Distilling Reflection Dynamics for Single-Image Reflection Removal

Quanlong Zheng[1], Xiaotian Qiao[2], Ying Cao[2], Shi Guo[3,4], Lei Zhang[3,4] and Rynson Lau[2]

[1]OPPO Research, [2]City University of Hong Kong,

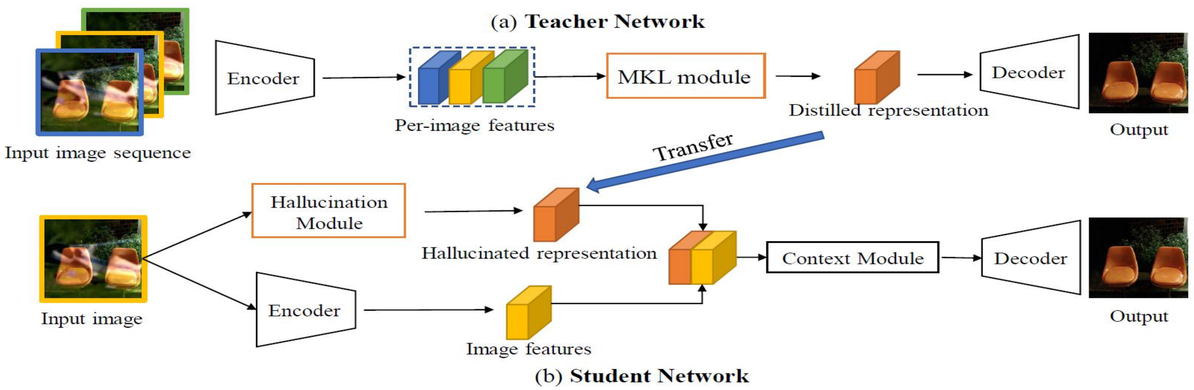
[3]The Hong Kong Polytechnic University, [4]DAMO Academy, Alibaba Group

1. Motivation

(a) Multi-view sequence (b) IBCLN [17] (c) Ours (d) GT

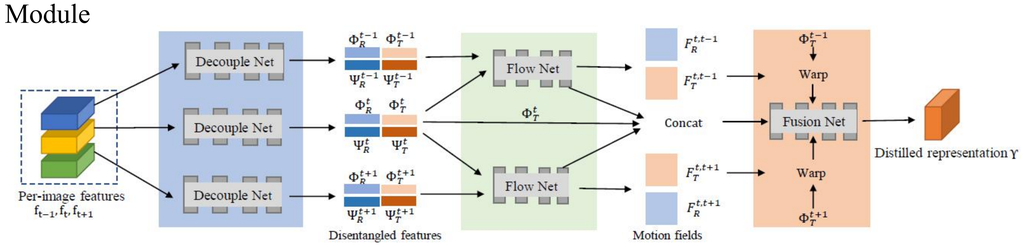
Given an image sequence of a static scene captured by moving a camera (a), observing the appearance change in the reflection regions (highlighted by red boxes) will make it easier to separate the transmission layer from the reflection layer. Such appearance change can be regarded as a result of moving the transmission layer and reflection layer along different trajectories. We learn such reflection dynamics from multi-view image sequences for single-image reflection removal (SIRR), enabling our method to remove the reflection more accurately (c) than the state-of-the-art method (b).

3. Learning Framework



Overview of our learning framework. (a) Teacher Network. The input is a multi-view image sequence and the output is the estimated transmission layer of the middle image of the input sequence. (b) Student Network. The input is a single image (or the middle image of the sequence during training) and the output is the estimated transmission image.

