2023 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-6891-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICME55011.2023.00332

COLLABORATIVE SPATIAL-TEMPORAL DISTILLATION FOR EFFICIENT VIDEO DERAINING

Yuzhang Hu[1], *Minghao Liu*[1], *Wenhan Yang*[2], *Jiaying Liu*[1*] *and Zongming Guo*[1]

[1]Peking University, Beijing, China
[2]Peng Cheng Laboratory, Shenzhen, China

ABSTRACT

In this paper, we propose a novel knowledge distillation framework to improve the efficiency of deep networks for video deraining. The knowledge is transferred from a large-scale powerful teacher network to a compact efficient student network via the proposed collaborative spatial-temporal distillation framework. The framework is equipped with three collaboration schemes of different granularities that make use of spatial-temporal redundancy in a complementary way for better distillation performance. First, the spatial alignment module applies distillation constraints at different spatial scales to achieve better scale invariance in transferred knowledge. Second, the temporal alignment module traces both temporal status between teacher and student separately and collaboratively, to comprehensively utilize inter-frame information. Third, these two alignment modules interact through a spatial-temporal adaptor, where spatial-temporal knowledge is transferred in a unified framework. Extensive experiments demonstrate the superiority of our distillation framework as well as the effectiveness of each module. Our code is available at: https://github.com/HuYuzhang/Knowledge-Distillation.

Index Terms— Video Deraining, Knowledge Distillation, Spatial Alignment, Temporal Alignment, Spatial-Temporal Adaptor

1. INTRODUCTION

As one of the most commonly seen adverse weather, rain can cause a series of visual degradation. Specifically, rain streaks change pixel intensities and fluctuate illumination severely, which leads to occlusions and blurriness of background scenes. What is more, these kinds of degradation might also lead to the failure of many outdoor vision applications like auto driving. As a result, video deraining has drawn great research attention in recent years.

*Corresponding Author. This work is supported by the National Natural Science Foundation of China under Contract No.62172020, and a research achievement of Key Laboratory of Science, Techonology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). This research work is also partially supported by the Basic and Frontier Research Project of PCL and the Major Key Project of PCL.

This problem is firstly explored by Garg and Nayar [1]. The following methods [2, 3] design more complex priors to reconstruct the clean frames. After stepping into the deep learning era, a series of end-to-end trainable deep learning-based approaches [4, 5, 6] are proposed. Most existing video deraining methods take the sliding-window way to make use of the temporal information. That is to say, to remove the rain in the t-th frame, the temporally adjacent frames are also fed into the network to reconstruct the clean t-th frame. The size of the sliding-window is the number of input frames. To improve the deraining performance of a video deraining network, there are two main branches of methods. First, the size of the sliding-window keeps growing because it is reasonable that a larger sliding-window can contain more temporal information in a larger temporal receptive field. Second, the parameter number keeps increasing, which can also equip the model with a stronger capacity. Although significant improvement has been achieved, these two classes lead to larger models and more computational costs. However, such drawbacks lead to significant performance bottlenecks during practical deployment, particularly in some real-time application scenarios. As a result, it is of great significance to build efficient video deraining networks for practical applications.

Knowledge distillation [7, 8] provides a feasible direction to address the issue. It is first proposed in [7] to transfer the knowledge from a powerful large network (teacher network) to a smaller network (student network). Compared with the supervision only under the ground truth labels, the intermediate feature of the teacher network also contains a wealth of useful information, which can guide the student network to also enrich its intermediate feature to achieve better performance. For simplicity, we denote the teacher network and student network as Teacher and Student, respectively. Recently, there are some explorations of knowledge distillation in low-level visions. In [9], it is first explored to distill a single image super-resolution network with a more powerful one. In [10], the affinity matrix is integrated into the distillation loss to model the spatial correlation of the intermediate feature. In [11], the temporal distillation is introduced to exploit the inter-frame correlation for video super-resolution. Better performance is achieved compared with pure spatial distillation. However, it only performs plain feature alignment,

1945-788X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICME55011.2023.00332

Authorized licensed use limited to: Peng Cheng Laboratory. Downloaded on August 29,2023 at 06:33:46 UTC from IEEE Xplore. Restrictions apply.

which is not effective in utilizing the rich information of the intermediate feature. Besides, the spatial and temporal distillation are applied independently, and the cross-domain correlation is ignored. What is more, the distillation for video deraining has not been explored. The above distillation frameworks might be not optimal for this problem due to the intrinsic gap between super-resolution and video deraining.

