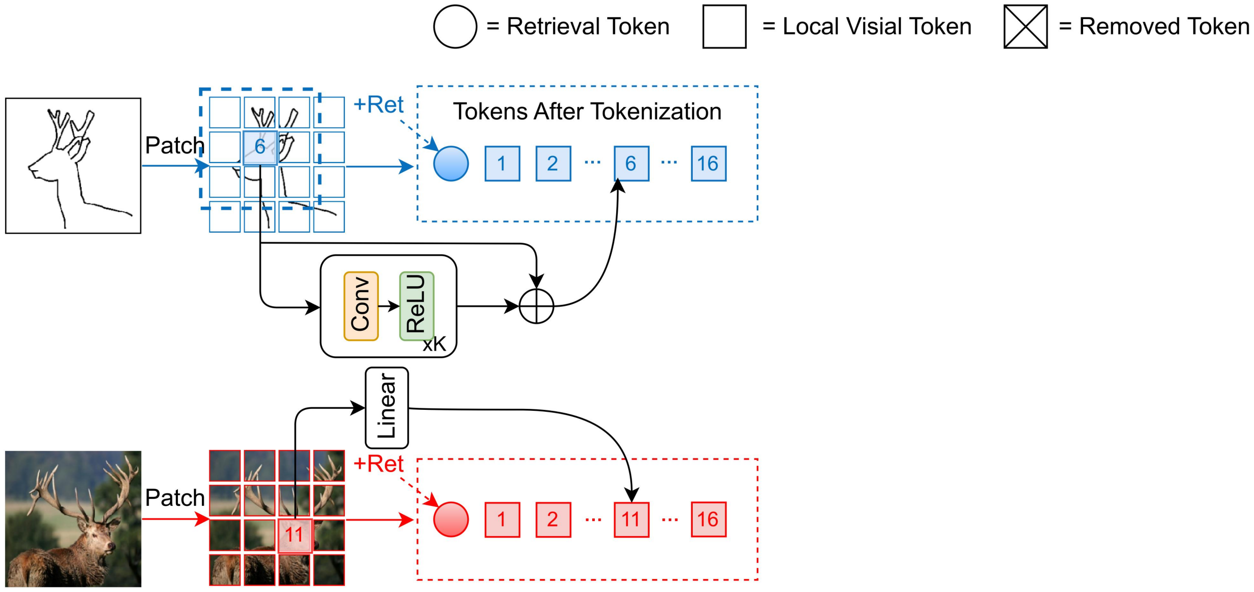
(c) An input sketch can be transformed into its image by the learned correspondence.

2. Methodology



2. Methodology

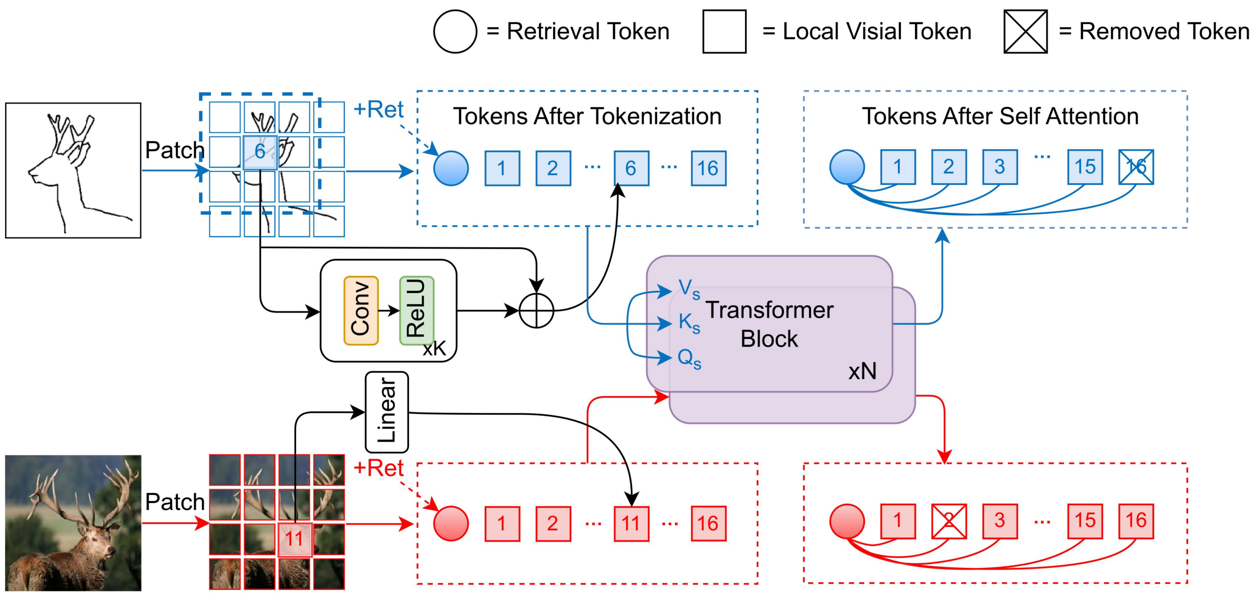
- Learnable Tokenization: $X = [\sigma(\mathrm{Conv}(\mathrm{S}))]_{\times 4}$ $X = X + S'$



