Image Harmonization with Attention-based Foreground-background Feature Map Modulation

Guoqing Hao, Satoshi Iizuka, Kazuhiro Fukui

Graduate School of Systems and Information Engineering, University of Tsukuba, Japan

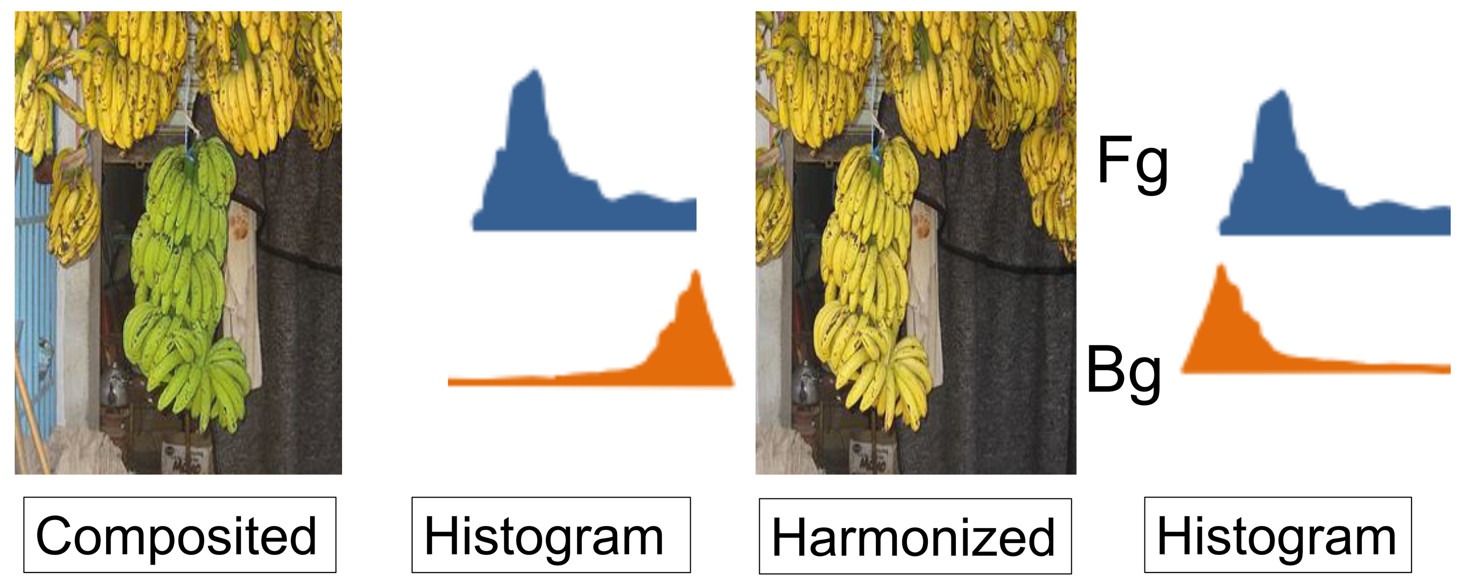
Overview

1. We present an <u>attention-based deep feature modulation</u> layer, which allows <u>modulating</u> the feature map of foreground <u>according to</u> those of similarity-weighted background, to <u>improve realism</u> of composites
2. Experimental results on the image harmonization dataset and real composite images show that our method <u>outperforms</u> existing methods both <u>quantitatively</u> and <u>qualitatively</u>

Background

1. <u>Inharmonious appearance</u> between foreground and background <u>degrades</u> quality of composite images
2. <u>Image harmonization</u> is often conducted <u>manually</u> by experts and requires a <u>significant amount</u> of time
3. Our goal is to generate realistic composites <u>automatically</u>

Background



Our work is motivated by the fact that the specific image statistics such as <u>histogram of luminance</u> between foreground and background <u>typically matches</u> in realistic composite images

