Semantic-Aware Spectral Reconstruction: A Spectral Library-Aided Unsupervised Method Based on the Diffusion Model

Supplementary Material

6. Implementation Details of the Lib-SRDM

The algorithm of Lib-SRDM is shown in Algorithm 1. The algorithm takes the RGB image $\mathcal{Y}$, the spectral response function $\mathbf{P}$, the number of switching steps N_{switch}, the number of low-rank steps N_{LR}, the noise prediction model $\mathcal{E}_\theta$, the number of spectra to be retrieved $N_{\mathbf{S}}$, and the spectral library $\mathcal{S}_{\text{Lib}}$ as inputs. The output is the reconstructed HSI $\mathcal{X}_0$.

To construct a better semantic-aware spectral distribution, we extend the retrieve function (Eq. (12)) as Algorithm 2 line 4 with the observed color of the RGB pixels. It is based on the observation that the color of the RGB pixels can provide additional information about the spectra. Therefore, spectra that are more consistent with the color of the RGB pixels are more likely to be the true spectra. Specifically, for the k-th instance in the RGB image $\mathcal{Y}$, we firstly obtain spectra $\mathbf{S}_{c_k}$ associated with the class c_k from the spectral library $\mathcal{S}_{\text{Lib}}$. Then, we calculate the color of these spectra using the spectral response function (SRF) $\mathbf{P}$ of the camera as

$$\mathbf{S}_{c_k}^{\text{RGB}} = \mathbf{S}_{c_k}\mathbf{P}. \tag{27}$$

Subsequently, we compute the distance between the color of the spectra and the color of the pixel in $\mathcal{Y} \odot \mathbf{M}_k$. For the i-th spectral $\mathbf{s}_i^{\text{RGB}} \in \mathbf{S}_{c_k}^{\text{RGB}}$, the distance is calculated as

$$d_i = \min_{\mathbf{y} \in \mathcal{Y} \odot \mathbf{M}_k} \|\mathbf{s}_i^{\text{RGB}} - \mathbf{y}\| \tag{28}$$

For better stability, the distance is normalized by the maximum distance of all spectra in $\mathbf{S}_{c_k}$, which is defined as

$$\bar{d}_i = d_i / \max_j d_j \tag{29}$$

The probability of the i-th spectral $\mathbf{s}_i$ to be selected is defined as

$$p_i = \frac{\exp(-\bar{d}_i^{\,2}/\tau)}{\sum_j \exp(-\bar{d}_j^{\,2}/\tau)} \tag{30}$$

where τ is a hyperparameter to control the sharpness of the distribution. In this paper, we set $\tau = 0.5$. The algorithm of the retrieval function is shown in Algorithm 2. The algorithm takes the spectral library $\mathcal{S}_{\text{Lib}}$, the RGB pixel $\mathcal{Y} \odot \mathbf{M}_k$, the number of spectra to be retrieved $N_{\mathbf{S}}$, the class c_k, and the spectral response function $\mathbf{P}$ as inputs. The output is the retrieved spectra $\mathbf{S}_k$.

Algorithm 1 Reverse Process of Lib-SRDM

Input: $\mathcal{Y}$, $\mathbf{P}$, N_{switch}, N_{LR}, $\mathcal{E}_\theta$, $N_{\mathbf{S}}$, and $\mathcal{S}_{\text{Lib}}$.
Output: Reconstructed HSI $\mathcal{X}_0$.

Identify the instances in $\mathcal{Y}$ using Grounded-SAM.
$\{c_k, \mathbf{M}_k\}_{k=1}^K = \text{Grounded-SAM}(\mathcal{Y})$
# Build SDM and LRM for each instance
for $k = 1$ **to** K **do**
  Retrieve the spectra from the spectral library.
  $\mathbf{S}_k = \text{Retrieve}(\mathcal{S}_{\text{Lib}}, \mathcal{Y} \odot \mathbf{M}_k, N_{\mathbf{S}}, c_k, \mathbf{P})$
  Construct the spectral DM $\text{SD}_k(\cdot, \cdot; \mathbf{S}_k)$ and the spectral low-rank model $\text{LR}_k(\cdot, \cdot; \mathbf{S}_k)$.
end for
# Running the reverse diffusion process
Initialize $\mathcal{X}_T$ with Gaussian noise.
for $t = T$ **to** 1 **do**
  if $t < N_{\text{switch}}$ **then**
    Using the HSI-DM for reverse diffusion.
    $\hat{\mathcal{X}}_{0|t} = \left(\mathcal{X}_t - \sqrt{1-\bar{\alpha}_t}\mathcal{E}_\theta(\mathcal{X}_t, t)\right)/\sqrt{\bar{\alpha}_t}$.
  else
    Using spectral DMs for reverse diffusion.
    $\hat{\mathcal{X}}_{0|t} = \sum_k \text{SD}_k(\mathcal{X}_t \odot \mathbf{M}_k, t; \mathbf{S}_k) \otimes \mathbf{M}_k$.
  end if
  $\mathcal{G}_t = \sum_k \text{LR}_k(\mathcal{X}_t \odot \mathbf{M}_k, t; \mathbf{S}_k) \otimes \mathbf{M}_k$.
  Calculate $\nabla_{\mathcal{X}_t} \log p(\mathcal{Y}|\mathcal{X}_t, \mathbf{P})$ using Eq. (7).
  $\mathcal{X}'_{t-1} = c_1\mathcal{X}_t + c_2\hat{\mathcal{X}}_{0|t} + \sigma_t\mathcal{E}$.
  $\mathcal{X}_{t-1} = \mathcal{X}'_{t-1} + s\sigma_t(\nabla_{\mathcal{X}_t} \log p(\mathcal{Y}|\mathcal{X}_t, \mathbf{P}) + \lambda\mathcal{G}_t)$.
end for

Algorithm 2 Retrieve Function

Input: $\mathcal{S}_{\text{Lib}}$, $\mathcal{Y} \odot \mathbf{M}_k$, $N_{\mathbf{S}}$, c_k, $\mathbf{P}$.
Output: Retrieved spectra $\mathbf{S}_k$.

1: Retrieve the spectra $\mathbf{S}_{c_k}$ from the spectral library $\mathcal{S}_{\text{Lib}}$ associated with the class c_k.
2: Calculate the color of the spectra using Eq. (27).
3: Calculate the normalized distance $\bar{d}_i$ using Eq. (28) and Eq. (29).
4: Calculate the probability p_i using Eq. (30).
5: Sample $N_{\mathbf{S}}$ spectra as $\mathbf{S}_k$ from $\mathbf{S}_{c_k}$ according to the probability distribution.