(a) Tokenization

2. Methodology

- Self-attention with Retrieval Token: $$\text{s-attn}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

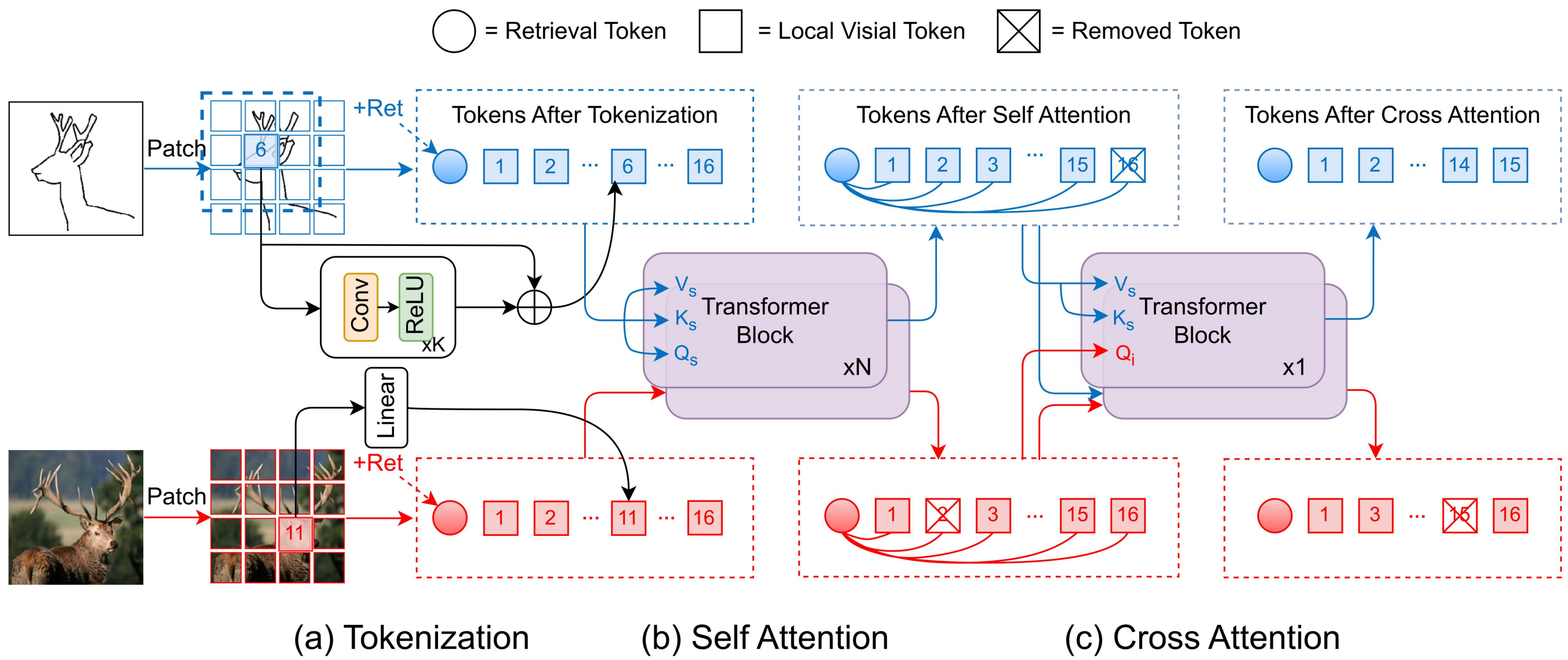


(a) Tokenization (b) Self Attention

2. Methodology

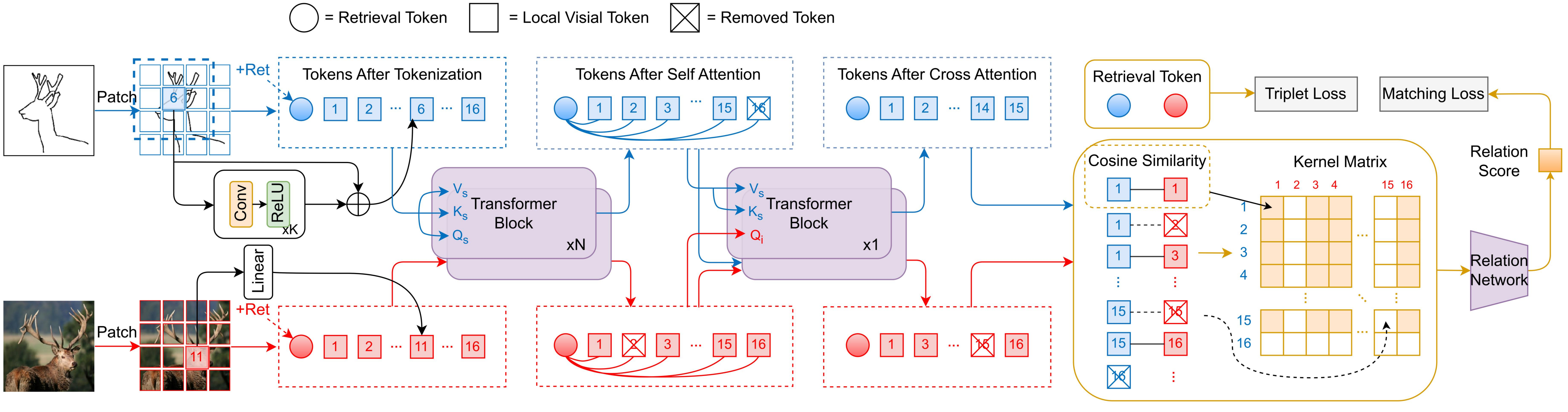
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- Cross-modal Attention: $\text{c-attn}(Q_I, K_S, V_S) = \text{softmax}\left(\frac{Q_I K_S^T}{\sqrt{d}}\right) V_S$



2. Methodology

- Kernel based Relation Network: $M_{i,j}^{S,I} = \frac{X_S^i \cdot X_I^{jT}}{\|X_S^i\| \|X_I^j\|}$ $\quad r(S,I) = \mathrm{sigmoid}(R_\psi(M^{S,I}))$



(a) Tokenization (b) Self Attention (c) Cross Attention (d) Token Level Relation

3. Experiments

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- Category-level ZS-SBIR:

Table 1. Category-level ZS-SBIR comparison results. "ESI" : External Semantic Information. "-" : not reported. The best and second best scores are color-coded in red and blue.

