Supplementary material: Look here! A parametric learning based approach to redirect visual attention

Youssef A. Mejjati[1], Celso F. Gomez[2], Kwang In Kim[3], Eli Shechtman[2], and Zoya Bylinskii[2]

[1] University of Bath, UK
[2] Adobe Research
[3] UNIST

1 Parameter definition

GazeShiftNet implements the following parametric transformations: sharpening, exposure, contrast, tone and color curve adjustment, as defined below.

Sharpness: Given an input image I and the predicted sharpness parameter $p_1 \in [-2, 2]$, the output image is obtained by first computing image edges using the Sobel filters f_1 and f_2 as follows: $I_{edge} = \sqrt{(I*f_1)^2+(I*f_2)^2}$, where $*$ is the convolution operation, and f_1,f_2 are the filters: $f_1 = \frac{1}{8}\begin{bmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{bmatrix}$,$f_2 = \frac{1}{8}\begin{bmatrix}-1 & -2 & -1\\ 0 & 0 & 0\\ 1 & 2 & 1\end{bmatrix}$.
Finally, the output image I' is calculated as $I' = I + p_1 I_{edge} I$.

Exposure: Given an input image I and the predicted exposure parameter $p_2 \in [-3,3]$, the output image is computed as $I' = I\exp(p_2 * \log(2))$, as in [5].

Contrast: Given an input image I and the predicted contrast parameter $p_3 \in [-1,1]$, the output image is obtained by the linear interpolation: $I' = (1-p_3)I + p_3 I''$, where $I'' = I\frac{\frac{1}{2}(1-\cos(\pi I_{lum}))}{I_{lum}}$, and $I_{lum} == 0.27R + 0.67G + 0.06B$, as in [5].

Tone and Color adjustment: We define color and tone adjustment using monotonic and piece-wise curve representations, in the same way as [5]. The curve is represented using L different parameters, e.g. for tone adjustment $\mathbf{p}^t = \{p_0^t, p_1^t, ..., p_L^t\}$. In this case, $I' = \frac{1}{\sum_{i=0}^{L} p_i^t} \sum_{i=0}^{L-1} \text{clip}(L.I - i, 0, 1) p_i^t$.

For tone adjustment, we define the same set of L parameters ($\in [0,3]$) for R, G, B. While, for color adjustment three distinct sets of L parameters ($\in [0,3]$) are defined for R and G and B. We set $L = 8$ for all our experiments.

In addition to the parameters defined above we experimented with additional parametric transformations, including white balance, blur, saturation, and gamma correction. For white balance and gamma correction, we found that similar effects could be reproduced by a combination of our other simpler parameters. We decided to focus on a smaller but sufficient subset, to avoid unnecessarily bulking up our model. Regarding saturation, we found

that it can have destructive artifacts, as discussed in our ablation study below. We also tried using Gaussian blur, but this parametric transformation was not used by our model. We believe this is due to the lack of examples in our training set containing blur, and thus blur being penalized by the discriminator. One solution would be to collect a large dataset with large defocus blur or add artificial blur to the background of our training set in a realistic manner. Both these solutions can be explored as an interesting future work direction.

2 User studies

2.1 Exploratory study: professional edits

Our computational approach to attention-aware photo editing is motivated by real, professional workflows. For this study, we selected 30 high-resolution Adobe Stock photos covering a wide range of themes/genres and containing objects of varying sizes. Object masks were then manually created by segmenting an image region that was not already the main subject of the photograph, and was off-center to the image, so that we could later detect shifts in attention. We ran initial studies on the crowdsourcing platform www.usertesting.com, asking participants familiar with professional photo editing software to load the provided images and masks into Adobe's Photoshop, and edit the images to make the selected objects "stand out (become more noticeable)" using any technique but encouraging "more subtle edits". Participants were presented with 3 image-mask pairs, and submitted their 3 final edits. Results from these studies were manually filtered out if they contained very strong and obvious effects, if they showed no signs of edits, or if they reduced the prominence of the masked object, thus failing to follow instructions. We ran studies until we obtained a total of 5 different valid edits per photo. We collected a total of 150 professional edits (5 participants $\times$ 30 photos, see some examples in Fig. 1). We call this the HighResClutter dataset.