MKL Module



Objective Functions

- Teacher network

$$\mathcal{L}_{teacher} = \alpha_t \mathcal{L}_{\mathcal{R}} + \beta_t \mathcal{L}_{\mathcal{P}} + \gamma_t \mathcal{L}_{\mathcal{FD}} + \zeta_t \mathcal{L}_{VS}.$$

- Student network

$$\mathcal{L}_{student} = \alpha_s \mathcal{L}_{\mathcal{R}} + \beta_s \mathcal{L}_{\mathcal{P}} + \gamma_s (\mathcal{L}_{\mathcal{G}} + \mathcal{L}_{\mathcal{D}}) + \zeta_s \mathcal{L}_{T}.$$

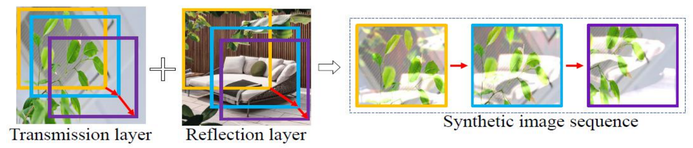
We first train the teacher model and freeze the weights of the teacher network. We then train the studernt work.

2. Contributions

We make the first attempt to distill reflection dynamics knowledge from *multi-view* images for *single image* reflection removal. We propose a framework that learns a representation of reflection dynamics from multiview image sequences and transfers it to single static images. we contribute a large-scale dataset of multi-view reflection image sequences, which can be used to train and evaluate both our model and traditional SIRR. Extensive evaluations on several benchmarks and our newly collected dataset show that our proposed method outperforms existing methods, achieving state-of-the-art results.

4. Datasets

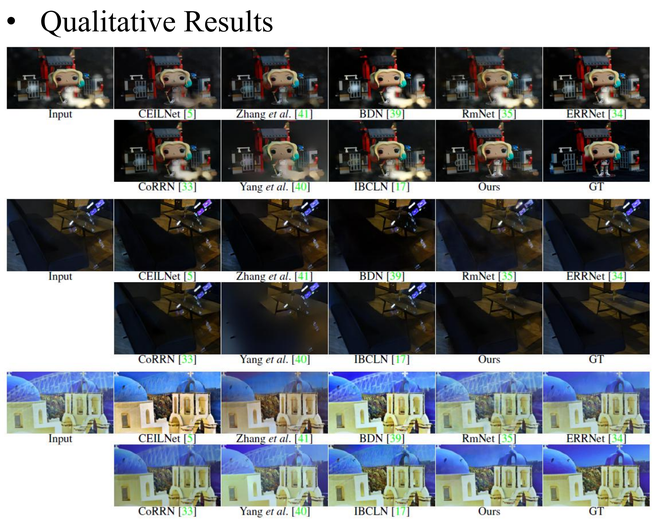
- Real world Multi-view Reflection Dataset
  - Cameras : Nikon D810 and Google Pixel2
  - Resolution : 1760 x 1160
  - # : 1,015 sequences
- Svnthetic Multi-view Reflection Dataset



