THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

ICPR 2020

Video Lightening with Dedicated CNN Architecture

Authors: Li-Wen Wang, Wan-Chi Siu, Zhi-Song Liu,

Chu-Tak Li, and Daniel P. K. Lun

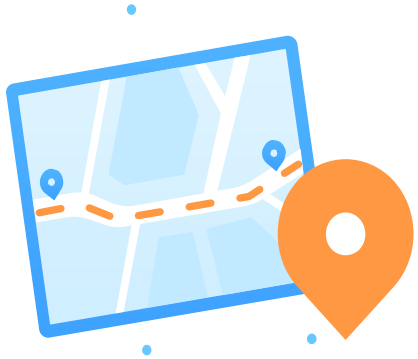
Department of Electronic and Information Engineering

The Hong Kong Polytechnic University

IAPR

IEEE COMPUTER SOCIETY

❖ Outline



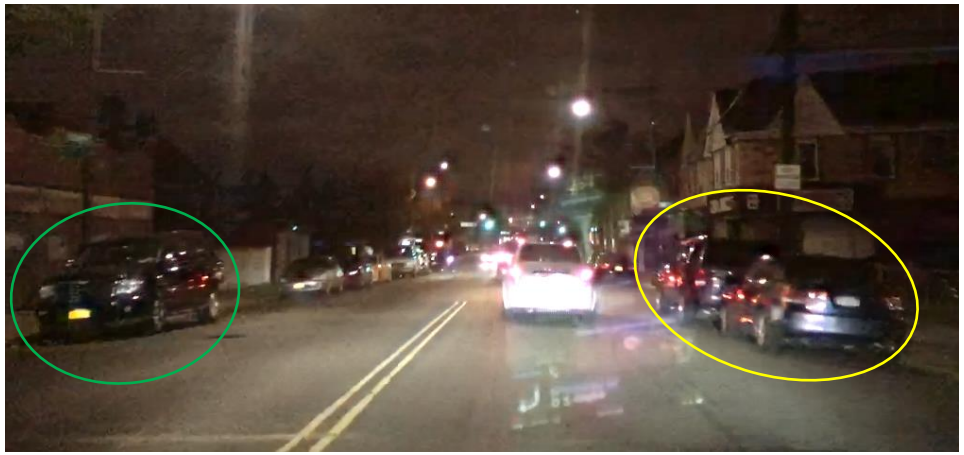
1. Background
2. The Proposed Method
3. Experiments
4. Conclusion

Background

❖ Background

- **Darkness** brings us uncertainty, worry and low-confidence, especially when we are walking or driving. We may use **visual aid** (including small devices such as **mobile phones**) which is more convenient compared with dedicated approaches such as **infrared detection**.
- We focus on video enhancement in low-light conditions and propose a new CNN structure, i.e., the **Video Lightening Network (VLN)**. It achieves remarkable enhancement for the low-light videos,

Low-light condition (night)



Our proposed method

❖ Background (cont'd)

To take good visual-quality images in dark environment, there are some possible solutions:

- **Flash:**

  It may **bother** others and is **not allowed** in some places (museum).

- **High ISO sensitivity:**

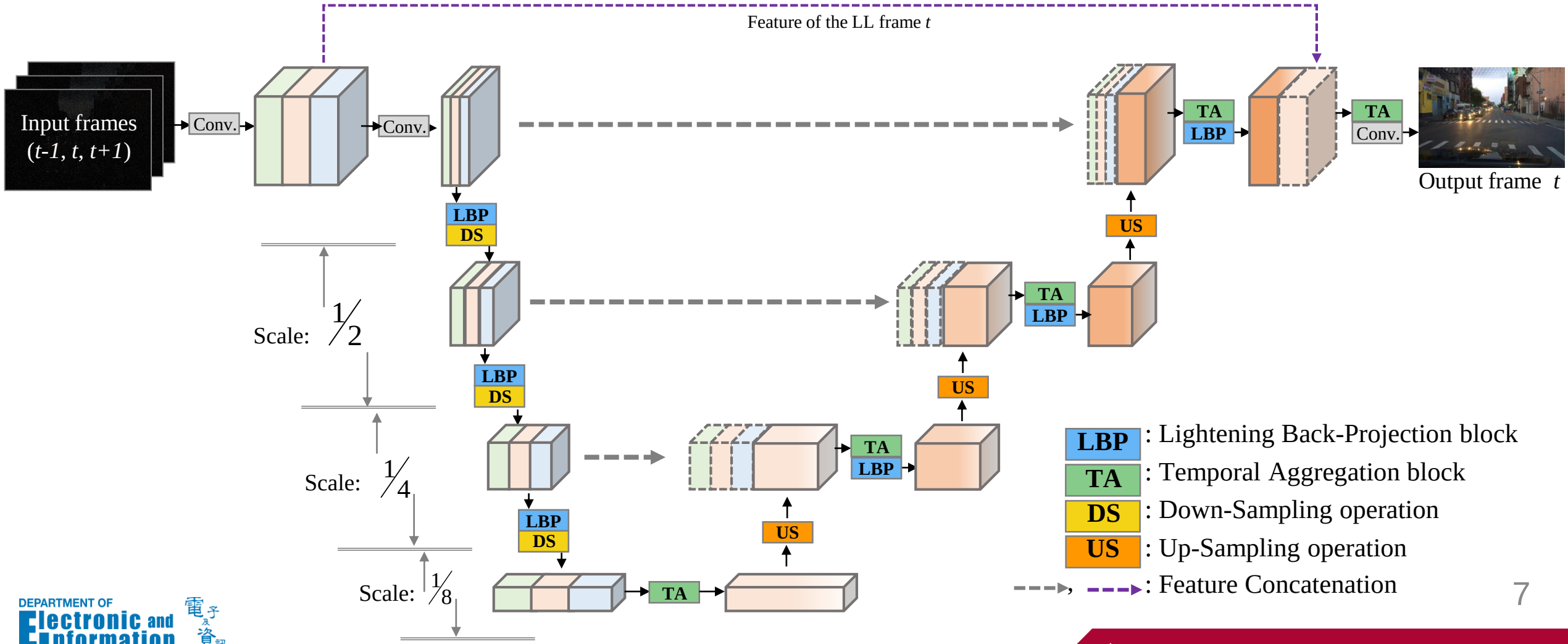
  It brings more **noise** and **overexposes** to the normal-light areas.

- **Longer exposure time:**

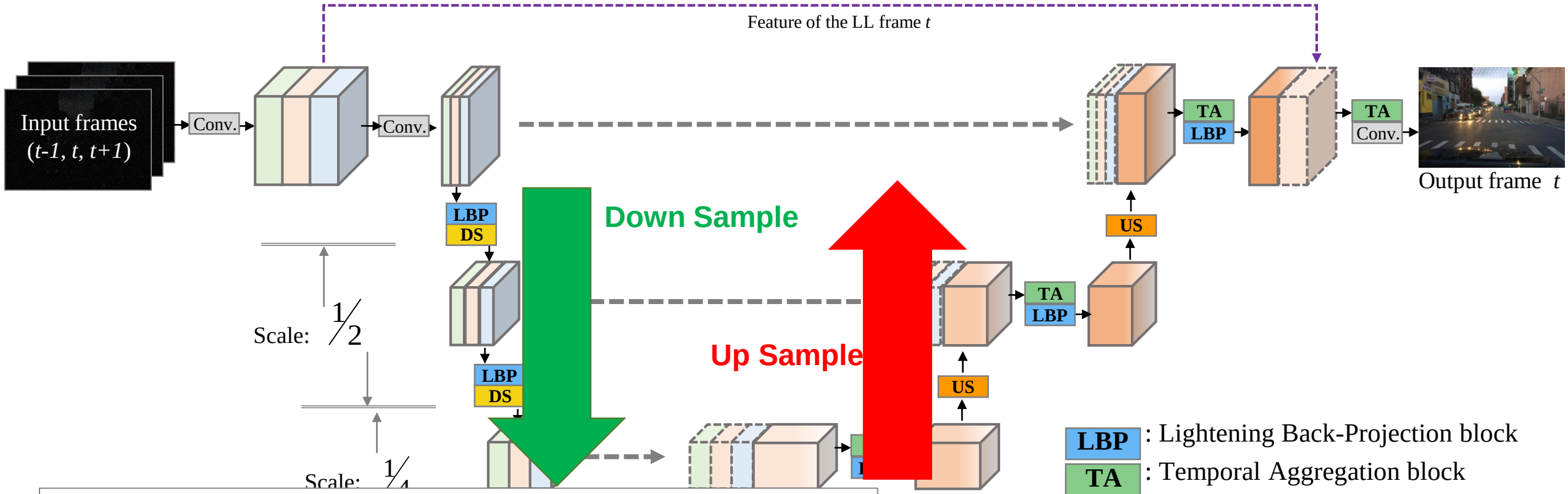
  It may suffer from **blur** problems and is not suitable for shooting **videos** (the interval of video frames may be too short).