In this paper, we propose a novel distillation framework consisting of two alignment modules to construct multiple collaboration schemes to address the above-mentioned issues. First, we propose a spatial alignment module, which takes the feature of different scales to construct the distillation loss to make the Student better at handling rain streaks of different scales. Different scales interact to collaboratively utilize both local and global information. Second, we propose a temporal alignment module to trace two kinds of temporal status of a rainy video. The features of Teacher and Student are aggregated both separately and collaboratively to comprehensively utilize inter-frame redundancy. Last, the spatial and temporal alignment modules mutually benefit through a Spatial-Temporal Adaptor. Specifically, except for transferring knowledge during the distillation process, the spatial/temporal alignment module will also output auxiliary information to assist the temporal/spatial alignment module alternately, to propagate the spatial-temporal information interactively. Our contributions are summarized as follows:

- We propose a Spatial Alignment Module with multiple scales to achieve better scale invariance in transferred knowledge and facilitate the student network to handle diverse kinds of rain streaks.
- We propose a Temporal Alignment Module with constraints on the temporal status of Student and Teacher separately and collaboratively, to guide the student network for more comprehensive temporal modeling.
- To further improve the distillation efficiency, a Spatial-Temporal Adaptor is proposed to make these two modules collaborate with cross-domain redundancy.

2. COLLABORATIVE SPATIAL-TEMPORAL DISTILLATION

Formulation. Denoting I_t as the t-th time-step rainy frame, a vanilla video deraining model takes consecutive rainy frames as inputs to predict the t-th time-step clean frame. Usually, I_t is the middle one of these input frames. The total number of input frames is denoted as $2k+1$, where k controls the size of the sliding window. To begin with, we decompose a video deraining network into two parts. The first one is the backbone module, where the input frames are mapped to the feature space. The second one is the reconstruction module, where the intermediate feature is mapped back to the image space to obtain the final deraining result. We use $\hat{F}_{T,t}$ and $\hat{F}_{S,t}$ to denote the intermediate features of the Teacher and the Student to recover the clean t-th frame, respectively. Following [9], the intermediate feature is aggregated to a single-channel one, which is denoted as $F_{T/S,t}$, namely $F_{T,t}$ or $F_{S,t}$. The following distillation loss is constructed on $F_{T/S,t}$.

Framework Overview. Our distillation framework consists of three modules, including the Spatial Alignment Module, the Temporal Alignment Module, and the Spatial-Temporal Adaptor as Fig. 1 (a) shows. The first two modules transfer the knowledge of the Teacher to the Student in the spatial domain and the temporal domain, respectively. The Spatial-Temporal Adaptor plays the role to exchange information between the first two modules in the distillation process. Specifically, the distillation process is performed recurrently across frames. For the t-th frame, the feature of the Teacher and Student $F_{T/S,t}$ is first fed to the Spatial Alignment Module to calculate the spatial distillation loss and obtain the spatial hint, which is the aggregated knowledge in the spatial domain. Then, the spatial hint is forwarded to the Temporal Alignment Module, which helps calculate the temporal distillation loss with $F_{T/S,t}$ and the previous temporal status h_{t-1}. Besides, the temporal hint is obtained, which is similarly the aggregated knowledge in the temporal domain and is forwarded to the Spatial Alignment Module for the distillation of the next time-step. This process is repeated progressively until all frames of a rainy video are processed. The detailed design of each module is described in the following sections.