7. Proof of Theoretical Analysis

These semantic-aware spectral models are incorporated via the approximation in Eq. (9), which can be reformulated as $p(\mathcal{X}|\mathcal{Y}, \mathbf{P}, \hat{\mathbf{C}})$ as a special case of $p(\mathcal{X}|\mathcal{Y}, \mathbf{P})$ assuming that $\mathbf{C}$ has a high probability at $\hat{\mathbf{C}}$. It is intuitive that the

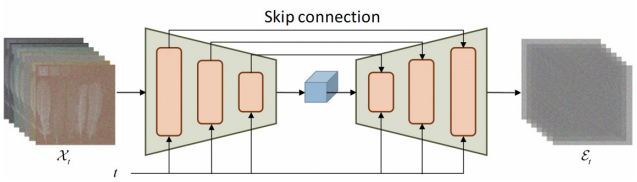

Figure 9. Illustration of the architecture of the HSI DM.

involvement of external semantics $\hat{\mathbf{C}}$ and library spectra can enhance the SR performance. In particular, we theoretically analyze the benefits of this approximation. In addition to Proposition 1 in the main paper, we provide two more propositions to demonstrate that the proposed method can reduce variance and increase the probability of obtaining the ground truth HSI. For simplicity, we **omit** $\mathbf{P}$ in the following analysis.

Proprosition 1 *The proposed method reduces the expected distance between the estimated and ground truth HSI* $\mathcal{X}^{GT}$*, i.e.,* $\mathbb{E}[D(\mathcal{X}, \mathcal{X}^{GT})|\mathcal{Y}, \hat{\mathbf{C}}] \leq \mathbb{E}[D(\mathcal{X}, \mathcal{X}^{GT})|\mathcal{Y}]$.

Proof: According to the total probability theorem, we have

$$\mathbb{E}[D(\mathcal{X}, \mathcal{X}^{\mathrm{GT}})|\mathcal{Y}] \tag{31}$$

$$= \sum_{\mathbf{C}} p(\mathbf{C}|\mathcal{Y}) \mathbb{E}[D(\mathcal{X}, \mathcal{X}^{\mathrm{GT}})|\mathcal{Y}, \mathbf{C}] \tag{32}$$

$$\geq \mathbb{E}[D(\mathcal{X}, \mathcal{X}^{\mathrm{GT}})|\mathcal{Y}, \hat{\mathbf{C}}]. \tag{33}$$

Eq. (33) leverages the inequality that $\mathbb{E}[D(\mathcal{X}, \mathcal{X}^{\mathrm{GT}})|\mathcal{Y}, \mathbf{C}] \geq \mathbb{E}[D(\mathcal{X}, \mathcal{X}^{\mathrm{GT}})|\mathcal{Y}, \hat{\mathbf{C}}]$. It can be derived from the following Proposition 2 with the assumption that $\mathbb{E}[\mathcal{X}|\mathcal{Y}, \mathbf{C}] \simeq \mathcal{X}^{\mathrm{GT}}$ and $\mathbb{E}[\mathcal{X}|\mathcal{Y}, \hat{\mathbf{C}}] \simeq \mathcal{X}^{\mathrm{GT}}$. Intuitively, when library spectra and semantic results are accurate, since different material classes often have distinct spectral characteristics, the spectra within the same class are more similar. Therefore, the semantic-aware spectral distribution provided by the true semantics $\hat{\mathbf{C}}$ is more likely to be close to the ground truth HSI $\mathcal{X}^{\mathrm{GT}}$ than wrong semantics $\mathbf{C} \neq \hat{\mathbf{C}}$.

Proprosition 2 $p(\mathcal{X}|\mathcal{Y}, \hat{\mathbf{C}})$ *has less variance than* $p(\mathcal{X}|\mathcal{Y})$.

Proof: According to the law of total variance, we have

$$\mathrm{Var}(\mathcal{X}|\mathcal{Y}) = \mathbb{E}_{\mathbf{C}|\mathcal{Y}}[\mathrm{Var}(\mathcal{X}|\mathcal{Y}, \mathbf{C})] + \mathrm{Var}_{\mathbf{C}|\mathcal{Y}}(\mathbb{E}[\mathcal{X}|\mathcal{Y}, \mathbf{C}]) \tag{34}$$

$$\geq \mathbb{E}_{\mathbf{C}|\mathcal{Y}}[\mathrm{Var}(\mathcal{X}|\mathcal{Y}, \mathbf{C})] \tag{35}$$

$$\simeq \mathrm{Var}(\mathcal{X}|\mathcal{Y}, \hat{\mathbf{C}}). \tag{36}$$

Eq. (34) shows that the method reduces the sampling variance by mitigating uncertainty due to incorrect semantics ($\mathbf{C} \neq \hat{\mathbf{C}}$).

Proprosition 3 *If the semantic result and library spectra are accurate, i.e.,* $p(\mathcal{X}^{GT}|\hat{\mathbf{C}}) \geq p(\mathcal{X}^{GT}|\mathbf{C})$*, then our method has a higher probability of obtaining the ground truth HSI* $\mathcal{X}^{GT}$*, i.e.,* $p(\mathcal{X}^{GT}|\mathcal{Y}, \hat{\mathbf{C}}) \geq p(\mathcal{X}^{GT}|\mathcal{Y})$.

Proof: According to Bayes's rule, we have

$$p(\mathcal{X}^{\mathrm{GT}}|\mathcal{Y}) = \sum_{\mathbf{C}} p(\mathcal{Y}|\mathcal{X}^{\mathrm{GT}}) p(\mathcal{X}^{\mathrm{GT}}|\mathbf{C}) p(\mathbf{C}|\mathcal{Y}) \tag{37}$$

$$\leq p(\mathcal{Y}|\mathcal{X}^{\mathrm{GT}}) p(\mathcal{X}^{\mathrm{GT}}|\hat{\mathbf{C}}) \sum_{\mathbf{C}} p(\mathbf{C}|\mathcal{Y}) \tag{38}$$

$$= p(\mathcal{X}^{\mathrm{GT}}|\mathcal{Y}, \hat{\mathbf{C}}). \tag{39}$$

Since DM-based SR solvers need to sample from $p(\mathcal{X}|\mathcal{Y})$ for reconstruction, Propositions 2 and 3 are also valuable to improve the performance. Less variance and a higher probability at $\mathcal{X}^{\mathrm{GT}}$ indicate that $p(\mathcal{X}|\mathcal{Y}, \hat{\mathbf{C}})$ is more concentrated around $\mathcal{X}^{\mathrm{GT}}$ with less uncertainty, leading to a more accurate and stable SR.

