SnowMaster: Comprehensive Real-world Image Desnowing via MLLM with Multi-Model Feedback Optimization

Supplementary Material

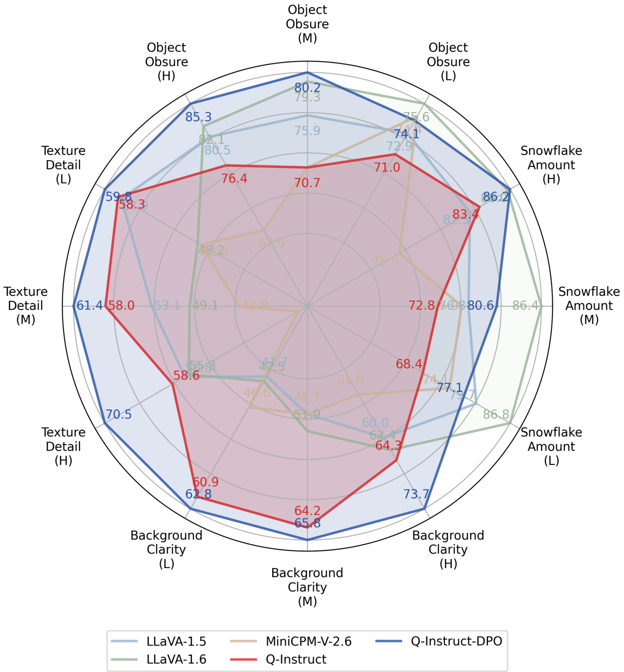


Figure 1. Detailed comparison of the accuracy of different models on various indicators

1. Contents

2. Further Experimental Details

In this section, we provide a more comprehensive overview of the experimental setup and procedures.

During the Direct Preference Optimization (DPO) fine-tuning stage of Q-instruct, we utilized the vision-LLM-Alignment framework. This adaptable framework supports the supervised fine-tuning (SFT) of various MLLMs, the training of reward models, and the implementation of both Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO) training protocols. Notably, among the compatible models is the LLava-1.5-hf model.

To ensure compatibility with the vision-LLM-Alignment framework, we employed the python-code-anls script to convert the original Q-instruct weights from the conventional LLava-1.5 format to the LLava-1.5-hf weight format. This conversion is essential as it enables the seamless integration and utilization of the transformed weights within the vision-LLM-Alignment framework, ensuring effective DPO alignment.

For the calculation of indicator scores, the Swift framework (Scalable LightWeight Infrastructure for Fine-Tuning) was employed, leveraging vllm inference acceleration to expedite result computation.

3. Supplementary Comparative Analysis

This section details additional comparative analyses, including comprehensive evaluations of accuracy metrics and enhanced visual comparisons. Experimental findings indicate that our proposed method outperforms existing techniques in both real snowfall image assessment and image desnowing tasks.

3.1. Expanded Quantitative Comparisons

Building on previous accuracy comparisons, we further delineated scene categories to derive more granulated quantitative results, as illustrated in Fig.1.

Utilizing snowfall severity scores assigned during dataset curation, we categorized the 1,047 test images into three levels: light snow, moderate snow, and heavy snow, comprising 266, 324, and 457 images with scores of 3-4.2, 4.2-4.8, and 4.8-5, respectively.

In Fig.1, it is evident that, across all 12 indicators, the Q-Instruct model following DPO fine-tuning leads in 9 indicators and consistently outperforms the pre-DPO Q-Instruct model. Furthermore, for all four snowfall image evaluation metrics, it is noted that while MLLMs generally exhibit lower accuracy for light snowfall images compared to heavy snowfall ones, they achieve higher accuracy rates for heavy snowfall scenarios.

In the context of heavy snowfall, where substantial snow removal is required, the DPO fine-tuned Q-instruct achieves an accuracy exceeding 85% for indicators such as ”Snowflake Amount in the Air” and ”Obscuring Objects with Snowflakes.” In addition, it shows an accuracy above 70% for ”Texture Detail” and ”Background Clarity,” providing a robust pseudo-label update signal for subsequent semi-supervised training.