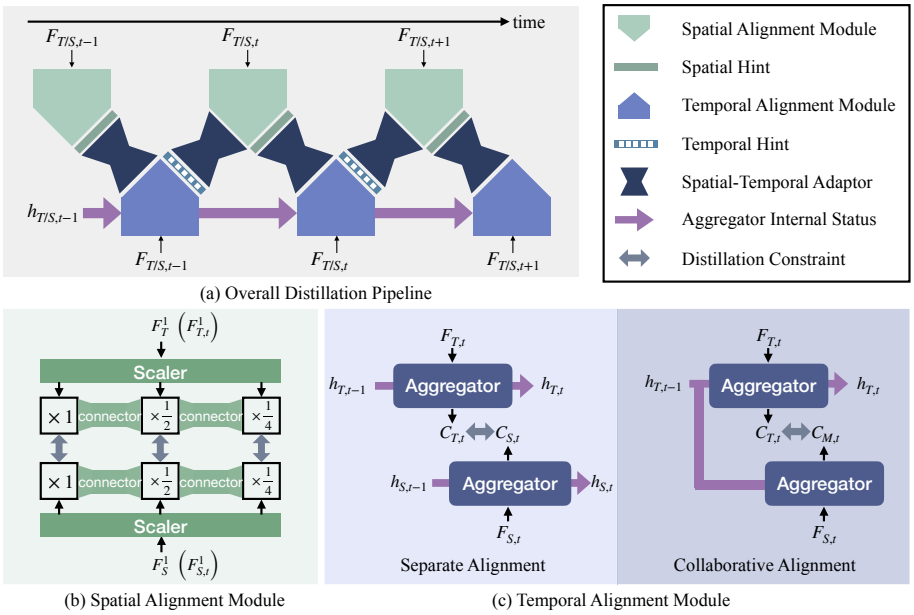
2.1. Spatial Alignment Module

This module transfers the knowledge from Teacher to Student by the feature of a single frame as shown in Fig. 1 (b). Considering the various directions, densities, and sizes of rain streaks, the original feature is down-sampled to 3 different scales to improve the invariance of the features distorted by different kinds of rain streaks. $F^p_{T/S,t}$ stands for the feature of the t-th time-step frame at the p-th scale. Here $p \in \{1, 2, 3\}$, corresponding to the original scale, the 1/2 scale, and the 1/4 scale, respectively. This module only involves the feature of the current t-th time-step frame, so we denote $F^p_{T/S,t}$ as $F^p_{T/S}$ for simplicity.

Rain streaks and background information are correlated among different scales, respectively, where this complementary information can benefit the distillation process For example, some rain streaks are of similar shape and direction across different scales. As a result, cross-scale collaboration can also contribute to more accurate rain removal. To this end, we connect the distillation process of different scales with a Connector for better utilization of cross-scale spatial information. The Connector merges the features at adjacent scales, and the spatial distillation loss is imposed on the merged feature. The Connector consists of the Local to Global (L2G) and Global to Local (G2L) connection passes. In the L2G connection pass, the input feature of a smaller scale is merged with the previously merged one of a larger-scale as follows:

$$C^{p,L2G}_{T/S} = f_\theta \left(F^p_{T/S}, C^{p-1,L2G}_{T/S} \right), \tag{1}$$

Authorized licensed use limited to: Peng Cheng Laboratory. Downloaded on August 29,2023 at 06:33:46 UTC from IEEE Xplore. Restrictions apply.



(a) Overall Distillation Pipeline

(b) Spatial Alignment Module

(c) Temporal Alignment Module

Fig. 1. The overall pipeline and module structure of our Collaborative Spatial-Temporal Distillation framework.

where $C_{T/S}^{p,L2G}$ is the merged feature at the p-th scale in the L2G pass and $f_\theta(\cdot)$ denotes the connector. Similarly, in the G2L pass, the input feature of a larger scale is merged with the previously merged one of a smaller-scale as follows:

$$C_{T/S}^{p,G2L} = f_\theta\left(F_{T/S}^{p}, C_{T/S}^{p+1,G2L}\right). \tag{2}$$

The overall spatial distillation loss is the summation of the absolute difference between the merged features of Teacher and Student at different scales as follows:

$$L_{Spatial} = \sum_{p=1}^{3} \sum_{d\in\{L2G,G2L\}} \left| C_T^{p,d} - C_S^{p,d} \right|. \tag{3}$$

2.2. Temporal Alignment Module

Temporal consistency and redundancy are critical for video modeling and provide useful guidance for video deraining. To make Student inherit the temporal modeling capacity from Teacher, we propose a Temporal Alignment Module. It consists of both the separate alignment and the collaborative alignment as shown in Fig. 1 (c). The key component, Aggregator, traces the temporal status of a rainy video. It progressively takes the feature of t-th frame as input to update the temporal status of the rainy video, which is used to calculate the temporal distillation loss. At the same time, the internal status of the Aggregator is also updated and propagated to the aggregation process of the next frame.

