RT-X NET: CROSS ATTENTION TRANSFORMER USING RGB AND THERMAL IMAGES FOR LOW-LIGHT IMAGE ENHANCEMENT

Raman Jha[1]*, Adithya Lenka*[1]*, Mani Ramanagopal*[2]*, Aswin Sankaranarayanan*[2]*, Kaushik Mitra*[1]

[1]Indian Institute of Technology, Madras, [2]Carnegie Mellon University

This supplementary material presents the following details which we could not include in the main paper due to space constraints. The additional references for this elaboration are also added here.

Contents

1. Visualization of Attention maps, and feature maps for RGB, and Thermal, and cross attention module.

2. Experimental Settings

2.1. V-TIEE dataset

2.2. Multiple exposures of V-TIEE dataset in high and low gain conditions.

2.3. Visual representation of noise in V-TIEE dataset.

2.4. Noise incorporation in simulated low-light LLVIP dataset

3. Additional Qualitative Results

1. VIZUALIZATION

1.1. Analysis of Attention and Feature Maps for Scene 1

Figure 1 and Figure 2 show the raw RGB and thermal images respectively for Scene 1. The RGB image captures a nighttime street scene with multiple cars and pedestrians, while the thermal image highlights the heat signatures, particularly making human figures more prominent.

In Figure 3, we visualize the self-attention maps for both RGB and thermal streams. The RGB self-attention map shows scattered focus across spatial patches, with notable activation around high-contrast areas like car boundaries and bright reflections. The thermal self-attention map, in contrast, places strong focus on human silhouettes and heat-reflective surfaces, which aligns with the thermal modality's sensitivity to temperature differences.

Figure 4 displays the feature maps extracted after the attention modules. The RGB features retain structural patterns, such as car contours and road edges, while the thermal features emphasize human figures with less environmental detail. The cross-attention features reveal a fusion where structural and thermal cues are combined, highlighting both human and vehicle zones, demonstrating the complementary enhancement achieved by the cross-modal design.

Observations: The RGB stream alone struggles with dark regions, while the thermal stream compensates by providing reliable human detection. The cross-attention fusion successfully integrates both modalities, enriching the representation and improving feature robustness under low-light conditions.

