High-Resolution Photorealistic Image Translation in Real-Time: A Laplacian Pyramid Translation Network

The Hong Kong Polytechnic University 香港理工大學
Opening Minds • Shaping the Future • 啟迪思維 • 成就未來
达摩院 Alibaba DAMO Academy

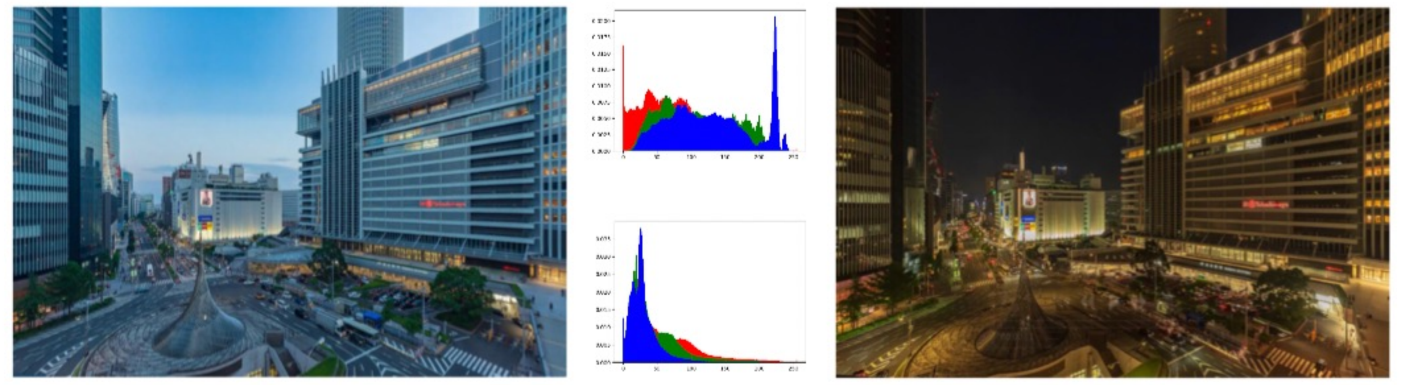
Jie Liang, Hui Zeng, Lei Zhang

Dataset and code: https://github.com/csjliang/LPTN Email: liang27jie@gmail.com

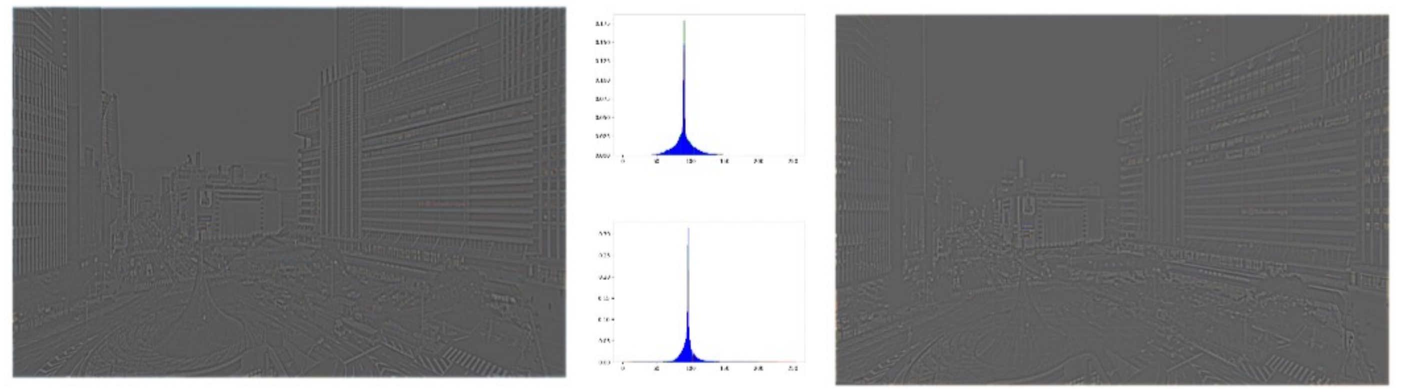
CVPR VIRTUAL JUNE 19-25

Introduction

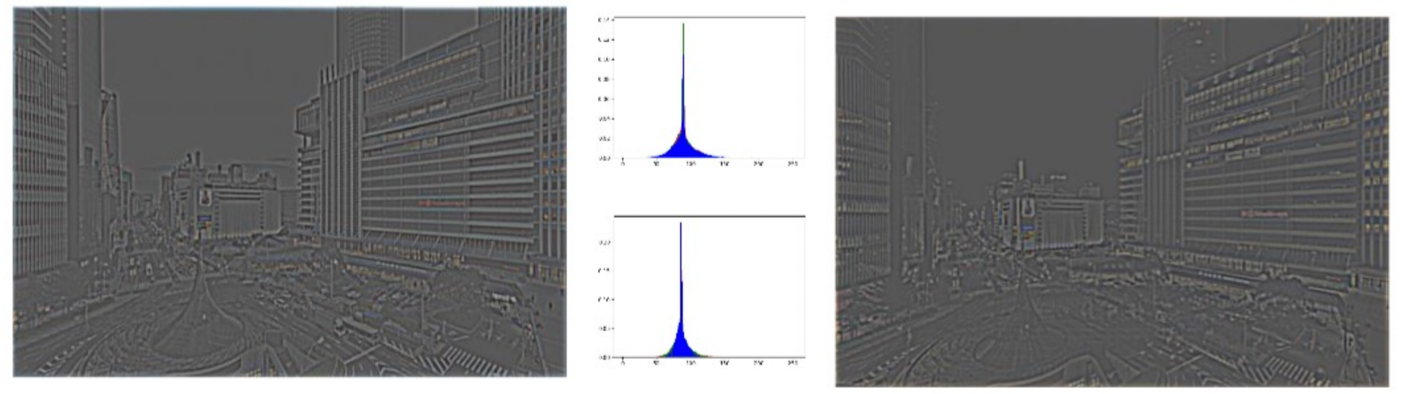
Existing image-to-image translation (I2IT) methods are either constrained to low-resolution images or long inference time due to their heavy computational burden on the convolution of high-resolution feature maps. Yet we reveal the attribute transformations in photorealistic I2IT, such as illumination and color manipulation, relate more to the low-frequency component in a Laplacian pyramid, while the content details can be adaptively refined on high-frequency components.