2.2 Fidelity study

In this study, participants were presented with pairs of images, an original photograph and an edited photograph (using one of multiple automated methods), and were asked to rate "Compared to the image on the left, how edited/manipulated/photoshopped does this image look?": "Not", "Slightly", "Moderately", or "Highly" (Fig. 2). Amazon's Mechanical Turk (MTurk) participants were recruited for the study, and assigned either 32 randomly selected images out of the 64 Mechrez dataset images, 25 randomly selected images out of the 50 CoCoClutter dataset images, or all 30 of the HighResClutter images. In all cases, participants were presented with an additional 5 sentinel images (the same for all datasets) which were randomly interspersed throughout the experiment. These sentinel images were chosen to be significantly different from the original images (as in the example in Fig. 2), and were used as a quality filter for participant data. The study took approximately 5 minutes to complete, and participants were paid $1 for their time. We kept track of screen size used, time spent per image pair, and selections made. Participants were filtered out if screen size was less than 1000×1000, if screen size was adjusted at any point throughout the experiment, if less than 2 seconds on average were spent per image pair, and if any of the sentinels were rated as not edited. These automatic criteria filtered out about 15% of the participant data, leaving an average of 25 participant responses per image pair, that

Fig. 1: Professional edits sourced from usertesting.com for the HighResClutter dataset.

we averaged together to obtain a fidelity score per image and automatic method. For the Mechrez dataset, we ran this study for each of OUR, MEC, HAG, and HOR models. For the CoCoClutter dataset, we ran this study for each of OUR, HAG, and HOR models. For the HighResClutter images, we ran the study on OUR, and 5 individual professional edits.

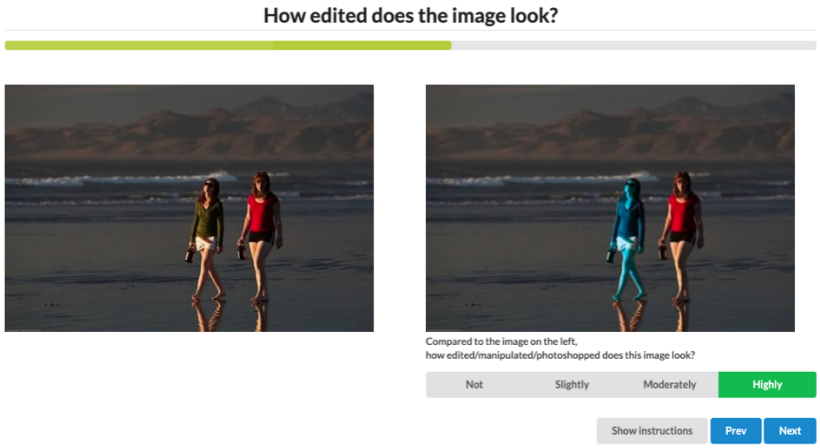

Fig. 2: Given a pair of images, an original and another variant, participants rated how edited the right image looked compared to the original.

2.3 Realism study

Participants were presented with a sequence of images and were asked to judge "Has the image been edited?", with possible answers: "Definitely not edited", "Probably not edited", "Probably edited", "Definitely edited" (Fig. 3). Half of the images in the sequence were edited, and the other half of the images were originals. Images were randomly sampled from either the Mechrez, CoCoClutter, or HighResClutter datasets (but all images in a single experiment came from a single dataset). All the images were randomly shuffled, and 5 sentinel images were randomly interspersed throughout the experiment. These sentinel images were chosen to be obviously edited. We collected an average of 30 MTurk participant responses for each image, which after automatic filtering, yielded an average of 20-25 raters per image. We filtered out participants who had screen sizes less than 1000×600, those who adjusted the screen size at any point throughout the experiment, spent less than 2 seconds per image on average, and rated any of the 5 sentinels as "Definitely not edited". Participants spent roughly 2-3 minutes on the task, and were compensated \$0.45-\$0.65 depending on the number of images shown. We tested the same images and models as in the fidelity study.

