Dual-Rain: Video Rain Removal using Assertive and Gentle Teachers

Tingting Chen[1*], Beibei Lin[1*], Yeying Jin[1], Wending Yan[2], Wei Ye[2], Yuan Yuan[2], and Robby T. Tan[1]

[1] National University of Singapore
[2] Huawei International Pte Ltd
{tingting.c,beibei.lin,e0178303}@u.nus.edu,
{yan.wending,yewei10,yuanyuan10}@huawei.com, {robby.tan}@nus.edu.sg

1 Synthetic Datasets Generation

Since our teachers must have an initial deraining ability to generate predictions with certain confidence levels, our two-teacher needs to have a pre-trained video deraining model w_{init}, which is trained on synthetics datasets $\mathbf{D}_{\text{syn}} = \{\mathbf{x}_i^{\text{syn}}, \mathbf{y}_i^{\text{syn}}\}_{i=1}^{N_{\text{syn}}}$, where $\mathbf{x}_i^{\text{syn}}$ and $\mathbf{y}_i^{\text{syn}}$ are the i-th input and ground-truth videos, respectively. N_{syn} is the number of supervised videos. These pre-trained parameters are then used to initialize the two teachers and one student.

1.1 Depth Maps Generation

The general model to generate hazy frames can be formulated as [2]:

$$F(x)_t = \alpha(x)_t B(x)_t + (1 - \alpha(x)_t) A_t, \tag{1}$$

where $F(x)$ is the frame with haze and $B(x)$ is the rain-free frame. $\alpha(x)$ is the atmosphere transmission. A is the atmospheric light. $t \in \{1, 2, ...N\}$ and t represents the time-step. Here, $\alpha = e^{-\beta d(t)}$, where β represents the depth map and and $d(t)$ represents the atmosphere scattering parameter. As the depth map is required to generate the hazy frame, method of [6] is used to estimate the depth map. The way of generating frames with rain accumulation is similar to the way of generating hazy frames. Therefore, depth maps are also required to generate frames with both rain streaks and rain accumulation.

1.2 Rain Frames Generation

We use the model from [3] to generate both rain streaks and rain accumulation. The model is formulated as:

$$F(x)_t = \alpha(x)_t (B(x)_t + S(x)_t) + (1 - \alpha(x)_t) A_t, \tag{2}$$

* Equal Contribution

where $F(x)$ is the frame with both rain streaks and veiling effect. $B(x)$ is the rain-free frame and $S(x)$ is the rain-streak frame. $\alpha(x)$ is the atmosphere transmission. A is the atmospheric light. $t \in \{1, 2, ...N\}$ and t represents the time-step as before. A_t usually have minor difference in one clip, and thus can be assumed to be consistent for all frames. To be more specific, β is set to be 2.5 and A is set to be $(1, 1, 1)$.

2 More Experiments Results

In this supplementary material, we compare our method with state-of-the-art methods on more synthetic videos and real-world videos, including Video Swin transformer [4], BIPNet [1], SAVD [7], ESTIL [8], as well as V-DiT. Note that V-DiT is a 3D version implemented by us, based on DiT [5]. The visual results on synthetic data are shown in Fig. 1 and Fig. 2. For each method, three frames are shown. Each column in the figure represents one frame. The visual results on real-world data are shown in Fig. 3 and Fig. 4. For each method, three frames are shown. Each column in the figure represents one frame. In addition, comparisons between the rain-removed videos generated by our method and those produced by other methods are also provided in the supplementary material.