8. Architecture of the HSI DM

We follow the UNet architecture used in the denoising diffusion probability model (DDPM) [17] to design the HSI DM by extending the input and output channels to B bands. The architecture is shown in Fig. 9. The input is the noisy HSI $\mathcal{X}_t \in \mathbb{R}^{B \times H \times W}$, and the output is the predicted noise $\mathcal{E}_\theta(\mathcal{X}_t, t) \in \mathbb{R}^{B \times H \times W}$. The architecture consists of five down-sampling blocks and five up-sampling blocks, with skip connections between corresponding down-sampling and up-sampling blocks. Each block contains one residual convolution block, with a timestep embedding input as an adaptive group norm. The down-sampling blocks use an average pooling layer to down-sample the feature maps, while the up-sampling blocks use a nearest interpolation layer to up-sample the feature maps. The output of the last up-sampling block is a convolution layer with B output channels. The key hyperparameters of the architecture are shown in Table 4.

9. RGB Space Relighting

For typical RGB space relighting methods, as the whole spectrum is not available, the imaging process is generally down-sampled to three channels as [13]

$$\mathcal{Y} = \mathcal{R}^{\mathrm{down}} \times_1 \mathrm{diag}(\boldsymbol{l}^{\mathrm{down}}) \tag{40}$$

Here, to distinguish from the full-spectrum illuminant $\boldsymbol{l}$ and reflectance $\mathcal{R}$, we denote the down-sampled illuminant and reflectance used in RGB space relighting as $\boldsymbol{l}^{\mathrm{down}} \in \mathbb{R}^3$ and $\mathcal{R}^{\mathrm{down}} \in \mathbb{R}^{3 \times H \times W}$, respectively.

Consider $\boldsymbol{l}^{\mathrm{down}}$ can be derived from $\boldsymbol{l}$ by the function $f : \mathbb{R}^B \to \mathbb{R}^3$. Combining Eq. (1) and Eq. (40), we have

$$\mathcal{R}^{\mathrm{down}} = \mathcal{R} \times_1 \mathrm{diag}(\boldsymbol{l}) \mathbf{P} \mathrm{diag}(f(\boldsymbol{l}))^{-1} \tag{41}$$

Hyperparameter	Value
in_channels	31
out_channels	31
num_channels	128
channel_mult	1, 1, 2, 2, 4, 4
num_res_blocks	1
learn_sigma	False
attension_reoslutoins	16
num_head_channels	64
use_scale_shift_norm	True
resblock_updown	True

Table 4. Key hyperparameters of the HSI DM architecture.

According to the definition of the reflectance, $\mathcal{R}^{\text{down}}$ should be independent of the illuminant. Therefore, we have

$$\nabla_{\boldsymbol{l}} \mathcal{R}^{\text{down}} \equiv 0 \tag{42}$$

Typically, there is no solution to Eq. (42), indicating the intrinsic error of the RGB space relighting. To obtain a statistically optimal solution, we assume the constant expectation of the reflectance $\mathbb{E}[\mathcal{R}] \equiv \text{const}$, and then we can derive the subsequent equation from Eq. (42),

$$\mathbf{P}\text{diag}(f(\boldsymbol{l})) = \nabla_{\boldsymbol{l}} f(\boldsymbol{l})^T \mathbf{P}^T \boldsymbol{l} \tag{43}$$

One of the solution can be obtained as

$$f(\boldsymbol{l}) = \mathbf{P}^T \boldsymbol{l} \tag{44}$$

Therefore, given the input RGB image $\mathcal{Y}'$, and input illuminant $\boldsymbol{l}$, the RGB space relighting (RGB-re) can be formulated as

$$\mathcal{Y} = \mathcal{Y}' \times_1 \text{diag}(\mathbf{P}^T \boldsymbol{l})^{-1} \text{diag}(\mathbf{P}^T \boldsymbol{l}^{\text{GT}}) \tag{45}$$

where $\boldsymbol{l}^{\text{GT}}$ is ground truth illuminant. However, Eq. (45) is a statistically optimal solution based on the constant expectation of the reflectance, which is not guaranteed to be true in practice, resulting in distortion in the reconstructed image. The color shift caused by RGB space relighting is illustrated in Fig. 10, taking a case from the NTIRE22 dataset as an example.

10. Recoloring in HSI Space

Given the input RGB image, recoloring aims to adjust the color of the image to a specified one. For RGB space recoloring [5], an RGB palette is extracted by a clustering algorithm (e.g., k-means) from the input RGB image, and then the weights of these RGB palette colors are calculated for each pixel (e.g., by radial basis function). However, the limited channels of RGB cannot fully represent the spectral

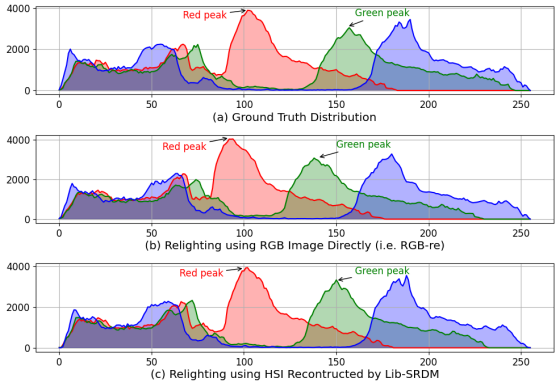

Figure 10. Ground truth RGB distribution and relighting results under the CIE A illuminant on the NTIRE22 dataset. RGB-re leads to severe color shifts, while Lib-SRDM maintains a distribution that closely aligns with the ground truth.

Figure 11. The raw RGB image from the NTIRE22 dataset used for recoloring.

information of different materials, leading to metamerism issues. It results in the difficulty of discovering the palette colors that can accurately represent the spectral information of all materials in the scene. In contrast, in HSI space recoloring, the full spectral information makes it easier to obtain an endmember matrix rather than an RGB palette, alleviating the metamerism problem. Furthermore, the high spectral resolution allows for more accurate weight calculation of each material for each pixel, enabling precise recoloring.