The Proposed Method

❖ **The Proposed Method** – Video Lightening Network (VLN)

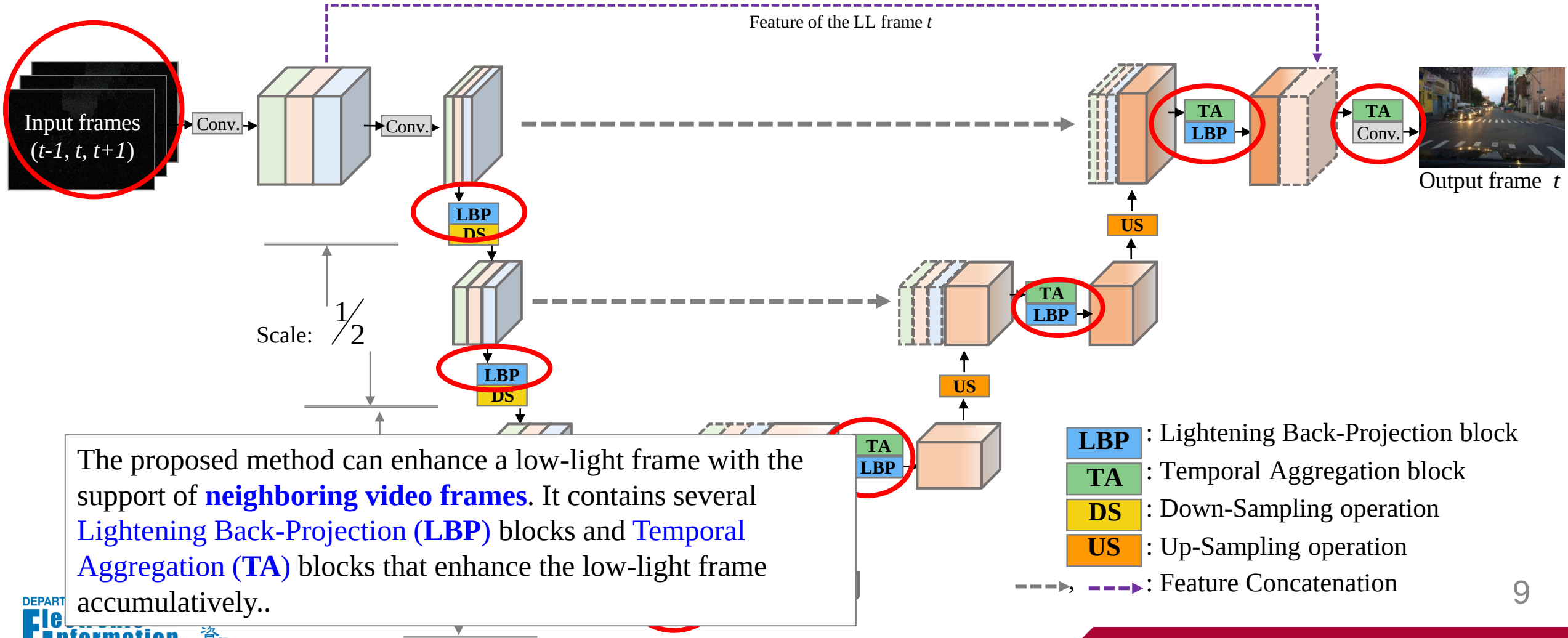


❖ **The Proposed Method** – Video Lightening Network (VLN)



The proposed **Video Lightening Network (VLN)** down- and up-samples the input frames that can obtain the context and localized information from the **global** and **local** features, which benefits the understanding and processing of the input frames.

❖ The Proposed Method – Video Lightening Network (VLN)



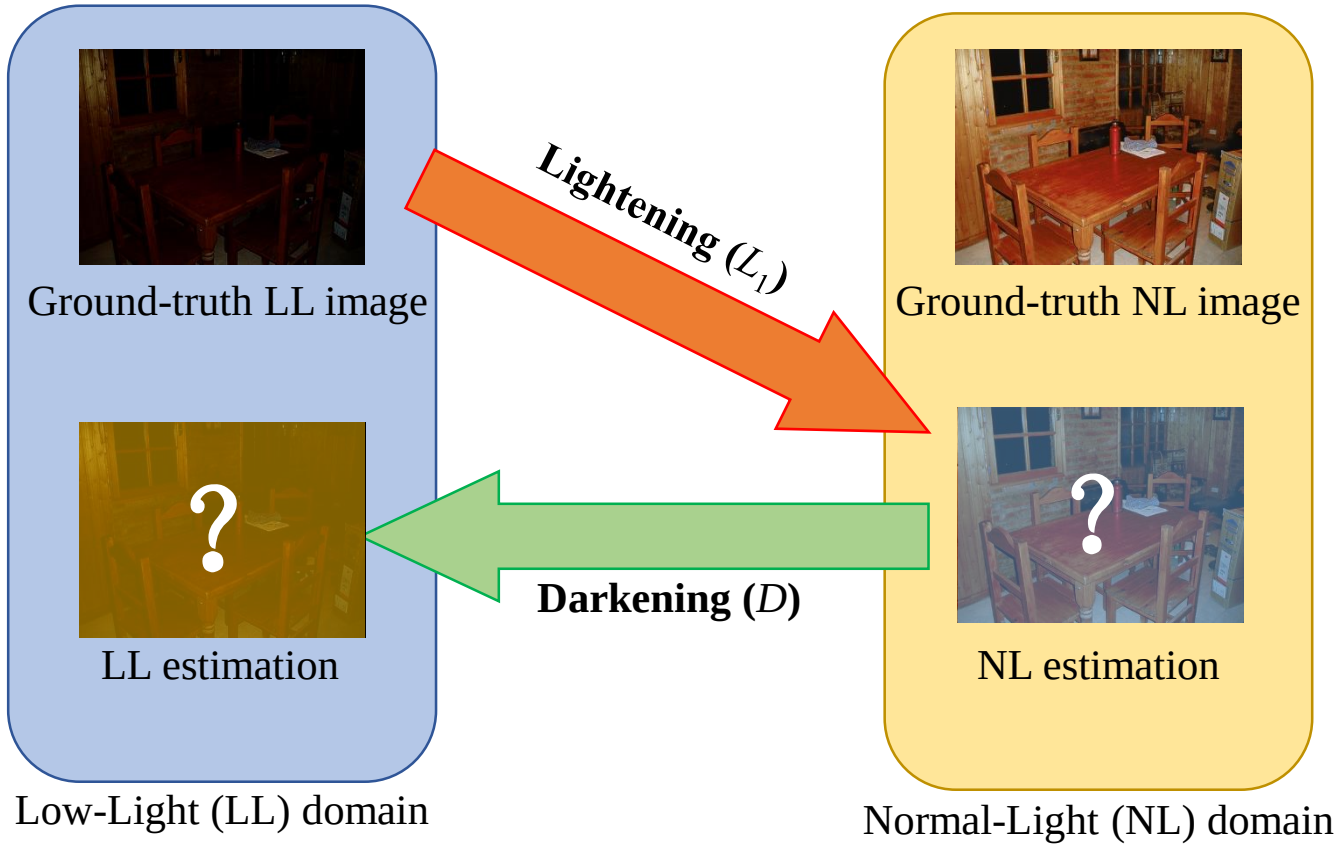
The proposed method can enhance a low-light frame with the support of **neighboring video frames**. It contains several Lightening Back-Projection (**LBP**) blocks and Temporal Aggregation (**TA**) blocks that enhance the low-light frame accumulatively..