3.2. Expanded Qualitative Comparisons

For additional qualitative comparisons, please refer to Fig. 2,3,4,5,6. The experimental results in various scenarios show that the SnowMaster method we proposed can better remove snow in real scenarios, especially some smaller snowflake particles.

4. Dataset Selection and Segmentation

This section provides an in-depth description of the dataset composition we have proposed, as well as the data employed for training and testing at each stage.

Stage	Snowfall imgs	Desnowed Pairs	Snow Accumulation	Total
Train	6406	1500(*2)	-	7906
Test	1047(*2)	-	-	1047
Extra	-	-	3723	3723

Table 1. RealSnow10k dataset composition.

Stage	Snow100k	Realsnow10k	Total
DPO-Train	-	1500(*2)	1500
DPO-Test	-	1047(*2)	1047
EMA-Train	50000(*2)	6406	56046
EMA-Test	1329	457	1786

Table 2. Dataset usage at different training stages.

Table.1 presents the detailed composition of the RealSnow10k dataset, where (*2) indicates the use of a pre-trained NAFNet model to construct paired data for image desnowing. It is noteworthy that the Desnowed Pairs section encompasses 1,500 images, consisting of 1,000 snowfall images and 500 snow accumulation images. This configuration is necessary for the model to learn image representation without snowfall during the DPO stage. The Extra section comprises 3,723 real snow images, serving as a foundation for future research endeavors, despite being excluded from the training and testing phases within this study. In total, we offer a comprehensive set of 12,676 real snow scene images, of which our proposed method utilizes 8953.

Table.2 provides a detailed account of dataset usage across the four stages. During the DPO training phase, we utilized 1,500 image pairs pre- and post-snow removal, resulting in the generation of 18,000 preference pairs aligned with two distinct indicators (36,000 pairs in total). In the testing phase of DPO, we processed 1,047 images for snow removal and acquired 2,094 images for accuracy evaluation. In the semi-supervised training framework, we utilized the complete set of 50,000 image pairs from Snow100K for supervised training, while incorporating 6,406 real-world snow images from RealSnow10K to participate in the semi-supervised training process. In the final evaluation phase, we selected all 1,329 images in Snow100k-realistic and 457 heavy snowfall images extracted from the 1,047 test images for quantitative metric assessment.