In the proposed Temporal Alignment Module, there are two kinds of temporal status. First, it traces the separate temporal status of Teacher and Student as follows:

$$\left(C_{T/S,t}, h_{T/S,t}\right) = f_\gamma\left(F_{T/S,t}, h_{T/S,t-1}\right), \tag{4}$$

where $f_\gamma(\cdot)$ stands for the Aggregator. Two internal status, denoted as $h_{T,t}$ and $h_{S,t}$, are maintained for the Teacher and the Student, respectively. $C_{T/S,t}$ is the separate temporal status of the rainy video in the current time step.

Second, the collaborative temporal status is traced as follows:

$$\left(C_{M,t}, h_{M,t}\right) = f_\gamma\left(F_{S,t}, h_{T,t-1}\right), \tag{5}$$

where $C_{M,t}$ is the collaborative temporal status of the rainy video in the current time step. Different with Eqn. (4), Student provides the input feature while Teacher provides the internal status of the Aggregator to calculate the collaborative temporal status. In the temporal distillation process, we constrain $C_{M,t}$ to be close to $C_{T,t}$. This design facilitates the joint optimization by enrolling the mixed aggregation inputs, which drives the Student feature $F_{S,t}$ to imitate the behavious of the Teacher feature $F_{T,t}$ in temporal modeling.

Finally, the overall temporal distillation loss is imposed on the traced temporal status as follows:

$$L_{Temporal} = \sum_{t=1}^{L} \left(\left|C_{T,t} - C_{S,t}\right| + \left|C_{T,t} - C_{M,t}\right|\right), \tag{6}$$

where L is the frame number in a video training clip, which is set to 5 in our implementation.

2.3. Spatial-Temporal Adaptor

The above-mentioned distillation schemes can well capture both the spatial and temporal characteristics of a video. However, up to now, these two modules still work separately. The correlation between the temporal domain and the spatial domain has not been fully exploited. We propose a Spatial-Temporal Adaptor to enable the information sharing between

Authorized licensed use limited to: Peng Cheng Laboratory. Downloaded on August 29,2023 at 06:33:46 UTC from IEEE Xplore. Restrictions apply.

Table 1. Performance Improvement with our distillation framework on the NTURain dataset. $\ominus$ denotes the performance before the distillation. $\oplus$ denotes the performance after the distillation.

Clip No.		a1	a2	a3	a4	b1	b2	b3	b4	average	ΔPSNR	ΔSSIM
Teacher	PSNR	37.60	34.94	36.58	40.86	38.25	38.90	40.07	38.33	38.19	—	—
	SSIM	0.9786	0.9734	0.9734	0.9853	0.9762	0.9742	0.9824	0.9761	0.9775		
Student#1$\ominus$	PSNR	35.49	31.28	34.49	37.63	35.92	34.67	36.93	34.96	35.17	1.53 ↑	0.0047 ↑
	SSIM	0.9723	0.9611	0.9651	0.9787	0.9704	0.9589	0.9741	0.9681	0.9686		
Student#1$\oplus$	PSNR	36.81	33.33	35.63	39.40	37.19	36.18	38.72	36.33	36.70		
	SSIM	0.9757	0.9686	0.9689	0.9829	0.9739	0.9645	0.9791	0.9726	0.9733		
Student#2$\ominus$	PSNR	35.35	31.91	34.36	37.78	35.80	34.60	36.73	35.01	35.19	0.95 ↑	0.0034 ↑
	SSIM	0.9721	0.9626	0.9644	0.9789	0.9699	0.9563	0.9739	0.9677	0.9682		
Student#2$\oplus$	PSNR	36.50	31.99	35.44	38.24	36.71	36.02	38.02	36.22	36.14		
	SSIM	0.9748	0.9645	0.9683	0.9816	0.9729	0.9607	0.9783	0.9722	0.9717		
Student#3$\ominus$	PSNR	28.78	25.71	28.93	30.47	31.21	27.41	26.29	28.80	28.45	3.23 ↑	0.0073 ↑
	SSIM	0.9442	0.9201	0.9305	0.9561	0.9522	0.9511	0.9468	0.9414	0.9428		
Student#3$\oplus$	PSNR	31.18	27.07	30.51	32.46	32.76	33.86	33.08	32.54	31.68		
	SSIM	0.9520	0.9305	0.9358	0.9618	0.9556	0.9600	0.9597	0.9455	0.9501		

