Single Image Reflection Separation via Interactive Dual-Stream Transformers

Qiming Hu Hainuo Wang Xiaojie Guo

College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
huqiming@tju.edu.cn hainuo@tju.edu.cn xj.max.guo@gmail.com

Motivation

- **Significance of Reflection Separation**: Reflection superposition commonly occurs in real-world applications, such as surveillance, autonomous driving, and mobile photography. Addressing this issue is critical not only for improving these fields but also for advancing general layer-decomposition tasks, including denoising and obstacle removal.
- **Challenges in Reflection Separation**:
  - Reflection and transmission layers are both natural images, which makes disentangling them challenging, especially when they interact in complex, nonlinear ways due to varying lighting and medium materials.
  - Existing dual-stream methods suffer from limitations in handling feature correlations and face difficulties with restricted receptive fields, which are crucial for capturing global context in reflection separation tasks.

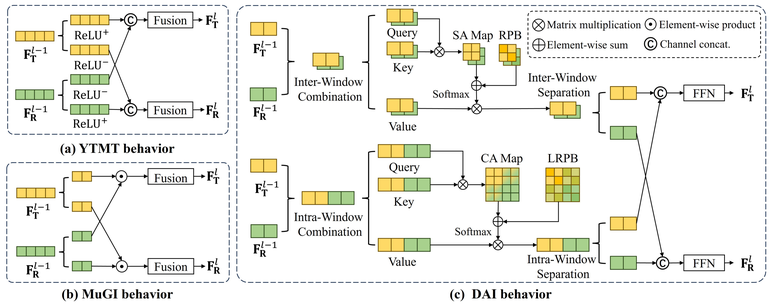

Figure 1: Illustration of dual-stream interactive behaviors, including YTMT, MuGI, and our proposed Dual-Attention Interaction (DAI) mechanisms.

- **Limitations of Previous Approaches**: As shown in Fig. 1, previous methods for dual-stream interactions, such as those used in YTMT and DSRNet, do not explicitly assess feature correlation during interactions. This often leads to the ineffective transmission of irrelevant information between layers, hindering the separation performance.

Contribution

- We propose a novel Dual-Attention Interaction (DAI) mechanism to energize Dual-Stream Interactive Transformers. DAI introduces the explicit correlation assessment within dual streams to effectively address the challenge of reflection separation;
- We customize a bridge, namely the Dual-Architecture Interactive Encoder (DAIE), to connect the pre-trained Transformer model with the task of layer decomposition, which alleviates the inherent ill-posedness of the problem;
- Through extensive experiments on multiple datasets, we demonstrate the efficacy of our design with superior performance over other SOTA competitors, both quantitatively and qualitatively. Moreover, the better generalizability compared to previous methods is also verified.

Dual-Stream Interactive Transformer

Overall architecture, shown in Fig. 2 (a), consists of a Dual-Architecture Interactive Encoder (DAIE) and a Dual-Stream Interactive Decoder (DSID).

DAIB: Takes transmission and reflection flows ($\mathbf{F}_\mathbf{T}^{\text{IN}}$, $\mathbf{F}_\mathbf{R}^{\text{IN}}$) as inputs, applies dual-stream self- and cross-attention (DS-SA, DS-CA) for inter- and intra-layer correlations, then outputs refined flows ($\mathbf{F}_\mathbf{T}^{\text{OT}}$, $\mathbf{F}_\mathbf{R}^{\text{OT}}$) through a locality-preserving block (DSLP).

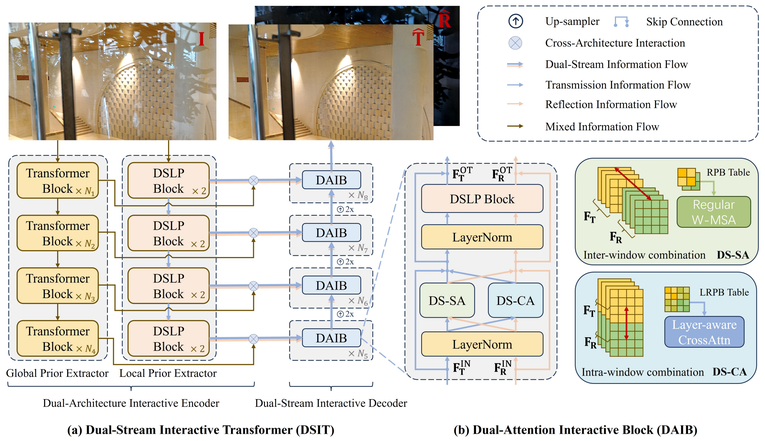

Figure 2: The overall architecture of our proposed Dual-Stream Interactive Transformer.

