Multinex: Lightweight Low-light Image Enhancement via Multi-prior Retinex

Supplementary Material

Content Summary: Section A explains the linear reconstruction analysis (LRA), which is used to visualize feature maps. Secion B includes further studies on additional analytic representation priors used by the luminance and reflectance guidance stacks, and explains the reason for choosing the used priors by Multinex. Section C provides additional details regarding to Multinex to complement the main paper, including its basic building blocks, the hybrid loss function, as well as the implementation and evaluation considerations. Section D presents extra experimental results and discussions on additional ablation studies of the adopted losses and model components, while Section E presents expanded qualitative visualizations.

A. Linear Reconstruction Analysis

Given an input feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times K}$ and a target signal $\mathbf{T} \in \mathbb{R}^{H\times W\times C}$ defined according to the analysis goal, LRA examines how much information in $\mathbf{T}$ is retained by $\mathbf{X}$. The core idea is to assess *what portion of the target signal can be reconstructed through a linear combination of the principal components of the feature map*.

Pre-processing. We pre-process the data before performing LRA. We firstly flatten the feature map into $\mathbf{X}_f \in \mathbb{R}^{N\times K}$, where $N = HW$. Each row of $\mathbf{X}_f$ corresponds to a pixel and each column to a feature type (channel). We then center each column by subtracting its mean, obtaining $\mathbf{X}_c = \mathbf{X}_f - \mu_{\mathbf{X}}$ with $\mu_{\mathbf{X}} = \frac{1}{N}\mathbf{1}^\top\mathbf{X}_f$, where $\mathbf{1}$ is a length-N column vector with all elements equal to 1. Similarly, we flatten the target signal into $\mathbf{T}_f \in \mathbb{R}^{N\times C}$, where C depends on the target definition (e.g., C=3 for RGB and C=1 for luminance), and center its columns by subtracting the mean vector $\mu_{\mathbf{T}} = \frac{1}{N}\mathbf{1}^\top\mathbf{T}_f$ to obtain $\mathbf{T}_c = \mathbf{T}_f - \mu_{\mathbf{T}}$.

LRA. It first applies principal component analysis (PCA), linearly projecting $\mathbf{X}_c$ onto its top D principal components through the orthogonal projection matrix $\mathbf{P}_{\text{PCA}} \in \mathbb{R}^{K\times D}$. This yields the reduced feature matrix $\mathbf{Z} \in \mathbb{R}^{N\times D}$,

$$\mathbf{Z} = \mathbf{X}_c\mathbf{P}_{\text{PCA}}, \tag{19}$$

which captures the dominant variation across the feature channels. PCA acts as a compact linear representation that removes redundancy in $\mathbf{X}$ while preserving its major structure. We then fit a linear model to map $\mathbf{Z}$ to the centered target $\mathbf{T}_c$ through ridge regression, resulting in the following mapped target:

$$\hat{\mathbf{T}}_c = \mathbf{Z}\mathbf{W}, \tag{20}$$

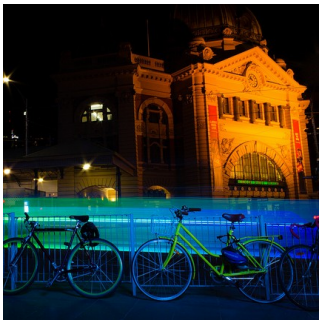

Figure 5. The example image used by Sec. B.

where $\mathbf{W} \in \mathbb{R}^{D\times C}$ is computed by

$$\mathbf{W} = (\mathbf{Z}^\top\mathbf{Z} + \lambda\mathbf{I}_D)^{-1}\mathbf{Z}^\top\mathbf{T}_c, \tag{21}$$

with $\mathbf{I}_D$ being the identity matrix of size D and $\lambda > 0$ a small ridge regularization parameter. Using reshape$(\cdot)$ to restore a flattened signal to spatial dimensions $H \times W \times C$, the final reconstructed target is given by

$$\hat{\mathbf{T}}(\mathbf{X}) = \text{reshape}(\mathbf{Z}\mathbf{W} + \mu_{\mathbf{T}}). \tag{22}$$

Feature Visualization. By setting $\mathbf{T}$ as the low-light image input, how different the reconstructed image $\hat{\mathbf{T}}(\mathbf{X})$ is from $\mathbf{T}$ provides a simple and interpretable measure of the information content carried by the feature map. Therefore, visualizing the reconstructed image $\hat{\mathbf{T}}(\mathbf{X})$ helps establish an understanding of the physical role of the feature map $\mathbf{X}$ for enhancing $\mathbf{T}$.

B. More Studies On Representation Prior

We have proposed the luminance and reflectance guidance stacks in Eqs. (4) and (9), serving as the representation prior. To arrive at this design, we initially considered a larger pool of analytic luminance and chrominance descriptors for computing the feature maps, including multiple linear and nonlinear ones. Strong non-linearity typically has the potential to encode richer or more distinctive relationships. Therefore, our selection strategy prioritizes features that are (1) inherently non-linear, and/or (2) less correlated with the other features. The goal is to enable the finally selected guidance stacks to offer the widest possible range of complementary lenses for the model to view the image. For each guidance stack, we explain below the additional feature candidates considered during model design. We then analyze and visualize the contribution of both the selected and non-selected descriptors by LRA, to validate our

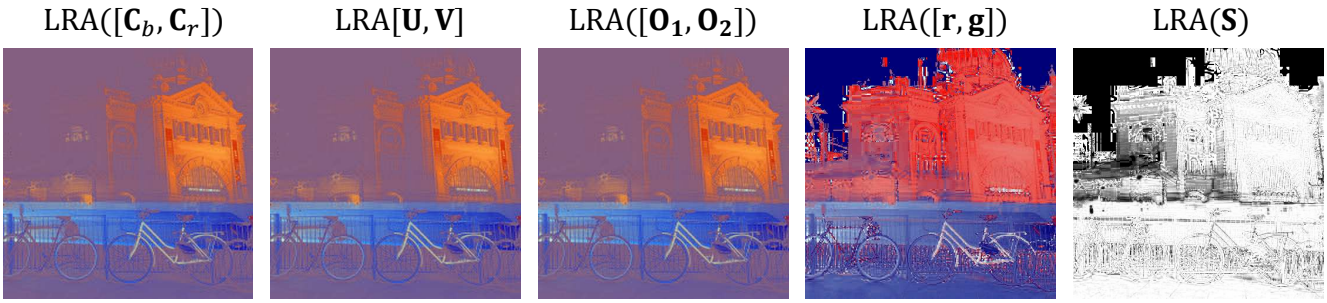

Figure 6. LRA visualization of chrominance candidates $\{[\mathbf{C}_b, \mathbf{C}_r], [\mathbf{U}, \mathbf{V}], [\mathbf{O}_1, \mathbf{O}_2], [\mathbf{r}, \mathbf{g}], \mathbf{S}\}$.

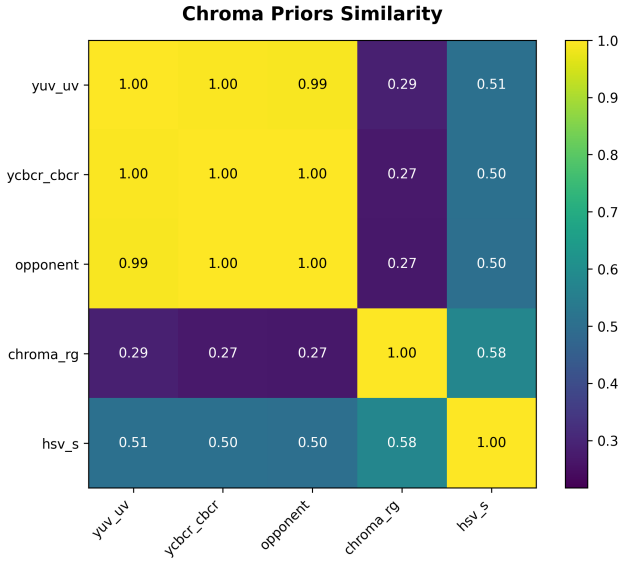

Figure 7. Similarity heatmaps between all candidate grouped reflectance-guidance components.

representation-prior design. The image shown in Fig. 5 is used throughout the analysis.