For an input RGB image $\mathcal{Y}$, the corresponding HSI $\mathcal{X}$ is first reconstructed by Lib-SRDM. Then, the endmember matrix $\mathbf{E} \in \mathbb{R}^{N_e \times B}$ is extracted by N-FINDR or manually selected. Next, the abundance map $\mathcal{A} \in \mathbb{R}^{N_e \times H \times W}$ is

(a) (b) (c) (d)

Figure 12. HSI space recoloring results using Lib-SRDM w/ $\mathcal{S}_{\text{all}}$ on the NTIRE22. (a) Recoloring the pink flowers to yellow. (b) Recoloring the blue sky to dusk. (c) Recoloring the dark green leaves to light green. (d) Recoloring the yellow flowers to blue.

(a) (b) (c) (d)

Figure 13. RGB space recoloring results using palette-based method on the NTIRE22. (a) Recoloring the pink flowers to yellow. (b) Recoloring the blue sky to dusk. (c) Recoloring the dark green leaves to light green. (d) Recoloring the yellow flowers to blue.

Figure 14. The raw RGB image from the CAVE dataset used for recoloring.

calculated by solving the following optimization problem:

$$\min_{\mathcal{A}} \|\mathcal{X} - \mathcal{A} \times_1 \mathbf{E}\|_F^2, \quad \text{s.t. } \mathcal{A} \geq 0 \tag{46}$$

To recolor the i-th material, we modify the corresponding endmember $\mathbf{e}_i$ to $\tilde{\mathbf{e}}_i$. The recolored HSI $\tilde{\mathcal{X}}$ is obtained as

$$\tilde{\mathcal{X}} = \mathcal{X} + \mathcal{A}_{i,:,:} \times_1 (\tilde{\mathbf{e}}_i - \mathbf{e}_i) \tag{47}$$

An RGB image from the NTIRE22 dataset is used for recoloring, as shown in Fig. 11. The HSI recoloring results using Lib-SRDM with the ideal library $\mathcal{S}_{\text{all}}$ are shown in Fig. 12. The RGB space recoloring results using the palette-based method [5] are shown in Fig. 13. It can be observed that the proposed HSI space recoloring method can accurately adjust the colors of different materials to the desired colors, while the RGB space recoloring method suffers from color distortion due to metamerism issues. Some raw RGB images from the CAVE and ICVL datasets used for recoloring are shown in Fig. 14 and Fig. 17, respectively. Accordingly, their HSI space recoloring results are shown in Fig. 15 and Fig. 18, respectively.

11. Computational Complexity

The computational complexity of the proposed and alternative methods is shown in Table 5. The GFLOPs are calculated based on the input size of 512×512 and the number of bands $B = 31$. The parameters are counted based on the number of trainable parameters in the model. Since

(a) (b) (c) (d)

Figure 15. HSI space recoloring results using Lib-SRDM w/ $\mathcal{S}_{\text{all}}$ on the CAVE. (a) Recoloring the orange toy to pink. (b) Recoloring the white background to red. (c) Recoloring the yellow cloth to light white. (d) Recoloring the blue trouser to green.

(a) (b) (c) (d)

Figure 16. RGB space recoloring results using palette-based method on the CAVE. (a) Recoloring the orange toy to pink. (b) Recoloring the white background to red. (c) Recoloring the yellow cloth to light white. (d) Recoloring the blue trouser to green.

Figure 17. The raw RGB image from the ICVL dataset used for recoloring.

Method	Parameters	GFLOPs
MST++	1.62M	201
MSFN	2.47M	130
CESST	1.54M	396
DDS2M	272M	509
MFormer	1.75M	361
UnGUN	21.9K	12
DDSR	93.5M	1582
Lib-SRDM	93.5M	1582

Table 5. Computational complexity. The GFLOPs are calculated based on the input size of 512×512 and the number of bands $B = 31$. The parameters are counted based on the number of trainable parameters in the model.

Lib-SRDM and DDSR use the same architecture of the pre-trained HSI DM, the parameters and GFLOPs of DDSR and Lib-SRDM are the same. It should be emphasized that although Lib-SRDM exhibits higher parameter complexity, Lib-SRDM enables direct adaptation from a pre-trained HSI DM to address SR under various input conditions (e.g., different illuminants, spectral libraries, and spectral response

(a) (b) (c) (d)

Figure 18. HSI space recoloring results using Lib-SRDM w/ $\mathcal{S}_{\text{all}}$ on the ICVL. (a) Recoloring the sky to blue. (b) Recoloring the green grass to yellow. (c) Recoloring the dust to green. (d) Recoloring the rock to green.

(a) (b) (c) (d)

Figure 19. RGB space recoloring results using palette-based method on the ICVL. (a) Recoloring the sky to blue. (b) Recoloring the green grass to yellow. (c) Recoloring the dust to green. (d) Recoloring the rock to green.

functions). Furthermore, with a proper N_{switch} and $N_{\mathbf{S}}$, Lib-SRDM can achieve satisfactory performance with fewer diffusion steps, thereby reducing the computational burden during inference.

12. Influence of the Low-Rank Model

To investigate the influence of the dimension N_{LR} of the low-rank model (LRM) on the SR performance, we conduct experiments on the CAVE dataset with N_{LR} ranging from 3 to 11. The results are presented in Fig. 20. $N_{\text{LR}} = 31$ indicates that the LRM is not used. It is found that a practical library $\mathcal{S}_{\text{part}}$ is robust to N_{LR}, and the performance is relatively stable when N_{LR} is greater than 6. However, when the library is ideal, there are obvious improvements in performance when N_{LR} is around 6 to 10. This is because the ideal library $\mathcal{S}_{\text{all}}$ provides more accurate spectra, allowing the LRM to better capture the spectral distribution with a moderate dimension. Therefore, we set $N_{\text{LR}} = 7$ in all experiments as a trade-off between ideal and practical libraries.

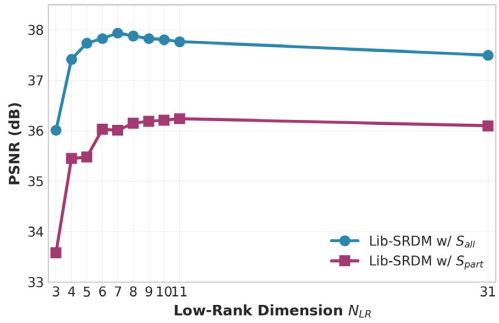

Figure 20. Influence of the dimension N_{LR} of the LRM on the CAVE.