Efficient Dual-Stream Cross-Attention Mechanism: We introduce a streamlined cross-attention mechanism for dual-stream Transformers. Given the feature streams $\mathbf{F_T}$ and $\mathbf{F_R}$, we concatenate them as $\mathbf{X}_{\text{CA}}$. Using $\mathbf{X}_{\text{CA}}$, we compute queries, keys, and values as $\mathbf{Q}_{\text{CA}}$, $\mathbf{K}_{\text{CA}}$, and $\mathbf{V}_{\text{CA}}$.
The cross-attention scores, $\mathbf{A}_{\text{CA}} = \text{Softmax}(\mathbf{Q}_{\text{CA}}\mathbf{K}_{\text{CA}}^\top)$, capture intra- and inter-layer interactions. The output $\mathbf{Y}_{\text{CA}}$ is given by $\mathbf{A}_{\text{CA}}\mathbf{V}_{\text{CA}}$, containing both intra-layer terms and inter-layer terms. Using a simplified function $\mathcal{G}(\mathbf{Z}_1, \mathbf{Z}_2) = \mathcal{S}(\mathbf{Z}_1\mathbf{W}_q\mathbf{W}_k^T\mathbf{Z}_2^\top)\mathbf{Z}_2\mathbf{W}_v$, we express $\mathbf{Y}_{\text{CA}}$ as:

$$\mathbf{Y}_{\text{CA}} = \begin{bmatrix} \mathcal{G}(\mathbf{F_T}, \mathbf{F_T}) + \mathcal{G}(\mathbf{F_T}, \mathbf{F_R}) \\ \mathcal{G}(\mathbf{F_R}, \mathbf{F_T}) + \mathcal{G}(\mathbf{F_R}, \mathbf{F_R}) \end{bmatrix} = \begin{bmatrix} \mathbf{F}_\mathbf{T}^{\text{CA}} \\ \mathbf{F}_\mathbf{R}^{\text{CA}} \end{bmatrix}.$$

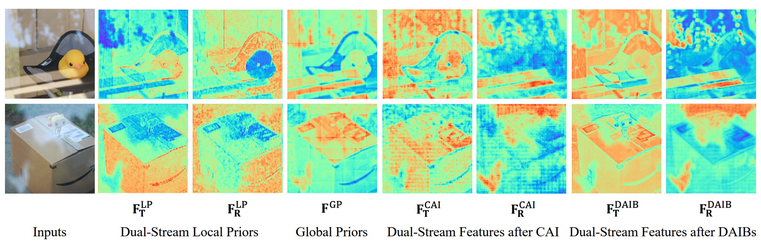

Figure 3: Visualization of extracted local priors, global priors, their cross-architecture-interacted dual-stream features and features after the DAIBs of two reflection-superimposed inputs.

Experiments & Results

Table 1: Quantitative results on real-world testing datasets of SIRS models.