B.1. Reflectance Guidance Stack

B.1.1. Additional Chrominance Candidates

For the reflectance guidance stack, besides $\{\mathbf{C}_b, \mathbf{C}_r, \mathbf{r}, \mathbf{g}, \mathbf{S}\}$ in Eq. (9), we considered two further pairs of chroma descriptors. These include the YUV chroma pair $[\mathbf{U}, \mathbf{V}]$, computed as

$$\mathbf{U} = -0.14713\mathbf{I}_R - 0.28886\mathbf{I}_G + 0.43600\mathbf{I}_B, \quad (23)$$

$$\mathbf{V} = \ \ 0.61500\mathbf{I}_R - 0.51499\mathbf{I}_G - 0.10001\mathbf{I}_B, \quad (24)$$

and the pair of the first two opponent channels $[\mathbf{O}_1, \mathbf{O}_2]$, as

$$\mathbf{O}_1 = \frac{1}{\sqrt{2}}(\mathbf{I}_R - \mathbf{I}_G), \quad (25)$$

$$\mathbf{O}_2 = \frac{1}{\sqrt{6}}(\mathbf{I}_R + \mathbf{I}_G - 2\mathbf{I}_B). \quad (26)$$

The pair $[\mathbf{U}, \mathbf{V}]$ encodes blue-yellow and red-cyan differences in a luminance-decoupled fashion, while $[\mathbf{O}_1, \mathbf{O}_2]$ is derived from the classical opponent color theory, spanning red-green and blue-yellow axes in a normalized manner. The above, together with Eq. (9), forms an initial candidate pool of chroma descriptors, i.e.,

$$\hat{\mathcal{S}}_{\mathcal{C}}(\mathbf{I}) = \{\mathbf{C}_b, \mathbf{C}_r, \mathbf{U}, \mathbf{V}, \mathbf{O}_1, \mathbf{O}_2, \mathbf{r}, \mathbf{g}, \mathbf{S}\}. \quad (27)$$

B.1.2. Comparative Visual Analysis

The three groups of chroma descriptors $[\mathbf{C}_b, \mathbf{C}_r]$, $[\mathbf{U}, \mathbf{V}]$ and $[\mathbf{O}_1, \mathbf{O}_2]$ are linear combinations of the underlying RGB channels, thus linearly correlated to each other. So we include only one group to the final chrominance stack. Below we visually validate our preference of using $[\mathbf{C}_b, \mathbf{C}_r]$ over $[\mathbf{U}, \mathbf{V}]$ and $[\mathbf{O}_1, \mathbf{O}_2]$, to complement the nonlinear descriptors $[\mathbf{r}, \mathbf{g}]$ and $\mathbf{S}$.

We produce the LRA visualization results with $D = 3$. Fig. 6 visualizes how well the chrominance candidates $[\mathbf{C}_b, \mathbf{C}_r]$, $[\mathbf{U}, \mathbf{V}]$, $[\mathbf{O}_1, \mathbf{O}_2]$, $[\mathbf{r}, \mathbf{g}]$, and $\mathbf{S}$ can reconstruct the RGB content of the low-light image $\mathbf{I}$. To quantify the redundancy among the candidate chroma priors, we compute their pairwise Pearson correlation. Because the descriptors vary in channel depth, we first compute the pixel-wise L_2 norm across the channel dimension for each prior. This reduces every candidate to a single spatial magnitude map representing its overall activation. We then flatten these magnitude maps and compute the Pearson correlation coefficient between them, yielding a unified similarity heatmap.

Fig. 6 shows that the three linear groups of $[\mathbf{C}_b, \mathbf{C}_r]$, $[\mathbf{U}, \mathbf{V}]$, and $[\mathbf{O}_1, \mathbf{O}_2]$ are able to achieve reconstructions with near-identical structure and color fidelity, which is consistent with their analytic formulations. The group analysis in Fig. 7 (**left**) confirms that $[\mathbf{C}_b, \mathbf{C}_r]$, $[\mathbf{U}, \mathbf{V}]$, and $[\mathbf{O}_1, \mathbf{O}_2]$ are highly correlated among themselves. However, the single-component heatmap in Fig. 7 (**right**) reveals that $[\mathbf{U}, \mathbf{V}]$ and $[\mathbf{C}_b, \mathbf{C}_r]$ are less correlated with the nonlinear descriptors $\mathbf{r}, \mathbf{g}, \mathbf{S}$ as compared to $[\mathbf{O}_1, \mathbf{O}_2]$. This suggests that $[\mathbf{U}, \mathbf{V}]$ and $[\mathbf{C}_b, \mathbf{C}_r]$ provide more complementary information when being combined with the normalized chromaticities and saturation. Between $[\mathbf{U}, \mathbf{V}]$ and $[\mathbf{C}_b, \mathbf{C}_r]$, we empirically observe that $[\mathbf{C}_b, \mathbf{C}_r]$ yields

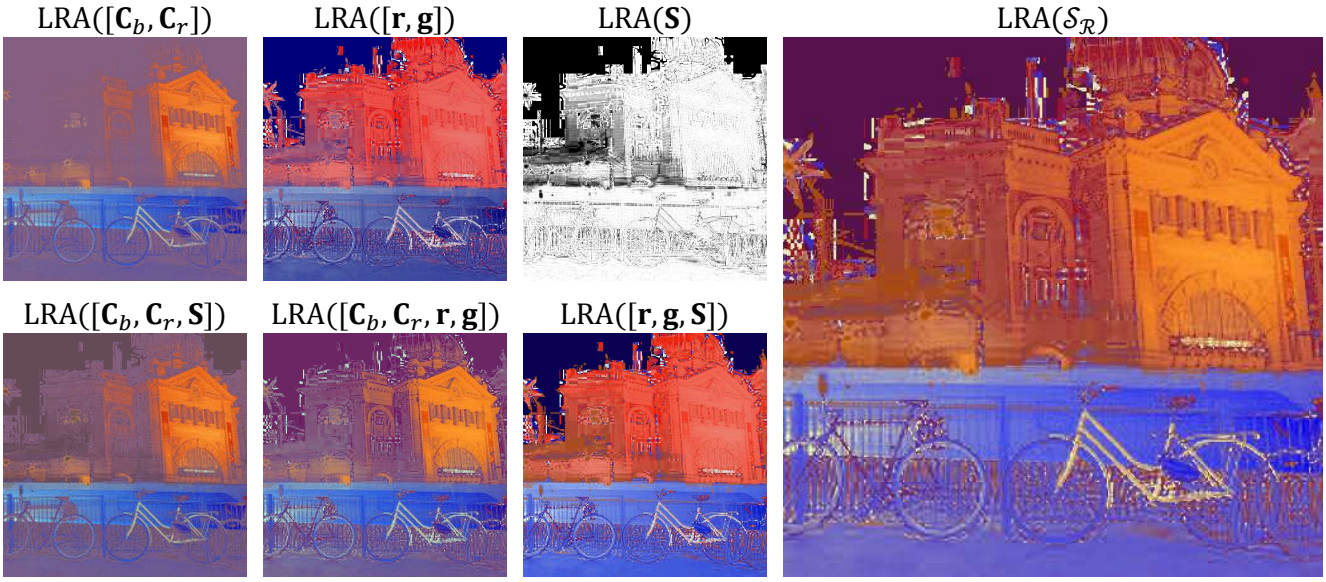

Figure 8. LRA visualization of combinations of the chosen feature maps for the reflectance guidance stack $\mathcal{S}_\mathcal{R}$.

slightly better discrimination of yellow-orange hues in the LRA reconstruction, which are particularly relevant to low-light scenes (e.g., street lights and indoor tungsten illumination). Based on these, for the reflectance guidance stack we retain $\mathbf{C}_b$ and $\mathbf{C}_r$, together with the non-linear maps $\mathbf{r}, \mathbf{g}$ and $\mathbf{S}$, leading to the final design in Eq. (9).

To visualize the effectiveness of $\mathcal{S}_\mathcal{R}$ as guidance for reflectance modeling, we use LRA on prior combinations of the proposed stack. As Fig. 8 shows, that $[\mathbf{C}_b, \mathbf{C}_r]$ successfully recovers general color of the input, $[\mathbf{r}, \mathbf{g}]$ finds regions that are uniform in reflectance, and $\mathbf{S}$ provides greater structural boundary and color control in dark areas.

B.2. Luminance Guidance Stack

B.2.1. Additional Luminance Candidates

Besides $\{\mathbf{Y}_{\text{Rec.709}}, \mathbf{Y}_{\text{vmax}}, \mathbf{Y}_{\text{lightness}}, \mathbf{Y}_{\text{L}_2}\}$ in Eq. (4), we also considered the following luminance candidates:

$$\mathbf{Y}_{\text{mean}} = \tfrac{1}{3}\left(\mathbf{I}_R + \mathbf{I}_G + \mathbf{I}_B\right), \tag{28}$$

$$\mathbf{Y}_{\text{YCgCo}} = 0.25\mathbf{I}_R + 0.50\mathbf{I}_G + 0.25\mathbf{I}_B. \tag{29}$$

Here, $\mathbf{Y}_{\text{mean}}$ is the simple arithmetic mean intensity, while $\mathbf{Y}_{\text{YCgCo}}$ is the luminance component of the YCgCo color space, which emphasizes more the green channel through a physically motivated transform. Together with Eq. (4), they form our initial candidate pool of six luminance descriptors:

$$\hat{\mathcal{S}}_\mathcal{L}(\mathbf{I}) = \{\mathbf{Y}_{\text{Rec.709}}, \mathbf{Y}_{\text{mean}}, \mathbf{Y}_{\text{YCgCo}}, \mathbf{Y}_{\text{vmax}}, \mathbf{Y}_{\text{lightness}}, \mathbf{Y}_{\text{L}_2}\}. \tag{30}$$

B.2.2. Descriptor Importance Analysis

To examine the importance of a luminance descriptor in capturing essential information, we introduce a leave-one-out approach that measures the loss of expressivity upon removing each descriptor from the full six-prior stack. As a result, a greater loss indicates a more important descriptor. In particular, we measure the loss across two orthogonal metrics computed pixel wise.

First, we apply a simple edge detection method of maximum gradient operator to highlight the direction along which the intensity changes the most for each pixel, e.g., by using a sober filter. Denote the output gradient map for each luminance feature map by $\mathbf{G}_c \in \mathbb{R}^{H\times W}$ for $c \in \{1, 2, \ldots, K_\mathcal{L}\}$. We focus on the strongest geometric boundary across all the prior candidates, which is computed by applying the max operator element-wise over the gradient maps, as

$$\mathbf{G}_{K_\mathcal{L}} = \max_{c=1}^{K_\mathcal{L}} \mathbf{G}_c. \tag{31}$$

The resulting gradient map extracts structural information from the stack.