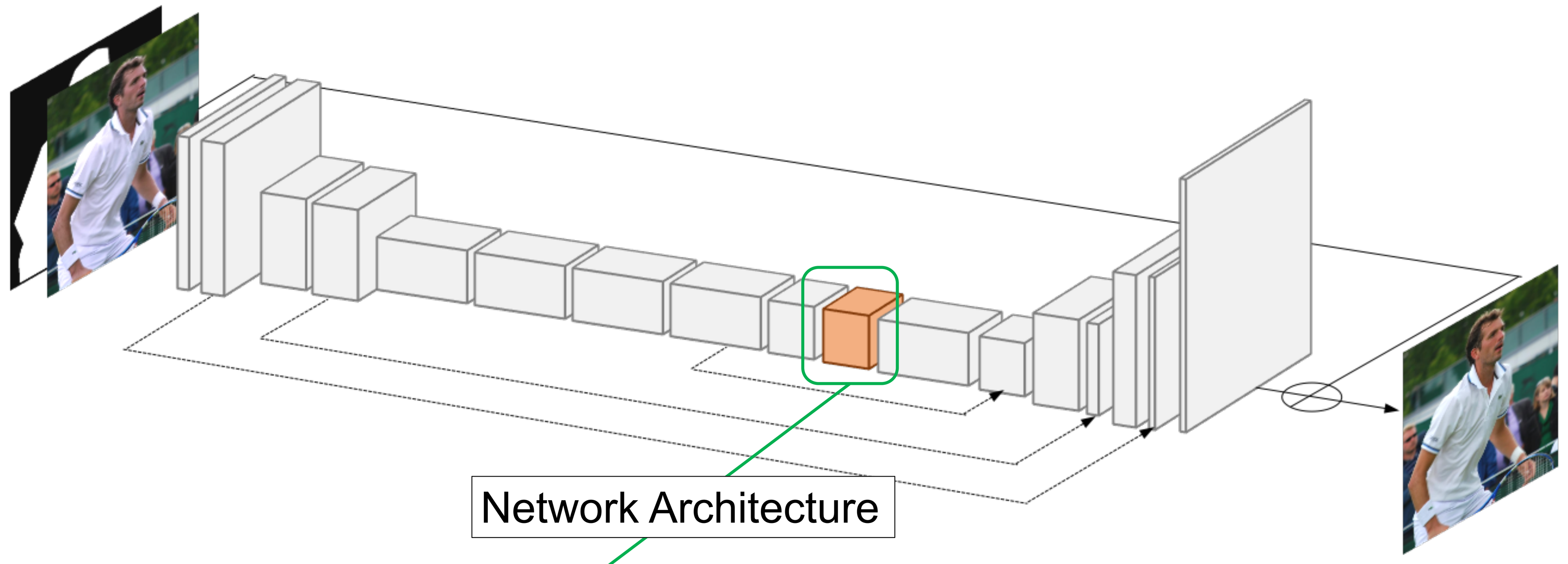
References

[1]. X. Cun and C. Pun. Improving the harmony of the composite image by spatial-separated attention module. IEEE Transactions on Image Processing. 2020

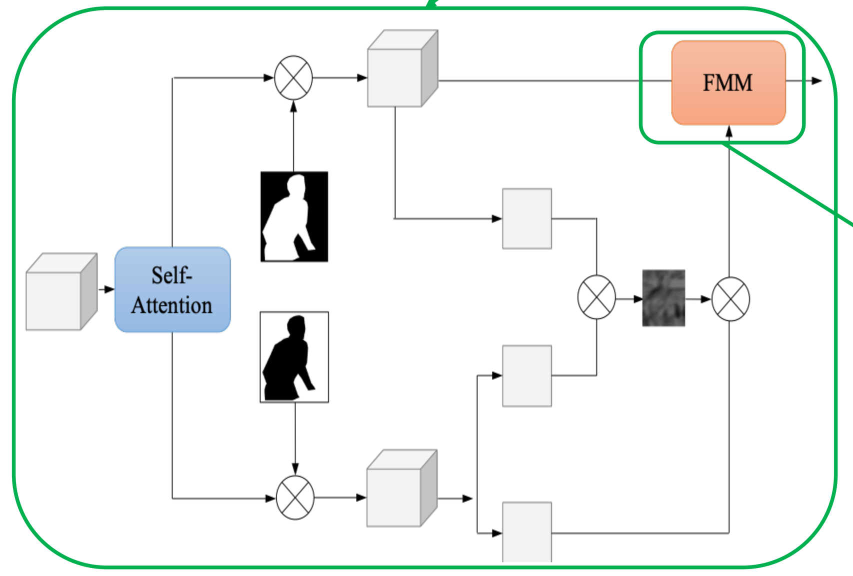
[2]. W. Cong, J. Zhang, L. Niu, L. Liu, Z. Ling, W. Li, and L. Zhang. Dovenet: Deep image harmonization via domain verification. In CVPR, 2020.

[3]. Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In ICML, 2019

Our Method



Network Architecture



Proposed Module

$$FMM(h_f, h_b^a) = \gamma \times \left(h_f \times \frac{\sigma_{h_b^a}}{\sigma_{h_f}}\right)$$

γ: learned parameter

h_f: foreground features

h_b^a: similarity-weighted background features

σ: standard deviation

Network:
- formed exclusively by convolutional layers

Proposed Module:
- adjust <u>high-level feature statistics</u> of foreground <u>according</u> to those of background
- capture <u>non-local dependencies</u> between foreground and background
- trained in an <u>end-to-end</u> fashion
- easily inserted into any convolutional neural networks with only a <u>small additional computational cost</u>

Results

Comparisons between our method against existing methods

Methods	PSNR	MSE
S^2AM [1]	34.35	59.67
DoveNet [2]	34.76	52.33
Ours	<u>**35.86**</u>	<u>**30.37**</u>

Ablation results of our proposed module. The "baseline" stands for the backbone net-work in our full method. The "A" stands for "remove self-attention block from our proposed module"

Methods	PSNR
Baseline	32.98
Baseline + Self-attention [3]	35.06
Baseline + A	35.17
Baseline + proposed module	<u>**35.86**</u>