Semantic-embedded Unsupervised Spectral Reconstruction from Single RGB Images in the Wild (Supplementary Materials)

Zhiyu Zhu[1], Hui Liu[1], Junhui Hou[1]*, Huanqiang Zeng[2], Qingfu Zhang[1]
[1] City University of Hong Kong, [2] Huaqiao University
zhiyuzhu2-c@my.cityu.edu.hk, hliu99-c@my.cityu.edu.hk, jh.hou@cityu.edu.hk,
zeng0043@hqu.edu.cn, qingfu.zhang@cityu.edu.hk.

In this document, we provide the supplementary information for the ICCV 2021 submission titled with "Semantic-embedded Unsupervised Spectral Reconstruction from Single RGB Images in the Wild," i.e., the detailed network architectures of all modules of our framework. Besides, we also submit the **source code** of our method and **video demo** as parts of the supplementary material.

1. HS Image Generation Network

Figure 1 shows the network architecture of the proposed HS image generation network, where we also propose a spectral zero-mean normalization layer (denoted as **SZM**) to regularize the intermediate feature maps $\mathbf{B} \in \mathbb{R}^{C \times HW}$ as

$$\widehat{\mathbf{B}}(i) = \mathbf{B}(i) - \overline{\mathbf{B}(i)}, \ \ (1 \leq i \leq HW)$$

where $\mathbf{B}(i) \in \mathbb{R}^{C \times 1}$ denotes the i-th column of the matrix; $\overline{\mathbf{B}(i)}$ is the mean value of $\mathbf{B}(i)$. We utilized such a layer because the $\mathsf{G}^{(t)}(\cdot)$ $(1 \leq t \leq N-1)$ mainly aims to reconstruct the high-frequency spectrum details to refine the coarse HS estimation.

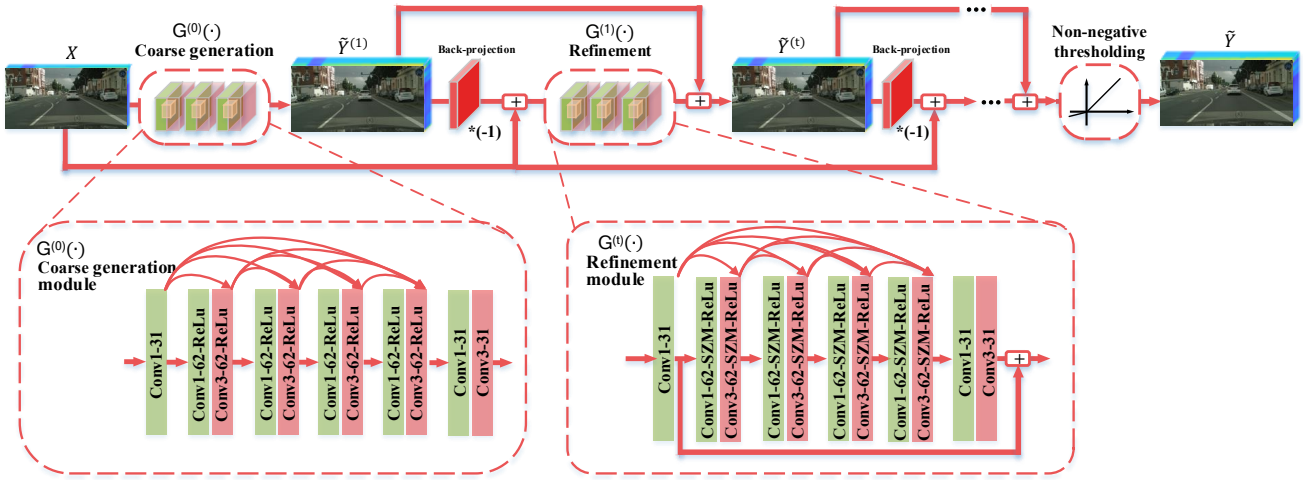

Figure 1. The detailed architecture of the proposed HS image generation network, where "Conv1-N" and "Conv3-N" denote the convolutional layer with the kernel size of 1×1 groups of 1 and 3×3 groups of 31, respectively; N is the number of output channels ; "ReLu" refers to the activation function—Rectified Linear Unit.

*Corresponding author. This work was supported by the Hong Kong Research Grants Council under grant CityU 11219019.

2. HS Image Discriminator

Figure 2 shows the network architecture of the HS image discriminator, which contains 3 stages: the first two stages are designed for feature extraction, and the final stage classifies the True or Fake.

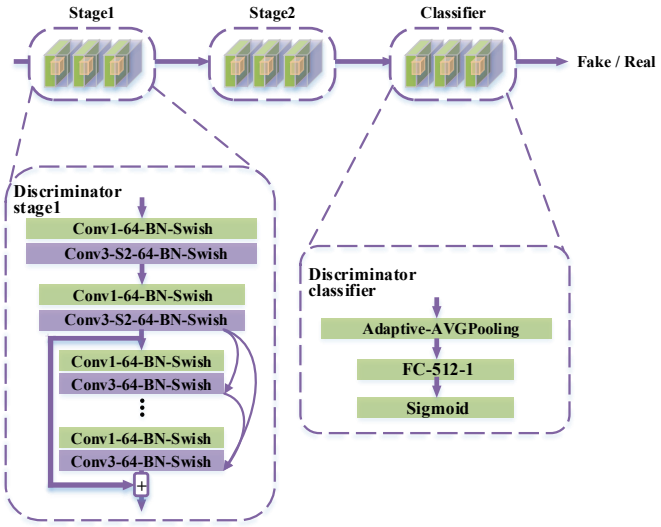

Figure 2. The detailed network architecture of the proposed HS image discriminator, where "swish" is an activation function [1]; "S2" denotes the convolutional layer with stride of 2; there are 7 densely-connected convolutional blocks in each stage; the "Stage2" has the same architecture as "Stage1" but all the output feature channels are changed to 128; in the classifier, we adopt adaptive-average pooling to change the spatial shape of features to 2×2 and reshape it to a vector of length 512; "FC-512-1" denotes the fully-connected layer, which accepts input with 512 channels and outputs 1 channel.