❖ The Proposed Method – domain transfer learning

Mapping between low- and normal-light domain:

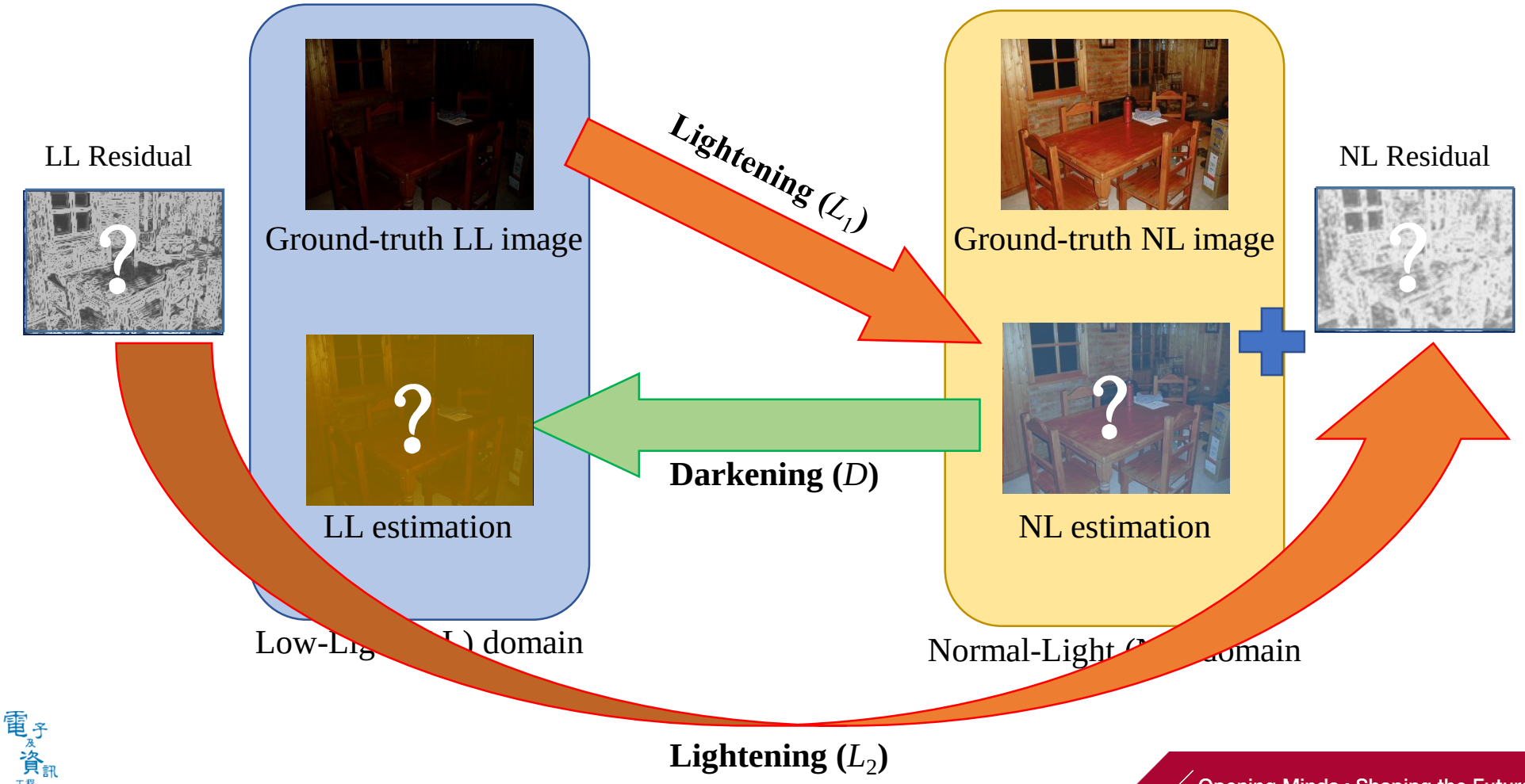
Our approach is to **make "Lightening" process and "Darkening" process iteratively,** which means to **learn the lightening mechanism gradually**, and finally obtain the normal-light image.

This process can also be roughly described as **domain transfer learning**.



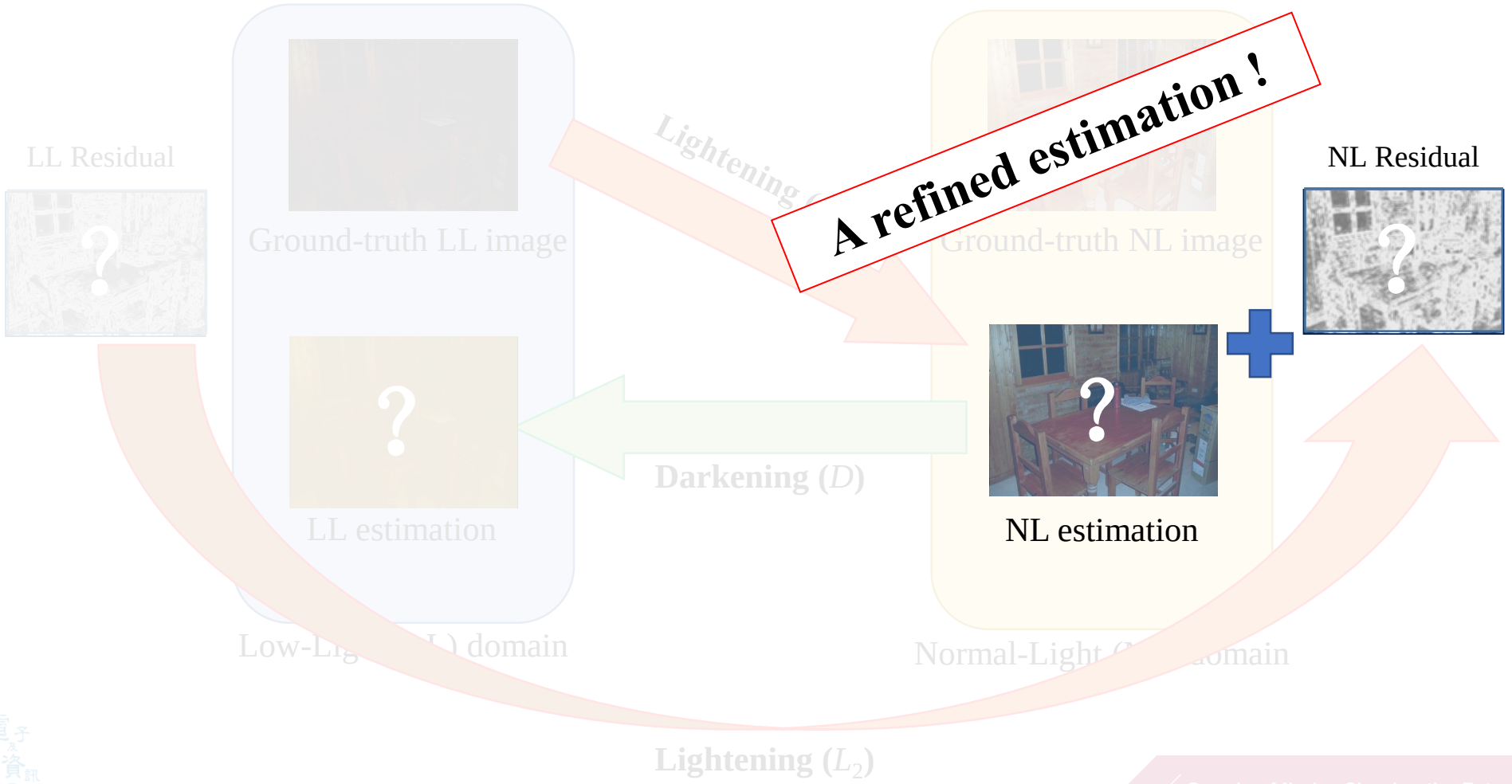
❖ **The Proposed Method** – domain transfer learning

Mapping between low- and normal-light domain:



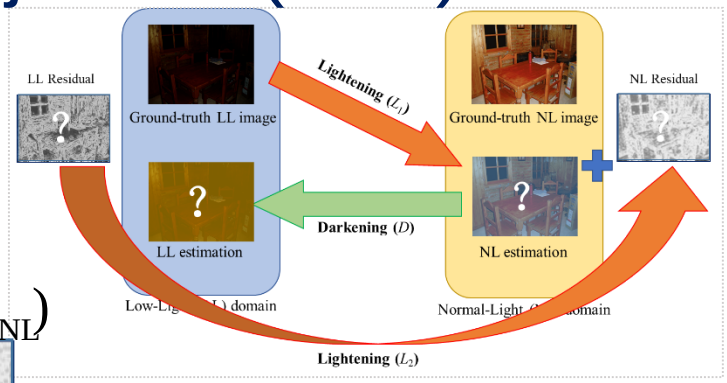
❖ **The Proposed Method** – domain transfer learning