Method	ESI	$\mathbb{R}^D$	TU-Berlin Ext		Sketchy Ext		Sketchy Ext [28] Split		QuickDraw Ext	
			mAP	Prec@100	mAP	Prec@100	mAP@200	Prec@200	mAP	Prec@200
ZSIH [50]	✓	64	0.220	0.291	0.254	0.340	-	-	-	-
CC-DG [40]	✗	256	0.247	0.392	0.311	0.468	-	-	-	-
DOODLE [16]	✓	256	0.109	-	0.369	-	-	-	0.075	0.068
SEM-PCYC [19]	✓	64	0.297	0.426	0.349	0.463	-	-	-	-
SAKE [34]	✓	512	0.475	0.599	0.547	0.692	0.497	0.598	0.130	0.179
SketchGCN [67]	✓	300	0.324	0.505	0.382	0.538	-	-	-	-
StyleGuide [20]	✗	200	0.254	0.355	0.376	0.484	0.358	0.400	-	-
PDFD [13]	✓	512	0.483	0.600	0.661	0.781	-	-	-	-
ViT-Vis [18]	✗	512	0.360	0.503	0.410	0.569	0.403	0.512	0.101	0.113
ViT-Ret [18]	✗	512	0.438	0.578	0.483	0.637	0.416	0.522	0.115	0.127
DSN [57]	✓	512	0.484	0.591	0.583	0.704	-	-	-	-
BDA-SketRet [8]	✓	128	0.375	0.504	0.437	0.514	0.556	0.458	0.154	0.355
SBTKNet [55]	✓	512	0.480	0.608	0.553	0.698	0.502	0.596	-	-
Sketch3T [44]	✓	512	0.507	-	0.575	-	-	-	-	-
TVT [54]	✓	384	0.484	0.662	0.648	0.796	0.531	0.618	0.149	0.293
Ours-RN	✗	512	0.542	0.657	0.698	0.797	0.525	0.624	0.145	0.216
Ours-Ret	✗	512	0.569	0.637	0.736	0.808	0.504	0.602	0.142	0.202

3. Experiments

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- Generalized ZS-SBIR:

Table 2. Generalized ZS-SBIR results.

Method	TU-Berlin Ext		Sketchy Ext	
	mAP	Prec@100	mAP	Prec@100
SEM-PCYC [19]	0.192	0.298	0.307	0.364
StyleGuide [20]	0.149	0.226	0.331	0.381
BDA-SketRet [8]	0.251	0.357	0.338	0.413
SBTKNet [55]	0.334	0.494	0.515	0.572
Ours-RN	0.432	0.460	0.634	0.651
Ours-Ret	0.464	0.485	0.656	0.670

- Zero-shot Fine-grained SBIR:

Table 5. Zero-shot FG-SBIR results (%). Note that all competitors are *not* zero-shot models, they are trained on Chair-V2.

Method	TripLet-SAN [62]	DSA [52]	TripLet-RL [2]
acc.@1	47.65	53.41	56.54
acc.@10	84.24	87.56	89.61
Method	StyleMeUp [45]	CC-DG [40]	Ours-RN/Ours-Ret
acc.@1	62.86	54.21	63.34/**64.31**
acc.@10	91.14	88.23	**94.53**/92.60

- Cross-Dataset category-level ZS-SBIR:

Table 6. Cross-dataset ZS-SBIR results. "S", "T" and "Q" denote Sketchy Ext, TU-Berlin Ext, and QuickDraw Ext, respectively. "(·)" denotes the number of test categories which are unseen to ensure the zero-shot setting. E.g., S→T(21) denotes that, we train on the training split of Sketchy Ext, then test on a subset (21 unseen classes) of the testing split of TU-Berlin Ext. Rows with a grey background indicate using ViT backbone for fair comparisons.

Method	S→T (21)		S→Q (11)		T→S (8)		T→Q (10)	
	mAP	Prec@100	mAP	Prec@100	mAP	Prec@100	mAP	Prec@100
CC-DG [40]	0.252	0.403	0.148	0.212	0.570	0.660	0.214	0.278
	0.308	0.434	0.156	0.227	0.624	0.693	0.231	0.296
DSN [57]	0.384	0.480	0.152	0.171	0.646	0.673	0.229	0.251
	0.356	0.469	0.149	0.178	0.613	0.654	0.218	0.246
SAKE [34]	0.421	0.549	0.183	0.250	0.657	0.722	0.248	0.340
	0.389	0.506	0.174	0.242	0.626	0.701	0.235	0.318
Ours-RN	**0.476**	**0.590**	**0.228**	**0.338**	**0.746**	**0.816**	**0.273**	**0.376**

3. Experiments

- Top 5 retrieval results:

Figure 3. Exemplar comparison retrieval results for the given query sketches and the top 5 retrieved images. Red box denotes false positive.