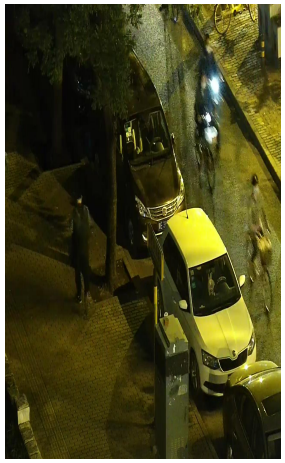

Fig. 1. RGB Image

Fig. 2. Thermal Image

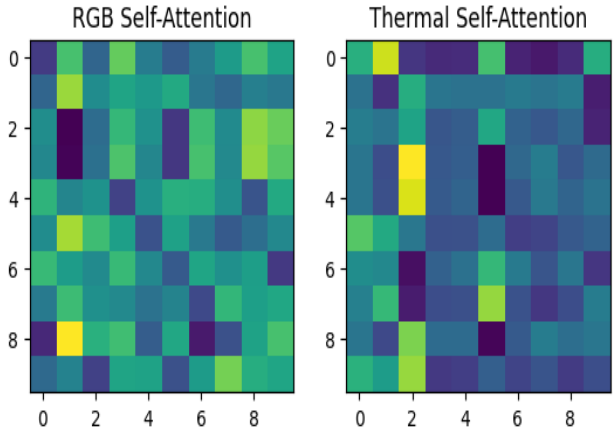

Fig. 3. Attention Maps of RGB, and Thermal Images

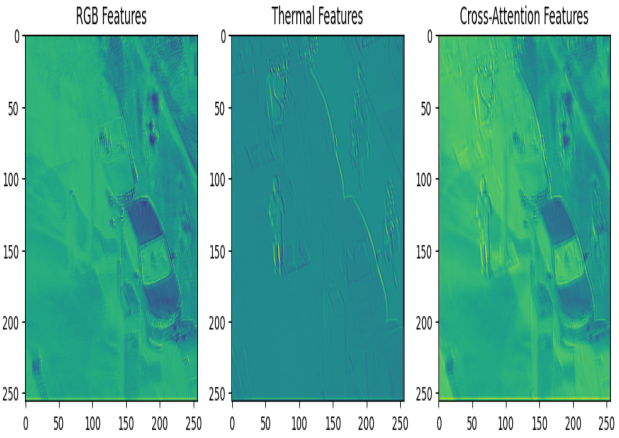

Fig. 4. Feature Maps of RGB, Thermal, and Cross Attention module

1.2. Analysis of Attention and Feature Maps for Scene 2

Figure 5 and Figure 6 display the RGB and thermal images of Scene 2. This scene features a night urban setting with pedestrians under streetlights, where shadows and reflections create visually challenging conditions in the RGB space. The thermal image effectively highlights the human heat signatures, especially in areas obscured or darkened in the RGB frame.

In Figure 7, the self-attention maps show distinct behavior: the RGB attention focuses mainly on high-brightness zones and sharp texture changes, while the thermal attention prioritizes human figures and their surroundings, demonstrating modality-specific selectivity. This suggests the attention mechanism leverages complementary information from each domain.

Figure 8 presents the post-attention feature maps. The RGB features accentuate road lines and lamp posts, while the thermal features isolate the heat-emitting pedestrians. The cross-attention fusion produces a richer representation, integrating geometric and thermal cues, which results in improved focus on pedestrians and their immediate vicinity.

Observations: In Scene 2, the RGB stream struggles with strong shadows and low contrast, while the thermal modality reliably highlights human activity. The fused cross-attention features demonstrate how combining these modalities strengthens both spatial and semantic representation, ultimately supporting better low-light scene understanding.