Transmission layer Reflection layer Synthetic image sequence

5. Results

- Qualitative Results



- Quantitative Results

	Real20		Nature		SIR^2		SIR^2 - Object		SIR^2 - Postcard		SIR^2 - Wild		Seq1K	
	SSIM ↑	LMSE ↓	SSIM ↑	LMSE ↓	SSIM ↑	LMSE ↓	SSIM ↑	LMSE ↓	SSIM ↑	LMSE ↓	SSIM ↑	LMSE ↓	SSIM ↑	LMSE ↓
CEILNet [5]	0.7185	0.0294	0.7048	0.0218	0.8520	0.0080	0.8764	0.0054	0.8388	0.0088	0.8113	0.0148	0.8130	0.0179
Zhang *et al.* [41]	0.7988	0.0204	0.7401	0.0159	0.8454	0.0072	0.8847	0.0054	0.8082	0.0073	0.8369	0.0132	0.8522	0.0123
BDN [39]	0.7464	0.0297	0.7449	0.0190	0.8616	0.0062	0.8637	0.0054	0.8686	0.0054	0.8290	0.0120	0.8176	0.0186
RmNet [35]	0.7194	0.0306	0.7433	0.0169	0.8328	0.0097	0.8246	0.0108	0.8375	0.0081	0.8459	0.0119	0.8273	0.0154
ERRNet [34]	0.8036	0.0210	0.7590	0.0150	0.8807	0.0062	0.8872	0.0040	0.8786	0.0056	0.8644	0.0169	0.8711	0.0104
CoRRN [33]	0.7140	0.0299	0.7400	0.0154	0.8392	0.0055	0.8678	0.0045	0.8119	0.0054	0.8343	0.0096	0.7809	0.0158
Yang *et al.* [40]	0.7084	0.0287	0.7415	0.0162	0.8570	0.0065	0.8502	0.0065	0.8651	0.0059	0.8526	0.0087	0.7885	0.0189
IBCLN [17]	0.7816	0.0224	0.7845	0.0126	0.8948	0.0050	0.9020	0.0038	0.8880	0.0052	0.8934	0.0085	0.8738	0.0105
CEILNET (F)	0.7284	0.0267	0.7506	0.0151	0.8689	0.0057	0.8737	0.0044	0.8737	0.0053	0.8339	0.0118	0.8525	0.0117
Zhang (F)	0.7998	0.0202	0.8026	0.0118	0.8952	0.0044	0.9009	0.0036	0.8884	0.0043	0.8994	0.0071	0.8873	0.0083
RmNet (F)	0.7559	0.0245	0.7649	0.0136	0.8862	0.0049	0.8845	0.0041	0.8868	0.0053	0.8904	0.0065	0.8682	0.0104
ERRNet (F)	0.8120	0.0184	0.7950	0.0121	0.8940	0.0046	0.9028	0.0032	0.8810	0.0057	0.9090	0.0057	0.8934	0.0074
IBCLN (F)	0.7759	0.0231	0.7746	0.0131	0.8946	0.0043	0.9026	0.0032	0.880	0.0048	0.8892	0.0046	0.8672	0.0106
Ours	0.8196	0.0187	0.8213	0.0104	0.9009	0.0041	0.9089	0.0031	0.8908	0.0045	0.9084	0.0064	0.9015	0.0072

- Results of Ablation Study

	Real20		Nature		SIR^2		Seq1K	
	SSIM ↑	LMSE ↓	SSIM ↑	LMSE ↓	SSIM ↑	LMSE ↓	SSIM ↑	LMSE ↓
w/o MKL	0.7936	0.0210	0.8055	0.0121	0.8861	0.0050	0.8911	0.0079
w/o Decouple Net	0.8039	0.0195	0.8080	0.0115	0.8945	0.0049	0.8953	0.0076
w/o Flow Net	0.8069	0.0221	0.8096	0.0113	0.8946	0.0046	0.8967	0.0078
w/o Fusion Net	0.8108	0.0203	0.8185	0.0111	0.8957	0.0043	0.8976	0.0075
w/o KT	0.7944	0.0205	0.8068	0.0129	0.8794	0.0052	0.8859	0.0079
w/o Hallucination	0.8075	0.0200	0.8153	0.0111	0.8918	0.0047	0.8966	0.0074
w/o Encoder	0.8153	0.0194	0.8162	0.0106	0.8920	0.0048	0.8968	0.0075
Ours	**0.8196**	**0.0187**	**0.8213**	**0.0104**	**0.9009**	**0.0041**	**0.9015**	**0.0072**

6. Failure Cases

Input Ours GT

Our method may fail when a transmission layer is textureless and the reflection layer has strong texture.

7. Conclusions

We propose a teacher-student framework, where the teacher network learns a reflection dynamics representation from multi-view image sequences with a newly proposed multiview knowledge learning module and teaches a student network to remove reflection from single images. We also construct a large-scale real-world dataset of multi-view reflection image sequences for reflection dynamics distillation and for SIRR evaluation. Extensive experiments demonstrate the effectiveness of our method and the usefulness of the newly collected dataset for SIRR.