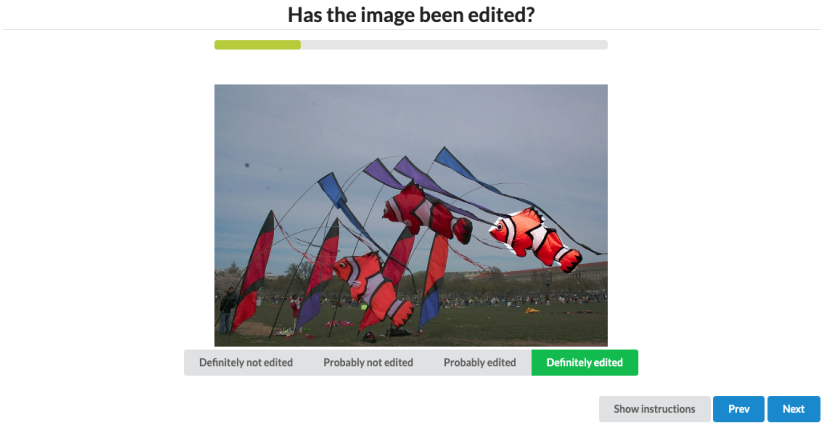

Fig. 3: Given a sequence of images, participants rated how edited an image looked. Half of the images shown to participants were originals, and the other half were edited images.

2.4 Attention study

We used the CodeCharts methodology [2][4] to collect ground truth attention data on all the original and edited images so that we could evaluate whether they successfully shifted viewer attention towards the desired image regions. Participants would be shown an image for 3 seconds, followed by a quickly-flashed chart of alphanumeric triplets (codechart), and they would be asked to report the last code they gazed at. This sequence

[4] Using the code provided at https://github.com/turkeyes/codecharts

repeated for many images in a row (30-60 in our experiments), and images were separated by fixation crosses to re-center participant gaze (Fig. 4). A total of 5 sentinels (validation trials), in the form of cropped faces against a white background, were spaced throughout the image sequence. If the code reported for a sentinel did not overlap with the location of the face, that was considered a failed trial. Any image was also classified as a failed trial if participants entered a code that was nowhere to be found on the corresponding codechart. Participant data was filtered out if any of the sentinels were a failed trial, and if invalid codes were entered on more than 25% of the other trials. A 6-image tutorial, including 3 natural images and 3 sentinel images preceded the rest of the experiment. Any participant that failed the tutorial, entering an invalid code for 2/6 images, was also filtered out. We filtered out a total of 15-20% of participants, depending on the experiment, based on these criteria. After filtering, an average of 45-50 gaze points remained per image, for further analysis. We collected attention data for the same images from the Mechrez, CoCoClutter, and HighResClutter datasets for which we collected realism and fidelity scores.

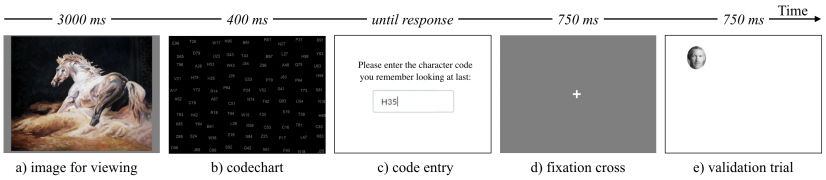

Fig. 4: CodeCharts user study experiment flow, adapted from [2].

2.5 Results

In Tables 1, 2, and 3 are results from the realism, fidelity, and attention user studies on the Mechrez, CoCoClutter, and HighResClutter datasets. Plots corresponding to these numbers for the Mechrez and CocoClutter studies can be found in the main paper, and the ones for the HighResClutter dataset are included in Fig. 5 here. We note that the differences across the methods were not found to be statistically significant across any of the measures (realism, fidelity, attention) due to the small number of images tested. However, the pattern of results shows that our approach consistently has small standard deviations of realism and fidelity scores, indicating that the produced results are more robust and consistent across different image types and datasets. Moreover, we achieve a balance between realism/fidelity and attention shift, whereas other methods trade off one for the other (this is easier to see from the scatter plots in the main paper than from the tables here). Surprisingly, compared to professionals we are able to shift attention more effectively, however this comes at the cost of reducing realism and fidelity. This suggests that the task we aim to solve is a challenging one even for humans, and that there is indeed a trade off between attention increase and realism.

3 Evaluations

3.1 Saliency model

We used a state-of-the-art saliency model [2] to evaluate the ability of each model to shift computational attention. At the time of submission, this model was second

Model	Realism	Fidelity	Attention increase		Similarity to mask	
			Absolute	Relative	WFB	CC
OUR	1.38 (0.58)	1.75 (0.52)	0.55	12.16	8.02	21.72
MEC [8]	1.47 (0.58)	1.34 (0.47)	0.42	12.22	7.40	19.30
HOR [7]	1.40 (0.70)	1.46 (0.74)	1.20	18.22	7.62	21.63
HAG [4]	1.46 (0.63)	1.86 (0.75)	-0.28	6.64	6.62	16.34

Table 1: Results from user studies on Mechrez dataset images. We include the average realism and fidelity scores (with standard deviations, in parentheses) across 25 human raters per image. All values are scaled by 100 apart from realism and fidelity.