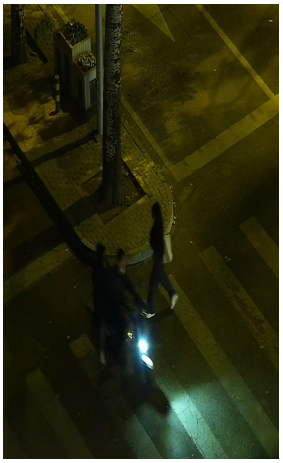

Fig. 5. RGB Image

Fig. 6. Thermal Image

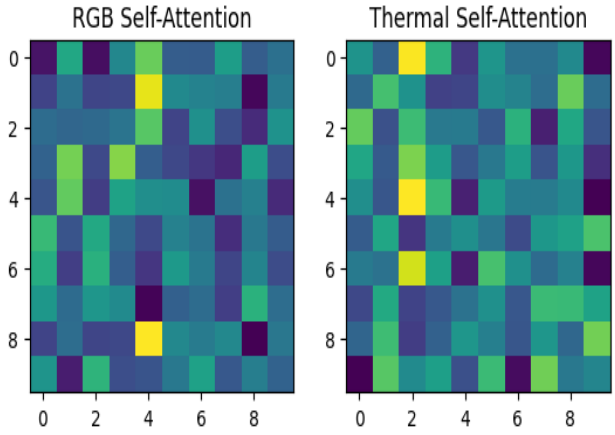

Fig. 7. Attention Maps of RGB, and Thermal Images

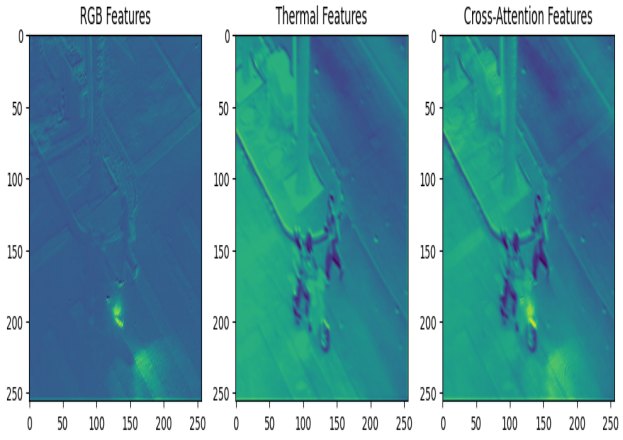

Fig. 8. Feature Maps of RGB, Thermal, and Cross Attention module

2. EXPERIMENTAL SETTINGS

2.1. V-TIEE dataset

In the V-TIEE dataset, we have systematically captured RGB and thermal images under varying gain and exposure conditions. Specifically, images were recorded at two distinct gain settings, each with five different exposure values. Furthermore, in four distinct scenes, we have expanded the exposure range to include ten different values for each gain setting. The gain values in our dataset span from 0 dB to 44.99 dB. Scenes include lower gain settings at 0, 4.99, 19.99, 23.99, and 24.99 dB, as well as higher gain settings at 23.99 and 44.99 dB. The exposure times range from as short as 1/20000 seconds to as long as 10 seconds. This extensive range of exposure settings can also facilitate the dataset's utility in generating High Dynamic Range (HDR) [1] images, providing a robust resource for HDR imaging research and applications.

2.2. Mutiple exposure of V-TIEE dataset in high and low gain conditions.

This extensive range of exposure settings can also facilitate the dataset's utility in generating High Dynamic Range (HDR) [1] research and applications in lot light image datasets. Since this dataset is not used for training, performance on this evaluation dataset would demonstrate the generalization of the models. In practical scenarios, noise levels increase as the exposure of a scene decreases. Our V-TIEE dataset captures images under various gain conditions, thus incorporating noise characteristics that are similar to real-world settings.

2.3. Visual representation of real-world V-TIEE dataset

In practical scenarios, noise levels increase as the exposure of a scene decreases. Our V-TIEE dataset captures images under various gain conditions, thus incorporating noise characteristics typical of real-world settings. The low-light input images, depicted in Fig. 10, were utilized in our experiments. For enhanced visual comprehension, amplified versions of these inputs are also presented. Additionally, we provide a well-exposed image of the same scene with the thermal image.

2.4. Noise incorporation in simulated low-light LLVIP dataset

Noise model: [2] To encapsulate the fundamental characteristics of noise, it is considered a variable with a mean of zero and variance from two independent sources. Specifically, the following representation applies to pixels below the saturation threshold:

$$Var(n) = \phi t/g^2 + \sigma^2_{read}/g^2 \qquad (1)$$

Here, g represents the sensor gain. The initial component describes the Poisson distribution of photon arrival, directly proportional to the accumulated photon count ϕt. The final component, representing the pre-amplification stage, accounts for noise from the sensor's readout process. We applied this method to add noise to the low-light synthetic LLVIP [3] dataset, ensuring real-world conditions. The amplified image in Fig.11 is provided solely for visual representation, of the noise simulation.