Second, we compute the orthogonal energy for each pixel, which is characterized by a $K_\mathcal{L}$-dimensional vector corresponding to the $K_\mathcal{L}$ candidate maps, measuring the global shading expressivity. We apply PCA to the set of $N = HW$ pixel vectors, and compute the energy using the non-principal components. Denote a centered pixel vector by $\mathbf{S}_{ij} \in \mathbb{R}^{K_\mathcal{L}}$, the resulting energy map of all pixels by $\mathbf{E}_{K_\mathcal{L}} \in \mathbb{R}^{H\times W}$, and the k-th principal direction by $\mathbf{p}_k \in \mathbb{R}^{K_\mathcal{L}}$. Each element of the energy map $\mathbf{E}_{K_\mathcal{L}}$ is then computed by

$$e_{ij} = \sqrt{\sum_{k=2}^{K_\mathcal{L}} \left(\mathbf{S}_{ij}^T \mathbf{p}_k\right)^2}, \tag{32}$$

summing over the non-dominant principal directions.

To quantify the unique contribution of each candidate, we remove each prior from the full stack $\mathcal{S}_\mathcal{L}$ and measure

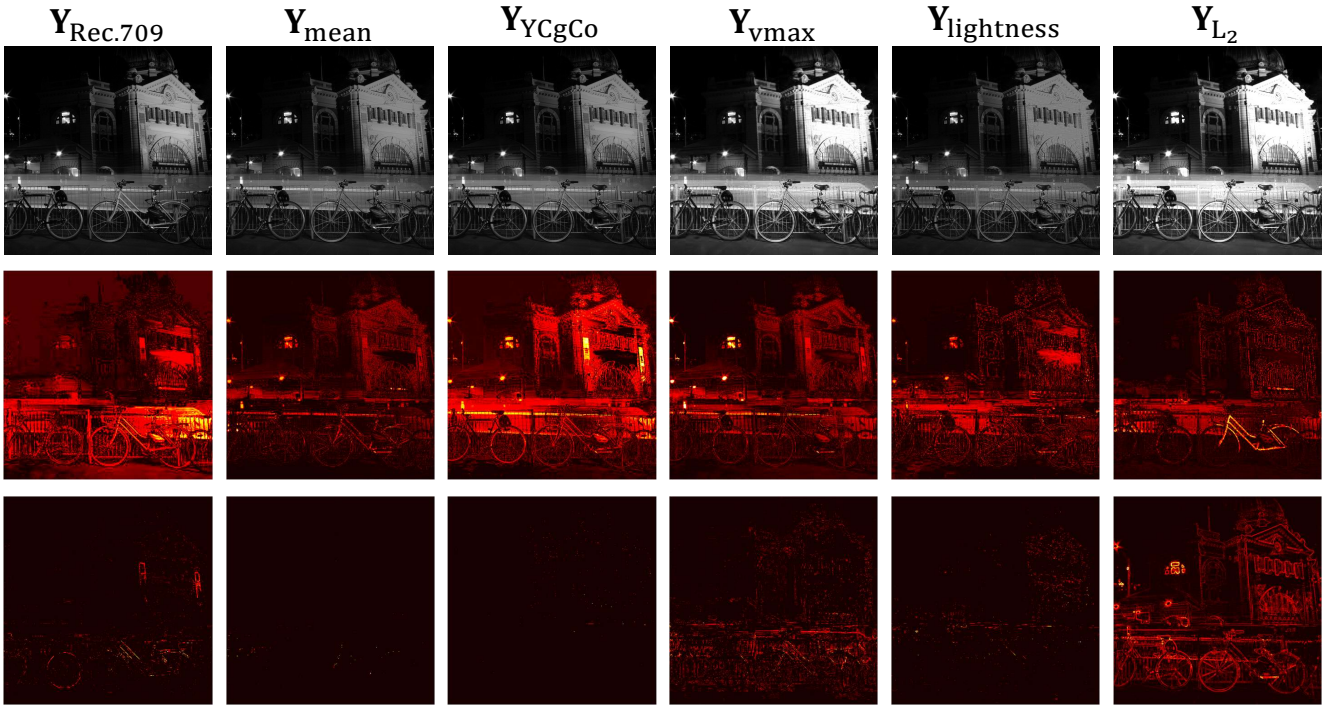

Figure 9. **Top Row:** Candidate illumination feature descriptors $\{\mathbf{Y}_{\text{Rec.709}}, \mathbf{Y}_{\text{mean}}, \mathbf{Y}_{\text{YCgCo}}, \mathbf{Y}_{\text{vmax}}, \mathbf{Y}_{\text{lightness}}, \mathbf{Y}_{\text{L}_2}\}$. **Middle Row:** $\boldsymbol{\Delta}_E(c)$ maps of the same candidate descriptors, where $c \in \{1, 2, \ldots, 6\}$. **Bottom row:** $\boldsymbol{\Delta}_G(c)$ maps of the same candidate descriptors.

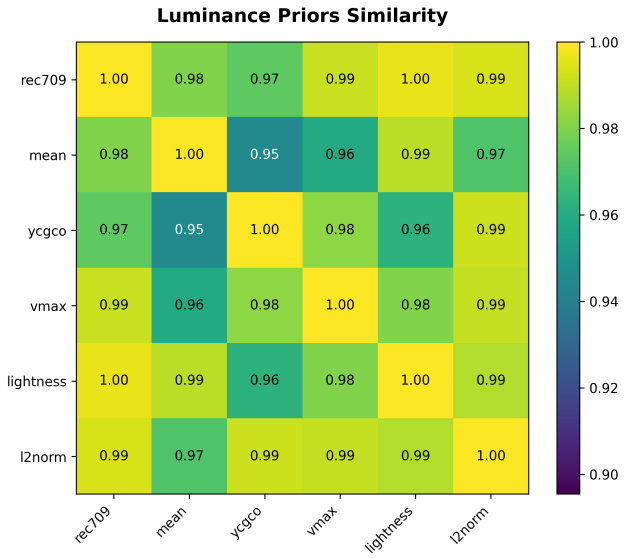

Figure 10. Similarity heatmap between all candidate individual illumination-guidance components.

the resulting reduction in the gradient and energy maps. This process is expressed as below:

$$\boldsymbol{\Delta}_G(c) = \max\left(0, \mathbf{E}_{K_{\mathcal{L}}} - \mathbf{E}_{K_{\mathcal{L}}\setminus\{c\}}\right), \tag{33}$$

$$\boldsymbol{\Delta}_E(c) = \max\left(0, \mathbf{G}_{K_{\mathcal{L}}} - \mathbf{E}_{K_{\mathcal{L}}\setminus\{c\}}\right). \tag{34}$$

The resulting maps are referred to as the energy and gradient importance maps, respectively, and are used for visualization and analysis.

B.2.3. Comparative Visual Analysis

The non-linear luminance descriptors of $\mathbf{Y}_{\text{vmax}}$, $\mathbf{Y}_{\text{lightness}}$, and $\mathbf{Y}_{\text{L}_2}$ provide distinct physical and perceptual interpretations, such as maximum channel response, HSL lightness, and RGB energy, to complement the linear descriptor $\mathbf{Y}_{\text{Rec.709}}$. We quantitatively and visually validate our selection of the final four-component stack below, specifically justifying the choice of $\mathbf{Y}_{\text{Rec.709}}$ over the alternative linear descriptors $\mathbf{Y}_{\text{mean}}$ and $\mathbf{Y}_{\text{YCgCo}}$.

The top row of Fig. 9 visualizes each candidate illumination map as it is individually. Furthermore, we compute the correlation matrix over the six maps, visualized in Fig. 10. As expected, the three linear descriptors of $\mathbf{Y}_{\text{Rec.709}}$, $\mathbf{Y}_{\text{mean}}$, $\mathbf{Y}_{\text{YCgCo}}$ are highly correlated with one another. Since they represent correlated transformations of the linear RGB space, we only adopt one linear descriptor for the guidance stack.

The middle and bottom rows of Fig. 9 visualize energy and gradient importance maps for each candidate descriptor. We also compute scalar energy and gradient importance scores each computed by averaging the corresponding importance maps over all pixels for each candidate descriptor. The two resulting scores are denoted by ΔE_c and ΔG_c. We report the scores and the descriptor rankings in Tab. 5, where higher scores indicate greater expressivity

Table 5. Importance ranking of candidate luminance priors. Higher values indicate greater unique contribution to the stack.

