WHFL: Wavelet-Domain High Frequency Loss for Sketch-to-Image Translation - Supplementary Material

Min Woo Kim, Nam Ik Cho
Department of ECE, INMC, Seoul National University, Seoul, Korea
{mwk0614, nicho}@snu.ac.kr

A. Dynamic and Adaptive Properties of High-Frequency Weight Matrix

Unlike the static matrix (*e.g.*, Laplacian operator), which imposes the constant value to the same spectral position, the weights of the high-frequency weight matrix (*HFWM*) are updated dynamically in the training progress. As shown in Figure 1, the distribution of the weights is adjusted adaptively to the outputs over training time. Hence, *HFWM* could dynamically impose weights on the frequencies, even in the fluctuating training condition. Also, to confirm the effect of *HFWM* on performance, we experiment in two settings with *ShoeV2* [9]; the case of using *WHFL* to the CycleGAN [10] baseline and the case of replacing *HFWM* of *WHFL* with the Laplacian operator. Consequently, we can observe that *HFWM* improves the performance, as shown in Table 1.

B. Quantitative Ablation Study

We conduct ablation studies to analyze how the proposed components, *HFWM* and multi-scale decomposition scheme, affect the evaluation metrics with three network baselines. We adopt five experimental settings, as illustrated in Table 2. With full *WHFL* (Table 2(d)), quantitative results are enhanced compared to the baseline with FFL [3] (Table 2(a)) in most cases. Also, we can observe that the performance is improved more consistently with *HFWM* (Table 2(b)) than with the multi-scale scheme (Table 2(c)). Therefore, we conclude that *HFWM* is a more crucial component in *WHFL*. Besides, since both metrics tend to be improved evenly with the multi-scale method, it seems that our scheme properly complements *HFWM*.

Table 1. Performance comparison between two settings. (a) indicates the result when *WHFL* is used, and (b) shows the case of replacing *HFWM* of *WHFL* to the Laplacian operator.

	(a)	(b)
FID ↓	**56.354**	76.786
IS ↑	**2.756±0.300**	2.593±0.336

Table 2. Quantitative results for ablation studies of five experimental settings. We check the performance on the network baseline, (a) FFL [3] (baseline+FFL), (b) replacing the weight matrix of FFL to *HFWM* (baseline+FFL+HFWM), (c) applying multi-scale decomposition scheme to FFL (baseline+FFL+Multi-Scale), and (d) our full *WHFL* (baseline+FFL+HFWM+Multi-Scale).

	Pix2Pix [2]		CycleGAN [10]		MUNIT [1]	
	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑
baseline	63.580	2.505±0.175	60.138	**2.787±0.354**	110.252	2.797±0.278
(a)	62.194	2.614±0.335	56.870	2.669±0.346	107.314	2.784±0.470
(b)	63.062	**2.652±0.262**	**54.958**	2.671±0.361	103.706	2.809±0.291
(c)	65.784	2.626±0.181	65.954	2.764±0.449	104.463	2.659±0.272
(d)	**61.744**	2.622±0.202	56.354	2.756±0.300	**102.203**	**2.925±0.344**