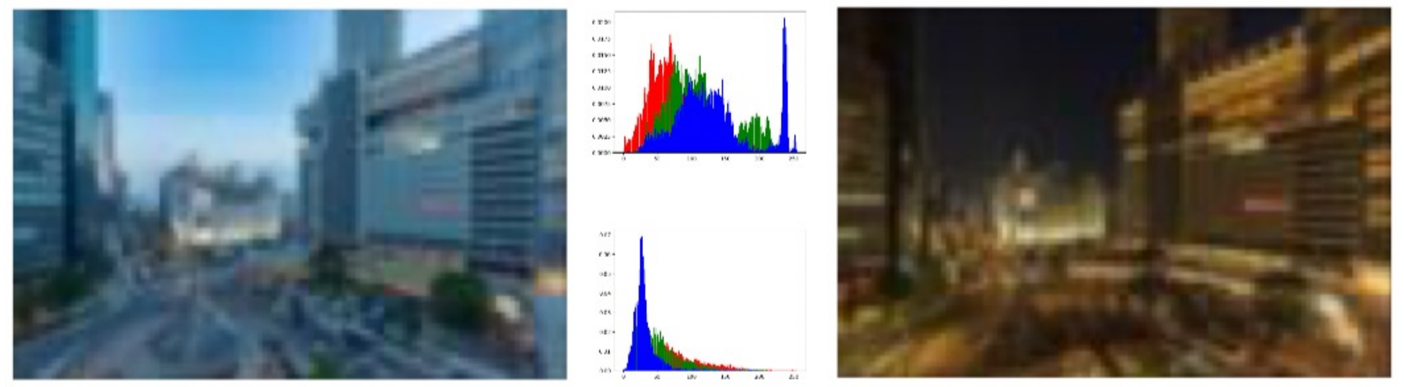
(a) Original Images, MSE=7853.9



(b) High Frequencies, Level=1, MSE=97.5

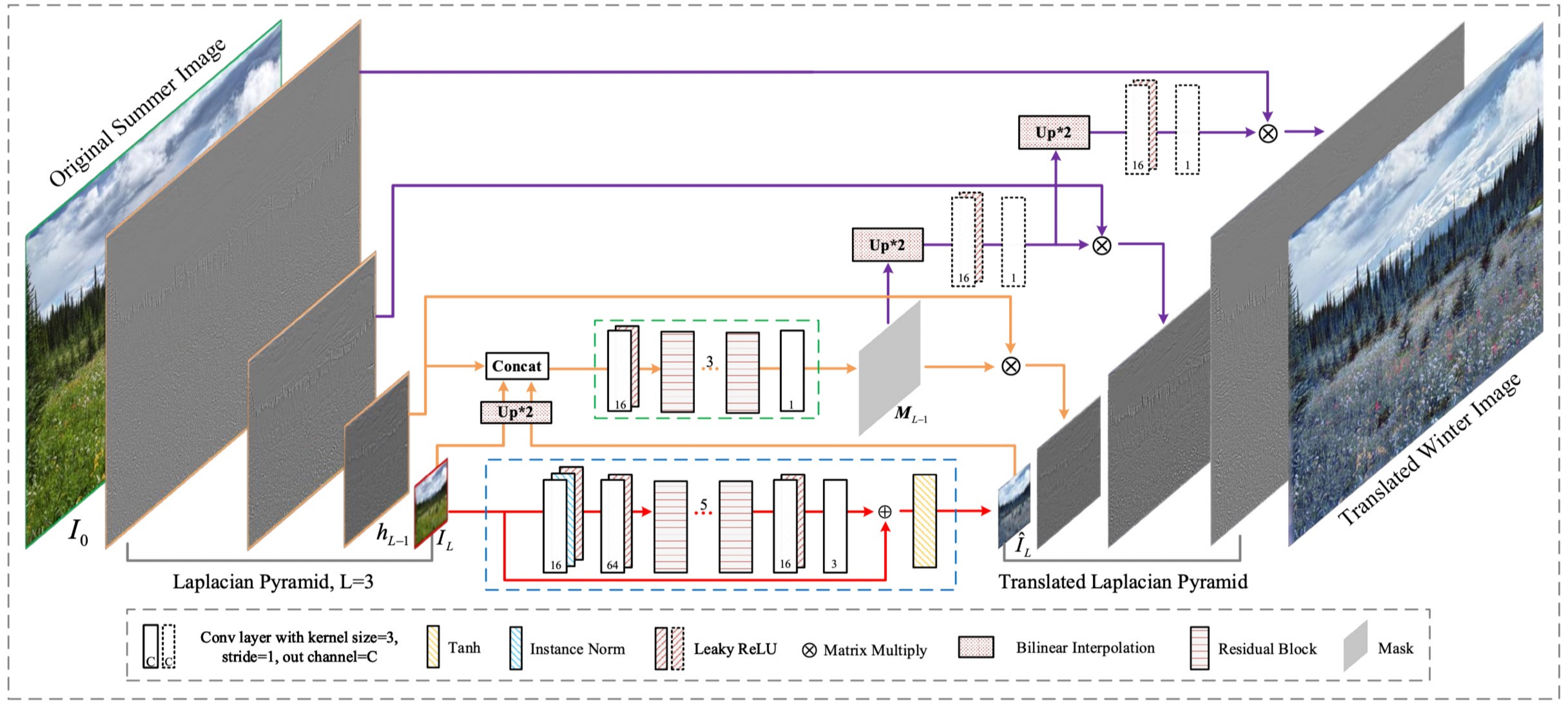


(c) High Frequencies, Level=2, MSE=107.7



(d) Low Frequencies, Level=3, MSE=6969.4

Methodology



Pipeline of the proposed LPTN algorithm. Given a high-resolution image $I_0 \in \mathbb{R}^{h \times w \times 3}$, we first decompose it into a Laplacian pyramid (e.g., $L = 3$). Rad arrows: For the low-frequency component $I_L \in \mathbb{R}^{\frac{h}{2^L} \times \frac{w}{2^L} \times 3}$, we translate it into $\hat{I}_L \in \mathbb{R}^{\frac{h}{2^L} \times \frac{w}{2^L} \times 3}$ using a lightweight network. Brown arrows: To adaptively refine the high-frequency component $h_{L-1} \in \mathbb{R}^{\frac{h}{2^{L-1}} \times \frac{w}{2^{L-1}} \times 3}$, we learn a mask $M_{L-1} \in \mathbb{R}^{\frac{h}{2^{L-1}} \times \frac{w}{2^{L-1}} \times 3}$ based on both high- and low-frequency components. Purple arrows: For the other components with higher resolutions, we progressively upsample the mask and finetune it with lightweight convolution blocks.

Methods	480p PSNR	480p SSIM	1080p PSNR	1080p SSIM	original PSNR	original SSIM
CycleGAN [33]	20.98	0.831	20.86	0.846	N.A.	N.A.
UNIT [23]	19.63	0.811	19.32	0.802	N.A.	N.A.
MUNIT [15]	20.32	0.829	20.28	0.815	N.A.	N.A.
White-Box [12]	21.32	0.864	21.26	0.872	21.17	0.875
DPE [4]	21.99	0.875	21.94	0.885	N.A.	N.A.
LPTN, $L = 3$	22.12	0.878	22.09	0.883	22.02	0.879
LPTN, $L = 4$	22.10	0.872	22.03	0.870	21.98	0.862
LPTN, $L = 5$	21.94	0.866	21.95	0.858	21.89	0.862

Quantitative comparison on the unpaired photo retouching task defined on the FiveK dataset. The LPTN performs favorably against the existing methods on various resolutions.