Prior	$\Delta_E(c) \uparrow$ / Rank	$\Delta_G(c) \uparrow$ / Rank	Avg. Rank
$\mathbf{Y}_{\text{Rec.709}}$	**0.0132** / 1	0.0007 / 3	2.0
$\mathbf{Y}_{\text{vmax}}$	0.0107 / 2	0.0019 / 2	2.0
$\mathbf{Y}_{\text{L}_2}$	0.0012 / 6	**0.0194** / 1	3.5
$\mathbf{Y}_{\text{lightness}}$	0.0029 / 4	0.0002 / 4	4.0
$\mathbf{Y}_{\text{YCgCo}}$	0.0038 / 3	0.0000 / 6	4.5
$\mathbf{Y}_{\text{mean}}$	0.0017 / 5	0.0000 / 5	5.0

loss, thus higher importance. The results demonstrate distinct roles among the candidate descriptors, justifying the necessity of a multi-prior stack. Specifically, $\mathbf{Y}_{\text{L}_2}$ dominates in gradient importance, proving critical for preserving high-frequency structural edges. But it contributes the least to global shading variance. Conversely, $\mathbf{Y}_{\text{vmax}}$ captures the most global illumination variance, excelling at distinguishing bright specularities from diffuse regions, but provides weaker structural guidance. Among the highly correlated linear candidates, $\mathbf{Y}_{\text{Rec.709}}$ significantly outperforms both $\mathbf{Y}_{\text{mean}}$ and $\mathbf{Y}_{\text{YCgCo}}$ across both metrics. Consequently, we retain $\mathbf{Y}_{\text{Rec.709}}$ as our sole linear luminance and discard the others, finalizing our four-dimensional illumination guidance stack as in Eq. 4.

C. Additional Details on Multinex

C.1. Basic Neural Building Blocks

We use a few basic neural building blocks in our lightweight fusion module. Taking as input a tensor $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, these include operations such as: depth-wise convolution $\text{DWConv} : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{H\times W\times C}$ and depth-wise separable convolution $\text{DSConv} : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{H\times W\times C}$. Both convolutions are designed to reduce the number of filter weights (i.e. parameters) to learn, which fits our goal of constructing a lightweight network. Additionally, two typical activation functions are employed, including the sigmoid activation $\sigma : \mathbb{R}^{H\times W\times C} \to [0,1]^{H\times W\times C}$ and ReLU activation $\sigma_{\text{ReLU}} : \mathbb{R}^{H\times W\times C} \to [0,+\infty)^{H\times W\times C}$.

Another used building block is the recent multi-stage squeeze & excite fusion (MSEF) module, i.e., a lightweight architecture with its effectiveness demonstrated particularly for LLIE [2], denoted by $\text{MSEF} : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{H\times W\times C}$. It generalizes the squeeze-and-excitation (SE) mechanism [13]. Given the input feature $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, it computes an output feature of the same size $\mathbf{Y} \in \mathbb{R}^{H\times W\times C}$:

$$\mathbf{Y} = \mathbf{X} + \text{DWConv} \circ \text{LN}(\mathbf{X}) \odot \mathbf{Z}. \tag{35}$$

where $f \circ g(x) = f(g(x))$ denotes the composition of two functions. In order to obtain $\mathbf{Z} \in \mathbb{R}^{H\times W\times C}$, a set of adaptive channel weights $\mathbf{w} = [w_1, w_2, \ldots, w_C] \in \mathbb{R}^C$ are computed by a two-layer excitation bottleneck, i.e.,

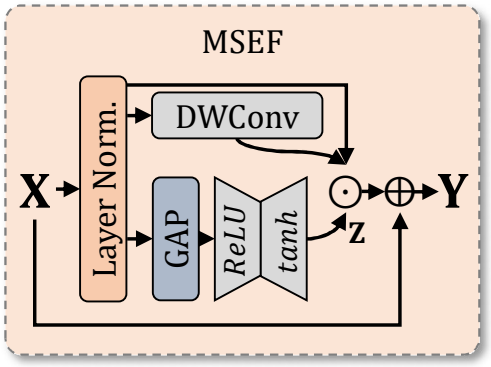

Figure 11. MSEF module architecture.

$$\mathbf{w} = \sigma_{\text{tanh}}\big(\mathbf{W}_2\sigma_{\text{ReLU}}(\mathbf{W}_1\text{GAP} \circ \text{LN}(\mathbf{X}))\big), \tag{36}$$

where the linear projection matrices $\mathbf{W}_1 \in \mathbb{R}^{d\times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C\times d}$ $(d < C)$ form a feature compression-expansion pair. Each adaptive weight w_i is then used to recalibrate the normalized features from the corresponding channel, denoted as $\text{LN}_i(\mathbf{X})$, by multiplication, i.e.,

$$\mathbf{Z}_i = w_i\text{LN}_i(\mathbf{X}), \text{ for } i = 1, 2, \ldots, C. \tag{37}$$

Such a design allows the MSEF module to capture both local fine texture (e.g., through convolution) and global semantics (e.g., through global pooling) in the input, with negligible computational cost.

C.2. Multinex Loss Function

To train the illumination and reflectance networks $f_{\mathcal{L}}$ and $f_{\mathcal{R}}$ derived from the fusion module, we adopt a hybrid loss that balances pixel-level fidelity, structural consistency, and perceptual quality, expressed as

$$\mathcal{L} = \lambda_{\text{MSE}}\mathcal{L}_{\text{MSE}} + \lambda_{\text{MS-SSIM}}\mathcal{L}_{\text{MS-SSIM}} + \lambda_{\text{Perc}}\mathcal{L}_{\text{Perc}}. \tag{38}$$

We adopted the existing hyper-parameter setting [3] of $\lambda_{\text{MSE}} = 1.0$, $\lambda_{\text{MS-SSIM}} = 0.2$, and $\lambda_{\text{Perc}} = 0.01$ in our implementation, weighting the contribution of each loss component. For the sake of convenience, we explain each individual loss term computed over one image pair $(\mathbf{I}_{\text{GT}}, \hat{\mathbf{I}})$, containing the predicted enhanced image $\hat{\mathbf{I}}$ and its corresponding ground-truth well-lit image $\mathbf{I}_{\text{GT}}$.

MSE Loss. The pixel-wise mean squared error (MSE) encourages numerically accurate reconstruction of the enhanced image, defined as

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\left\|\hat{\mathbf{I}} - \mathbf{I}_{\text{GT}}\right\|_2^2, \tag{39}$$

where N is the total number of pixels. It stabilizes training and provides strong gradients for correcting global brightness and color deviations.

MS-SSIM Loss. To better preserve structural consistency under varying illumination, we employ a multi-scale structural similarity (MS-SSIM) [41] loss, defined as

$$\mathcal{L}_{\text{MS-SSIM}} = 1 - \prod_{m=1}^{M} \text{SSIM}^{(m)}\big(\hat{\mathbf{I}}, \mathbf{I}_{\text{GT}}\big), \tag{40}$$

where $\text{SSIM}^{(m)}(\cdot,\cdot)$ is a structural similarity function computed at scale m, and M is the set of all scales. It captures contrast, luminance, and texture consistency between the prediction and ground truth at multiple spatial resolutions, which is particularly beneficial for low-light scenes where edges and fine details are difficult to recover.

Perceptual Loss. Finally, to enhance perceptual realism and encourage faithful recovery of semantic structures, we incorporate a perceptual loss computed in a deep feature space. Letting $\Phi(\cdot)$ denote a fixed VGG-based feature extractor, the perceptual term is defined as

$$\mathcal{L}_{\text{Perc}} = \frac{1}{N}\big\|\Phi(\hat{\mathbf{I}}) - \Phi(\mathbf{I}_{\text{GT}})\big\|_2^2. \tag{41}$$

It is beneficial to compare the feature activations of the prediction with the ground truth. This helps maintain natural textures and suppress color artifacts, two common failure points in low-light enhancement.

Together, the three losses form a complementary objective that encourages numerical accuracy, structural fidelity, and perceptual quality. In practice, we find that the hybrid loss significantly improves the visual consistency and robustness of Multinex compared to using a single loss component alone. We perform ablation study on loss terms in Section D.1 and refer to results in Tab. 6.

C.3. Multinex Implementation

All experiments are conducted using the PyTorch framework [30]. Training data are augmented via random cropping, horizontal and vertical flipping, and random rotation. Model parameters are optimized with the Adam optimizer [19], using a cosine annealing learning-rate schedule [24] that decays from 2×10^{-4} to 1×10^{-6}. Multinex is trained from scratch for 150K iterations with a batch size of 8 and patch size of 256×256, using the designated training splits of each dataset.

C.4. Discussion on GT-Mean

GT-Mean is a post-processing step used by some LLIE works [12, 40, 45, 52] when evaluating their approaches on small paired datasets, e.g., LOL-v1. We do not perform GT-Mean in our assessment, as it removes the global brightness errors from the evaluation that actually is a core part of the LLIE performance. In more detail, GT-Mean rescales the output of the enhanced image $\hat{\mathbf{I}}$ to match the mean grayscale of the ground truth image $\mathbf{I}_{\text{GT}}$, before computing PSNR/SSIM. The rescaling is defined as $\hat{\mathbf{I}}_{\text{GT-Mean}} = q\hat{\mathbf{I}}$ with $q = \frac{\text{mean}(\mathbf{I}_{\text{GT}})}{\text{mean}(\hat{\mathbf{I}})}$, and such a processing removes the global brightness errors. We do not perform this rescaling, as brightness correction is a core part of LLIE. Enforcing matched luminance defeats the purpose of measuring enhancement accuracy. Consequently, GT-Mean can inflate the performance metrics, for instance, by several dB of PSNR on LOL-v1, e.g., CIDNet rises from 23.81dB to 28.20dB, while RetinexFormer from 25.15dB to 27.14dB. We report all the results *without* GT-Mean to ensure that our evaluation reflects the true quality of LLIE enhancement.