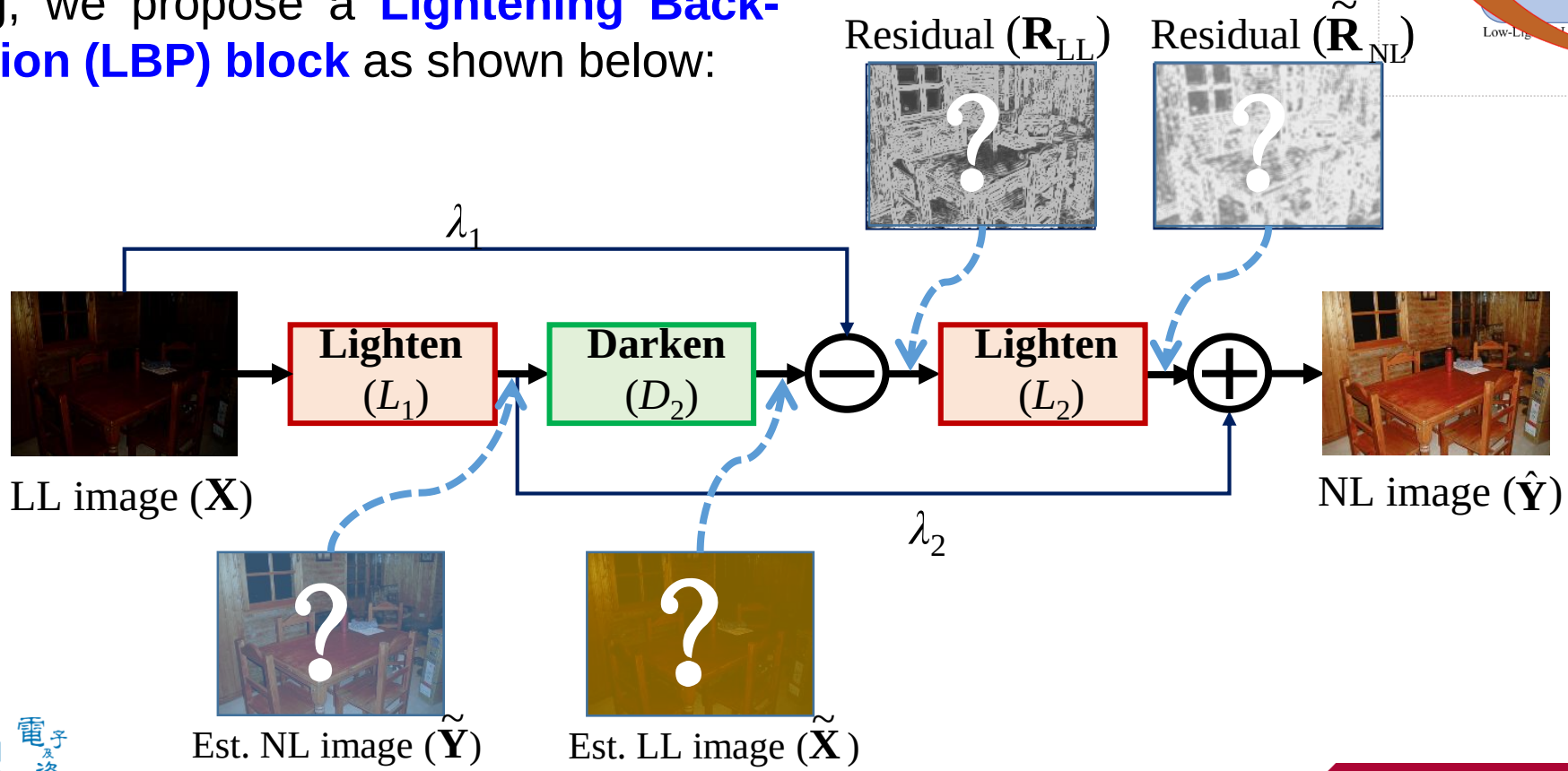
Mapping between low- and normal-light domain:



❖ The Proposed Method – Lightening Back-Projection (LBP)

Based on the ideas of the domain transfer learning, we propose a **Lightening Back-Projection (LBP) block** as shown below:





❖ The Proposed Method – Lightening Back-Projection (LBP)

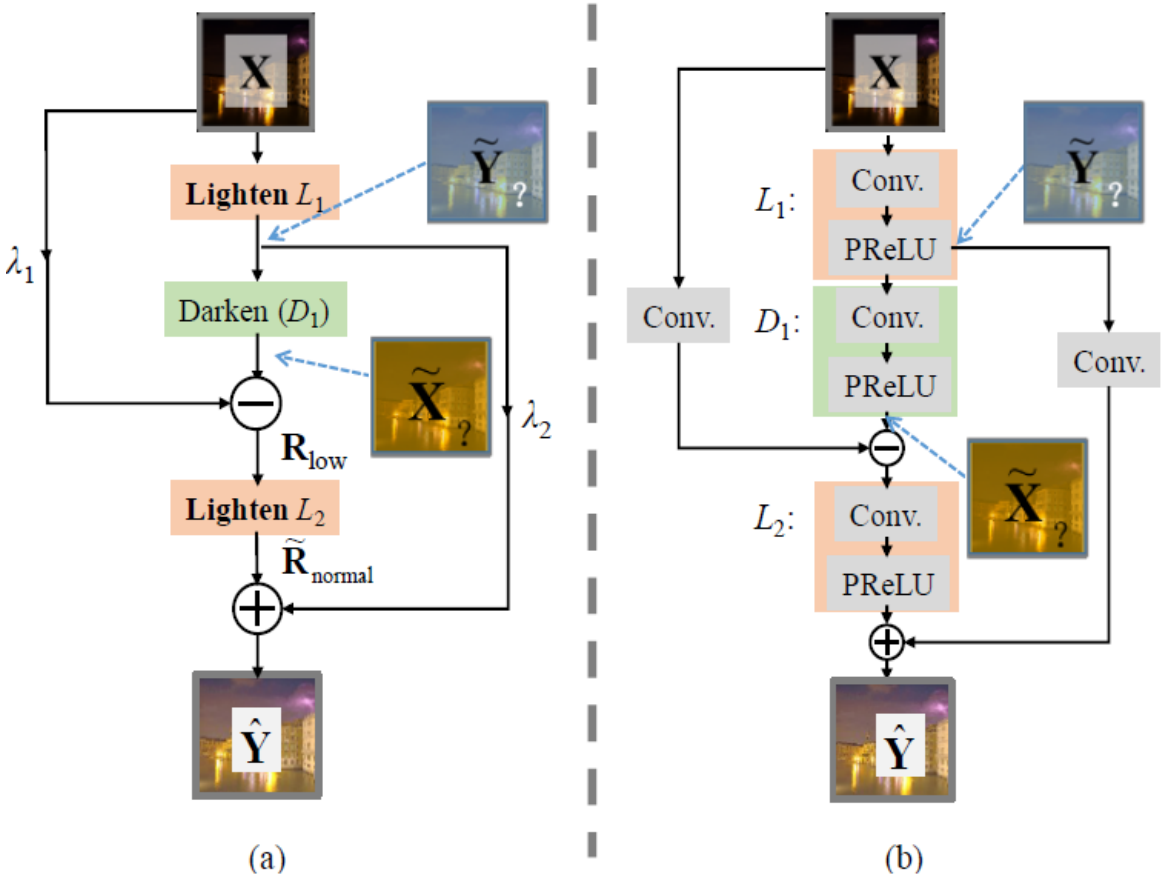
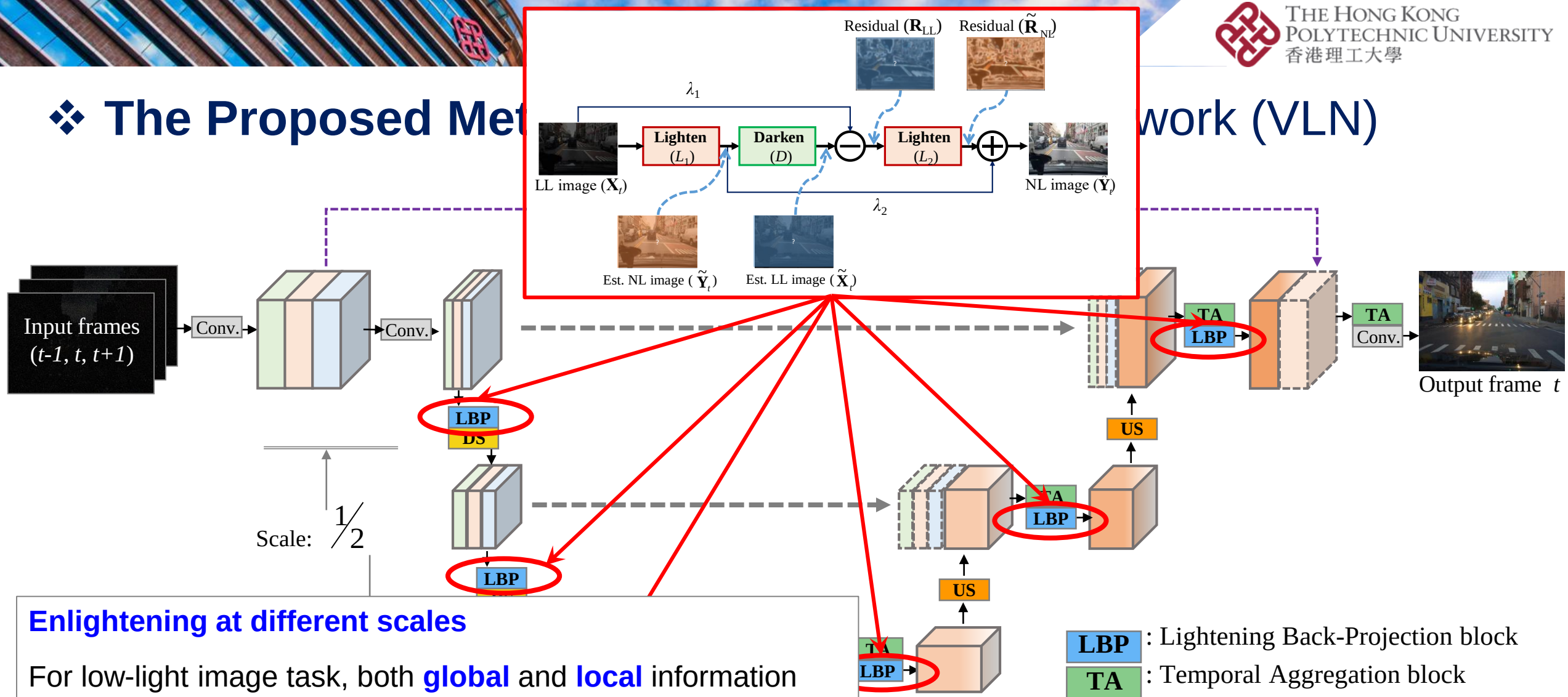