Model	Realism	Fidelity	Attention increase		Similarity to mask	
			Absolute	Relative	WFB	CC
OUR	1.67 (0.54)	2.01 (0.34)	0.44	17.08	9.57	13.43
HOR [7]	0.98 (0.73)	1.03 (0.79)	1.16	30.66	12.05	21.07
HAG [4]	1.50 (0.61)	2.08 (0.66)	0.43	12.12	9.04	13.46

Table 2: Results from user studies on the CoCoClutter dataset. All values are scaled by 100 apart from realism and fidelity.

Model	Realism	Fidelity	Attention increase		Similarity to mask	
			Absolute	Relative	WFB	CC
OUR	1.07 (0.50)	1.67 (0.42)	11.79	149.86	7.12	17.93
PROF	1.42 (0.39)	1.94 (0.47)	4.27	134.47	5.27	11.39

Table 3: Results from user studies on HighResClutter, where PROF refers to professionals recruited via the www.usertesting.com crowdsourcing platform. All values are scaled by 100 apart from realism and fidelity.

Model	LPIPS ↓			Saliency increase ↑		Similarity to mask ↑	
	Full	BG	FG	Absolute	Relative	WFB	CC
OUR	6.16	5.35	0.82	4.19	38.82	9.13	25.64
PROF	6.93	6.58	0.36	1.74	15.77	8.0	21.42
PROF-range	[1.15, 16.89]	[0.73, 16.74]	[0.04, 0.83]	[-0.46, 4.31]	[-4.33, 35.39]	[7.03, 9.12]	[17.72, 25.53]

Table 4: Scores (scaled by 100) from computational measures on HighResClutter.

overall on the LSUN 2017 challenge leaderboard[5] as of 09/24/2019, and leading in terms of NSS score. Our model choice criteria included having a top-performing model with a small footprint, so that our final model, in which saliency would be one of multiple sub-components, would be fit for practical use and not bulky. In contrast, other top-performing models are quite bulky: SAM [1] at 70M parameters, DeepGaze II [6] at 40M parameters. MD-SEM [2] has 30M parameters while outperforming these models on the LSUN challenge, making it more attractive as a building block within larger systems.

For completeness, we also provide the absolute and relative saliency increases of all methods using the DeepGaze model [3] in Table 5. Under this alternative saliency model, our approach still outperforms the alternatives on both datasets evaluated.

[5] https://competitions.codalab.org/competitions/17136results

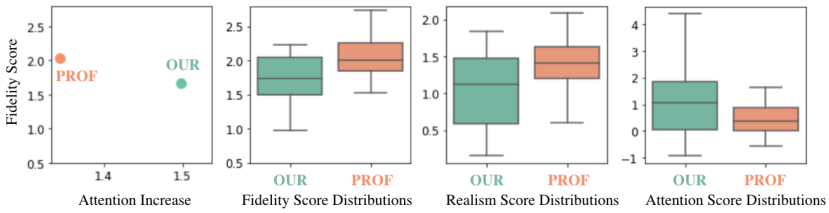

Fig. 5: Results of user studies run on three separate crowdsourcing tasks - measuring image fidelity, realism, and human attention - on 30 high-resolution images. The fidelity, realism, and attention score distributions are visualized as box plots. We compare to the results of professionals (PROF) recruited via the www.usertesting.com crowdsourcing platform. Compared to professionals, we are able to increase attention more effectively, but at the cost of reducing realism and fidelity.

Model	Saliency increase	
	Absolute	Relative
OUR	31.96	2.96
MEC	16.60	1.52
HAG	17.68	16.06
HOR	3.12	0.27
GAT	10.41	0.95

Model	Saliency increase	
	Absolute	Relative
OUR	17.20	15.11
HAG	15.28	1.31
HOR	1.30	0.09
GAT	8.13	0.68

Table 5: Absolute and relative saliency increase (scaled by 100) using DeepGaze as a measure of computational saliency on the Mechrez (left) and CoCoClutter (right) datasets.