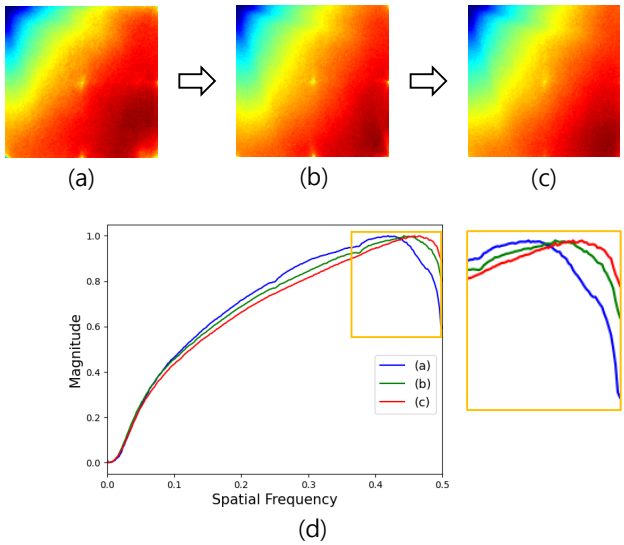

Figure 1. From (a) to (c), they visualize the sequential change of high-frequency weight matrix (*HFWM*) over training time (the upper left corner is (0,0) and the lower right corner is (π, π)). Note that we use a *jet colormap* where the reddish colors represent high values, and the blue ones indicate the lower. (d) plots the averaged magnitude of the weights along the diagonal direction (0.5 corresponds to π), and the figure on the right is an enlarged view of the yellow box in the left plot. Also, all graphs in (d) are normalized to [0,1].

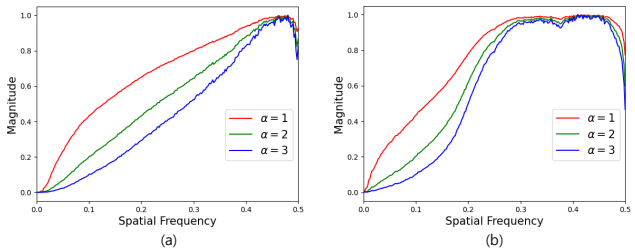

Figure 2. Plot of the averaged magnitude of the weights along the diagonal direction from $(0, 0)$ to (π, π) according to the weight control factor for (a) *ShoeV2* and (b) *edges2shoes*.

C. Ablation Studies for Weight Control Factor

To determine which weight control factor is appropriate, we experiment with three cases of $\alpha = 1, 2, 3$. As shown in Figure 2, we find that as α increases, the weights tend to be concentrated in the narrow frequency bands, which is not desirable. Although we confirm no meaningful difference in performance according to the factors, we set $\alpha = 1$ based on the above observation and Jiang *et al*. [3].

D. Spectral Bias of Traditional Methods

Generative models that adopt convolutional neural networks tend to learn low frequency in a biased way. As shown in Figure 3, when we calculate the absolute value of the difference between the averaged log-magnitudes of DFT from real and fake images, we can observe that the values at high frequencies are larger than those at low ones.

E. Network Details

E.1. Pix2Pix

Pix2Pix [2] was proposed for image-to-image translation tasks for general purposes. The framework consists of a generator adopting U-Net [6] and a patch-based discriminator. For the objective functions, the conditional GAN loss and L1 loss between the generated and real images are used to train the network. Therefore, to complement a loss function defined in the spatial domain, we apply *WHFL* to L1 loss.

E.2. CycleGAN

CycleGAN [10] learns two mapping networks between domains with the unpaired dataset. One generator translates an image from domain A to domain B, and the other does it in the opposite direction. The generators use the architecture suggested in [4], and the discriminators adopt PatchGAN [2]. The cycle consistency loss and adversarial loss [5] are utilized for the objective functions. The cycle consistency loss penalizes the difference between outputs and real images using L1 loss in both directions, and *WHFL* complements the loss.

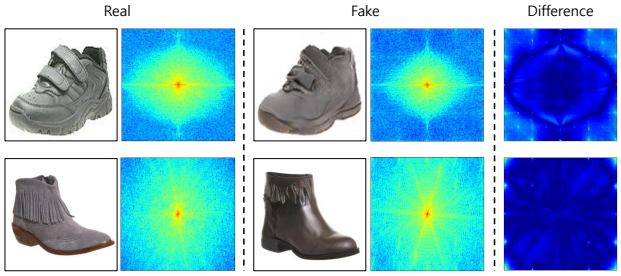

Figure 3. Real and fake images with the corresponding log-magnitudes of DFT, and the difference of the log-magnitudes. Note that the log-magnitudes are averaged on multiple samples. The first row displays *edges2shoes*, and the second *ShoeV2*.