Figure 4. Lightening Back-Projection (LBP) Block: (a) framework, (b) CNN structure

An advantage of CNN is that the convolutional layers contain trainable parameters (weight *w* and bias *b*), which makes CNN be adaptive to different tasks. Therefore, we can use the same CNN structure for both Lighten and Darken processes.

A reasonable inference of the intermediate processes is that: The input of L_2 is the residual information of the low-light image, which makes the output of L_2 also be residual. Considering the LBP predicts normal-light image at the output, the output of L_2 should be the residual $\tilde{\mathbf{R}}_{normal}$ of normal-light condition, so L_2 **lightens** the residual from the low- to normal-light domain. To obtain the normal-light image after an addition process, it needs another information $\tilde{\mathbf{Y}}$ that is from the output of L_1. The input of L_1 is low-light image $\mathbf{X}$ and output is the normal-light image $\tilde{\mathbf{Y}}$, hence L_1 achieves the **Lighten** process. Because the input of L_2 is the residual $\mathbf{R}_{low}$, D_1 should act **darken** process that transfer the normal-light estimation $\tilde{\mathbf{Y}}$ to low-light's $\tilde{\mathbf{X}}$.

❖ The Proposed Met…work (VLN)



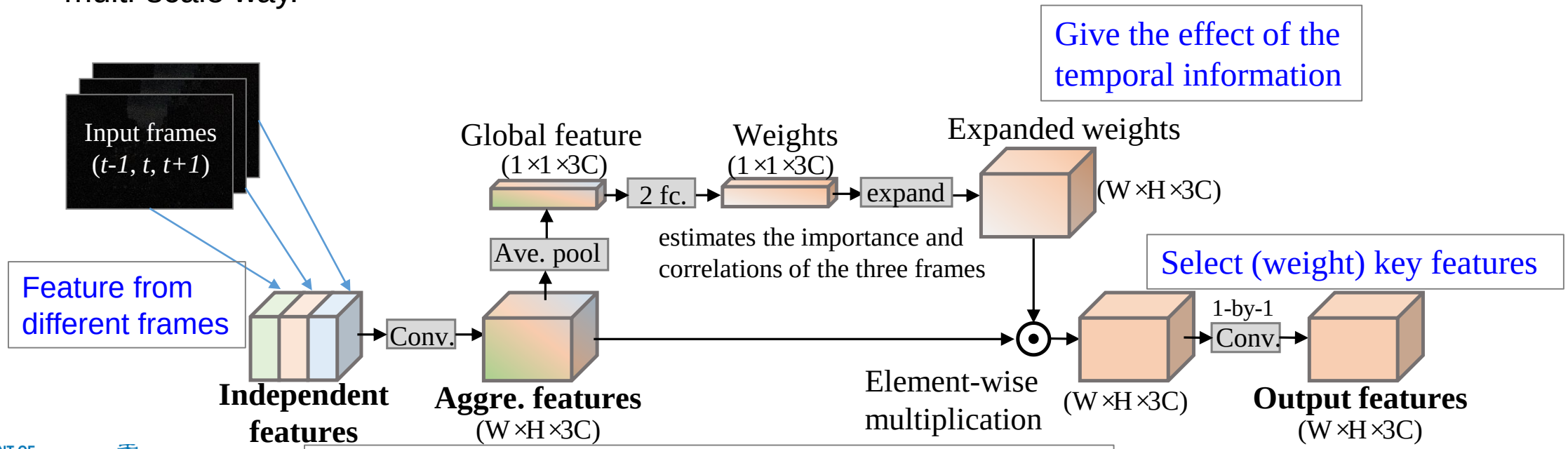
Enlightening at different scales

For low-light image task, both **global** and **local** information are important, as we need global information to evaluation the **light condition** of the whole image, and local information to **enrich the details**. Hence, lightening image at multi-level features can benefit the performance.

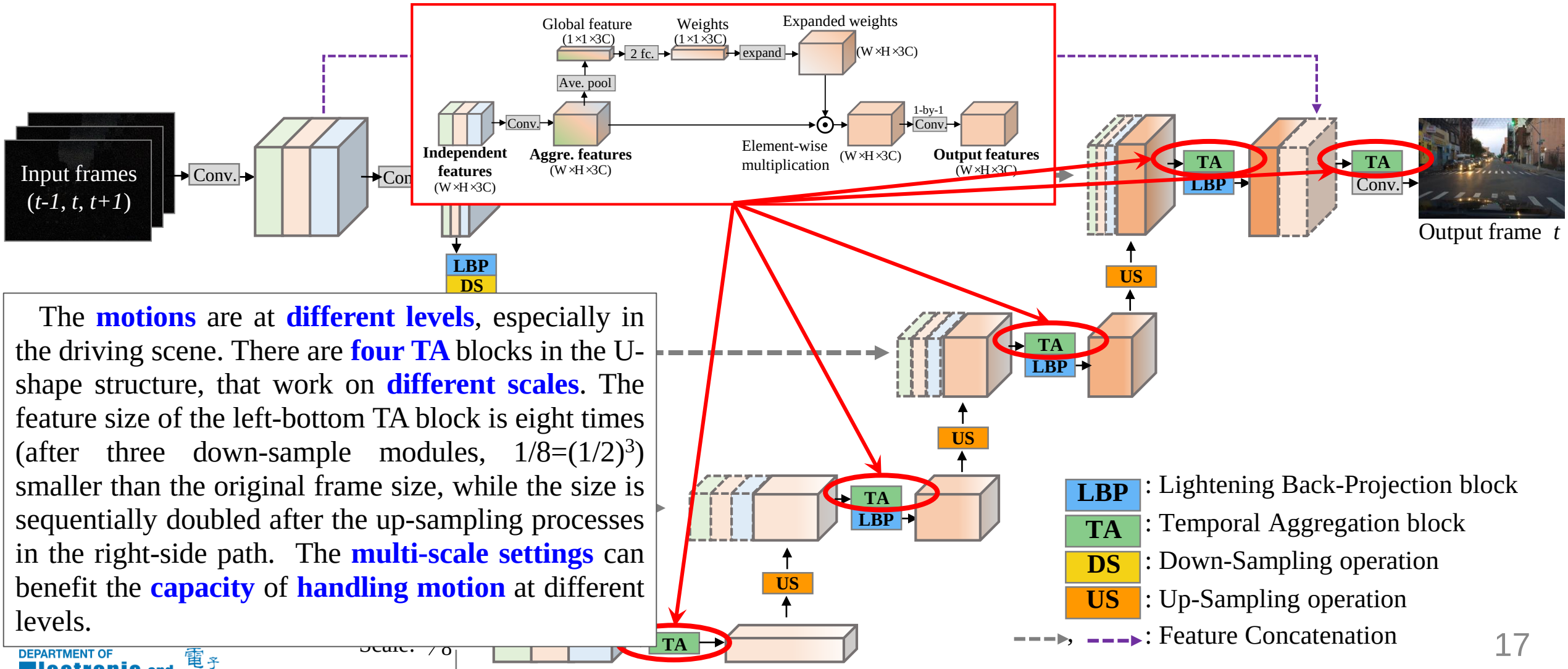
LBP : Lightening Back-Projection block
TA : Temporal Aggregation block
DS : Down-Sampling operation
US : Up-Sampling operation
: Feature Concatenation