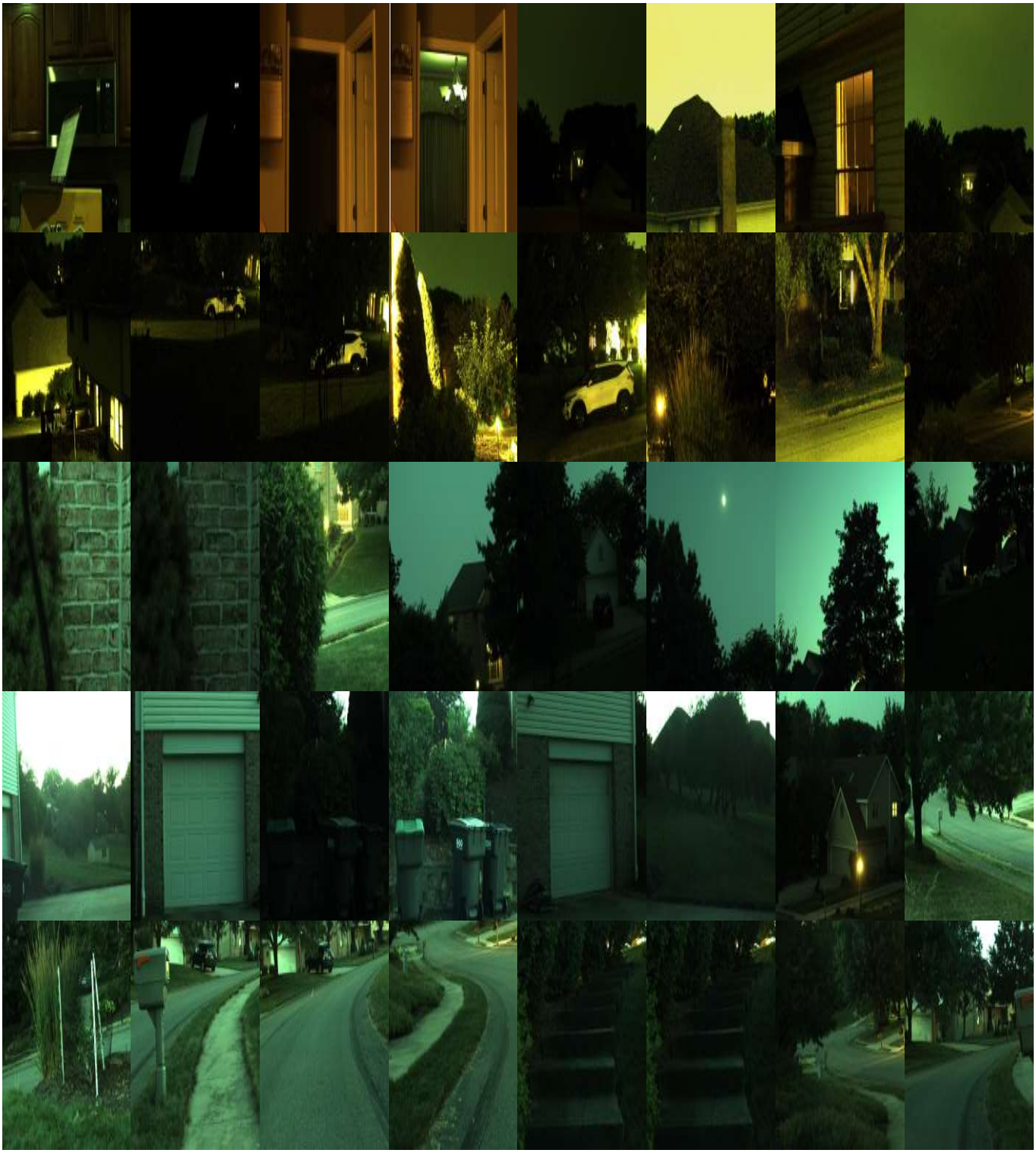

Fig. 9. The figure shows all the indoor and outdoor scenes of RGB and Thermal images in the V-TIEE dataset which we captured in real-time for the low light image enhancement. For space constraints, we are only showing RGB images.

Fig. 10. The figure shows noise in our real-world V-TIEE dataset, showing low-light input, amplified input, well-lit, and corresponding thermal images. Amplified images, enhanced by a factor of 10, highlight the noise in low-light images.

3. ADDITIONAL QUALITATIVE RESULTS

4. REFERENCES

[1] Paul E. Debevec and Jitendra Malik, "Recovering high dynamic range radiance maps from photographs," USA, 1997, ACM Press/Addison-Wesley Publishing Co.

[2] Samuel W. Hasinoff, Frédo Durand, and William T. Freeman, "Noise-optimal capture for high dynamic range photography," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 553–560.