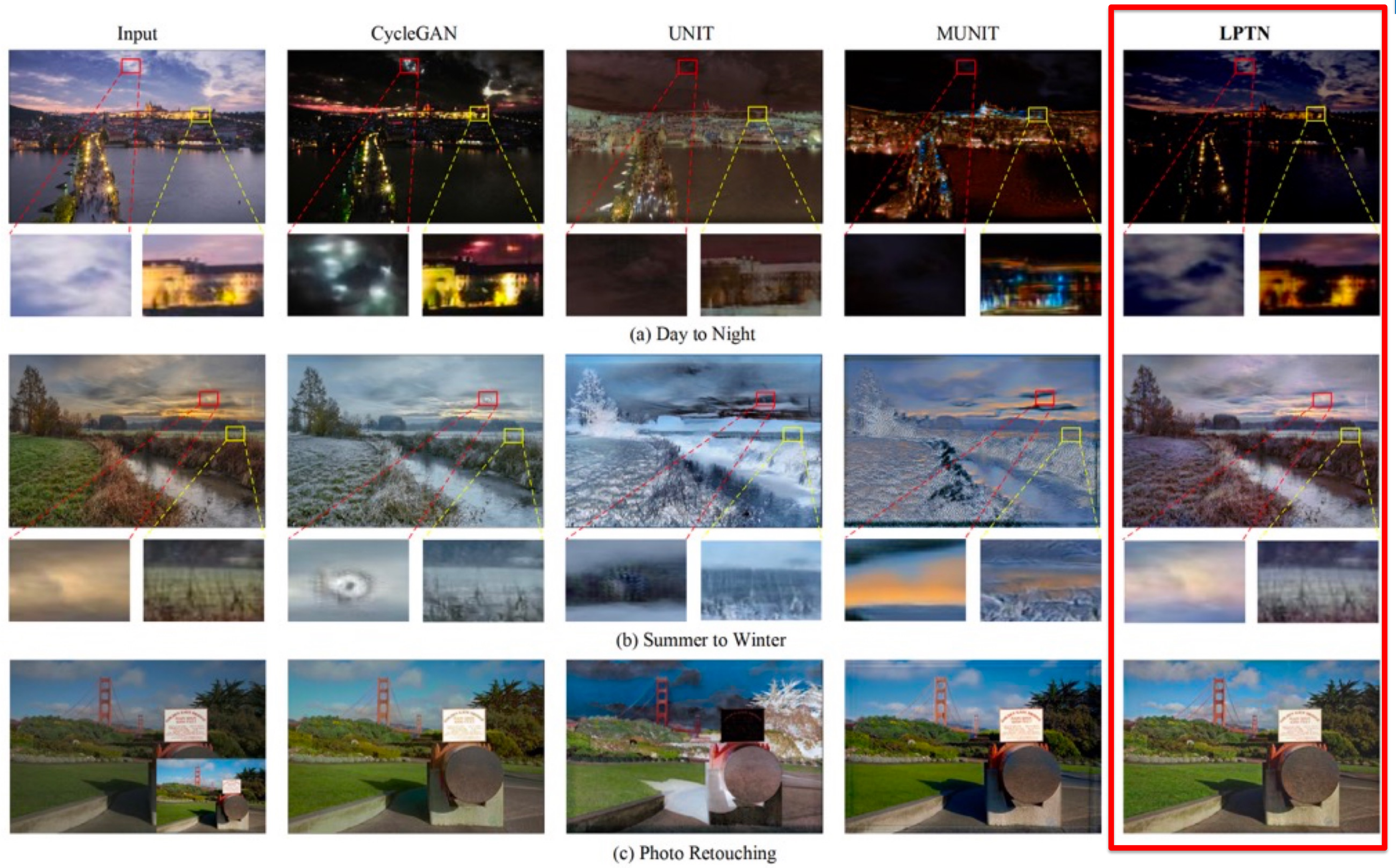
Methods	480p	1080p	2K	4K
CycleGAN [33]	0.325	0.562	N.A.	N.A.
UNIT [23]	0.294	0.483	N.A.	N.A.
MUNIT [15]	0.336	0.675	N.A.	N.A.
White-Box [12]	2.846	5.123	6.542	9.785
DPE [4]	0.032	0.091	N.A.	N.A.
LPTN, $L = 3$	0.003	0.012	0.043	0.082
LPTN, $L = 4$	0.002	0.007	0.015	0.033
LPTN, $L = 5$	0.0008	0.005	0.011	0.016

Comparison about the time consumption (in seconds) of different inference models. The LPTN runs orders of magnitude faster than others!

Comparable or superior performance, yet are orders of magnitude faster than other methods!

Experiments

Ablation study toward the model structures on the photo retouching task. The PSNRs are the average of 500 test images under the specific setting.



Visual comparisons among different I2IT methods on three tasks.

Conclusion

An efficient framework LPTN is proposed for the photorealistic I2IT problems, where the translation is mainly conducted on low-frequency components. The LPTN exhibits comparable or superior translation performance on three practical tasks, and can run at real-time on 4K resolution images by using a desktop GPU.