Methods	Real20 (20)		Objects (200)		Postcard (199)		Wild (55)		Average	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Zhang *et al.*	22.55	0.788	22.68	0.879	16.81	0.797	21.52	0.832	20.08	0.835
BDN	18.41	0.726	22.72	0.856	20.71	0.859	22.36	0.830	21.65	0.849
ERRNet	22.89	0.803	24.87	0.896	22.04	0.876	24.25	0.853	23.53	0.879
IBCLN	21.86	0.762	24.87	0.893	23.39	0.875	24.71	0.886	24.10	0.879
RAGNet	22.95	0.793	26.15	0.903	23.67	0.879	25.53	0.880	24.90	0.886
DMGN	20.71	0.770	24.98	0.899	22.92	0.877	23.81	0.835	23.80	0.877
Zheng *et al.*	20.17	0.755	25.20	0.880	23.26	0.905	25.39	0.878	24.19	0.885
YTMT	23.26	0.806	24.87	0.896	22.91	0.884	25.48	0.890	24.05	0.886
RobustSIRR	23.30	0.827	24.90	0.917	19.91	0.868	23.67	0.884	22.59	0.889
DSRNet	24.23	0.820	26.28	0.914	24.56	0.908	25.68	0.896	25.40	0.905
PromptRR	24.11	0.813	24.17	0.859	23.03	0.895	26.43	**0.930**	23.95	0.880
Ours	**25.06**	**0.836**	**26.81**	**0.919**	**25.63**	**0.924**	**27.06**	0.910	**26.27**	**0.917**
Dong *et al.*†	23.34	0.812	24.36	0.898	23.72	0.903	25.73	0.902	24.21	0.897
DSRNet†	23.91	0.818	26.74	0.920	24.83	0.911	26.11	0.906	25.75	0.910
RRW†△	21.83	0.801	26.67	0.931	24.04	0.903	26.49	0.915	25.34	0.912
Zhong *et al.*†*	24.05	0.824	26.51	0.927	25.02	0.915	26.23	**0.925**	25.75	0.917
Ours†	**25.22**	**0.836**	**27.27**	**0.932**	**25.58**	**0.922**	**27.40**	0.918	**26.49**	**0.922**

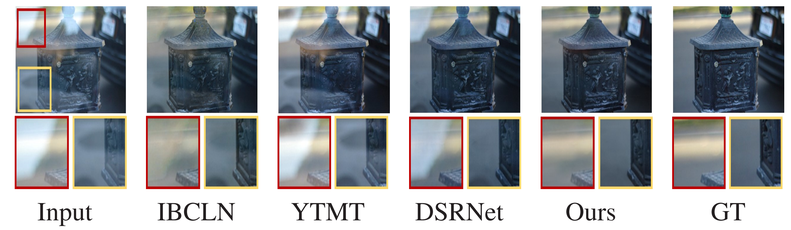

Figure 4: Visual comparison of transmission layer predictions between previous state-of-the-arts and ours on samples from Real20 and SIR^2 datasets.

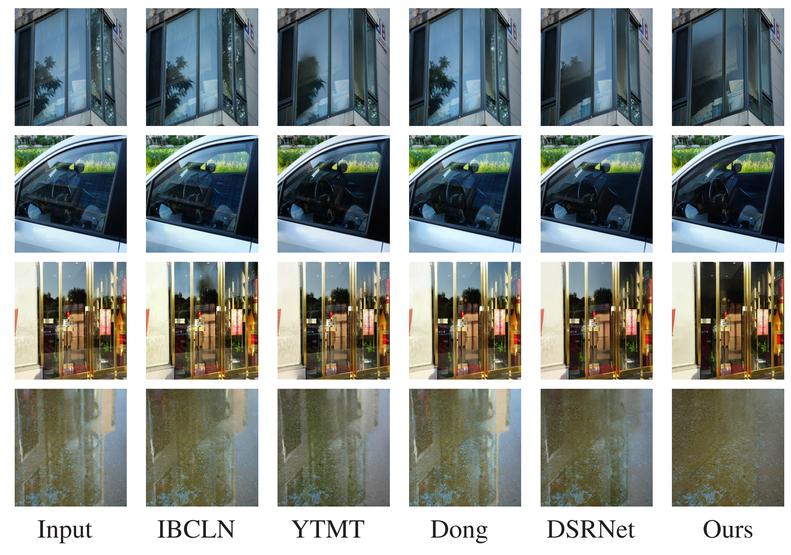

Figure 5: Visual comparison of transmission predictions between previous state-of-the-arts and ours in real-world scenarios additionally captured in this paper. The broad advantages demonstrated by our method across these diverse conditions highlight its superior generalization capability.