❖ The Proposed Method – Temporal Aggregation (TA)

- The **adjacent frames** are strongly correlated. They can provide rich information for low-light enhancement. E.g., the noise at different frame may locate at different position.
- Most of the additional information is **redundant**, and even some information will degrade the result as **noise** and **distorting** motion.
- To effectively utilize the information among the neighboring frames, we propose a novel Temporal Aggregation (TA) block which aims to work for **spatial-temporal feature aggregation** through a multi-scale way.



❖ The Proposed Method – Multi-scale Motion Compensation



The **motions** are at **different levels**, especially in the driving scene. There are **four TA** blocks in the U-shape structure, that work on **different scales**. The feature size of the left-bottom TA block is eight times (after three down-sample modules, $1/8=(1/2)^3$) smaller than the original frame size, while the size is sequentially doubled after the up-sampling processes in the right-side path. The **multi-scale settings** can benefit the **capacity** of **handling motion** at different levels.

❖ The Proposed Method – loss function

- **For Training**, we regard the low-light video enhancement as a supervised learning task. For each low-light input, there is a normal-light target that is known to us.
- We need to measure the difference between the estimation and the ground-truth target. The loss is defined as two parts:
  - The training loss contains two parts: **L1-norm Loss** measures the difference between the estimation $\hat{\mathbf{Y}}_t$ and its ground truth target $\mathbf{Y}_t$. Considering that the L1-norm averages the errors of all locations, which may cause the blur problem. We include a refinement term makes uses of the **content perceptron loss** $cp(\cdot)$.

$$Loss(\hat{\mathbf{Y}}_t, \mathbf{Y}_t) = \|\hat{\mathbf{Y}}_t - \mathbf{Y}_t\|_1 + \lambda \cdot \|cp(\hat{\mathbf{Y}}_t) - cp(\mathbf{Y}_t)\|_2$$

λ is a balanced coefficient (it was set to 0.01 in our experiment)

where $cp(\cdot)$ is defined as the feature maps from relu3_3 layer of the VGG-16. In other words, the estimation $\hat{\mathbf{Y}}_t$ and the ground truth $\mathbf{Y}_t$ are processed by the VGG-16 network. We then measure the L2-norm distance between their features from the layer relu3_3. The **VGG-16** network was trained **at ImageNet** for image classification, where the extracted **feature** has **discrimination power** for recognizing **different contents.**

Experiments

❖ Experiments

- Cameras **cannot** capture **two videos simultaneously** under **different illuminations**, which makes it impossible to obtain the low- and normal-light video pairs.
- The normal-light images usually contain more information and less noise than the low-light ones. It is **feasible** to synthesize the low-light images from the normal-light ones.
- Following the analysis, **a LL image can be simulated from an NL image** through the following simulation equation [18]:

$$\overline{\mathbf{X}}^{(i)} = \beta(\alpha \mathbf{Y}^{(i)})^{\gamma}$$

Where $\overline{\mathbf{X}}^{(i)}$ represents a simulated LL image. The piel value of the NL image $\mathbf{Y}$ is compressed to [0, 1]. i denotes the R, G, or B channel of the image. $\alpha \sim U(0.9, 1)$, $\beta \sim U(0.5, 1)$ and $\gamma \sim U(1.5, 5)$ which control the effect of low-light simulation.

[18] Feifan Lv and Feng Lu, "Attention-guided Low-light Image Enhancement," *arXiv preprint arXiv:1908.00682*, 2019.

❖ Simulation of different illumination levels

$$\overline{\mathbf{X}}^{(i)} = \beta(\alpha\mathbf{Y}^{(i)})^{\gamma}$$

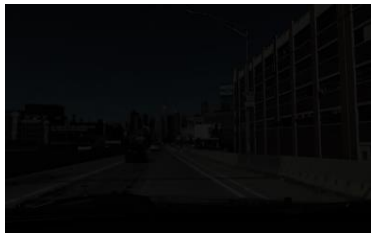
Different scenes may under **different illuminations**, which could produce inconsistent brightness of video sequences. We can use **different parameters** to simulate different levels of illumination

	Slight dark	Middle dark	Extreme dark
β	0.9	0.8	0.6
α	0.98	0.95	0.93
γ	2	3	4
Saturation	0.8	0.6	0.4
Contrast	0.8	0.6	0.4


Highway (NL)
(ave. brightness: 101.86)

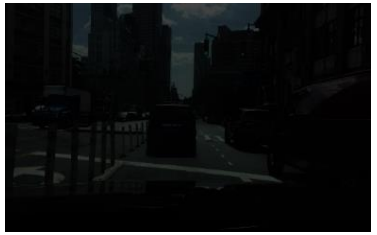

Slight dark (LL)
(ave. brightness: 58.49)

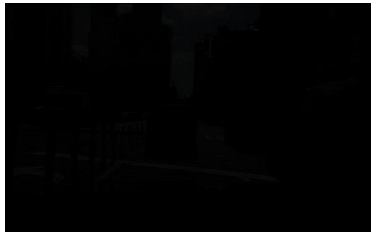

Middle dark (LL)
(ave. brightness: 23.56)


Extreme dark (LL)
(ave. brightness: 6.52)


Cityscape (NL)
(ave. brightness: 84.13)


Slight dark (LL)
(ave. brightness: 37.54)


Middle dark (LL)
(ave. brightness: 12.36)


Extreme dark (LL)
(ave. brightness: 2.40)

❖ Experiments

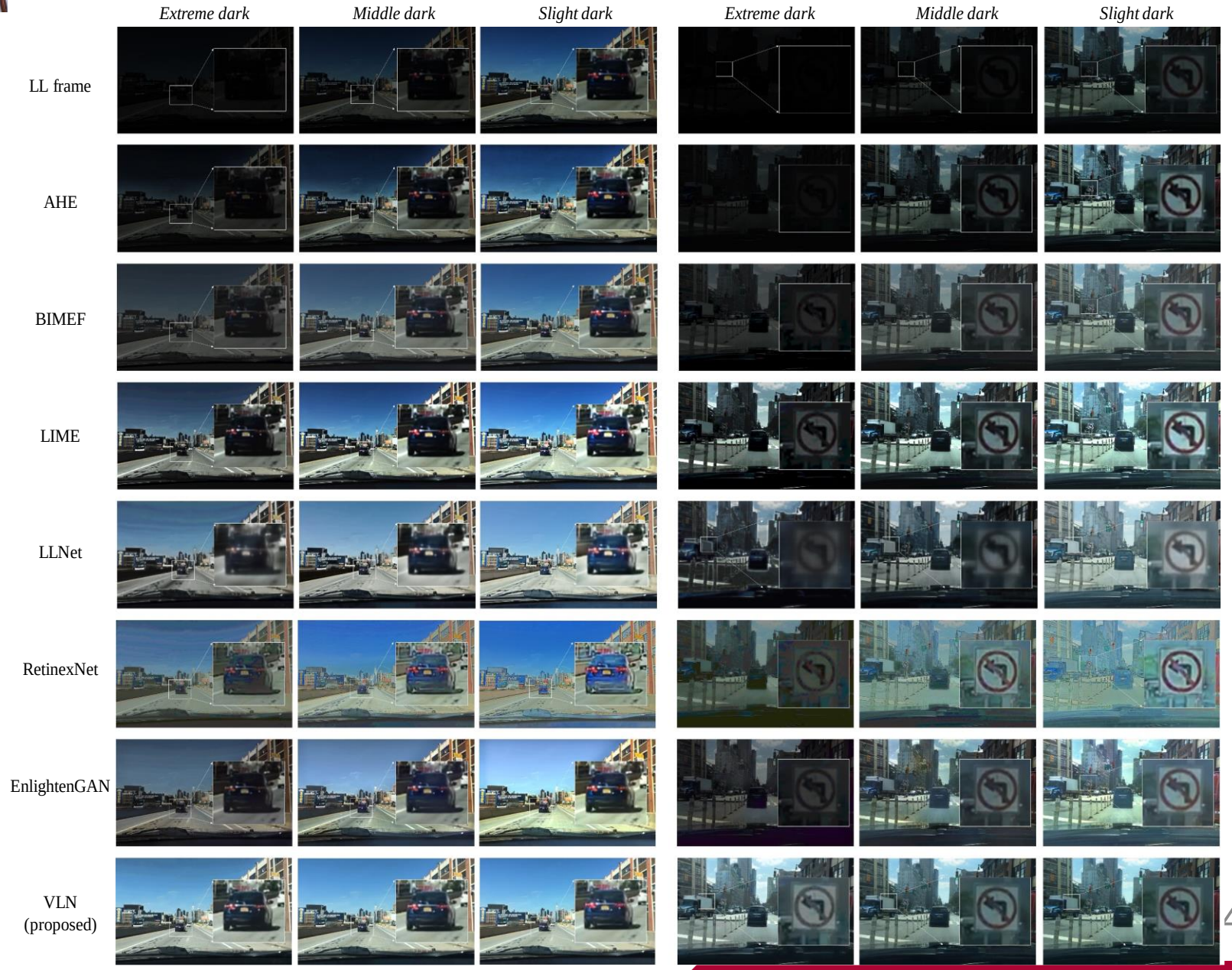
- **Dataset**: Berkeley Deep Drive (BDD) dataset is a widely used dataset that contains many HD video sequences of driving experience. We selected seven video sequences (**five sequences for training and two sequences for testing**) in ideal NL conditions, where each sequence lasts for **40 seconds** (about **1,200 frames**, **30 fps**).
- **Implementation Details**: We **randomly initialized** the weights of the VLN with normal distribution and biases equal to zero. We adopted the **Adam optimization** method with momentum of 0.9, weight decay of 0.0001. The **learning rate** was set as 0.0001. We randomly cropped 256*256 **patches** from the LL and NL frames as the training pair. For each iteration, the **mini-batch size** was set to 20, and the model was trained for **500 epochs**. All experiments were conducted through a PC with two NVIDIA GTX2080Ti GPUs.

