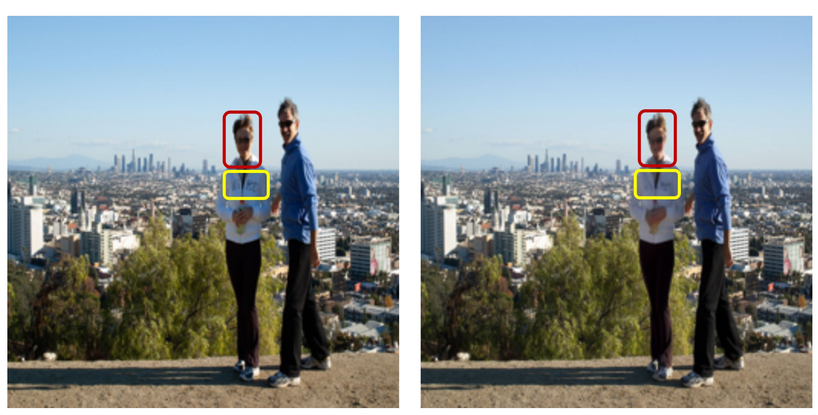
DiffHarmony: Latent Diffusion Model Meets Image Harmonization

Pengfei Zhou, Fangxiang Feng, Xiaojie Wang
Beijing University of Posts & Telecommunications

ICMR 2024
June 10-14, 2024
Dusit Thani Laguna Phuket, Thailand

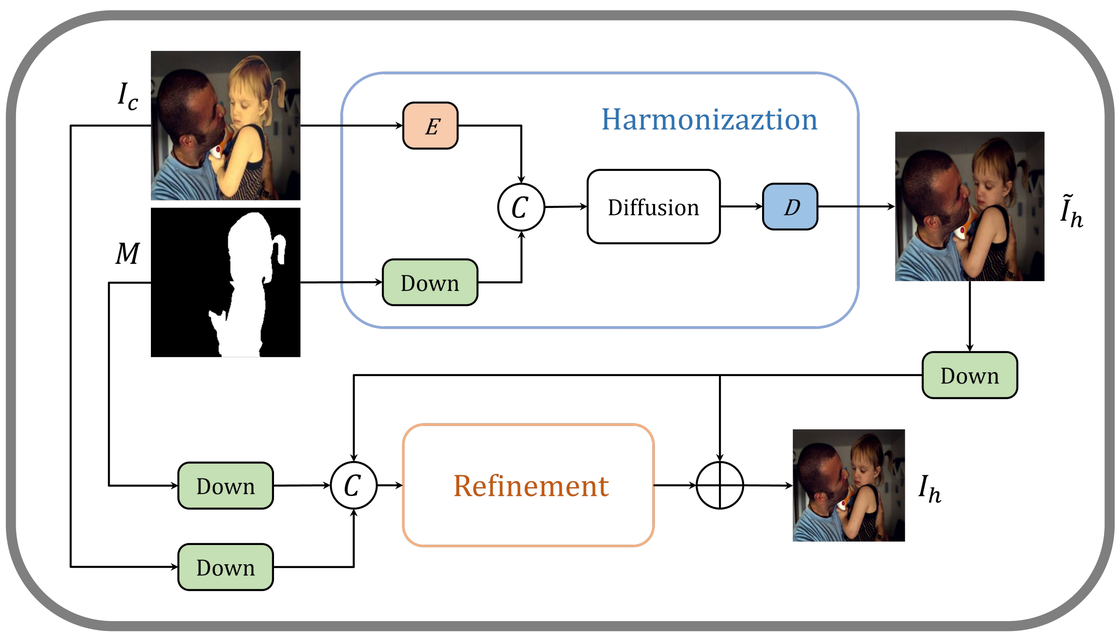
Introduction

- Motivation: leverage pretrained diffusion model to do image harmonization task.
- Problem: severe distortion in reconstructed/decoded image when using latent diffusion model.