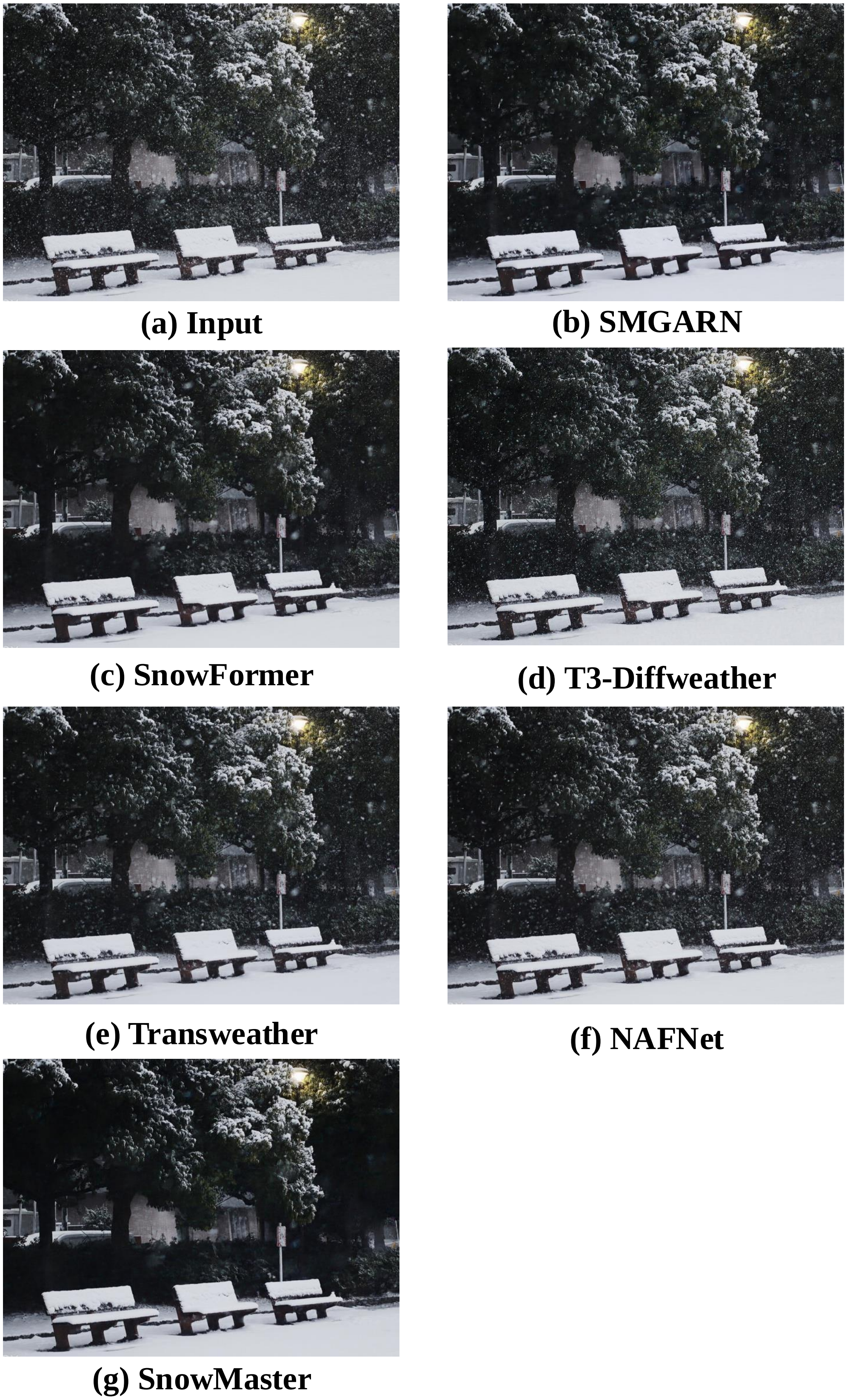
5. Future Work

In this section, we outline and discuss potential enhancements for the current methodology.

Integration of Advanced Multimodal Models. During the experimental and manuscript preparation phases, several novel multimodal large language models, such as LLama-3.2-vision, Qwen2-VL, InternVL2, etc., were introduced and demonstrated exceptional performance across various MLLM benchmarks. Incorporating feedback from these more sophisticated models in snowfall image evaluation to construct a preference dataset might enhance the efficacy of direct preference optimization.

Expansion and Diversity of Real-world Datasets. Each phase of our training protocol necessitates an extensive collection of real snowfall images, yet our dataset predominantly comprises images of light to moderate snowfall, with a notable scarcity in heavy snowfall instances. We hypothesize that augmenting the dataset with a larger volume of heavy snowfall images could more accurately capture the intricate morphology of snowflakes in real-world conditions, thereby further enhancing the model's generalization capabilities.

Enhanced Snow Removal Techniques. Presently, previous efforts in image snow removal, along with our proposed SnowMaster method, primarily concentrate on eliminating snowflakes, yielding significant improvements in visibility restoration and texture enhancement. However, analysis of real snowfall images reveals that scenes with abundant snowflakes frequently include accompanying phenomena such as snow fog and snow accumulation, which also affect image quality. We believe that our proposed approach of using a multimodal large language model to assess snowfall can also be easily extended to assess snow fog and snow accumulation. By carefully designing snow removal frameworks and models, the impact of snow fog and snow accumulation may be mitigated to some extent, which is a promising avenue for future exploration.