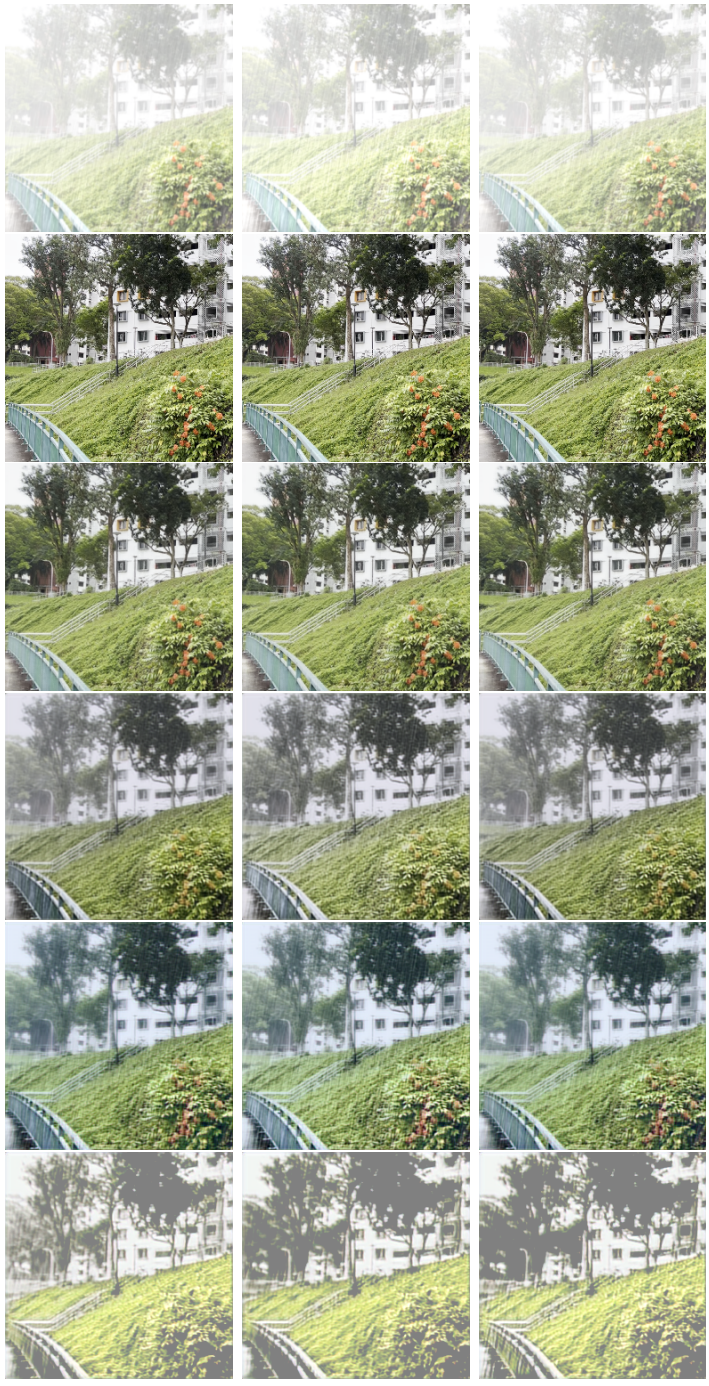

Fig. 1: Qualitative comparison on Synthetic data. Each column represents one frame. Zoom in for better visualisation.