$\mathcal{L}_{\text{MSE}}$	$\mathcal{L}_{\text{MS-SSIM}}$	$\mathcal{L}_{\text{Perc}}$	PSNR↑	SSIM↑
✓	✗	✗	22.31	0.815
✗	✓	✗	19.04	0.820
✗	✗	✓	18.80	0.793
✓	✓	✗	22.43	0.830
✓	✗	✓	22.80	0.821
✗	✓	✓	19.74	0.838
✓	✓	✓	**23.19**	**0.843**

Table 6. Ablation of loss functions.

D. Additional Ablation Studies

To complement Sec. 4.3, we conduct additional ablation experiments to further validate the Multinex design. Unless otherwise stated, we use the same dataset and configuration as in Sec. 4.3, and report the performance in terms of both the PSNR and SSIM metrics.

D.1. On Loss Function

Tab. 6 evaluates contributions of the three individual loss terms and their combinations to the final enhancement quality. The MSE loss on its own already provides a strong baseline, reaching around 22dB PSNR with a moderate SSIM of about 0.82. The MS-SSIM loss alone preserves quite well the structural similarity by offering a slightly higher SSIM, but it yields a noticeably lower PSNR, i.e., around 19dB. This indicates that MS-SSIM emphasizes contrast consistency over pixel-wise accuracy. The perceptual loss alone performs the worst in terms of distortion. This is expected since it optimizes high-level features rather than low-level fidelity. Pairwise combinations are able to improve the enhancement performance. For instance, the combination of MSE and MS-SSIM slightly strengthens the structural consistency, while MSE and Perceptual together increase brightness and color realism. A full combination of all the three losses yields the best results, reaching roughly 23dB PSNR with an SSIM of around 0.84. This suggests that the

$\mathcal{S}_\mathcal{L}$ Components				#Ch	PSNR↑	SSIM↑
$\mathbf{Y}_{\text{Rec.709}}$	$\mathbf{Y}_{\text{vmax}}$	$\mathbf{Y}_{\text{lightness}}$	$\mathbf{Y}_{\text{L}_2}$			
✓	✗	✗	✗	1	18.80	0.753
✗	✓	✗	✗	1	19.10	0.744
✗	✗	✓	✗	1	19.23	0748
✗	✗	✗	✓	1	19.05	0.757
✓	✓	✗	✗	2	19.83	0.762
✓	✗	✓	✗	2	20.12	0.776
✓	✗	✗	✓	2	19.65	0.788
✗	✓	✓	✗	2	20.84	0.795
✗	✓	✗	✓	2	20.21	0.770
✗	✗	✓	✓	2	20.55	0.782
✓	✓	✓	✗	3	22.05	0.825
✓	✓	✗	✓	3	21.89	0.808
✓	✗	✓	✓	3	22.23	0.813
✗	✓	✓	✓	3	22.74	0.829
✓	✓	✓	✓	4	**23.19**	**0.843**

Table 7. Ablation studies on feature maps of luminance guidance stack, supported by a complete Multinex architecture.

pixel-level, structural, and perceptual cues are all important, while being complementary, for low-light enhancement.

D.2. On Luminance Guidance Stack

Tab. 7 analyzes the contribution of each individual illumination feature map of $\mathbf{Y}_{\text{Rec.709}}$, $\mathbf{Y}_{\text{vmax}}$, $\mathbf{Y}_{\text{lightness}}$, and $\mathbf{Y}_{\text{L}_2}$, also their combinations. When used alone, they provide limited benefit, remaining in the range of 18-19dB PSNR with SSIM values just below 0.76. With pairwise combination, the performance consistently improves to the 20dB range, with the pair $(\mathbf{Y}_{\text{vmax}}, \mathbf{Y}_{\text{lightness}})$ performing slightly better than the others. This indicates that exposure adjustment can be stabilized by mixing physically grounded and perceptually aligned brightness cues. The three-component combinations increase the performance further to around 22dB, showing that the different feature maps contribute useful complementary information, instead of being just equivalent variants of each other. The full combination of four maps leads to the best result, reaching roughly 23dB PSNR and 0.84 SSIM. This confirms that a multi-view luminance prior helps the Multinex illumination network $f_\mathcal{L}$ infer the required luminance adjustment more accurately.

D.3. On Reflectance Guidance Stack

Tab. 8 presents a similar ablation for feature maps of the reflectance guidance stack. The individual and combined contribution of the chromaticity pair $[\mathbf{r}, \mathbf{g}]$, YCbCr pair, and saturation $\mathbf{S}$ are examined. For individual contribution, the chromaticity pair $[\mathbf{r}, \mathbf{g}]$ performs the best, offering a PSNR around 22dB and SSIM slightly above 0.80, likely because it provides illumination-invariant color ratios. The

$\mathcal{S}_\mathcal{R}$ Components			#Ch	PSNR↑	SSIM↑
$[\mathbf{C_b}, \mathbf{C_r}]$	$[\mathbf{r}, \mathbf{g}]$	$\mathbf{S}$			
✓	✗	✗	2	21.55	0.792
✗	✓	✗	2	21.90	0.805
✗	✗	✓	1	20.62	0.779
✓	✓	✗	4	22.98	0.835
✓	✗	✓	3	22.12	0.819
✗	✓	✓	3	22.65	0.826
✓	✓	✓	5	**23.19**	**0.843**

Table 8. Ablation studies on feature maps of reflectance guidance stack, supported by a complete Multinex architecture.

Placement	**Formulation** $f(\mathcal{S}) = \text{Conv}_{1\times1} \circ \ldots$	PSNR↑	SSIM↑
Before	$\text{FB}^{2T}\left(\text{Conv}_{1\times1}(\mathcal{S}) \odot \text{CWA}(\mathcal{S})\right)$	22.78	0.831
Between	$\text{FB}^{T}\left(\text{CWA}(\mathcal{S}) \odot \text{FB}^{T} \circ \text{Conv}_{1\times1}(\mathcal{S})\right)$	**23.19**	**0.843**
After	$\text{CWA}(\mathcal{S}) \odot \text{FB}^{2T} \circ \text{Conv}_{1\times1}(\mathcal{S})$	22.41	0.823

Table 9. Ablation on where to place CWA in fusion networks.

YCbCr pair offers slightly lower performance, while saturation alone is the weakest due to its sensitivity to noise. Pairwise combination yields clear improvements. For example, $[\mathbf{C}_b, \mathbf{C}_r]$ together with $[\mathbf{r}, \mathbf{g}]$ achieves close to 23dB PSNR with the highest SSIM within the pairwise combination group. Combinations involving saturation improve slightly less on SSIM, due to its noisier behavior. The complete three-way combination yields the best performance with a PSNR around 23dB and SSIM around mid-0.84, showing that the full stack provides a balanced and complementary chroma representation for the Multinex reflectance network to learn effectively the reflectance adjustment.

D.4. On CWA Placement

In this section, we examine the effect of where to place CWA within the fusion networks, by experimenting with three ways to insert the CWA mechanism as listed in Tab. 9, where the "between" setting corresponds to the proposed design in Eq. (18). In the other two settings of "before" and "after", we also increase the layer depth by applying the FB modules $2T$ times instead of T. By applying CWA early, i.e., before the main fusion blocks, we can obtain good result of roughly 23dB PSNR and 0.83 SSIM. By placing CWA after deeper FBs, we obtain slightly worse performance, indicating that a late attention is less effective when the features have already been heavily mixed. Our adopted design inserts CWA between the projection and the stacked FBs, which yields the best performance of a similar 23dB PSNR while higher SSIM close to 0.84. Overall,

despite having deeper layers, the other two ways of placing CWA decrease the model performance. This suggests that the proposed approach of weighting the analytic feature maps early, followed by a lightweight spatial refinement, is an effective strategy for leveraging multi-view representation priors under tight parameter constraints.

E. More Qualitative Examples

We provide additional qualitative comparisons and results on various datasets. It can be seen from Figs. 12 and 13 that Multinex achieves better color fidelity and detail recovery as compared to other previous lightweight and micro approaches, while also attain overall better brightness recovery. In Fig. 14, Multinex-Nano shows significantly better level of detail and illumination correction compared to prior micro-sized approaches. In Fig. 15, Multinex shows strong performance, approaching (sometimes outperforming) heavier models. While level of detail is lower in some cases, illumination and color correction are stronger, despite having a significantly smaller model size. Fig. 16 shows that Multinex exhibits greater color fidelity with stronger brightness improvements. However, it produces less-detailed outputs as compared to GLARE, due to the massive parameter scale difference. Finally, we demonstrate more examples of low-light images and their enhanced images predicted by Multinex in Figs. 17 18 19 20.

For a matter of interest, we also show in Fig. 21 a few very challenging low-light images. It can be seen that Multinex produces light distortions and noise on the DICM examples, and exhibits color loss on the MEF examples due to over-exposure. This is primarily due to its low parameter count which hinders detail reconstruction and color balancing in challenging scenarios. However, it can be seen from Fig 21 that images enhanced by those mid-sized approaches like RetinexFormer [4] and CIDNet [45] also show artifacts.

Figure 12. Qualitative comparison on reference dataset LOL-v1 [42] between Multinex and state-of-the-art lightweight and micro scale models RetinexNet [42], PairLIE [7], ZeroDCE [8], LYT-Net [2], RUAS [21], RSF [33], SCI [27].