(a) Input

(b) SMGARN

(c) SnowFormer

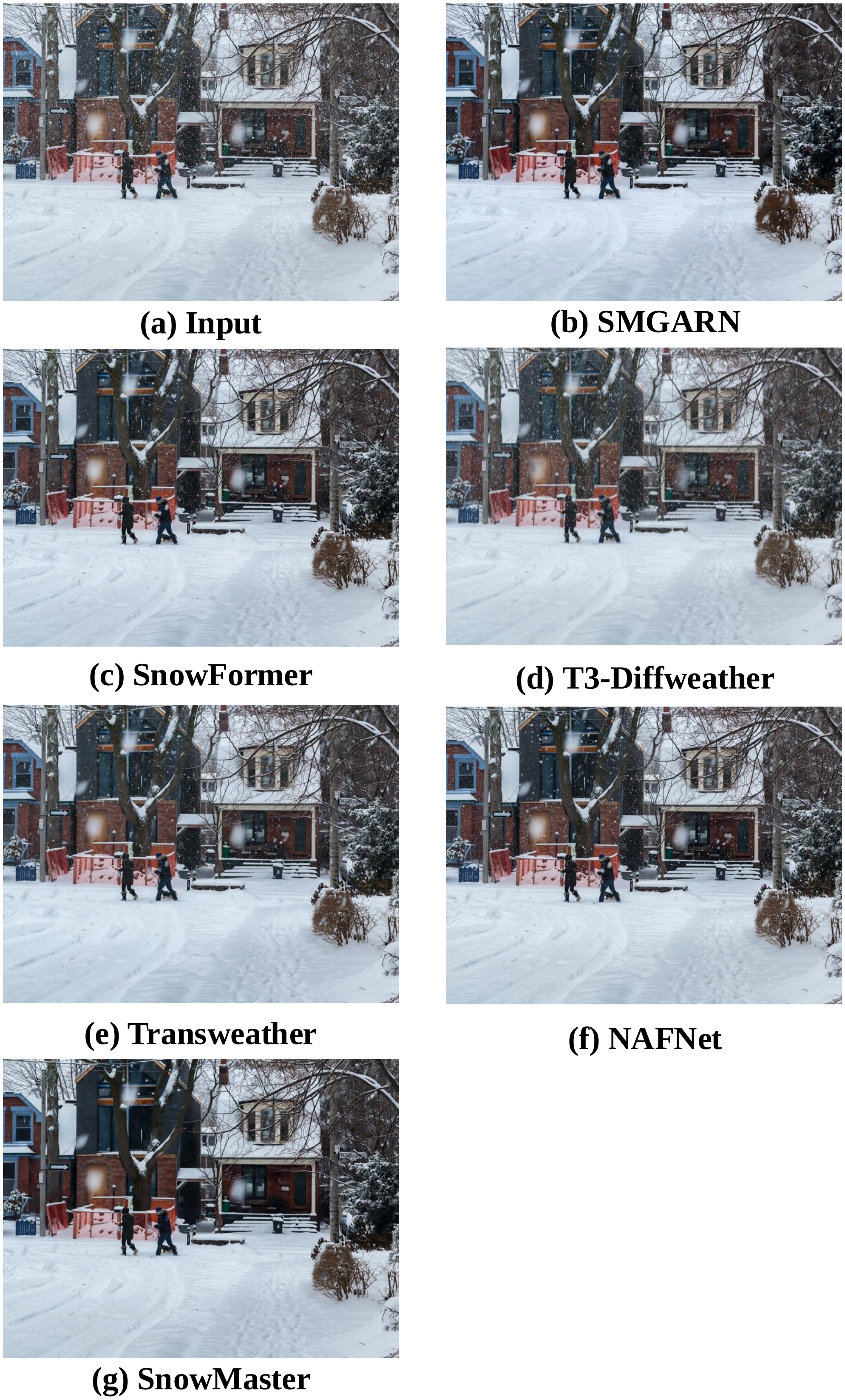
(d) T3-Diffweather

(e) Transweather

(f) NAFNet

(g) SnowMaster

Figure 2. Expanded Visual comparison of different models for snow removal on real images.