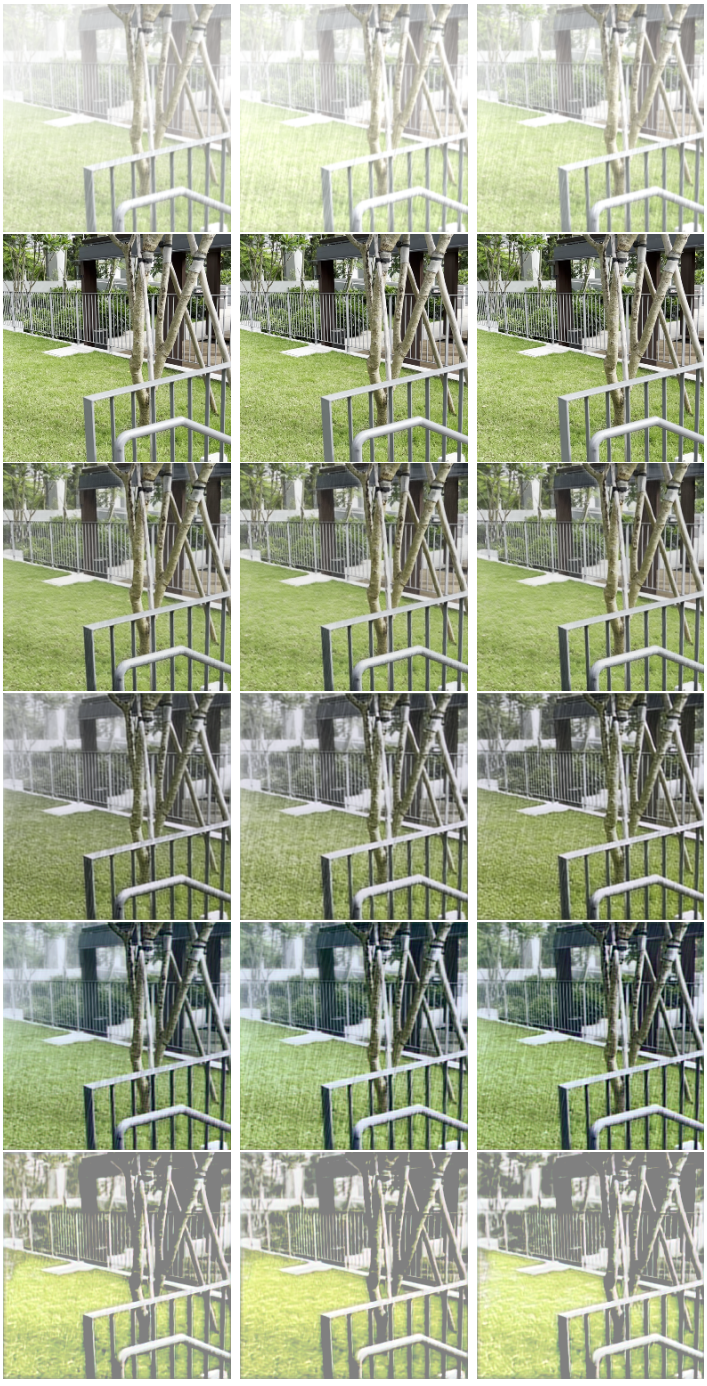

Fig. 2: Qualitative comparison on Synthetic data. Each column represents one frame. Zoom in for better visualisation.

Fig. 3: Qualitative comparison on real-world data. Red boxes show the region where other methods fail to remove rain streaks. Each row represents three consecutive frames and each column represents the same frame. Zoom in for better visualisation.

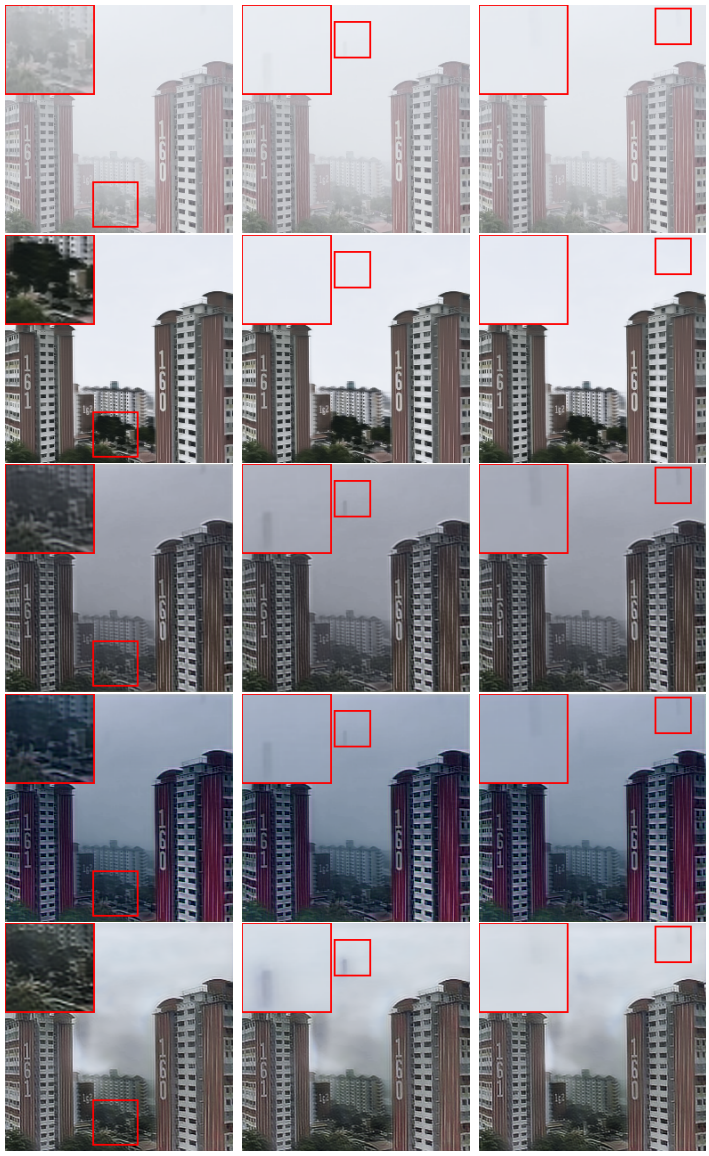

Fig. 4: Qualitative comparison on real-world data. Red boxes show that other methods fail to remove rain streaks or rain accumulation. Each row represents three consecutive frames and each column represents the same frame. Zoom in for better visualisation.

References

1. Dudhane, A., Zamir, S.W., Khan, S., Khan, F.S., Yang, M.H.: Burst image restoration and enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5759–5768 (2022) 2, 3, 4, 5, 6
2. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE transactions on pattern analysis and machine intelligence **33**(12), 2341–2353 (2010) 1
3. Li, R., Cheong, L.F., Tan, R.T.: Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1633–1642 (2019) 1
4. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022) 2, 3, 4, 5, 6
5. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023) 2
6. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020) 1
7. Yan, W., Tan, R.T., Yang, W., Dai, D.: Self-aligned video deraining with transmission-depth consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11966–11976 (2021) 2, 3, 4
8. Zhang, K., Li, D., Luo, W., Ren, W., Liu, W.: Enhanced spatio-temporal interaction learning for video deraining: faster and better. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 1287–1293 (2022) 2, 5, 6