Cross-Modal Scene Semantic Alignment for Image Complexity Assessment

YuQing Luo[1], Yixiao Li[1], Wei Zhou[1,*], and etc

([1]School of Computer Science and Informatics, Cardiff University, UK)

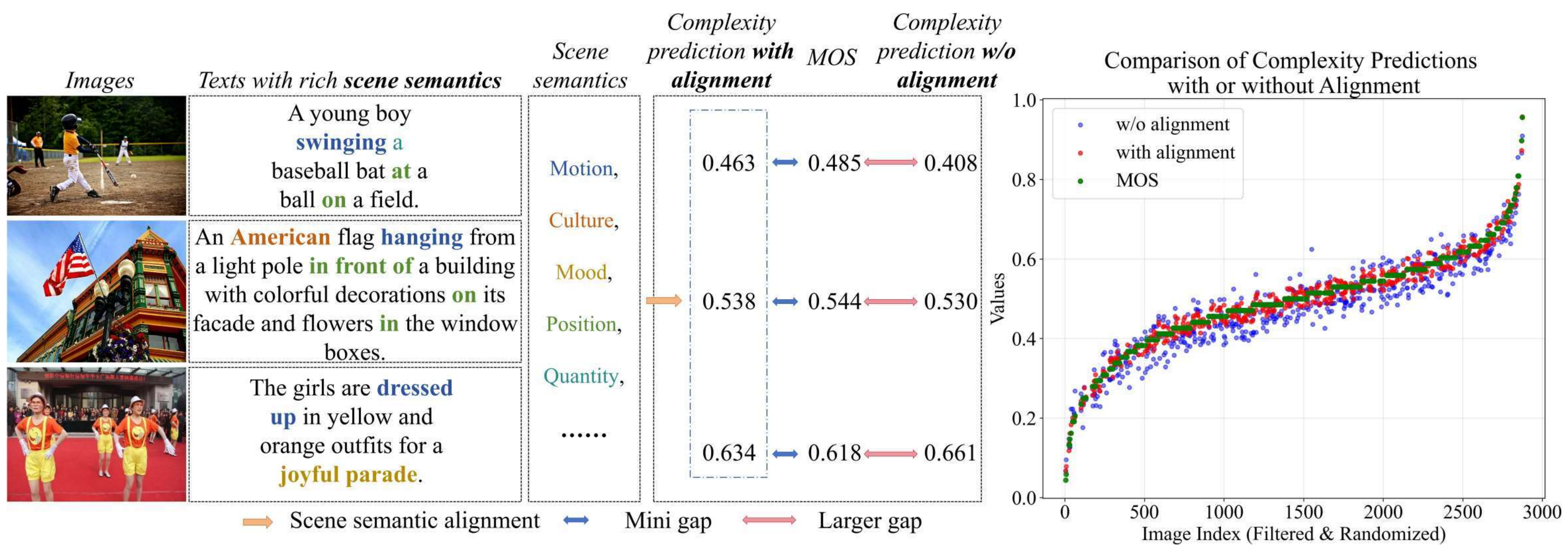
*zhouw26@cardiff.ac.uk

Image Complexity Assessment

■ **Goal**

- Complexity predictions that highly align with human perception.

➢ What Role Does Scene Semantics Play in Complexity Evaluation?