13. Robustness to SRF

To evaluate the robustness of Lib-SRDM to the spectral response function (SRF), we conduct experiments on the CAVE dataset. During training, all necessary RGB images are generated using the SRF of the Nikon D700 camera, and the model is subsequently tested under both the SRF of the D700 and the SRF from NTIRE22 [2]. The results are presented in Table 6. The performance of all methods declines

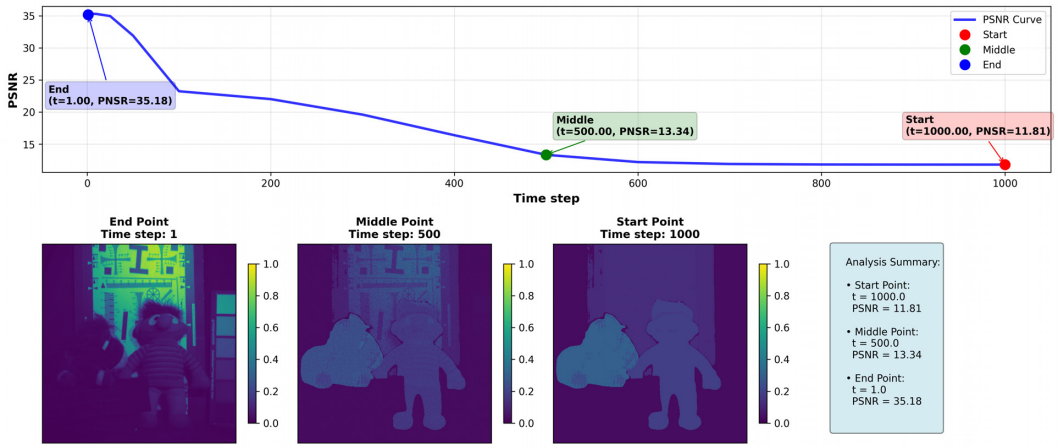

Figure 21. Visualization of intermediate results of Lib-SRDM w/ $\mathcal{S}_{\text{part}}$.

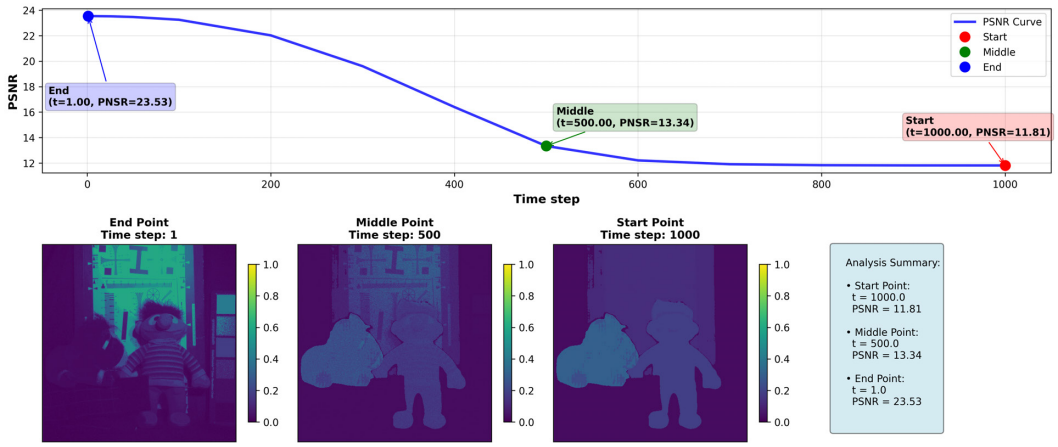

Figure 22. Visualization of intermediate results of Lib-SRDM w/o HSI DM.

when the SRF is switched from the D700 to the NTIRE22 SRF, indicating that the SRF significantly affects SR. Notably, Lib-SRDM with $\mathcal{S}_{\text{part}}$ achieves the best performance under both SRFs, demonstrating its robustness to variations in SRF.

14. Analysis of the Reverse Diffusion Process

To gain deeper insights into the reverse diffusion process of Lib-SRDM and to better understand the roles of the semantic-aware spectral models and the pre-trained HSI DM, we visualize intermediate results at different diffusion steps and compare them with two ablated versions as described in Sec. 4.1: 1) Lib-SRDM w/o Lib, and 2) Lib-SRDM w/o HSI DM. The "Lib-SRDM w/o Lib" version removes the spectral library and semantic-aware spectral models, relying solely on the pre-trained HSI DM for reconstruction. In contrast, "Lib-SRDM w/o HSI DM" eliminates the pre-trained HSI DM, utilizing only the semantic-aware spectral models for reconstruction. The results are shown in Fig. 21, Fig. 22, and Fig. 23, respectively.

As illustrated in Fig. 22, the high-level semantic structures are quickly recovered in the early stages of the reverse diffusion process when only the spectral models are used. However, due to the limitations of the semantic-aware spectral distribution, the finer details are not adequately reconstructed, even at the final step. In Fig. 23, the coarse structures are initially recovered, but the details remain noisy during the early and middle stages of the process, as the noise level is still high. As the noise level decreases, the details are gradually refined. Nevertheless, compared to Fig. 22, which uses only the spectral models, the overall structural reconstruction is slower; for instance, the edges remain blurry and noisy at time step 500, leading to a less accurate final result.

In contrast, as shown in Fig. 21, Lib-SRDM effectively integrates the strengths of both components. In the early stages, the semantic-aware spectral models quickly recover the high-

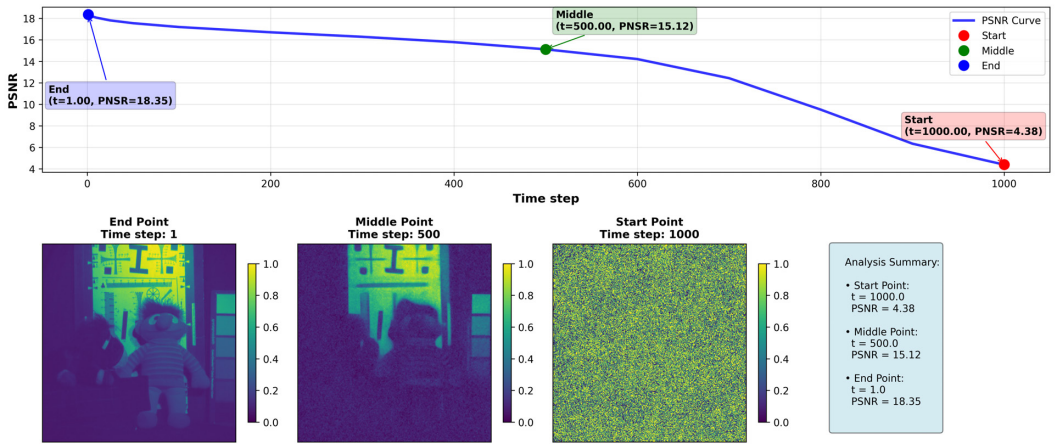

Figure 23. Visualization of intermediate results of Lib-SRDM w/o Lib.