❖ Experimental result - visualization

Methods	*Highway* Slight dark PSNR	*Highway* Slight dark SSIM	*Highway* Middle dark PSNR	*Highway* Middle dark SSIM	*Highway* Extreme dark PSNR	*Highway* Extreme dark SSIM	*Cityscape* Slight dark PSNR	*Cityscape* Slight dark SSIM	*Cityscape* Middle dark PSNR	*Cityscape* Middle dark SSIM	*Cityscape* Extreme dark PSNR	*Cityscape* Extreme dark SSIM	Average PSNR	Average SSIM
AHE [2]	13.578	0.812	9.332	0.482	6.581	0.161	17.465	0.800	11.089	0.398	8.184	0.111	11.038	0.461
BIMEF [4]	17.414	**0.920**	13.119	0.722	8.732	0.343	**21.869**	**0.964**	14.784	0.737	9.911	0.224	14.305	0.652
LIME [5]	**21.438**	**0.921**	17.102	**0.861**	12.599	0.628	18.391	0.862	**19.611**	**0.815**	**14.343**	0.446	**17.247**	**0.756**
LightenNet [6]	16.154	0.867	13.742	0.823	13.868	**0.759**	15.966	0.814	17.160	0.760	13.024	0.383	14.986	0.734
LLNet [8]	**20.451**	0.896	**21.932**	**0.860**	**15.409**	0.697	**19.380**	**0.869**	17.967	0.723	13.221	**0.505**	**18.060**	**0.758**
Retinex-Net [9]	15.147	0.810	17.237	0.851	**14.911**	**0.726**	12.996	0.652	18.387	0.788	**14.482**	**0.468**	15.527	0.716
EnlightenGAN [13]	17.888	0.847	**19.493**	0.850	12.205	0.599	17.840	0.814	**20.030**	**0.855**	12.083	0.417	16.590	0.730
VLN(proposed)	**29.750**	**0.974**	**29.822**	**0.964**	**27.120**	**0.917**	**23.759**	**0.922**	**23.816**	**0.907**	**25.759**	**0.863**	**26.671**	**0.924**
Average	18.977	0.881	17.722	0.802	13.928	0.604	18.458	0.837	17.855	0.748	13.876	0.427	16.803	0.716

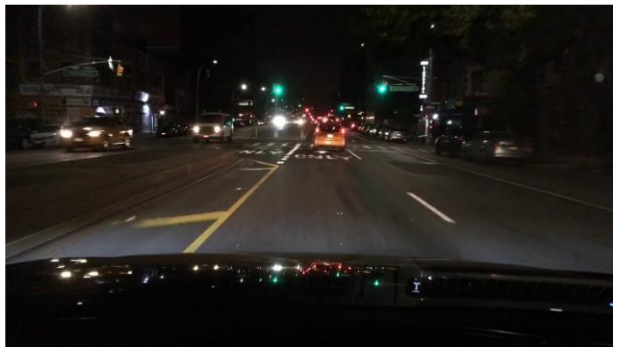
For the evaluation indices (i.e. PSNR and SSIM), a larger value means the estimation is closer to the ground truth. Video frames with darker scene usually faces more substantial content degradation. The performance in **extreme-dark** cases is worse than the slight-dark ones (e.g., the average SSIM is 0.604 at the extreme-dark *highway*, but it is 0.881 at the slight-dark *highway*). It can be seen from the table that AHE and BIMEF are more sensitive to the illumination changes. For example, the BIMEF achieves 0.964 SSIM in slight-dark cityscape, but the performance reduces significantly that it obtains 0.224 SSIM in the extreme-dark case. Learning-based approaches (LightenNet and Retinex-Net) were usually trained with huge datasets that contain scenes under different illuminations. It significantly improves the robustness of the methods. It can be seen from the table that the learning-based methods give a better estimation for the extreme cases compared with the conventional approaches (LightenNet achieves 0.759 SSIM at the extreme-dark *highway*). Our method is **more robust** and obtains **consistent results** in **different illumination conditions**.

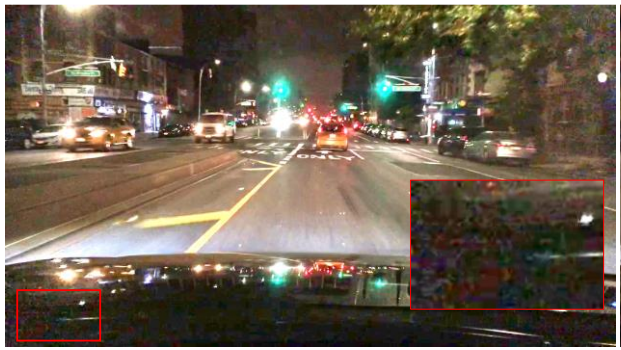
❖ Experimental resu

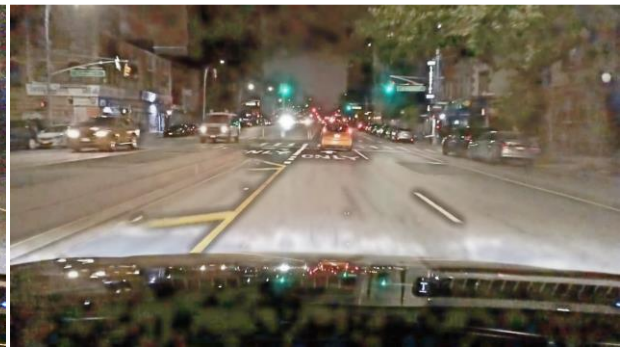


The proposed method has a **consistent enhancement** for **different levels of illuminations** (see the cars and signs of different illuminations), which suggests the robustness of the method.