these two modules. First, the merged feature $C_{T/S}^{1,G2L}$ in Eqn. (2) is collected as the spatial hint, which is concatenated with the internal status of the Aggregator in the Temporal Alignment Module to obtain the temporal status of the rainy video. Second, the updated internal status of the Aggregator $h_{T/S}^{t}$ is viewed as the temporal hint to calculate the merged feature at the 1-st scale. Formally, the temporal hint plays the role of $C_{T/S}^{p-1,L2G}$ in Eqn. (1). More implementation details of the Spatial-Temporal Adaptor are provided in the supplementary material.

2.4. Overall Loss Function

The distillation losses on the intermediate feature in previous sections play the role in transferring the knowledge from the Teacher to the Student. Besides, the intermediate feature is also fed into the subsequent convolution layer to obtain the deraining result. The reconstruction loss is calculated between the deraining result and the ground truth as follows:

$$L_{Recon} = \sum_{p=1}^{3} \lambda_p SSIM(\hat{I}_t^p, I_t^p), \tag{7}$$

where $\hat{I}_t^p$ denotes the t-th deraining frame of the p-th scale and I_t^p is the corresponding ground truth. $SSIM(\cdot)$ stands for the Structural Similarity Index Measure [12] metric. λ_p is the weight to balance the importance of different terms. In our implementation, λ_1, λ_2, and λ_3 are set to 0.6, 0.2, and 0.2, respectively.

The Student is trained with the total loss as follows:

$$L_{total} = L_{Recon} + \alpha L_{Spatial} + \beta L_{Temporal}, \tag{8}$$

where α and β are the weights to balance three items, which are both set to 0.1 in our implementation.

Table 2. Configuration of the networks in the distillation experiment.

Network	Teacher	Student#1	Student#2	Student#3
#input frames	7	3	3	1
#channel	64	16	16	32
#network depth	22	22	13	22
#parameter	2.02M	136K	81K	186K
#FLOPs	127.60G	5.76G	2.81G	3.06G

3. EXPERIMENT

3.1. Implementation Details

Network Structure. We use the fully-supervised version of SLDNet [13] as our Teacher. The EHNet for detail compensation in the original SLDNet is removed and we only retain the PredNet for deraining, which consists of cascaded 3D convolution layers. It takes 7 successive rainy frames as input and the channel number of the intermediate feature is set to 64, leading to more than two million parameters. We build multiple lightweight Students by modifying the size of the sliding window, the channel number of the intermediate feature and the network depths. The detailed configurations of the Teacher and Students are shown in Tab. 2.

Dataset. We choose the NTURain dataset [4] for training and testing. The training set contains 24 synthetic rainy videos and the corresponding ground truth. The testing set contains 8 synthetic rainy videos and 7 real rainy videos. We use the training set to train our network in the training stages.

Training Details. There are two training stages, including the pre-training stage and the distillation stage. We first pre-train both the Teacher and Students without the distillation losses. In the second stage, the pre-trained Students are distilled with guidance from the pre-trained Teacher. After finishing the training, the distilled Student is evaluated solely

Authorized licensed use limited to: Peng Cheng Laboratory. Downloaded on August 29,2023 at 06:33:46 UTC from IEEE Xplore. Restrictions apply.