(a) Input

(b) SMGARN

(c) SnowFormer

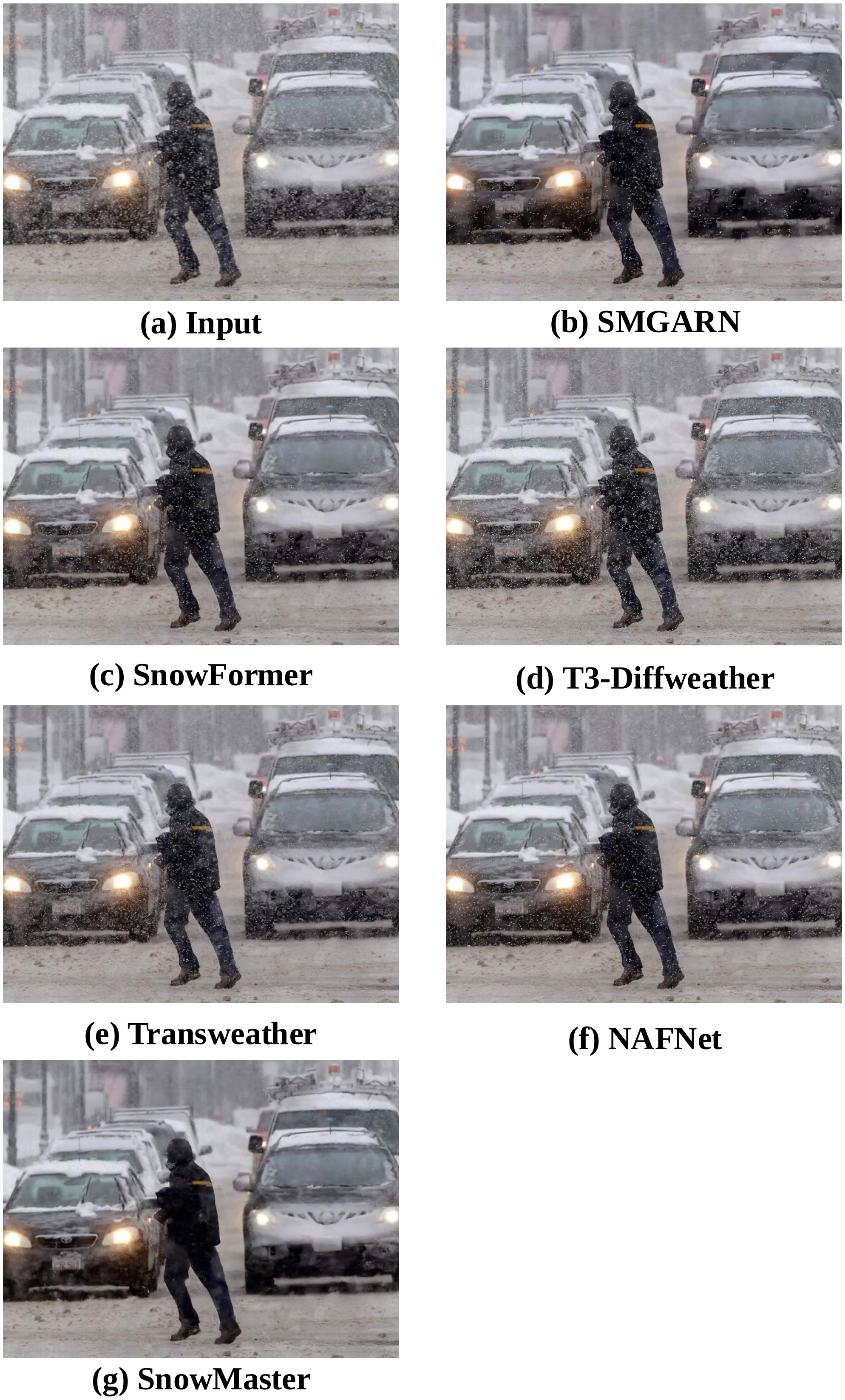
(d) T3-Diffweather

(e) Transweather

(f) NAFNet

(g) SnowMaster

Figure 3. Expanded Visual comparison of different models for snow removal on real images.