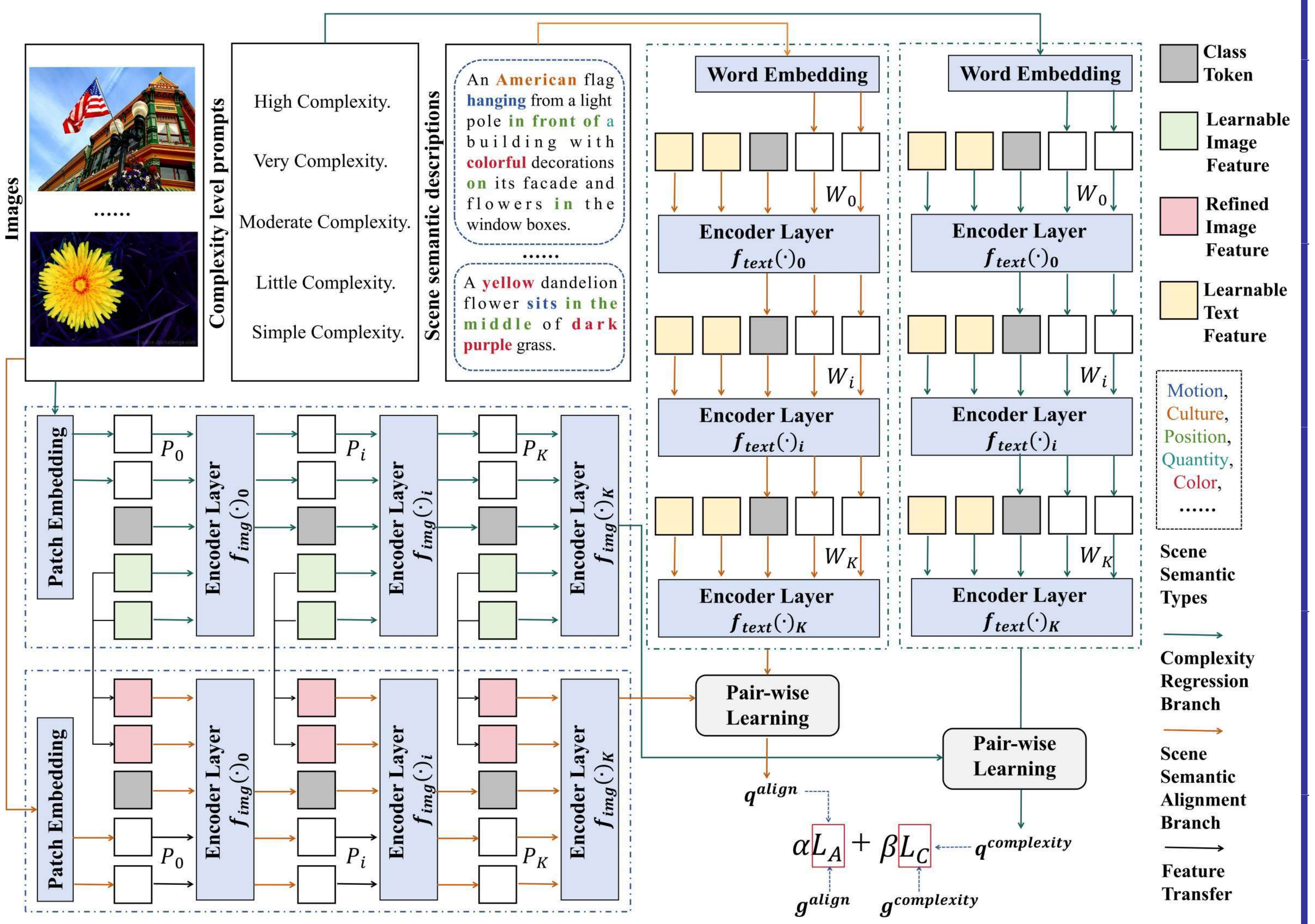


The Complexity Predictions guided by Scene Semantic Alignment are more align with human visual perception !

DATASET

- IC9600: 9600 images across eight semantic categories.
- VISC-C: 800 images across eight semantic categories.
- SAVOIAS: 1,400 images from seven categories.

CM-SSA: Cross-Modal Scene Semantic Alignment



$$[Z_i^{\text{Img}}, Z_k^{\text{Complexity}}] = [f_{\text{img}}(X_i^{\text{Img}}), f_{\text{text}}(X_k^{\text{Complexity}})],$$

$$s_{ik}^{complexity} = \frac{Z_i^{\text{Img}} \odot Z_k^{\text{Complexity}}}{\|Z_i^{\text{Img}}\| \cdot \left\|Z_k^{\text{Complexity}}\right\|},$$

$$q_i^{complexity} = \frac{e^{s_{i1}^{complexity}}}{\sum_{k=1}^{5} e^{s_{ik}^{complexity}}}, i \in \{1,2,\ldots,N\}, k \in \{1,2,3,4,5\}.$$

$$L_C = \frac{1}{N}\sum_{i=1}^{N}\left\|q_i^{complexity} - g_i^{complexity}\right\|_2^2,$$

$$[Z_i^{\text{Img}}, Z_i^{\text{Scene}}] = [f_{\text{img}}(X_i^{\text{Img}}), f_{\text{text}}(X_i^{\text{Scene}})], \quad i = 1,2,\ldots,N.$$

$$s_i^{scene} = \frac{Z_i^{\text{Img}} \odot Z_j^{\text{Scene}}}{\|Z_i^{\text{Img}}\| \cdot \left\|Z_j^{\text{Scene}}\right\|},$$

$$q_i^{align} = \frac{e^{s_i^{scene}}}{\sum_{j=1}^{N} e^{s_j^{scene}}}, i, j \in \{1,2,\ldots,N\}.$$

$$L_A = \frac{1}{N}\sum_{i=1}^{N}\left\|q_i^{align} - g_i^{align}\right\|_2^2,$$

RESULTS

Table 1: Performance comparisons on the IC9600, VISC-C, and Savoias datasets. The best-performing results are highlighted in bold. The calculation of model Flops is based on the input image shape $3 \times 512 \times 512$, and the inference speed is taken from an average of 100 inferences. * The CLICv2 is pretrained on IC1M dataset and obtained by linear probing on the IC9600. * The Savoias results of DINO-v2 are pretrained on IC9600.