3. Experiments

- Self-attention map:

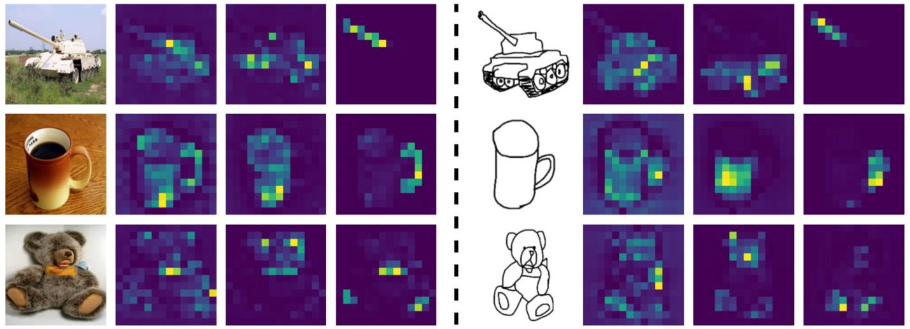


Figure 5. Attention maps of self-attention module on unseen categories. Given the tensors (heads) of the last layer of the self-attention module, we display the attention maps by using the retrieval token [Ret] as query. Original inputs are in the first column, followed by attention maps from multiple heads.

- Cross-modal visual correspondence:

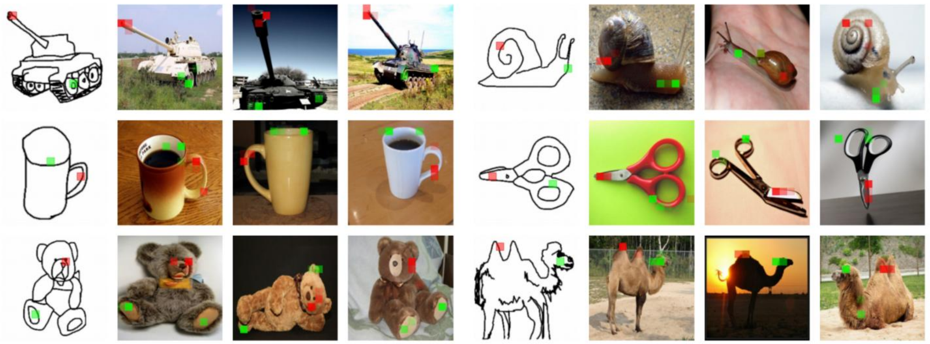


Figure 6. Visual correspondence across two modalities. Given a query sketch with two manually selected key regions (color-coded in red and green), we show the retrieved images with the corresponding matched regions (Top 3) in the same color.

3. Experiments

- Sketch-to-photo synthesis:

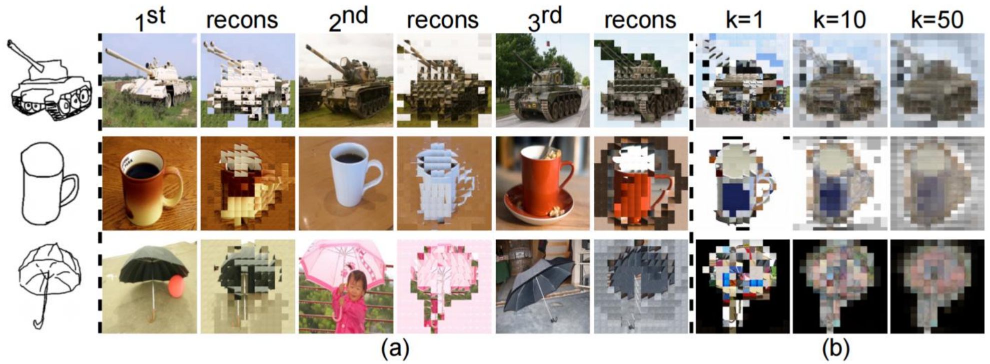


Figure 7. Cross-modal patch replacement. Given a sketch, (a) "recons" images are obtained by replacing sketch patches with the closest image patches of the top-3 retrieved images. (b) Reconstructed images using the k-nearest patches of the whole gallery.