3.2 Ablation studies

Here we report on the influence of different components of our model to its final performance. First, we consider which set of parametric transformations to use. Second, we consider what happens if we only predict parametric transformations for the foreground or background of the image, instead of both. Third, we discuss how the order of parameter application influences the results. As it is not immediately obvious how to balance the trade-off between realism and attention shift across our different ablations, we identify the five best models according to three criteria (LPIPS, absolute and relative saliency increase), and choose a high-performing model across all criteria. Results for all the following ablations refer to Table 1d in the main paper, where the best performing models are highlighted in green (darker is better), and the least performing models are highlighted in red (darker is worst).

Parameter ablations: Tone and color curve adjustments are two of our most powerful transformations as each is defined by multiple parameters describing a piecewise linear function. If excluded ('sharp+exp+cont' in Table 1d in the main paper), the model achieves only a small increase in saliency and small LPIPS values, indicating that the generated image is very similar to the original. On the other hand, a model that only uses tone and color adjustment ('tone+color') produces a significant increase in both saliency and LPIPS values. Generated images from this approach often look unrealistic as the network overuses these parameters to minimize the attention loss. A similar but smaller effect is noticeable when using only color adjustment ('color'). This suggests that combining tone and color with subtler transformations such as contrast, exposure and sharpening ('our') gives more freedom to the network for achieving high saliency

shifts while maintaining image realism. We also consider adding saturation to our model ('our+saturation'), and find an increase in saliency corresponding to increased flexibility for our network. However, saturation often introduces destructive artifacts [7], resulting in very high LPIPS scores. Figure 6 displays examples of such artifacts.

Finally, we compute the mean of each predicted parameter value over the entire CoCoClutter validation set, and apply the resulting set of parameters to the input images ('Fixed parameters'). The motivation is to evaluate if our network learns content-aware transformations, or if similar performance is achievable by applying the same transformation independently of the image content. This approach indeed achieves a shift in saliency, showing such a set of transformations can generalize (e.g., brightening the foreground and darkening the background). However, this saliency shift is significantly weaker than 'our', suggesting that our network is able to adapt a set of transformations to each image.

FG/BG ablations: Our network predicts two sets of parameters, one for the foreground, and one for the background. Is similar performance achievable using only one set of parameters? Table 1d (fg/bg ablations in the main paper) shows that only applying parameters to the background ('bg-only') leads to marginal saliency increase. In contrast, modifying the foreground while leaving the background untouched, performs better, but still fails to outperform the setting with both sets of parameters. Note that a lower LPIPS value is obtained by 'fg-only' because when the background, which covers a larger area in the image, is left untouched, the final result is more similar to the input image.

Order ablations: We test 4 configurations in terms of order of parameter application, presented in Table 1d in the main paper as 'order of application ablations', where 'sha' is sharpening, 'exp' is exposure, 'con' is contrast, 'ton' is tone adjustment and 'col' is color adjustment. Changing the order that parameters are applied in affects the final images generated. Some orders achieve higher saliency increases compared to others, but compromise in terms of image realism (higher LPIPS values). We chose the ordering that achieves the best trade off between saliency shift and realism. While we did not extensively test all possible order combinations, we leave the task of automatically finding such an optimal ordering for future work.

3.3 Model architectures

Tables 6,7 summarize our generator and discriminator architectures.

3.4 Qualitative results

Additional qualitative results on the Mechrez dataset are provided in Figures 7, 8 and on the CoCoClutter dataset in Figures 9, 10. Figure 11 includes additional results using the stochastic image generation model. Figure 12 provide additional results for increasing and decreasing saliency.