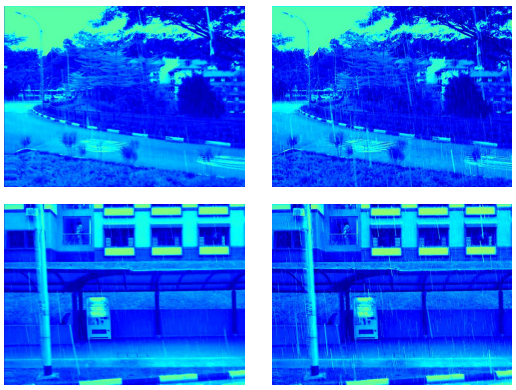


Fig. 2. Visualization of the intermediate feature. It can be observed that the intermediate feature after the distillation owns richer information to facilitate the subsequent reconstruction of the clean image.

without the enrollment of the Teacher. More training details are provided in the supplementary materials.

3.2. Quantitative Evaluation

Tab. 1 shows the results of our distillation framework. It can be observed that significant improvement can be achieved for Students with different scales, which proves the generality of our framework. Specifically, an improvement of more than 2dB is obtained for Student#1 in the a2 video.

Our Student network is not only lightweight to be more practical for real applications but also achieves satisfactory deraining performance after the distillation. To prove it, we compare the distilled Student#1 with some existing deraining methods in Tab. 3. We also provided the visual comparison in Fig. 3. PReNet [14] is a single image deraining method. For a fair comparison, we re-train it with the NTURain dataset under the same training setting. It mistakenly removes the background contents that are not rain streaks as Fig. 3 (b) shows. SpacCNN [4] is a video deraining method, equipped with a super-pixel segmentation-based alignment scheme according to the background context. On the contrary, S2VD [5] explicitly models the rain layers of different frames with a statistical model. Better performance is obtained with these two methods qualitatively. While there is also a significant increase in the parameter number and some visual artifacts are also introduced in their results as Fig. 3 (c-d) shows. Compared with these methods, our distilled Student network achieves the best visual results with the smallest number of parameters.

3.3. Feature Map Visualization

We also visualize the feature map of the Student in Fig. 2. It can be observed that before distillation, the rain patterns are of blurred shape, which means that the rain can not be accurately distinguished from the background. On the contrary, the rain patterns can be well distinguished in the feature map of the distilled network.

Table 3. Comparisons with Existing De-raining Methods in PSNR and SSIM. The best and second best results are highlighted in red and blue, respectively.

Clip No.	Network	PReNet	SpacCNN	S2VD	Student#1
	#parameter	169K	477K	525K	136K
1	PSNR	31.40	30.57	36.39	36.81
	SSIM	0.9505	0.9334	0.9658	0.9757
2	PSNR	28.58	31.29	33.06	33.33
	SSIM	0.9291	0.9356	0.9519	0.9686
3	PSNR	30.57	30.63	35.75	35.63
	SSIM	0.9375	0.9247	0.9564	0.9689
4	PSNR	33.62	35.30	39.53	39.40
	SSIM	0.9658	0.9620	0.9779	0.9829
average	PSNR	31.04	31.95	36.18	36.29
	SSIM	0.9457	0.9389	0.9630	0.9740

3.4. Comparison with Existing Distillation methods

We compare our distillation framework with two distillation methods for low-level vision. SRKD [9] is a distillation method to distill image super-resolution networks. This method only aligns the spatial feature of each image and the temporal correlation is not taken into consideration. STD [11] proposes to use both the spatial and temporal guidance of Teacher to distill video super-resolution networks, while the guidance of the spatial domain and temporal domain is independent. We re-implement these two methods and use them to distill the Student#1. Tab. 4 shows the comparison result. It can be observed that our distillation method brings in the most performance improvement with our well-designed modules.

Table 4. De-raining Results on the NTURain Testing Set with Different Distillation Methods.

Distillation Method	PSNR	SSIM
Baseline	35.17	0.9686
SRKD [9]	35.58	0.9709
STD [11]	35.74	0.9710
Ours	**36.70**	**0.9733**

3.5. Ablation Studies

We study the impact of proposed modules on performance to verify their effectiveness. All ablation studies are performed on Student#1 on the testing set of NTURain. Tab. 5 shows the ablation results. The corresponding proposed module of the abbreviation in the table is shown below:

- MS: Multi-Scale Spatial Distillation.
- CSC: Cross-Scale Connection.
- ITA: Independent Temporal Alignment.
- CTA: Collaborative Temporal Alignment.
- STA: Spatial-Temporal Adaptor.