(a) Input

(b) SMGARN

(c) SnowFormer

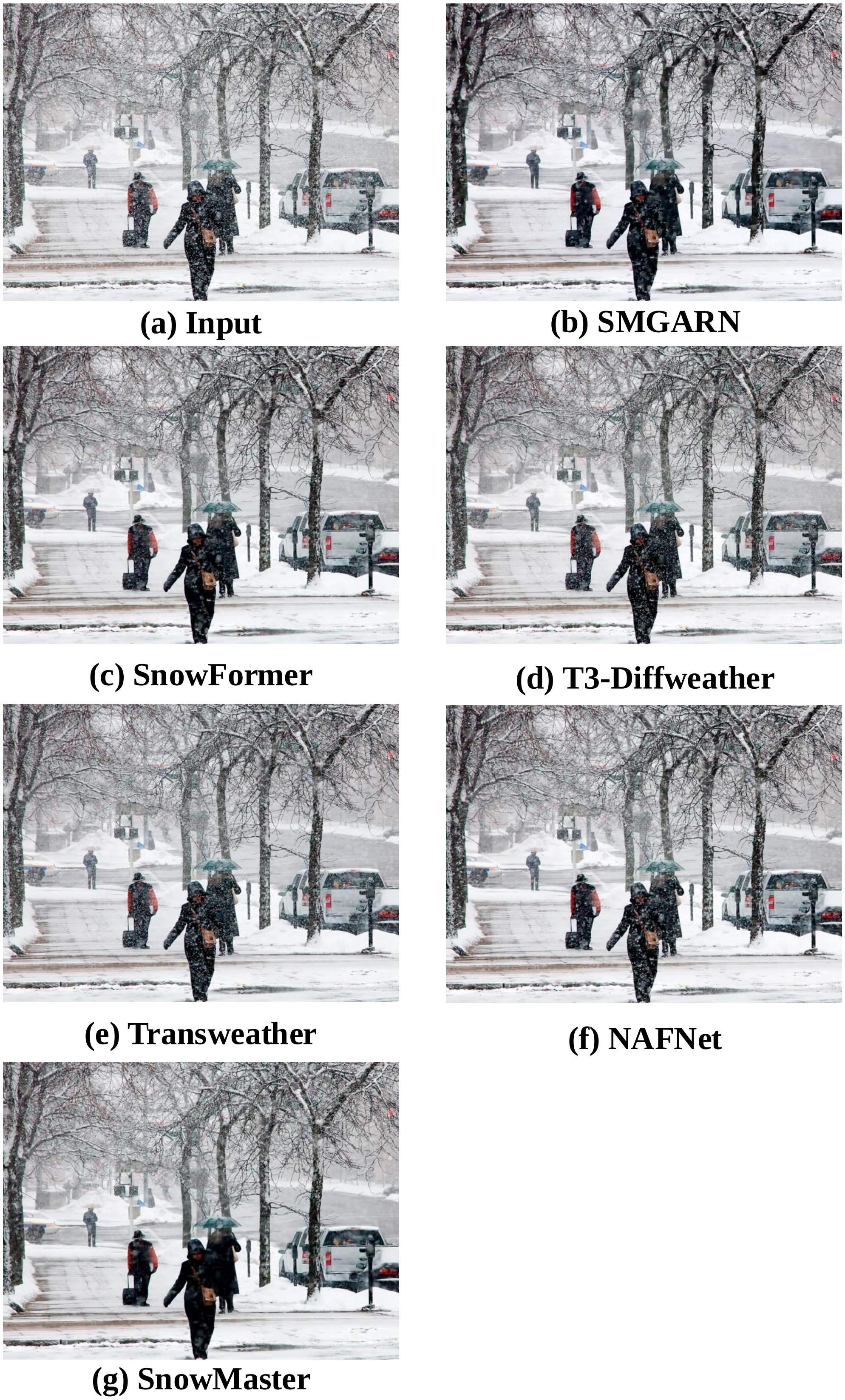
(d) T3-Diffweather

(e) Transweather

(f) NAFNet

(g) SnowMaster

Figure 4. Expanded Visual comparison of different models for snow removal on real images.