Dataset	IC9600			VISC-C			Savoias			Complexity		
Model	SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓	Params/M	Flops/G	Speed/ms
CR [29]	0.314	0.228	0.196	-	-	-	0.305	0.271	0.257	-	-	-
ED [16]	0.491	0.569	0.226	-	-	-	0.449	0.467	0.273	-	-	-
DBCNN [50]	0.871	0.879	0.071	0.779	0.783	0.086	0.768	0.770	0.147	15.31	86.22	23.69
NIMA [44]	0.838	0.555	0.194	0.810	0.803	0.125	0.781	0.771	0.210	54.32	13.18	25.00
HyperIQA [42]	0.926	0.933	0.204	0.734	0.739	0.181	0.801	0.798	0.293	27.38	107.83	29.64
CLIPIQA [46]	0.897	0.898	0.078	0.781	0.796	0.122	0.779	0.794	0.101	-	61.07	21.76
TOPIQ [4]	0.938	0.944	0.049	0.803	0.811	0.079	0.838	0.832	0.123	45.20	37.26	14.43
CNet [21]	0.870	0.873	-	-	0.716	-	-	-	-	-	-	-
*CLICv2 [28]	0.927	0.933	-	-		-	-	-	-	-	-	-
*DINO-v2 [3]	0.847	0.851	0.086	-		-	0.675	0.689	-	-	-	-
ICCORN [17]	0.951	0.954	0.048	0.766	0.784	0.332	-	-	-	-	-	-
ICNet [12]	0.937	0.946	0.049	0.818	**0.814**	0.079	0.865	0.849	0.121	20.33	28.40	4.56
CM-SSA	**0.958**	**0.961**	**0.009**	**0.823**	0.805	**0.018**	**0.883**	**0.875**	**0.026**	205.45	52.66	35.43

Table 2: Cross-dataset validation. The best performances are highlighted in bold.

Train	IC9600				VISC-C				Savoias			
Test	VISC-C		Savoias		IC9600		Savoias		IC9600		VISC-C	
Model	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
DBCNN	0.685	0.674	0.606	0.626	0.742	0.750	0.525	0.538	0.773	0.787	0.644	0.652
NIMA	0.566	0.297	0.603	0.240	0.714	0.390	0.575	0.540	0.818	0.725	0.630	0.588
HyperIQA	0.711	0.688	0.669	0.669	0.668	0.668	0.550	0.554	0.761	0.748	0.643	0.629
CLIPIQA	0.702	0.686	0.600	0.608	0.406	0.399	0.487	0.481	0.789	0.775	0.550	0.568
TOPIQ	0.759	0.730	0.659	0.648	0.695	0.700	0.594	0.597	0.806	0.804	0.674	0.685
ICNet	0.753	0.712	**0.728**	**0.716**	**0.718**	**0.739**	**0.664**	**0.660**	0.805	0.807	**0.712**	**0.709**
CM-SSA	**0.760**	**0.740**	0.710	0.712	0.577	0.556	0.541	0.486	**0.838**	**0.828**	0.702	0.681

Table 3: Performance comparison of different branch configurations. The text prompts are coarse-grained complexity prompts when the single complexity regression branch Branch_C is utilized, and the text prompts are fine-grained textual prompts when the single text-driven scene semantic branch Branch_A is utilized. The best results are highlighted in bold.

Branch_C	Branch_A	SRCC ↑	PLCC ↑	RMSE↓	RMAE↓
✓	×	0.947	0.951	0.010	0.008
×	✓	0.915	0.922	0.013	0.010
✓	✓	**0.958**	**0.961**	**0.009**	**0.007**

Table 4: Ablation study on the prompt levels in Branch_C. 3 levels refer to ["Simple Complexity, Moderate Complexity, High Complexity"], while 7 levels refer to ["Very Low Complexity, Fairly Low Complexity, Slightly Low Complexity, Moderate Level Complexity, Slightly High Complexity, Fairly High Complexity, Very High Complexity"].

Class Name	SRCC ↑	PLCC↑	RMSE↓	RMAE ↓
3 levels	0.951	0.954	0.010	0.008
5 levels	**0.958**	**0.961**	**0.009**	**0.007**
7 levels	0.955	0.960	**0.009**	**0.007**

CONCLUSION

✓ We propose the first cross-modal metric, namely Cross-Modal Scene Semantic Alignment (CM-SSA), for image complexity assessment.

✓ We introduce a dual-branch framework, including the complexity regression branch and the scene semantic alignment branch, for exploiting cross-modal scene semantic information to assist the complexity prediction.

✓ Experiments on several ICA datasets, including IC9600, VISC-C, and SAVOIAS, demonstrate that our proposed CM-SSA outperforms state-of-the-art methods. Ablation studies further validate the effectiveness of each proposed key component.

The 36th British Machine Vision Conference

24th - 27th November 2025, Sheffield, UK