Method	SRF of D700			SRF from NTIRE22		
	PSNR	SSIM	SAM	PSNR (Δ)	SSIM (Δ)	SAM (Δ)
MST++	35.20	0.9841	9.05	31.84 (-3.36)	0.9603 (-0.0238)	15.23 (+6.18)
MSFN	35.88	0.9837	9.32	31.27 (-4.61)	0.9634 (-0.0203)	14.27 (+4.95)
CESST	34.80	0.9810	10.26	31.26 (-3.54)	0.9633 (-0.0177)	14.52 (+4.26)
DDS2M	24.87	0.7445	31.56	17.25 (-7.62)	0.6435 (-0.1010)	41.62 (+10.06)
MFormer	25.59	0.8892	33.65	25.43 (-0.16)	0.8782 (-0.0110)	34.39 (+0.74)
UnGUN	29.60	0.9234	24.62	28.32 (-1.28)	0.9187 (-0.0047)	25.69 (+1.07)
DDSR	32.24	0.9464	19.91	27.94 (-4.30)	0.8907 (-0.0557)	21.10 (+1.19)
Lib-SRDM w/ $\mathcal{S}_{\text{all}}$	36.01	0.9745	12.28	31.65 (-4.36)	0.9619 (-0.0126)	14.43 (+2.15)
Lib-SRDM w/ $\mathcal{S}_{\text{part}}$	**37.94**	**0.9829**	**10.59**	**34.56** (-3.38)	**0.9685** (-0.0144)	**13.18** (+2.59)

Table 6. Robustness to Spectral Response Function (SRF). The results are evaluated on the CAVE dataset. The Δ values are computed as the difference between the results obtained using the SRF of the Nikon D700 camera and those using the SRF from NTIRE22. The best results are highlighted in bold.

level structures, while the pre-trained HSI DM progressively refines the finer details in the later stages. This synergistic approach results in a more accurate and detailed reconstruction of the HSI. Furthermore, as the spectral models show strong capability to reconstruct high-level structures, we can accelerate the inference by reducing the inference steps at the early and middle stages without sacrificing performance, thereby enhancing the efficiency of Lib-SRDM.

15. Spectral Reconstruction Visual Results

The SR results of Lib-SRDM and alternative methods on the CAVE and NTIRE22 datasets are shown in Fig. 24 and 25, respectively. The first column is the ground truth, and the rest are the SR results of Lib-SRDM and alternative methods. The error maps are also shown for better visualization.

16. Relighting Visual Results

The relighting results of Lib-SRDM and alternative methods on the CAVE, ICVL, and NTIRE22 datasets are shown in Fig. 26-31. The first column is the ground truth, and the rest are the relighting results of Lib-SRDM and alternative methods. The relighting results are shown under CIE A1 and CIE F6 illuminants.

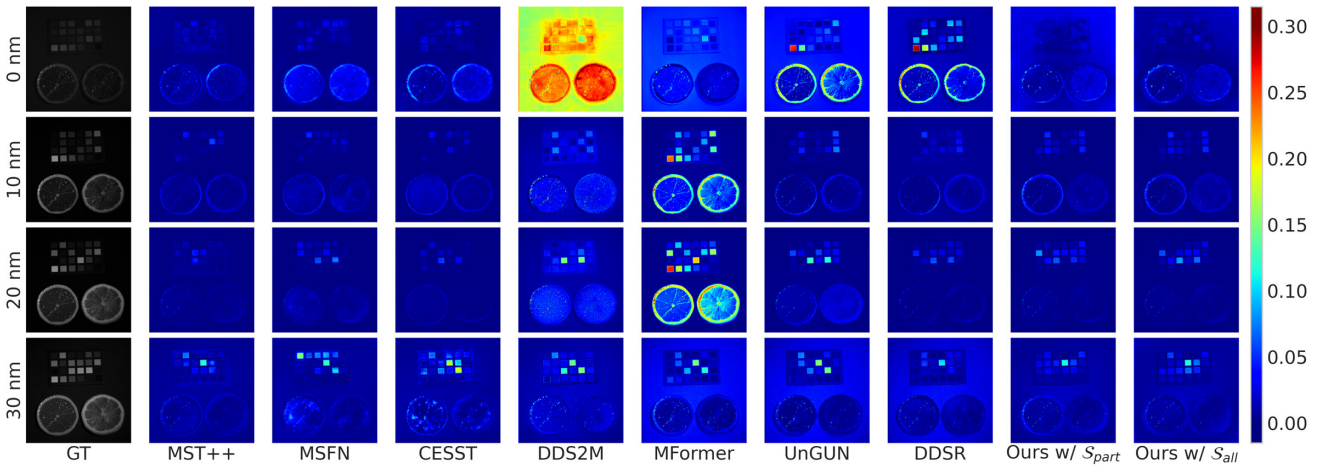

Figure 24. Illustration of error maps of SR results on the CAVE dataset. Four different bands are shown, including 400 nm, 500 nm, 600 nm, and 700 nm. The first column is the ground truth, and the rest are the error maps of the proposed method and alternative methods.

