RVDNET: A TWO-STAGE NETWORK FOR REAL-WORLD VIDEO DESNOWING WITH DOMAIN ADAPTATION

Tianhao Xue, Gang Zhou, Runlin He, Zhong Wang, Juan Chen, Zhenhong Jia*
Key Laboratory of Signal Detection and Processing, Xinjiang University, Urumqi, China

ABSTRACT

Video snow removal is an important task in computer vision, as the snowflakes in videos reduce visibility and negatively affect the performance of outdoor visual systems. However, due to the complexity of real snowy scenarios, it is difficult to apply existing supervised learning-based methods to process real-world snowy videos. In this paper, we propose a novel two-stage video desnow network for the real world, called **RVDNet**. The first stage of RVDNet utilizes Spatial Feature Extraction Modules (SFEM) to extract the spatial features of the input frames. In the second stage, we design Spatial-Temporal Desnowing Modules (STDM) to remove snowflakes via spatio-temporal learning. Furthermore, we introduce the unsupervised domain adaptation module, which is embedded for aligning the feature space of real and synthetic data in the spatial and spatio-temporal domains, respectively. Experiments on the proposed SnowScape dataset prove that our method has superior desnow performance not only on synthetic data, but also in the real world.

Proposed Method

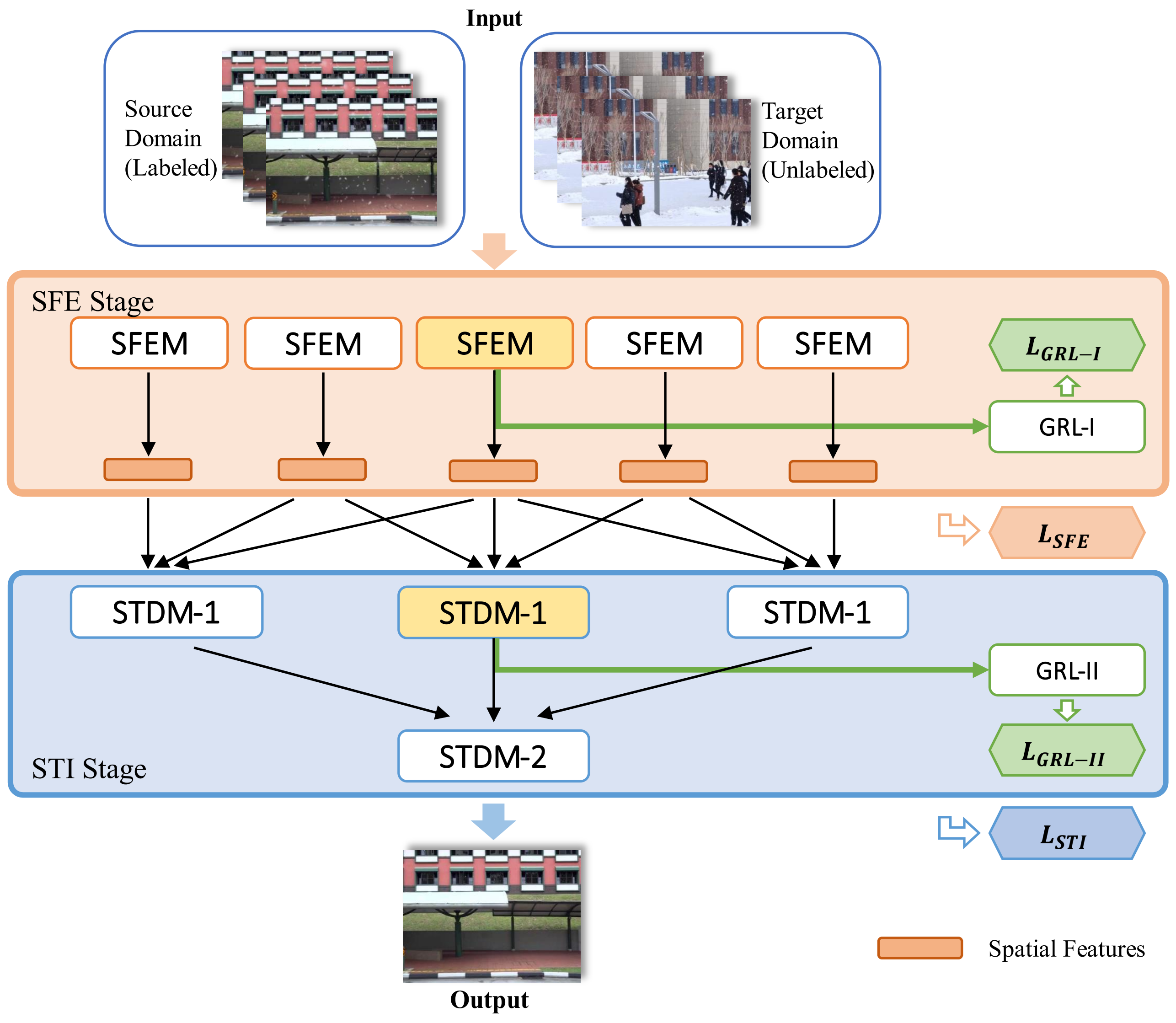

Fig. 1. Overall architecture of RVDNet

In order to remove snowflakes from real-world videos, we propose a novel **Two-stage Network for Real-world Video Desnowing with Domain Adaptation (RVDNet).**

- The first stage of RVDNet is the **Spatial Feature Extraction (SFE)** Stage, which inputs the adjacent frames into the network in parallel to obtain the spatial domain features of each frame.
- The second stage is the **Spatio-Temporal Interaction (STI)** Stage, which concatenate the results of the first stage as inputs and processes them through two layers to obtain a snow free video.
- Furthermore, RVDNet embeds **Domain Adaptation (DA) modules** into the two stages to achieve the goal of multi-stage feature alignment.