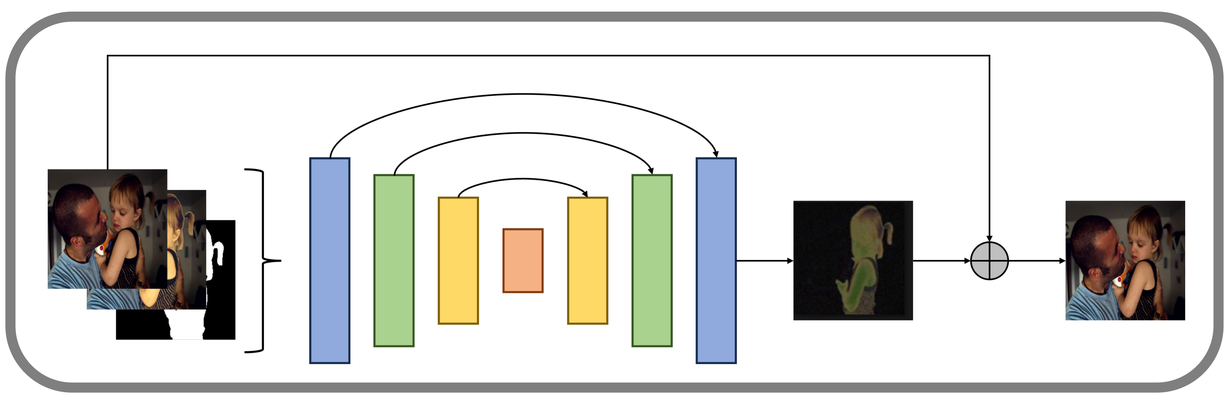
Method

■ Overall illustration



- Additional input I_c and M to Stable Diffusion.
- Give null string as text input.
- Use higher resolution in training and inference to alleviate distortion.
- Refinement module further improve image quality.

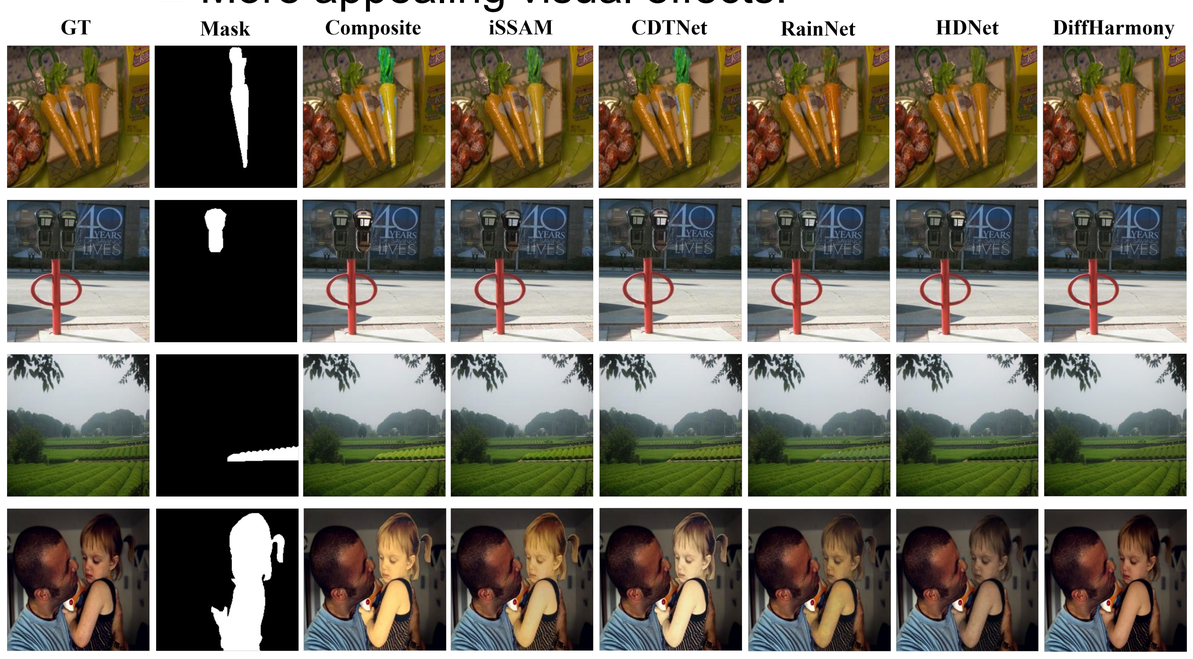
■ Details of refinement module



- U-Net architecture internally.
- Only learn the residual to facilitate convergence.
- Operating on lower (evaluation ready) resolution.

Results

■ More appealing visual effects.