3. Semantic-embedded Regularization Network

Figure 3 shows the detailed architecture of our semantic-embedded regularization network, where $\mathsf{E}_1(\cdot)$ encodes spatial features from RGB images, $\mathsf{E}_2(\cdot)$ learns spectral information from reconstructed HS images, and $\mathsf{SE}(\cdot)$ finally comprehensively explores the spatial-spectral information.

4. SRF Weight Learning Sub-network

Figure 4 shows the detailed architecture of the sub-network for learning SRF weights, which consists of 5 stages. The first two stages utilize convolutional layers with kernel size of 5×5 for a large receptive field. We progressively reduce its spatial resolution with a ratio of 0.5 at each stage and the spatial resolution decreases to 1 in the final stage.

References

[1] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[2] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[3] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019.

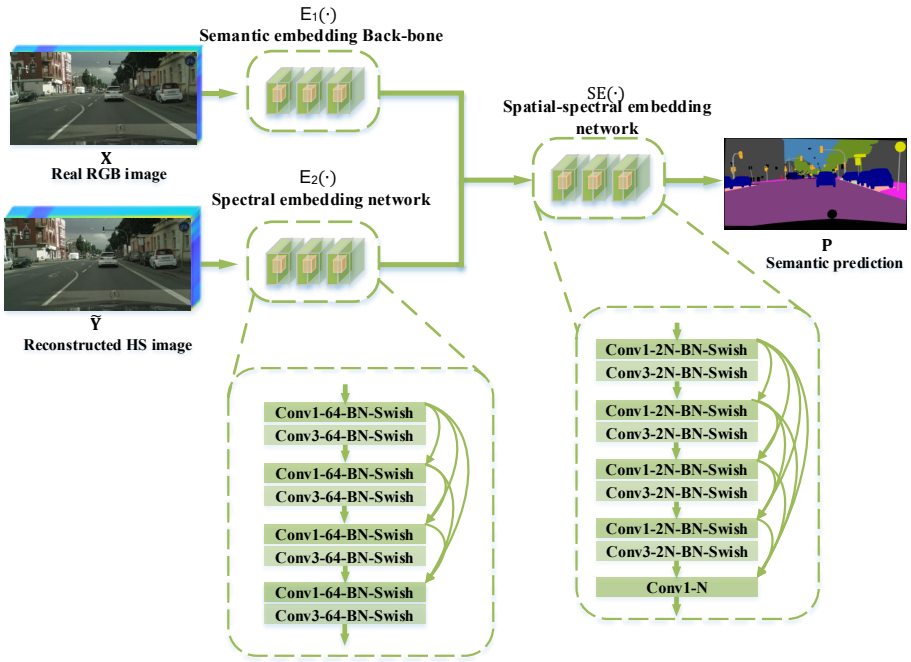

Figure 3. The detailed architecture of the proposed semantic-embedded regularization network, where the semantic embedding network $\mathsf{E}_1(\cdot)$ denotes the HRNetV2-W48 [2, 3]; N in $\mathsf{SE}(\cdot)$ is the number of categories, e.g., 59 for PASCAL and 19 for Cityscapes.

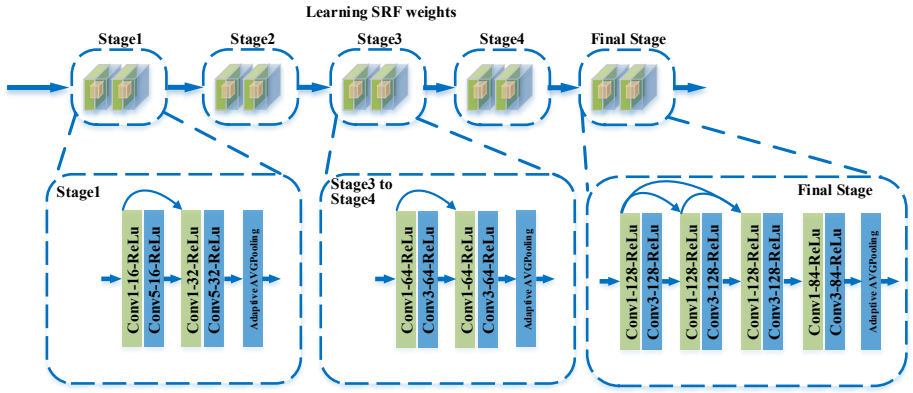

Figure 4. The detailed architecture of the SRF weight learning sub-network, where each "Adaptive AVGPooling layer" in the first four stages shrinks the spatial resolution as half of input of each stage, while the final one pools the feature maps to 1×1.