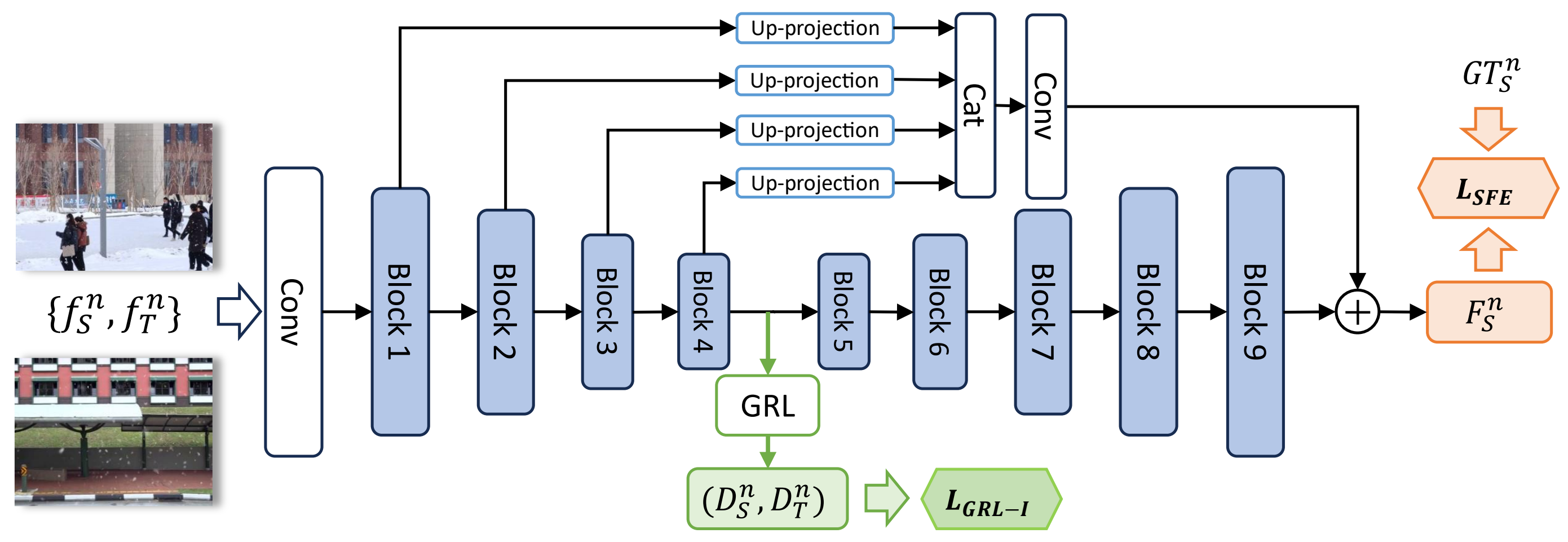

Fig. 2. SFEM with DA module

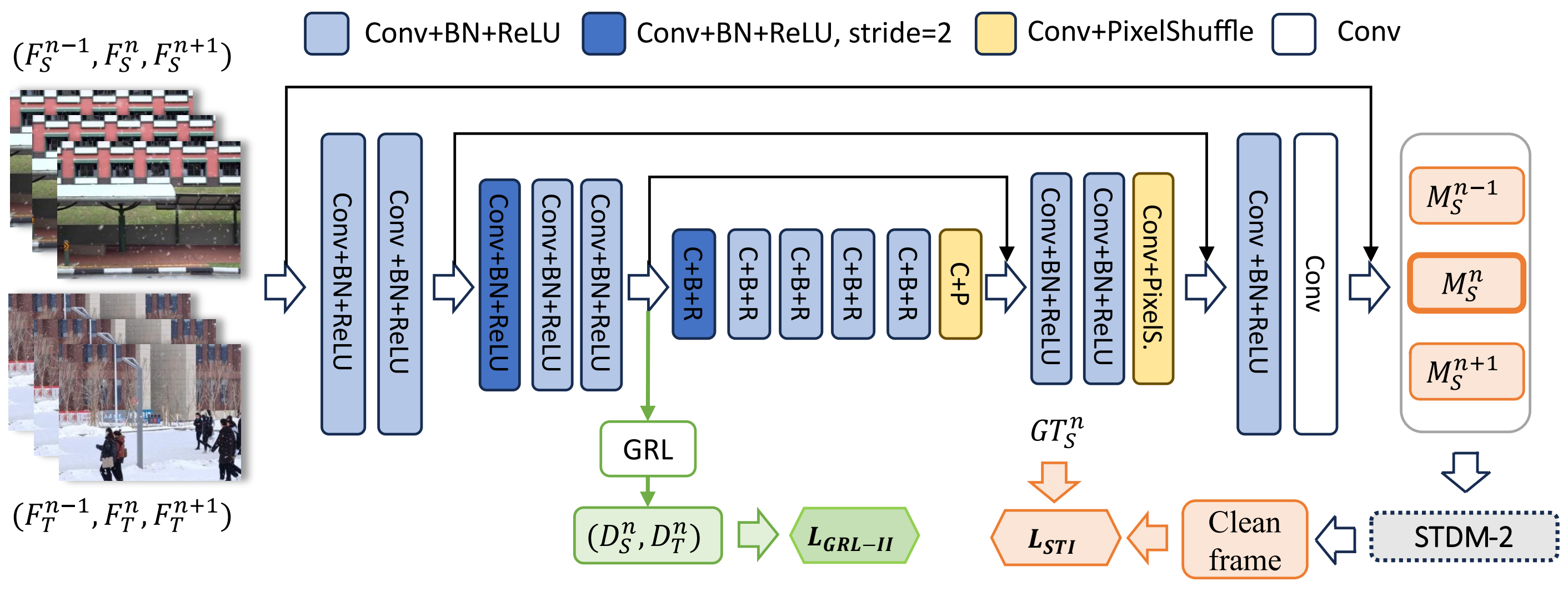

Fig. 3. STDM with DA module