■ SOTA qualitative performance.

Dataset	Metric	Composite	DIH[3]	S^2AM[24]	DoveNet[4]	BargainNet[25]	Intrinsic[26]	RainNet[27]	iS^2AM[7]	D-HT[6]	SCS-Co[28]	HDNet[10]	Li[19] *et al.*	Ours
HCOCO	PSNR↑	33.94	34.69	35.47	35.83	37.03	37.16	37.08	39.16	38.76	39.88	41.04	34.33	41.25
	MSE↓	69.37	51.85	41.07	36.72	24.84	24.92	29.52	16.48	16.89	13.58	11.60	59.55	9.22
	fMSE↓	996.59	798.99	542.06	551.01	397.85	416.38	501.17	266.19	299.30	245.54	-	-	153.60
HAdobe5k	PSNR↑	28.16	32.28	33.77	34.34	35.34	35.20	36.22	38.08	36.88	38.29	41.17	33.18	40.29
	MSE↓	345.54	92.65	63.40	52.32	39.94	43.02	43.35	21.88	38.53	21.01	13.58	161.36	17.78
	fMSE↓	2051.61	593.03	404.62	380.39	279.66	284.21	317.55	173.96	265.11	165.48	-	-	107.04
HFlickr	PSNR↑	28.32	29.55	30.03	30.21	31.34	31.34	31.64	33.56	33.13	34.22	35.81	29.21	36.99
	MSE↓	264.35	163.38	143.45	133.14	97.32	105.13	110.59	69.97	74.51	55.83	47.39	224.05	29.68
	fMSE↓	1574.37	1099.13	785.65	827.03	698.40	716.60	688.40	443.65	515.45	393.72	-	-	199.59
Hday2night	PSNR↑	34.01	34.62	34.50	35.27	35.67	35.69	34.83	37.72	37.10	37.83	38.85	34.08	38.35
	MSE↓	109.65	82.34	76.61	51.95	50.98	55.53	57.40	40.59	53.01	41.75	31.97	122.41	24.94
	fMSE↓	1409.98	1129.40	989.07	1075.71	835.63	797.04	916.48	590.97	704.42	606.80	-	-	502.40
Average	PSNR↑	31.63	33.41	34.35	34.76	35.88	35.90	36.12	38.19	37.55	38.75	40.46	32.70	40.44
	MSE↓	172.47	76.77	59.67	52.33	37.82	38.71	40.29	24.44	30.30	21.33	16.55	141.84	14.29
	fMSE↓	1376.42	773.18	594.67	532.62	405.23	400.29	469.60	264.96	320.78	248.86	-	-	151.42

Table 1: Quantitative comparison across four sub-datasets of iHarmony4 and in general. Top two performance are shown in red and blue. ↑ means the higher the better, and ↓ means the lower the better.

■ Further analysis.

inf res	refine	PSNR↑	MSE↓	fMSE↓
512px	✘	37.65	26.14	290.66
512px	✔	39.47	19.59	205.07
1024px	✘	40.12	15.56	166.19
1024px	✔	40.44	14.29	151.42

Table 2: Ablation study on using different input resolution and w/wo refinement stage.

PSNR↑	MSE↓	fMSE↓
37.66 ± 0.02	25.44 ± 0.31	291.03 ± 2.08

Table 3: Randomness analysis. Our generative method can also produce stable harmonized results.

Model	0% ~ 5%	5% ~ 15%	15% ~ 100%
$HDNet_{512}$	**PSNR: 45.64** **MSE: 3.16** **fMSE: 143.93**	**PSNR: 39.97** **MSE: 11.33** fMSE: 129.87	PSNR: 34.59 MSE: 47.19 fMSE: 152.01
Ours	PSNR: 43.28 MSE: 4.46 fMSE: 173.10	PSNR: 39.55 MSE: 11.90 **fMSE: 126.69**	**PSNR: 34.80** **MSE: 40.47** **fMSE: 128.45**

Table 4: Comparison between $HDNet_{512}$ and our method. $HDNet_{512}$ is trained with 512px images, and the inputs are 1024px images during inference. This is exactly the same as the experimental setting of our method.

Conclusion

- The strategies we design to solve image distortion problem are proven to be effective.
- Our method shows great advantages when foreground area is large enough.
- Diffusion models are capable of doing low-level image processing task.