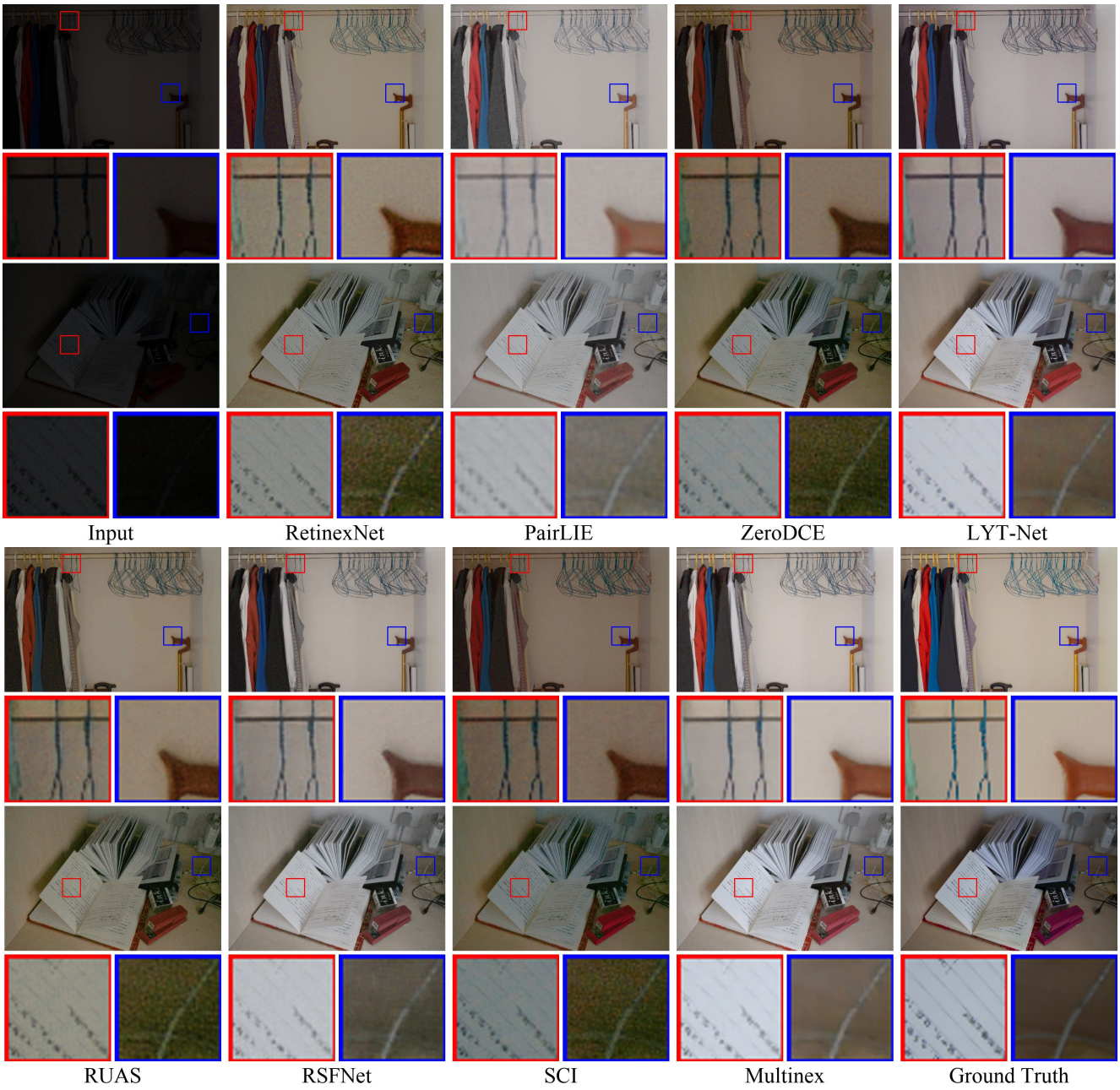

Figure 13. Qualitative comparison on reference dataset LOL-v1 [42] between Multinex and state-of-the-art lightweight and micro scale models RetinexNet [42], PairLIE [7], ZeroDCE [8], LYT-Net [2], RUAS [21], RSF [33], SCI [27].

Figure 14. Qualitative comparison on reference dataset LOL-v1 [42] between Multinex-Nano and state-of-the-art micro scale models RUAS [21], RSF [33], SCI [27].

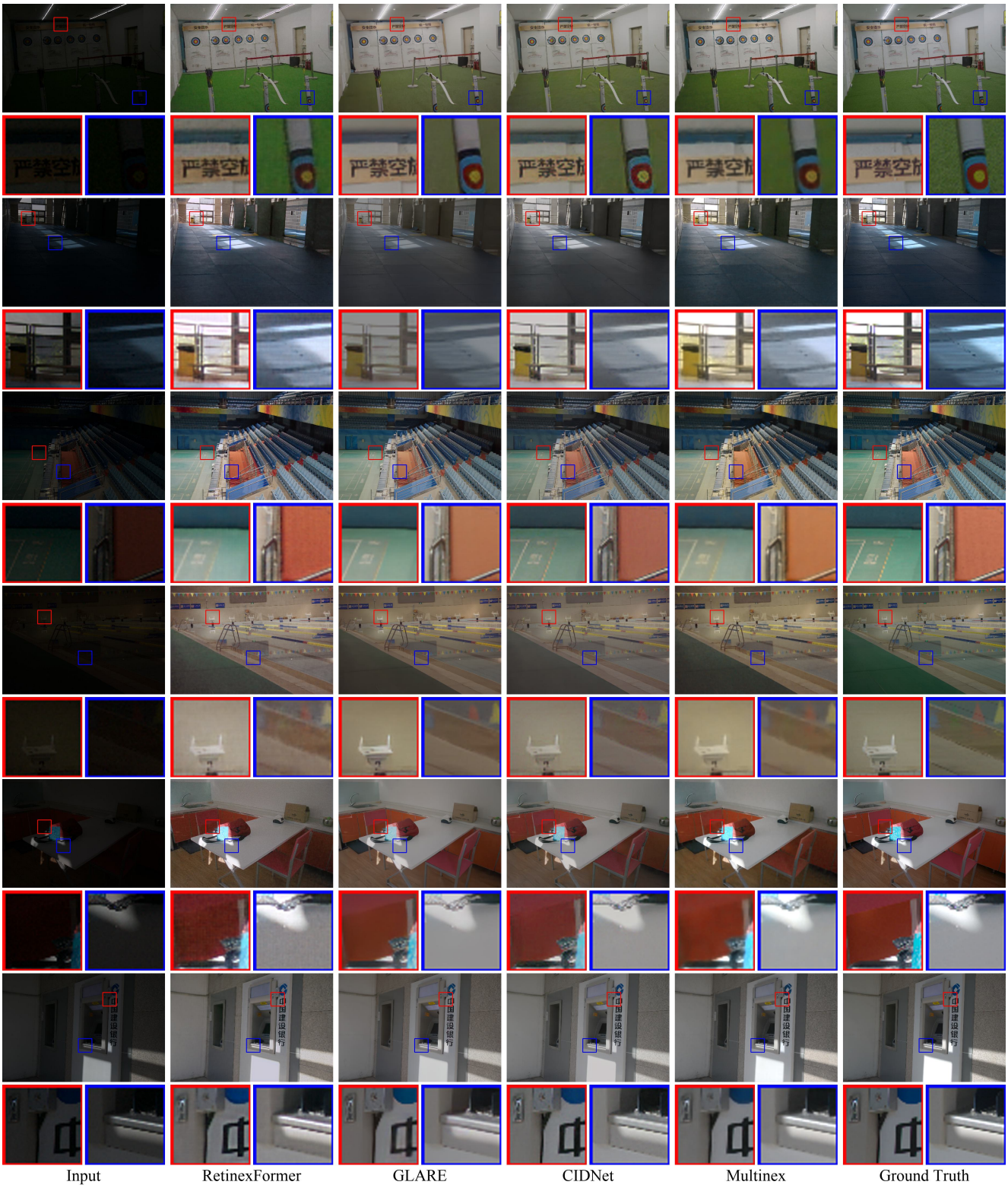

Figure 15. Qualitative comparison on reference dataset LOL-v2-real [46] between Multinex and state-of-the-art mid-sized (1-10 M param.) models RetinexFormer [4] and CIDNet [45], and heavy (>10 M param.) model GLARE [52].