Experiments

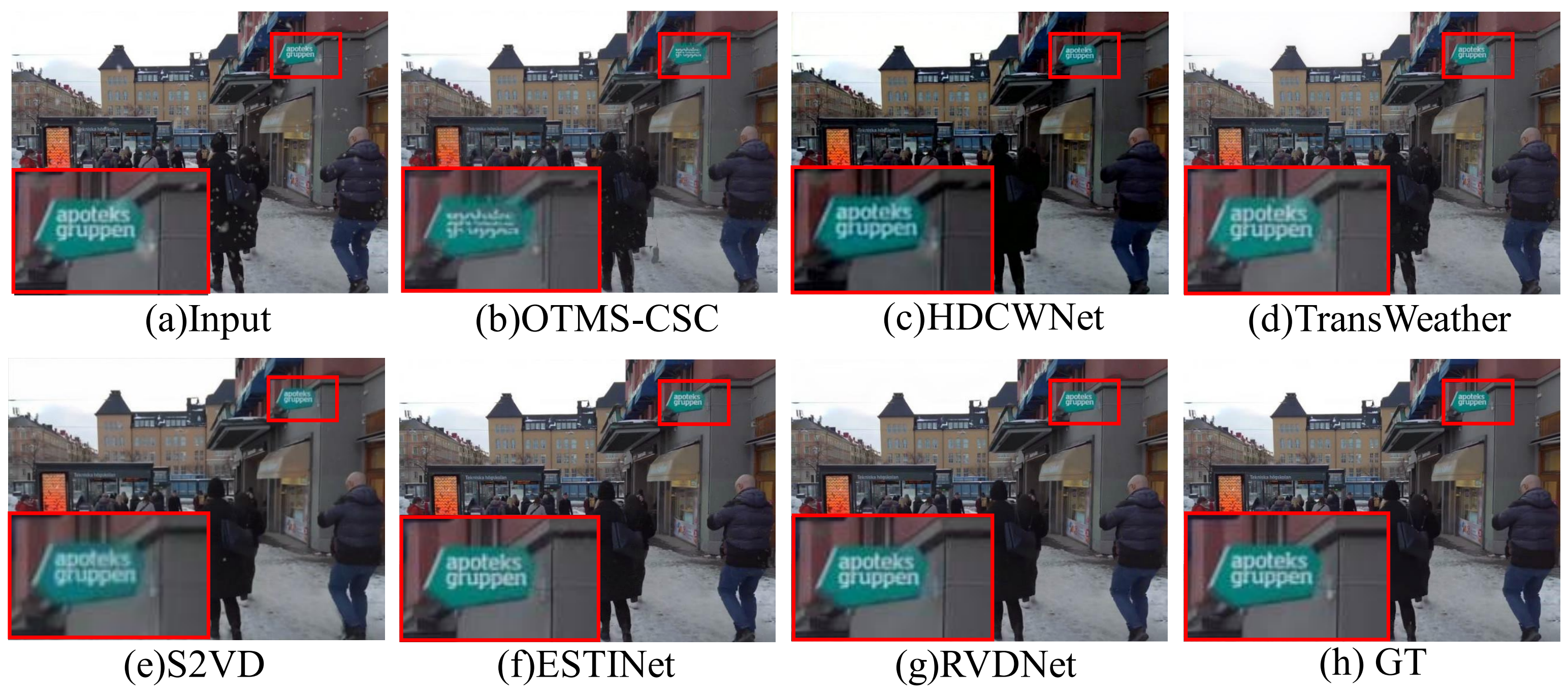
(a)Input (b)OTMS-CSC (c)HDCWNet (d)TransWeather
(e)S2VD (f)ESTINet (g)RVDNet (h) GT