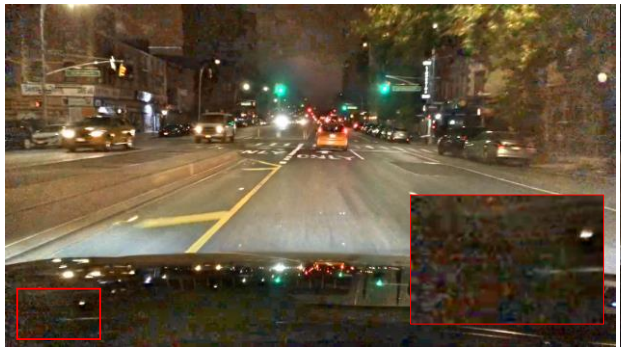
❖ Experimental result – Real dataset

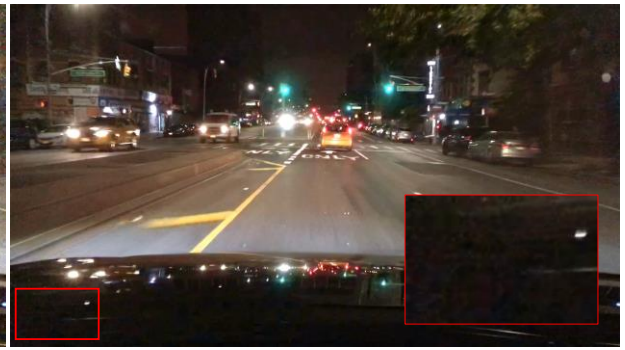

Original Video (night)


LIME


LLNet


Retinex-Net

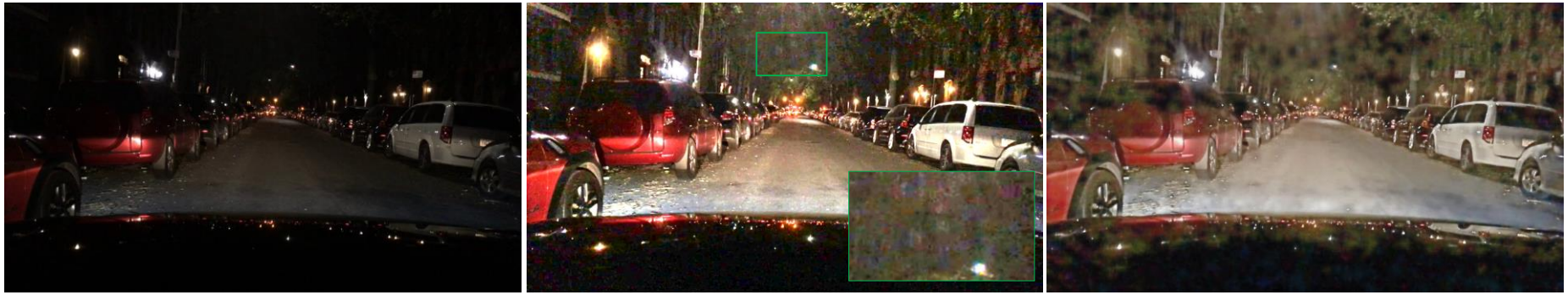

EnlightenGAN


VLN (proposed)

distorts the color information

The result of our method contains **less noise** and **reasonable brightness**, which provides a **better view** for driving.

❖ Experimental result – Real dataset

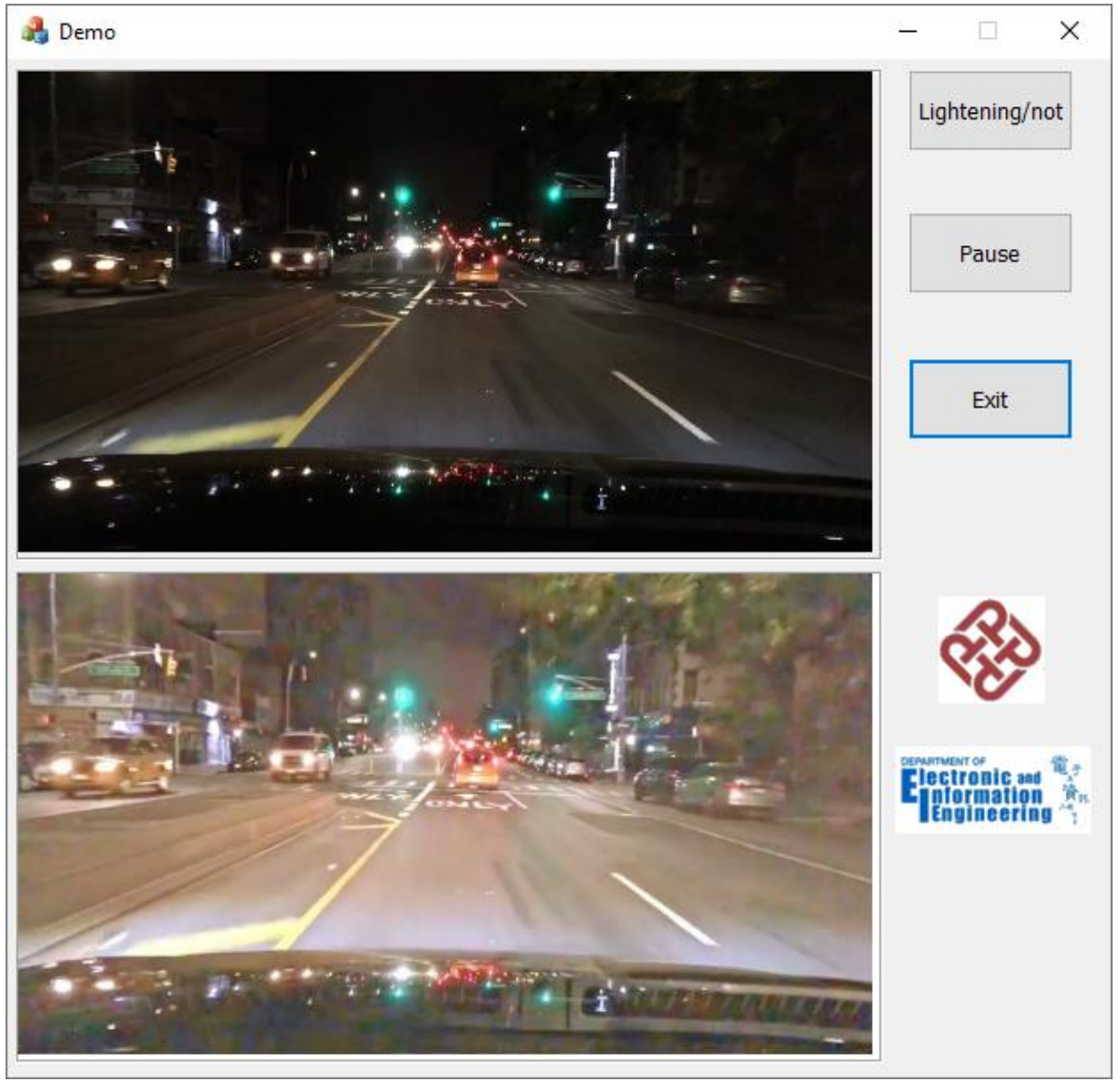


Original Video (night) | LIME | LLNet

Retinex-Net | EnlightenGAN | VLN (proposed)

The result of our method contains **less noise** and **reasonable brightness**, which provides a **better view** for driving.

❖ Experimental result – Video Demo



Demo

❖ Conclusion

- In this paper, we have introduced our proposed **Video Lightening Network (VLN)** for low-light **video enhancement**.
- We have proposed our **Lightening Back-Projection (LBP)** as a basic enhancing module that iteratively learns the mappings between low- and normal-light domain.
- To utilize the **temporal information** among adjacent frames, we have also proposed a novel Temporal Aggregation (TA) block, which investigates the spatial and temporal relations of a small region. Based on the **hierarchically multi-scale features**, the TAs can handle the **motion of different levels**.
- Extensive experimental results show that the proposed method outperforms others (both conventional and learning-based) in quantitative and qualitative aspects.

The end,
thank you!

Appendix

❖ Reference

Conventional approaches:

[1] Etta D Pisano, Shuquan Zong, Bradley M Hemminger, Marla DeLuca, R Eugene Johnston, Keith Muller, M Patricia Braeuning and Stephen M Pizer, "Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms," *Journal of digital imaging*, vol. 11, no. 4, pp. 193, 1998.

[17] Xiaojie Guo, Yu Li and Haibin Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE transactions on image processing (TIP)*, vol. 26, no. 2, pp. 982-993, 2016.

Learning-based approaches:

[11] Chongyi Li, Jichang Guo, Fatih Porikli and Yanwei Pang, "Lightennet: A convolutional neural network for weakly illuminated image enhancement," *Pattern recognition letters*, vol. 104, pp. 15-22, 2018.

[10] Chen Wei, Wenjing Wang, Wenhan Yang and Jiaying Liu, "Deep retinex decomposition for low-light enhancement," arXiv preprint arXiv:1808.04560, 2018.

[12] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou and Zhangyang Wang, "EnlightenGAN: Deep Light Enhancement without Paired Supervision," *arXiv preprint arXiv:1906.06972*, 2019