[3] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou, "Llvip: A visible-infrared paired dataset for low-light vision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3496–3504.

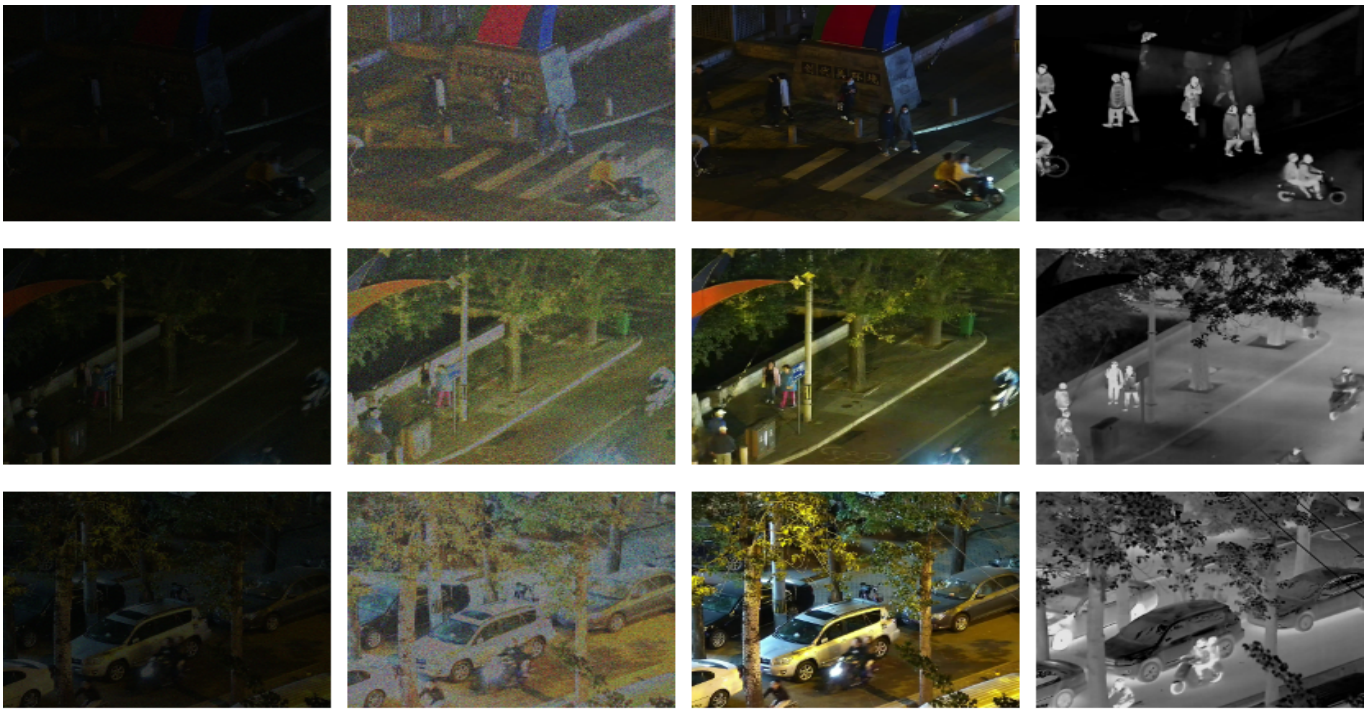

Fig. 11. The figure shows noise incorporation in the simulated low-light LLVIP dataset, low-light input, amplified input, well-lit, and corresponding thermal images. Amplified images, enhanced by a factor of 10, highlight the noise in low-light images.