E.3. MUNIT

MUNIT [1] is an unsupervised framework that can translate the image from a source domain to multiple target domains. The framework uses an approach to decompose an image into content and style code. Therefore, the framework comprises encoders that map an image to the content and style code and a decoder generating an image with the codes. Reconstruction losses and adversarial loss are used as objective functions. L1 loss is adopted to penalize reconstructions of the image and latent codes (content and style). Moreover, the adversarial loss [5] utilizes the decoded and real images from the target domain. We select the reconstruction loss for an image to *WHFL* to be applied.

F. Dataset Details

F.1. Edges2shoes

For a paired dataset, we choose edges2shoes [8] in which an edge map and the corresponding photo are coupled. The photos are binarized to edge maps by HED [7] and post-processing. The training and test sets consist of 49,825 and 200 images, respectively.

F.2. ShoeV2

To train CycleGAN and MUNIT, we exploit ShoeV2 [9]. The dataset comprises photos and free-hand sketches, but the images are not fed to the network in pairs during training. The number of images in the training and test sets is 6,648 and 2,000. Besides, we use the same number of sketches and photos during the test.

G. Additional Visualization

We display the comparisons between the weight matrix in FFL [3] and *HFWM* (Figure 4). Also, we show more results of Pix2Pix (Figure 5), CycleGAN (Figure 6), and MUNIT (Figure 7). We also exhibit edge maps extracted by our *HFWM* for training samples of edges2shoes (Figure 8)

and ShoeV2 (Figure 9). Finally, we specify the limitation by visualizing flat regions of failure cases (Figure 10).