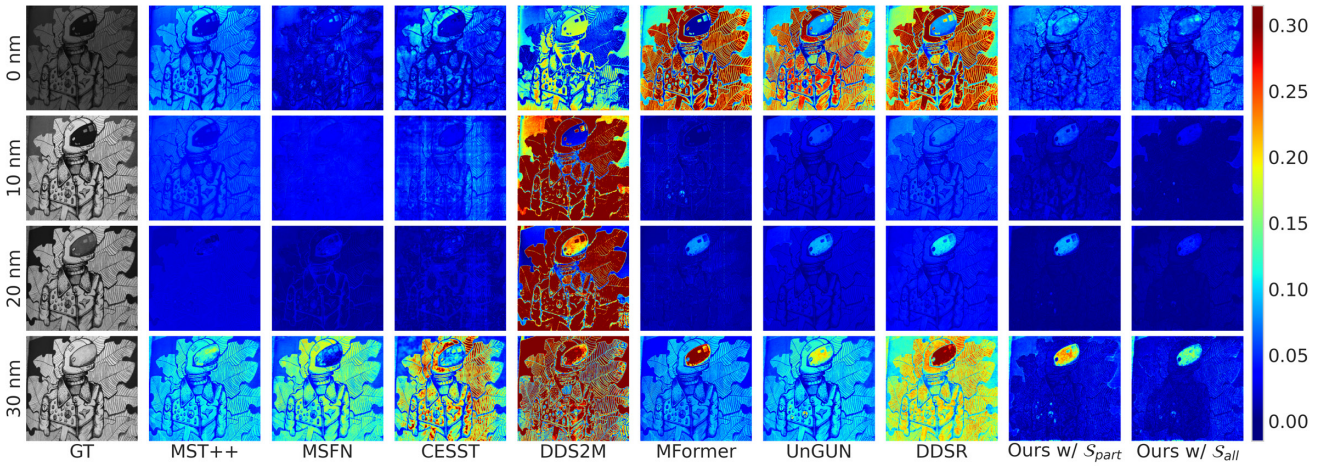

Figure 25. Illustration of error maps of SR results on the NTIRE22 dataset. Four different bands are shown, including 400 nm, 500 nm, 600 nm, and 700 nm. The first column is the ground truth, and the rest are the error maps of the proposed method and alternative methods.

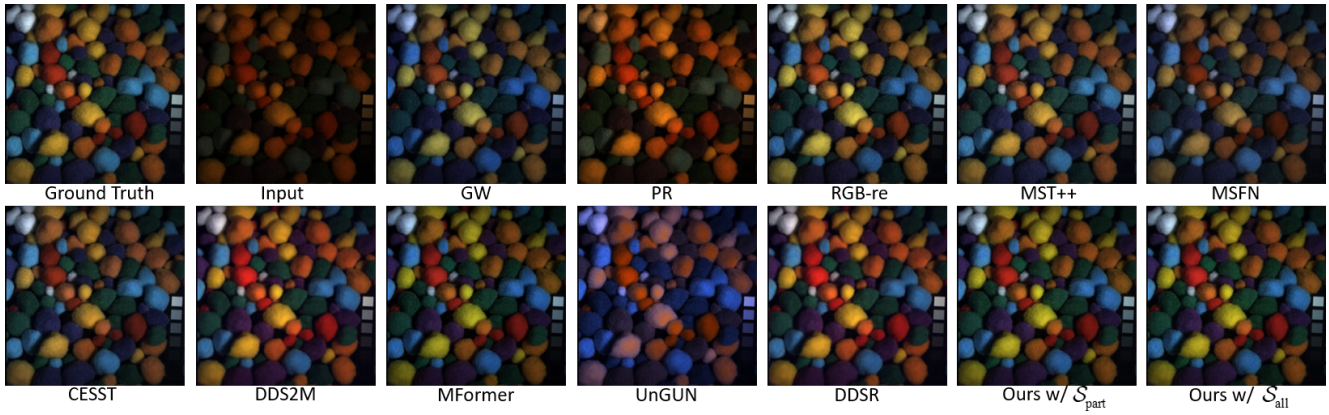

Figure 26. Illustration of the ground truth and relighting results under CIE A illuminant on the CAVE dataset.

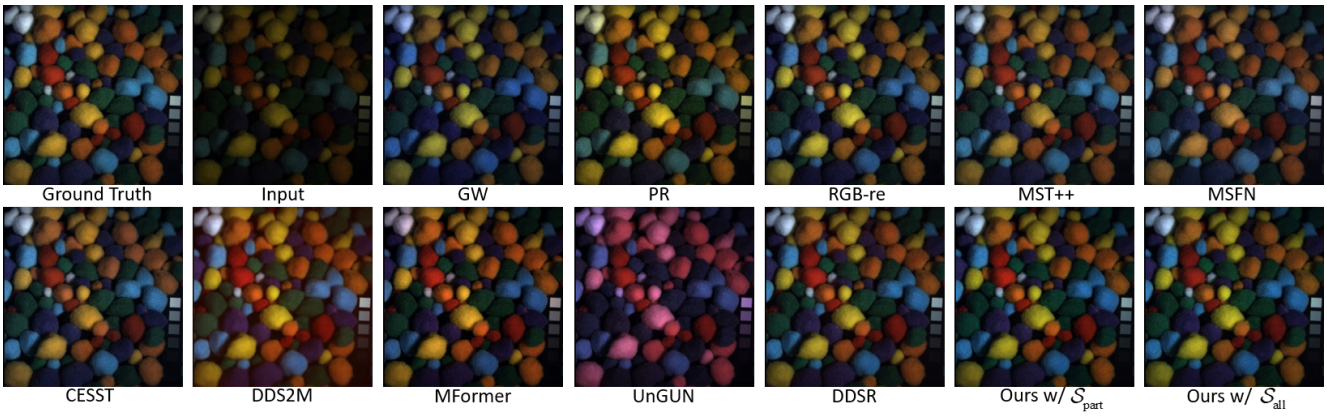

Figure 27. Illustration of the ground truth and relighting results under CIE F6 illuminant on the CAVE dataset.

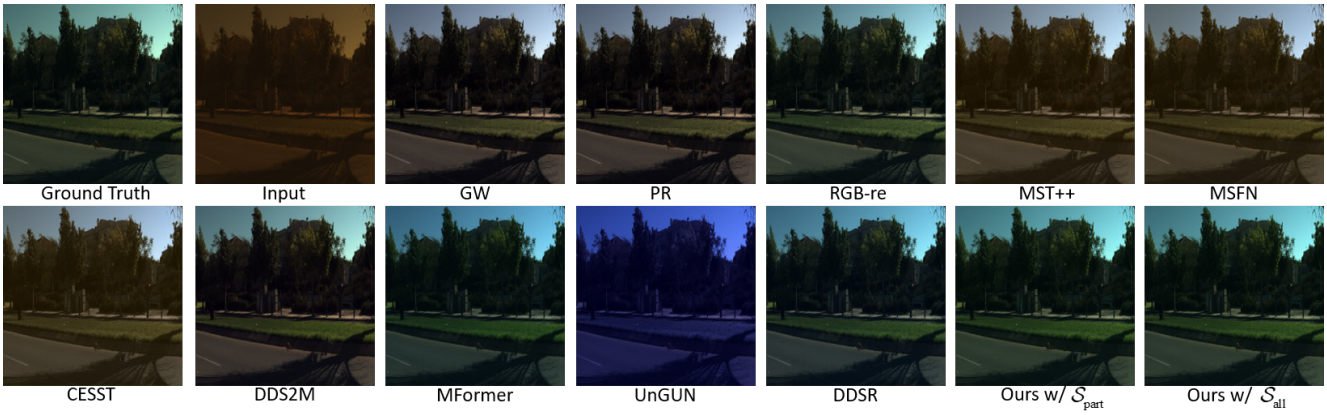

Figure 28. Illustration of the ground truth and relighting results under CIE A illuminant on the ICVL dataset.