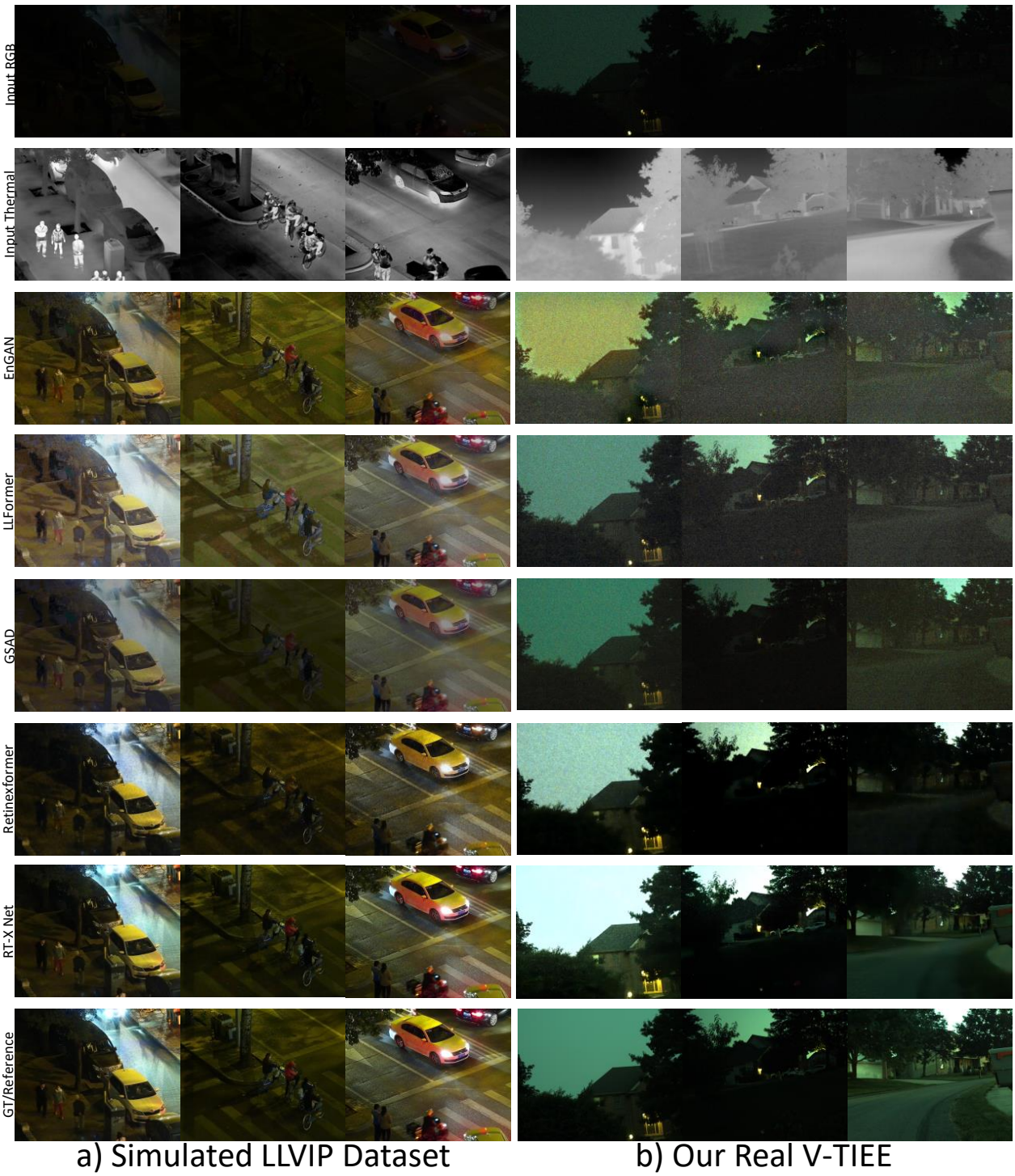

Fig. 12. Qualitative results on the synthetic LLVIP dataset and real-world V-TIEE dataset. Columns denote different scenes. The first two rows show the input visible and thermal images. The next five rows are the outputs from RT-X Net and state-of-the-art visible image enhancement algorithms. The last row shows the reference well-exposed image.