- How transfer happens:

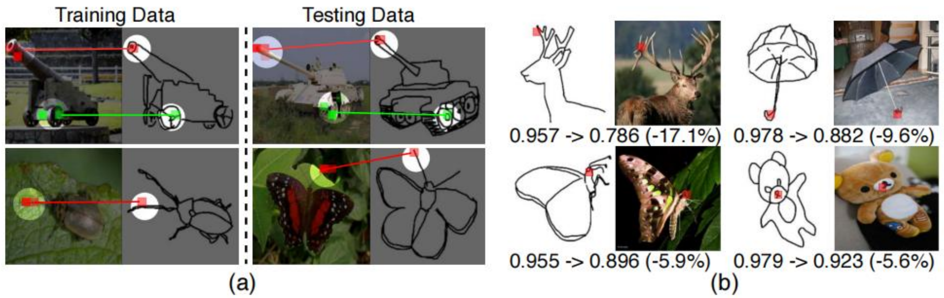


Figure 8. (a) Example of shareable local matches. The observed visual correspondences in training data show up again in testing data. (b) Example of most important token pair (red) which led maximum reduction of the matching score. Zoom in for best view.

3. Experiments

- Ablation study:

Table 3. Ablation study results on manifesting importance of each key *component*, and using different *token selection rates*.

	Model	Keep Rate		TU-Berlin Ext		Sketchy Ext		RPM
		r_S^{SA}/r_I^{SA}	r^{CA}	mAP	Prec@100	mAP	Prec@100	(ms)
Components	w/o CA	-	-	0.294	0.352	0.295	0.346	-
	w/o SA	-	-	0.256	0.388	0.286	0.381	-
	w/o Cos-K	-	-	0.342	0.419	0.390	0.481	-
	w/o RN loss	-	-	0.497	0.610	0.656	0.744	-
	w/o [Ret]	-	-	0.519	0.623	0.681	0.767	-
	w/o L-Tok	-	-	0.514	0.621	0.672	0.767	-
	Ours-full	-/-	-	0.542	0.657	0.698	0.797	0.148
Token Selection	Ours-full	0.9/0.9	1.0	0.523	0.634	0.682	0.786	0.108
	Ours-full	0.7/0.7	1.0	0.509	0.619	0.671	0.778	0.056
	Ours-full	0.5/0.5	1.0	0.432	0.571	0.596	0.743	0.028
	Ours-full	0.7/0.9	1.0	0.519	0.628	0.678	0.782	0.082
	Ours-full	0.9/0.7	1.0	0.512	0.622	0.673	0.779	0.082
	Ours-full	0.7/0.7	0.9	0.510	0.618	0.668	0.774	0.055
	Ours-full	0.7/0.7	0.7	0.497	0.604	0.653	0.762	0.052

- Token selection:

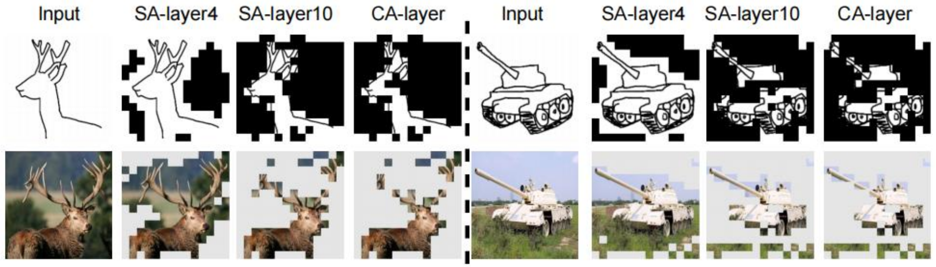


Figure 4. Visualization of token selection at different layers by setting keep rate for SA layers to 0.7 and the CA layer to 0.9.

- Computational cost analysis:

Table 4. Comparison of computational cost.

	SAKE [34]	SEM-PCYC [19]	Ours-RN (SA+CA)	Ours⋆
# Params (M)	27.6	137.9	102.2(87.8+14.4)	102.2
GFLOPs	3.90	15.5	19.5 (17.8+1.7)	12.6 (12.0+0.6)
RPM (ms)	0.138	0.070	0.148 (0.118+0.030)	0.056 (0.048+0.008)

4. Conclusion

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- A transformer-based cross-modal network that sources local patches independently in each modality, and establishes patch-to-patch correspondences across two modalities.

- A kernel-based relation network to aggregate the correspondences and calculate a similarity score between each sketch-photo pair.

- Explainability offered as per tradition in terms of visualizing patch correspondences, and by replacing all patches in a sketch with their photo correspondences.

Paper

Source Code
https://github.com/buptLinfy/ZSE-SBIR

Thank you!

Poster Session THU-PM. Poster Location: West Building Exhibit Halls ABC 262