Fig. 4. Visual comparison of different methods on synthetic snowy sequences

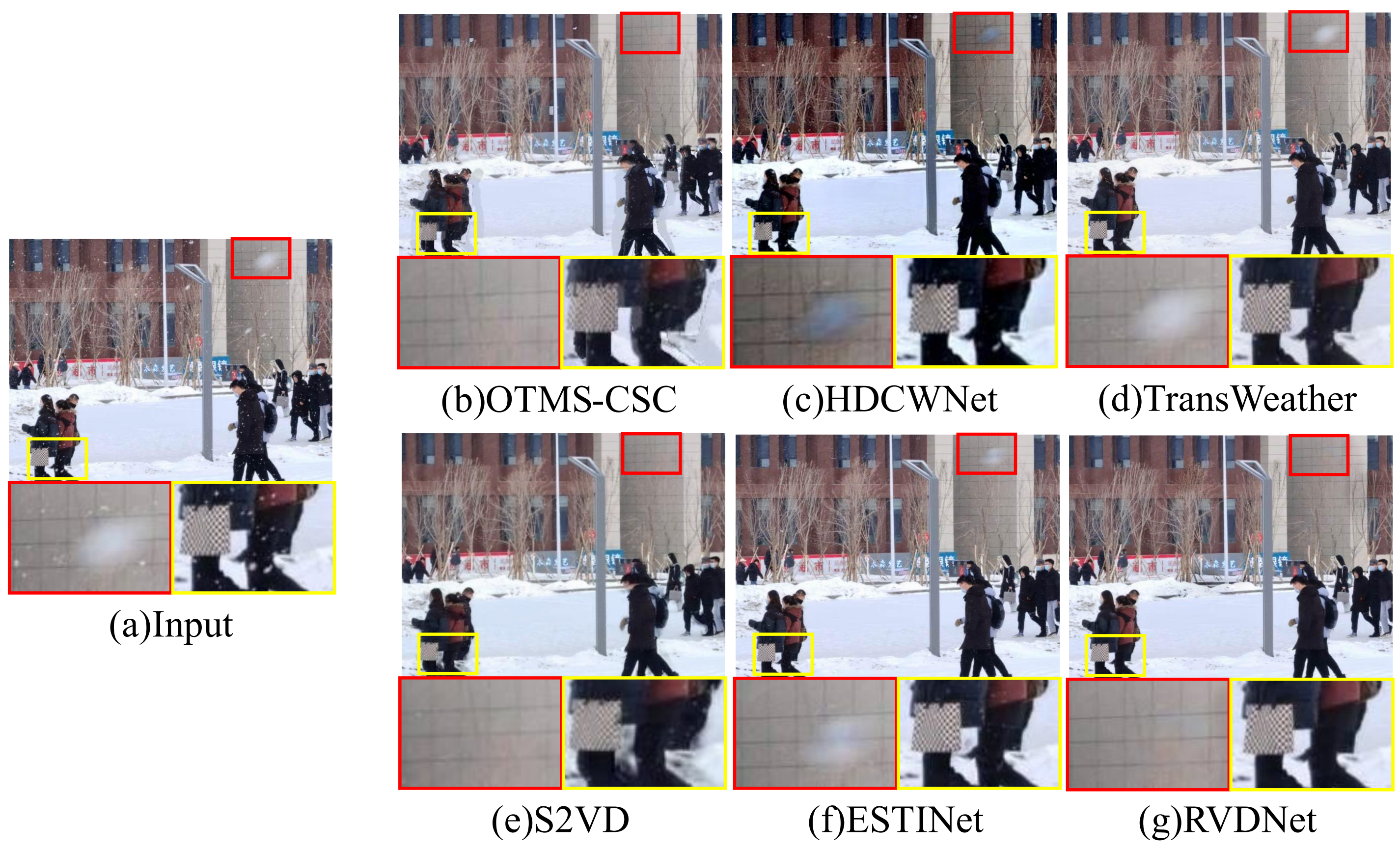
(a)Input (b)OTMS-CSC (c)HDCWNet (d)TransWeather
(e)S2VD (f)ESTINet (g)RVDNet

Fig. 5. Visual comparison of different methods on real-world snowy sequences

Table 1. Performance comparison with state-of-the-art methods.

Methods	Synthetic Dataset		Real Dataset		
	PSNR↑	SSIM↑	NIQE↓	BRISQUE↓	PIQE↓
OTMS-CSC	29.75	0.9122	2.6380	28.084	39.069
HDCWNet	21.59	0.8493	2.7235	33.189	36.721
TransWeather	34.72	0.9638	2.7318	30.044	39.701
S2VD	35.56	0.9644	2.9460	20.470	42.696
ESTINet	34.97	0.9547	2.3391	18.756	37.794
RVDNet	**37.89**	**0.9773**	**2.2143**	**18.546**	**35.676**

Table 2. Ablation study of different architectures in our work.

Methods	Synthetic Dataset		Real Dataset		
	PSNR↑	SSIM↑	NIQE↓	BRISQUE↓	PIQE↓
SFE	26.92	0.8462	2.9985	31.719	40.884
STI	35.77	0.9477	3.4886	31.703	37.363
SFE + STI	**38.01**	**0.9805**	2.7693	28.641	38.626
SFE + STI + GRL-I	37.36	0.9745	2.2538	19.785	36.917
SFE + STI + GRL-II	37.39	0.9734	2.2639	21.063	37.807
RVDNet	37.89	0.9773	**2.2143**	**18.546**	**35.676**

Conclusion

We propose **RVDNet** as a solution to real world video snow removal. Through the two-stages structure, RVDNet makes full use of the information in both spatial domain and spatio-temporal domain, so that snowflakes are effectively removed from the video. Furthermore, the unsupervised domain adaptation module embedded in RVDNet tackles the problem of domain shift between the synthetic and real data, and improves the desnowing performance in real scenes. Experiments on the proposed SnowScape dataset show that RVDNet has superior desnow performance not only on synthetic data, but also in the real world. In the future, we will continue our efforts to improve the efficiency of video desnowing and seek more appropriate ways to evaluate video desnowing results, such as utilizing high-level vision tasks.

Acknowledgment This work was supported in part by the National Natural Science Foundation of China under Grant 62166040, Grant 62137002 and Grant 62261053, and in part by the Natural Science Foundation of XinJiang under Grant 2021D01C057 and Grant 2022D01C.