Authorized licensed use limited to: Peng Cheng Laboratory. Downloaded on August 29,2023 at 06:33:46 UTC from IEEE Xplore. Restrictions apply.

(a) Input (b) PReNet (c) SpacCNN (d) S2VD (e) Ours

Fig. 3. Visual comparison among different deraining methods. Compared with our distilled network, the deraining results of other methods contain more remaining rain streaks as shown in the first line and obvious artifacts as shown in the second line.

Table 5. Ablation Results of the Proposed Modules.

Setting No.	MS	CSC	ITA	CTA	STA	PSNR	SSIM
1	✓					35.85	0.9706
2	✓	✓				36.23	0.9724
3			✓			35.63	0.9710
4				✓		35.80	0.9707
5			✓	✓		35.63	0.9712
6	✓	✓	✓	✓		36.21	0.9720
7	✓	✓	✓	✓	✓	**36.70**	**0.9733**

Setting 1-2 and 3-5 analyze the Spatial Alignment Module and the Temporal Alignment Module, respectively. In Setting 6, these two modules are combined while the Spatial-Temporal Adaptor is removed. A performance drop can be observed due to the lack of cross-domain information. The final distillation framework, which is equipped with all proposed modules, achieves the best performance as Setting 7 shows.

4. CONCLUSION

In this paper, we propose a Collaborative Spatial-Temporal Distillation framework to construct lightweight and efficient video deraining networks. Both the spatial and temporal knowledge are efficiently transferred with the proposed alignment modules. Multiple collaboration schemes of different granularities lead to better distillation results. Experimental results not only prove the performance improvement with our proposed distillation framework but also show the efficiency of the distilled networks compared with existing video deraining methods.

5. REFERENCES

[1] Kshitiz Garg and Shree K Nayar, "Detection and removal of rain from videos," in *CVPR*, 2004.

[2] Tai-Xiang Jiang, Ting-Zhu Huang, Xi-Le Zhao, Liang-Jian Deng, and Yao Wang, "A novel tensor-based video rain streaks removal approach via utilizing discriminatively intrinsic priors," in *CVPR*, 2017.

[3] Yi-Lei Chen and Chiou-Ting Hsu, "A generalized low-rank appearance model for spatio-temporally correlated rain streaks," in *ICCV*, 2013.

[4] Jie Chen, Cheen-Hau Tan, Junhui Hou, Lap-Pui Chau, and He Li, "Robust video content alignment and compensation for rain removal in a CNN framework," in *CVPR*, 2018.

[5] Zongsheng Yue, Jianwen Xie, Qian Zhao, and Deyu Meng, "Semi-supervised video deraining with dynamical rain generator," in *CVPR*, 2021.

[6] Jiaying Liu, Wenhan Yang, Shuai Yang, and Zongming Guo, "D3R-Net: Dynamic routing residue recurrent network for video rain removal," *IEEE Trans. on Image Processing*, vol. 28, no. 2, pp. 699–712, 2018.

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[8] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *CVPR*, 2017.

[9] Qinquan Gao, Yan Zhao, Gen Li, and Tong Tong, "Image super-resolution using knowledge distillation," in *ACCV*, 2018.

[10] Zibin He, Tao Dai, Jian Lu, Yong Jiang, and Shu-Tao Xia, "FAKD: Feature-affinity based knowledge distillation for efficient image super-resolution," in *ICIP*, 2020.

[11] Zeyu Xiao, Xueyang Fu, Jie Huang, Zhen Cheng, and Zhiwei Xiong, "Space-time distillation for video super-resolution," in *CVPR*, 2021.

[12] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.

[13] Wenhan Yang, Robby T. Tan, Shiqi Wang, and Jiaying Liu, "Self-learning video rain streak removal: When cyclic consistency meets temporal correspondence," in *CVPR*, 2020.

[14] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng, "Progressive image deraining networks: A better and simpler baseline," in *CVPR*, 2019.

Authorized licensed use limited to: Peng Cheng Laboratory. Downloaded on August 29,2023 at 06:33:46 UTC from IEEE Xplore. Restrictions apply.