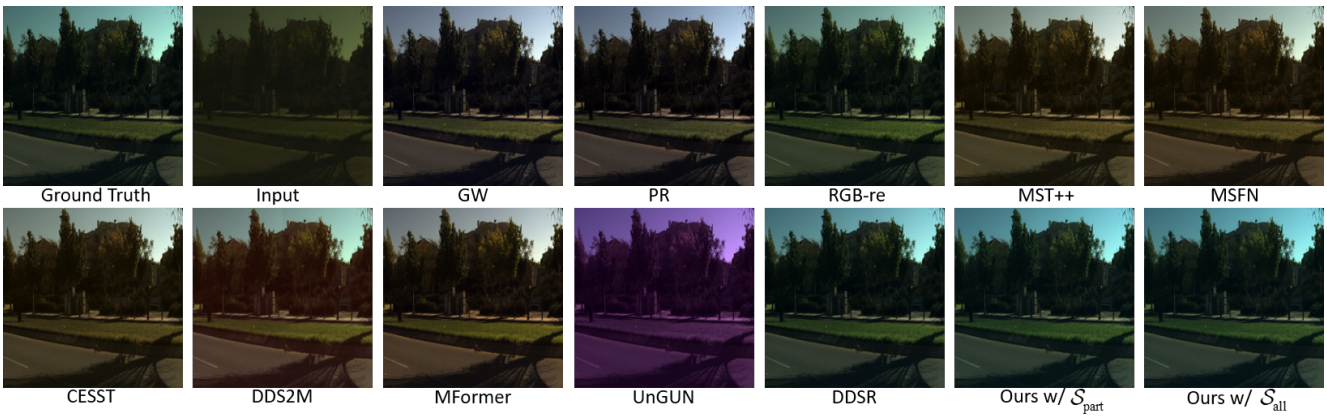

Figure 29. Illustration of the ground truth and relighting results under CIE F6 illuminant on the ICVL dataset.

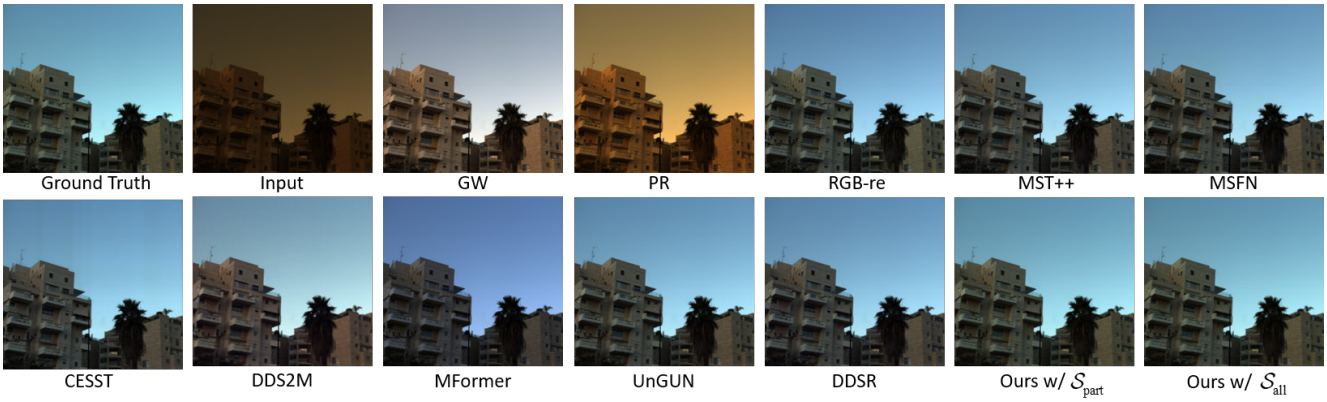

Figure 30. Illustration of the ground truth and relighting results under CIE A1 illuminant on the NTIRE22 dataset.

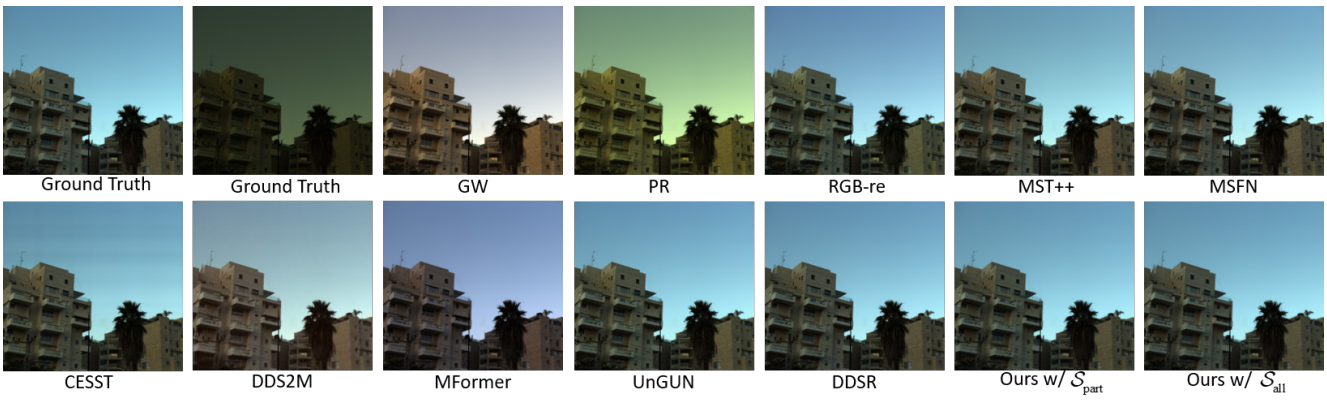

Figure 31. Illustration of the ground truth and relighting results under CIE F6 illuminant on the NTIRE22 dataset.