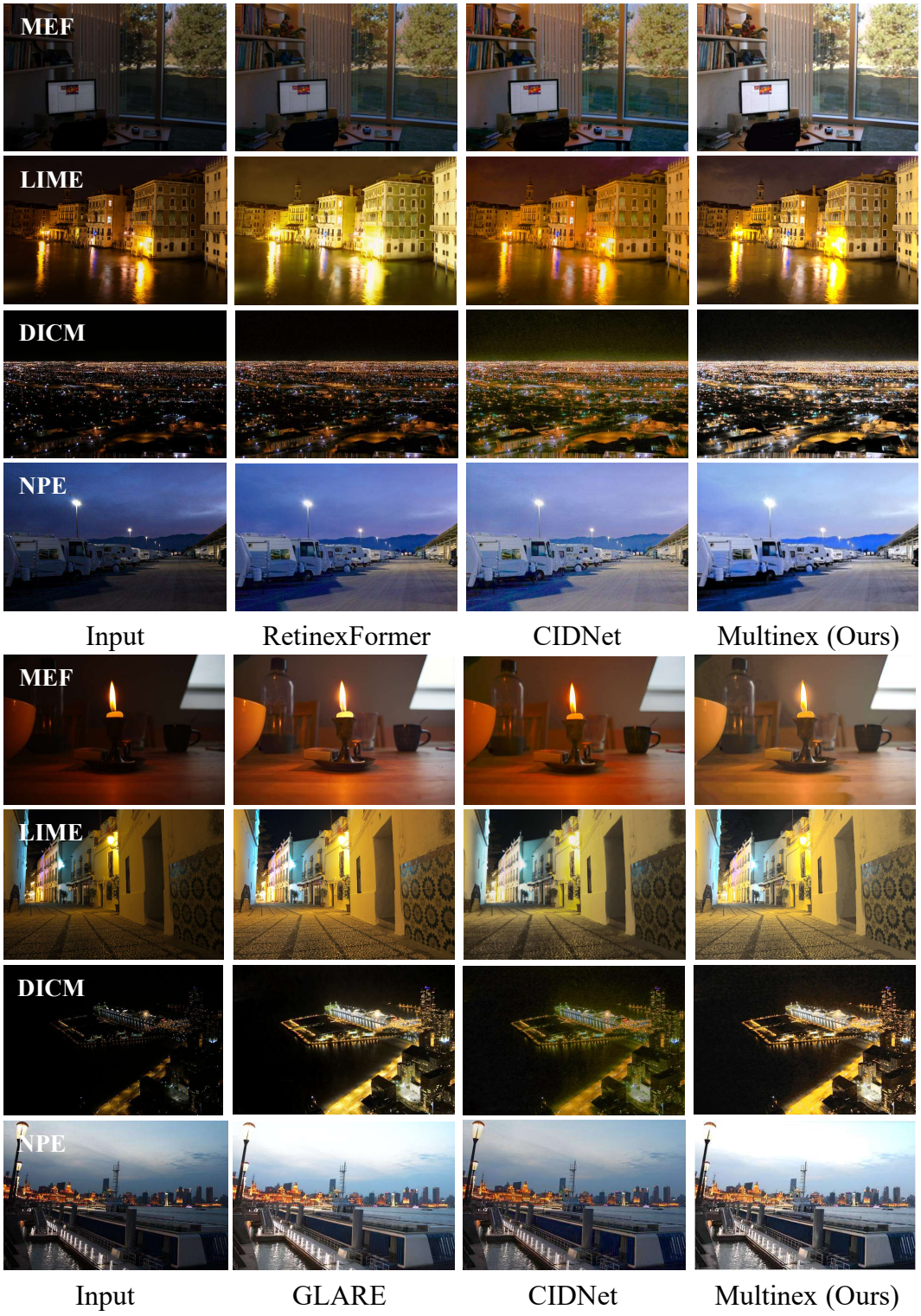

Figure 16. Qualitative comparison on no-reference datasets MEF [26], LIME [10], DICM [20], NPE [37] between Multinex and state-of-the-art mid-sized (1-10M param.) models RetinexFormer [4] and CIDNet [45] and heavy-weight (>10M param.) GLARE [52].

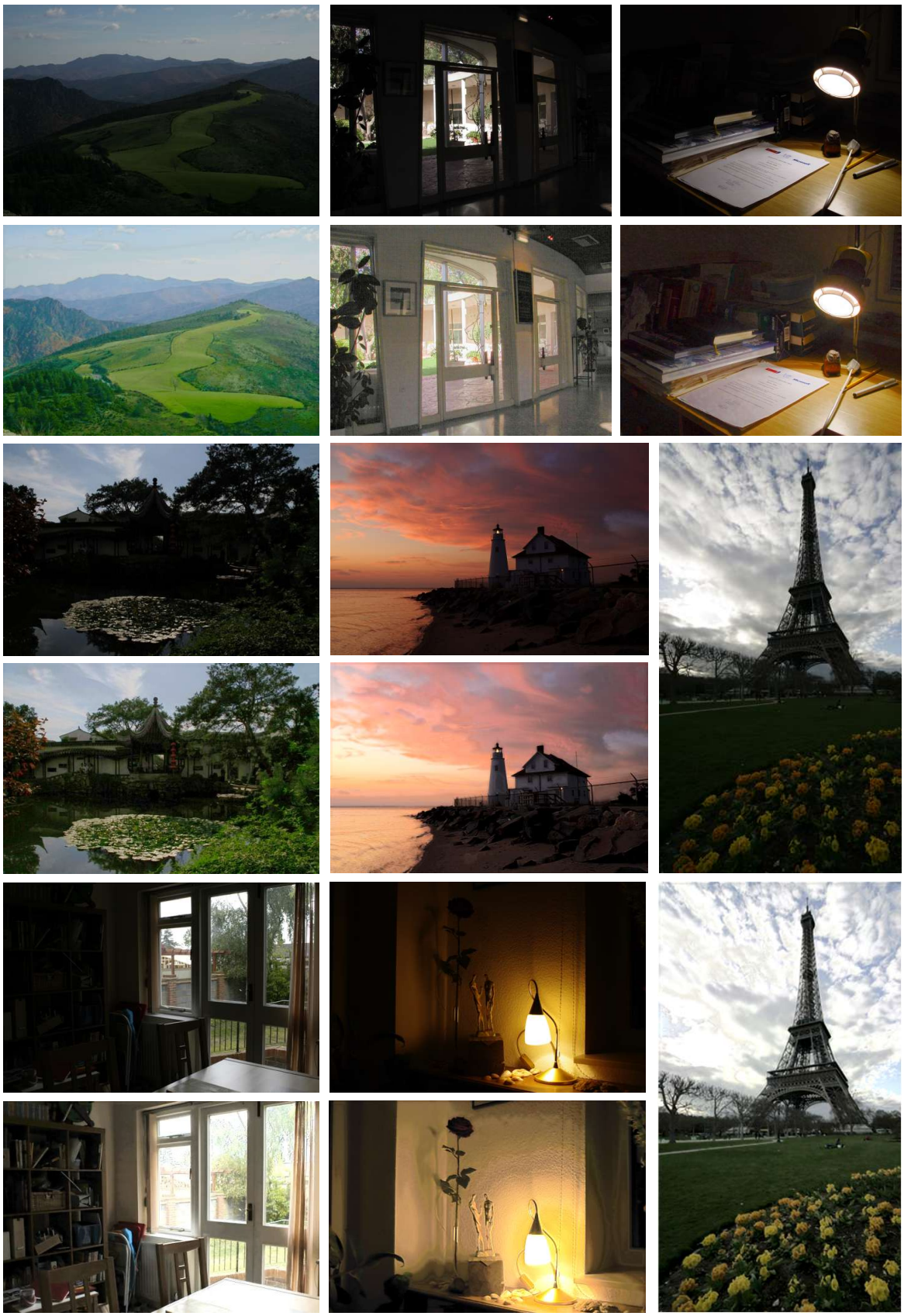

Figure 17. Additional results on no-reference dataset MEF [26]. For corresponding images, top is input, and bottom is Multinex output.

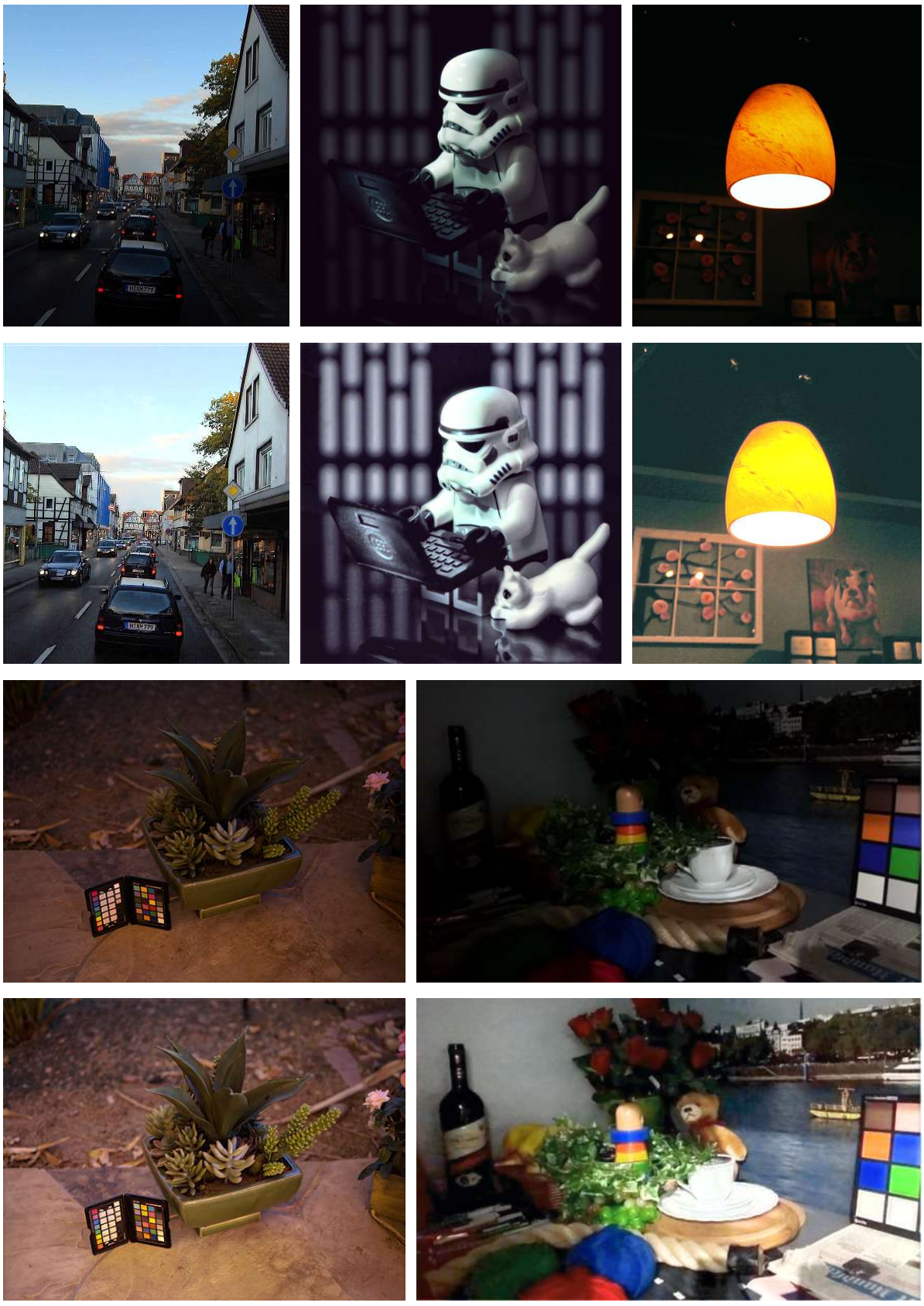

Figure 18. Additional results on no-reference dataset LIME [10]. For corresponding images, top is input, and bottom is Multinex output.