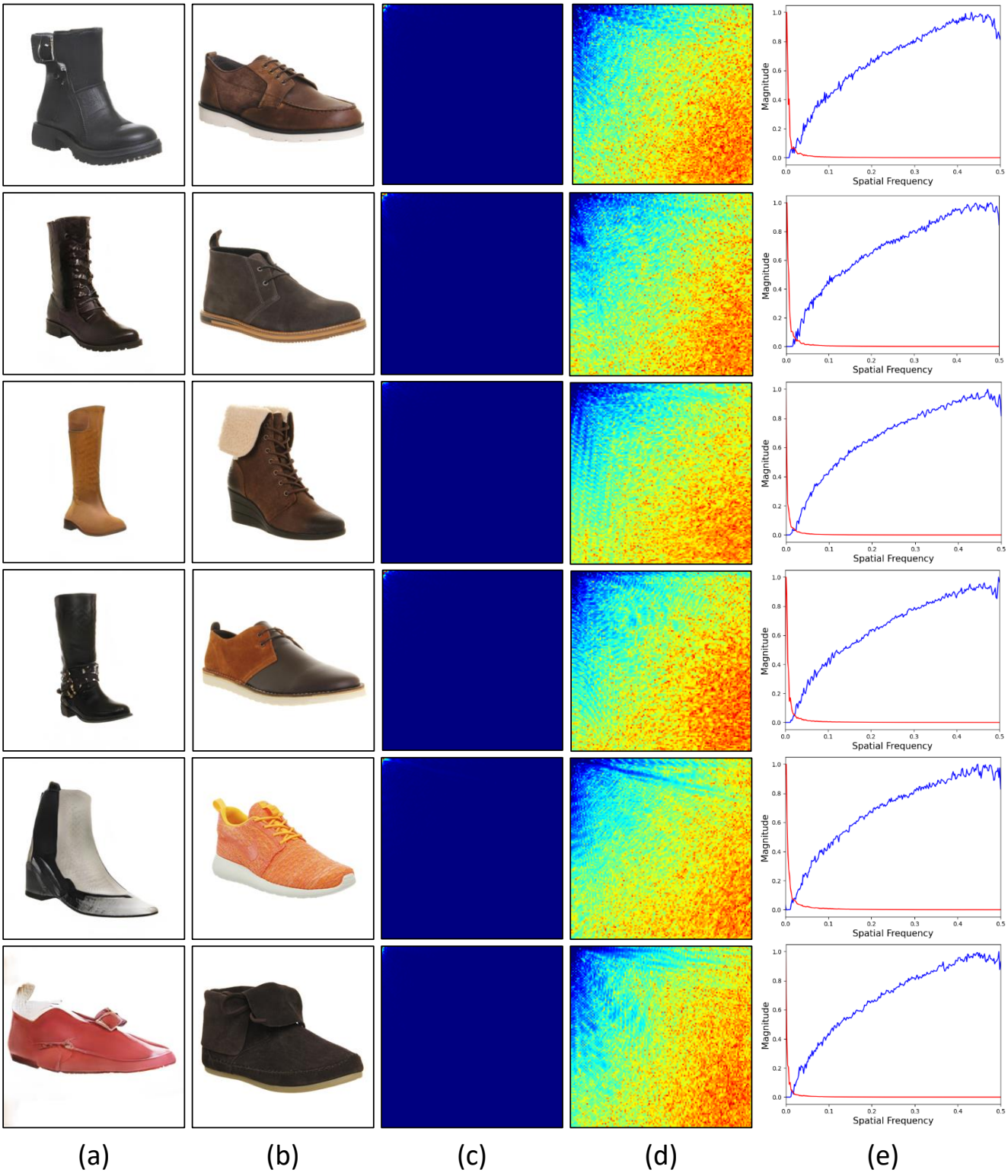

Figure 4. More visualizations for the comparison between the weight matrix in FFL [3] and *HFWM*, which consists of (a) fake image, (b) real image, (c) the weight matrix in FFL, (d) *HFWM*, and (e) the graph plotting the averaged magnitude of the weights along the diagonal direction from $(0, 0)$ to (π, π) (the red line means (c) and the blue one indicates (d)). Note that an output image and ground-truth do not need to be paired in CycleGAN [10] and MUNIT [1].

Figure 5. Additional visualization for Pix2Pix [2]

Figure 6. Additional visualization for CycleGAN [10]

Figure 7. Additional visualization for MUNIT [1]

Figure 8. Additional visualization for edge maps extracted by *HFWM* for training samples of *edges2shoes*.

Figure 9. Additional visualization for edge maps extracted by *HFWM* for training samples of *ShoeV2*.

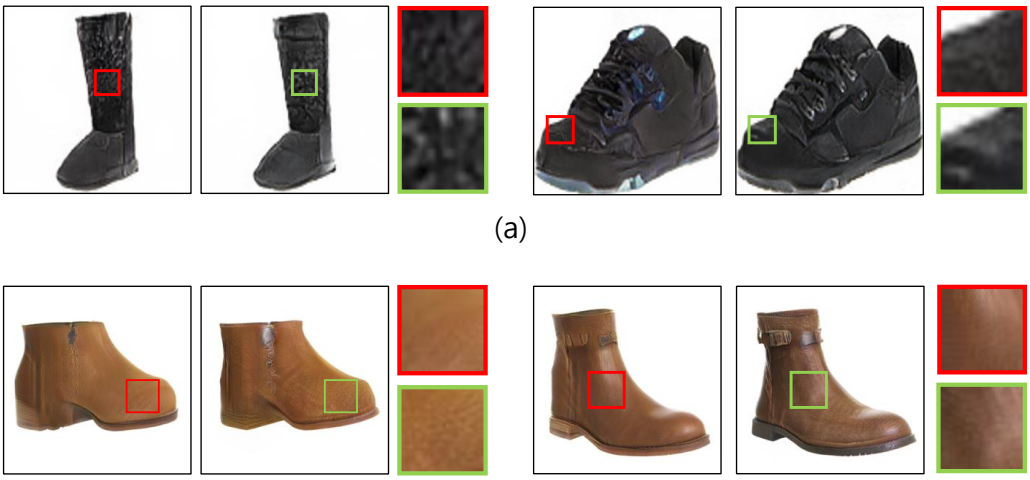

Figure 10. Visualization for failure cases (*i.e.*, flat regions) for (a) *edges2shoes* and (b) *ShoeV2*. There is no significant improvement in the texture between the fake images from the vanilla method (red) and *WHFL* applied (green).

References

[1] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[3] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13919–13929, 2021.

[4] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[5] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[7] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

[8] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 192–199, 2014.

[9] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 799–807, 2016.

[10] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.