(a) Input

(b) SMGARN

(c) SnowFormer

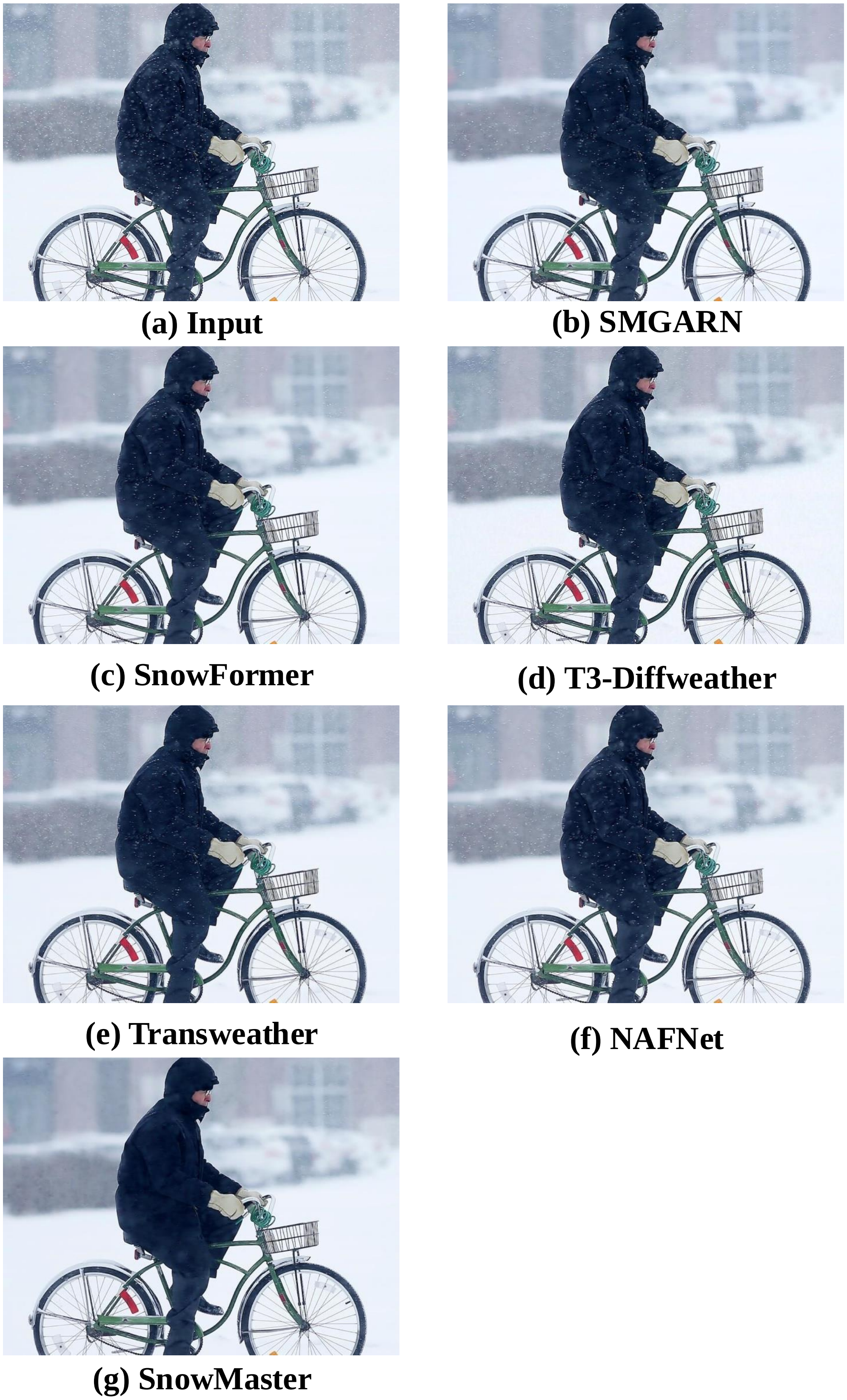
(d) T3-Diffweather

(e) Transweather

(f) NAFNet

(g) SnowMaster

Figure 5. Expanded Visual comparison of different models for snow removal on real images.



(a) Input

(b) SMGARN

(c) SnowFormer

(d) T3-Diffweather

(e) Transweather

(f) NAFNet

(g) SnowMaster

Figure 6. Expanded Visual comparison of different models for snow removal on real images.