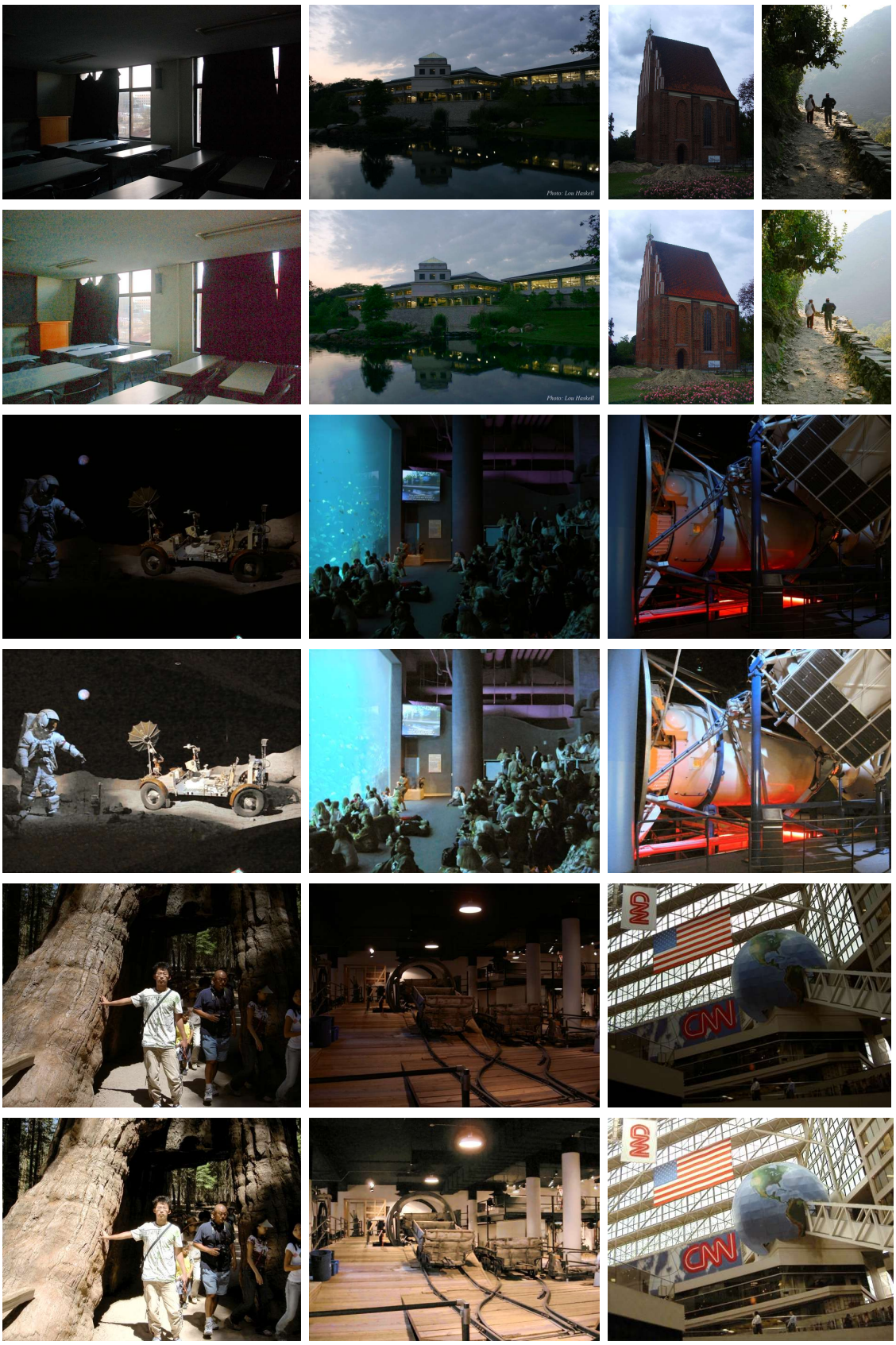

Figure 19. Additional results on no-reference dataset DICM [20]. For corresponding images, top is input, and bottom is Multinex output.

Figure 20. Additional results on no-reference dataset NPE [37]. For corresponding images, top is input, and bottom is Multinex output.

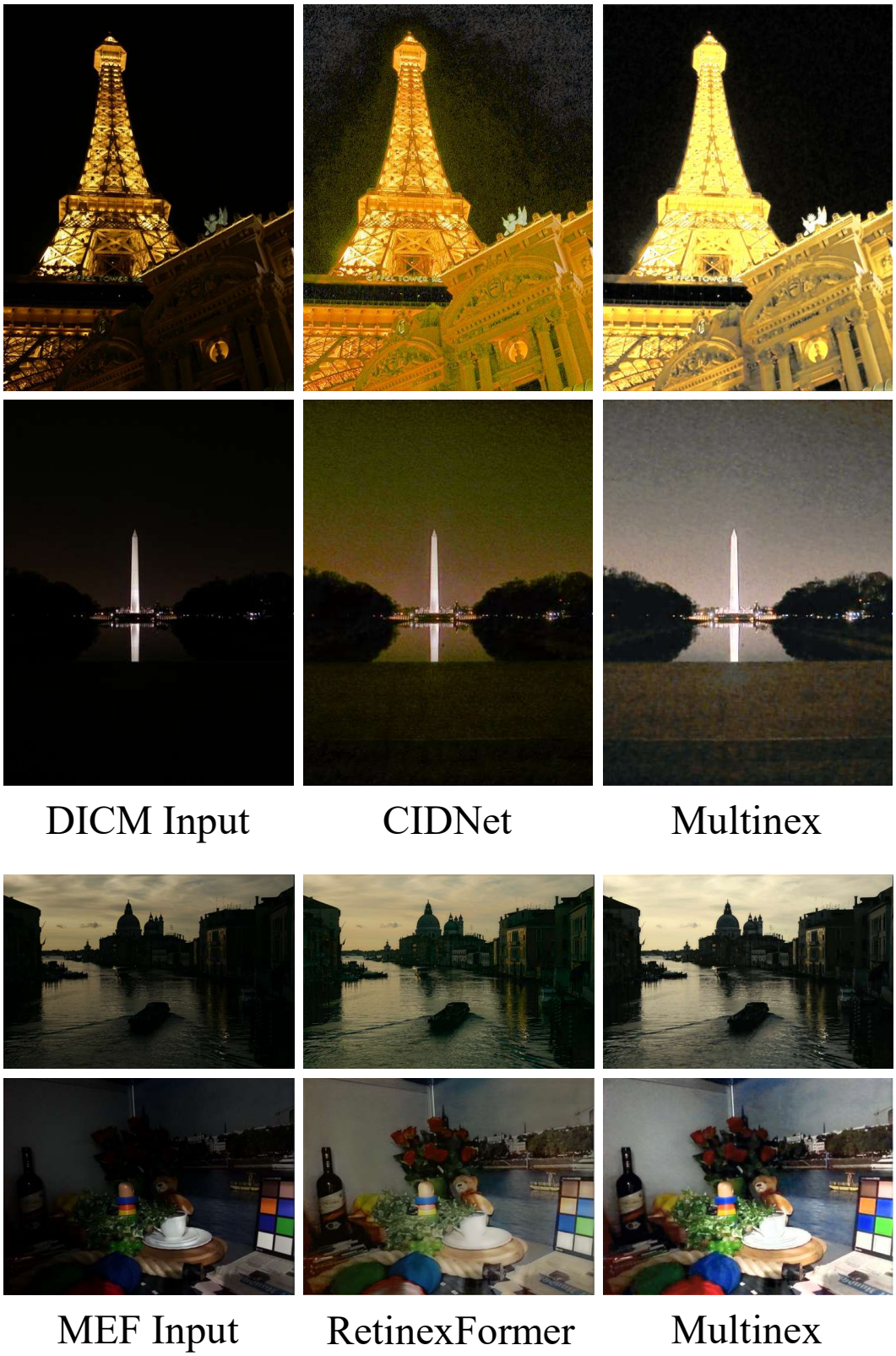

Figure 21. A few challenging cases from DICM and MEF datasets.