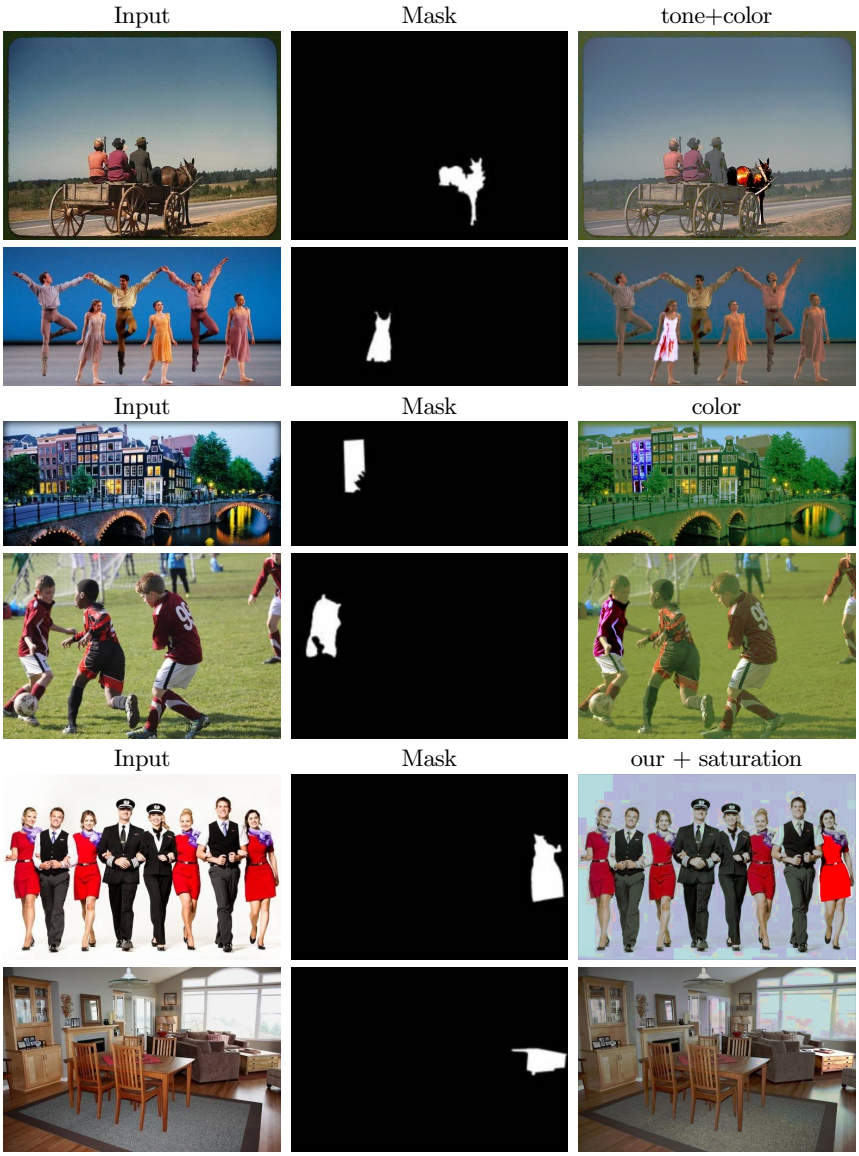

Fig. 6: Artifacts created by some of the parameter ablations experiments.

Layer	#Filters	Size	Stride	InstNorm	Act.
Conv.	64	7×7	2	✓	LReLU
Conv.	128	3×3	2	✓	LReLU
Conv.	256	3×3	2	✓	LReLU
Conv.	512	3×3	2	✓	LReLU
Conv.	1024	3×3	2	-	LReLU
AvgP.	-	-	-	-	-

Layer	#Neurons	Act.
FC.	128	LRelu
FC.	128	LRelu
FC.(Sharp)	1	Tanh
FC.(Exposure)	1	Tanh
FC.(Contrast)	1	Tanh
FC.(Tone curve)	L	Sigmoid
FC.(Color curve)	$3L$	Sigmoid

Table 6: (Left): Shared convolutional part of G_E. (Right): Specialized densely connected head for predicting foreground and background parameters. 'Conv.' is convolutional layer; 'FC.' is fully connected layer; 'AvgP.' is global average pooling; 'InstNorm' is instance normalization; 'Act.' is activation function. 'LReLU' denotes Leaky ReLU with a factor of 0.2.

Layer	#Filters/Neurons	Size	Stride	Act.
Conv.	64	4×4	2	LReLU
Conv.	128	4×4	2	LReLU
Conv.	256	4×4	2	LReLU
Conv.	256	4×4	1	LReLU
FC.	128	-	-	LReLU
FC.	1	-	-	None

Layer	#Neurons	Act.
FC.	256	LRelu
FC.	256	LRelu
FC.	256	LRelu
FC.	256	LRelu
FC.	10	None

Table 7: (Left): Architecture of our discriminator. We use a Multi-Scale discriminator [9] of scale 3 as it has proven to provide better results with GANs. 'Conv.' is convolutional layer; 'FC.' is fully connected layer; 'LReLU' denotes Leaky ReLU with a factor of 0.2. (Right): Architecture of the encoder ENC used to reconstruct the latent vector z in the multi-style setting.

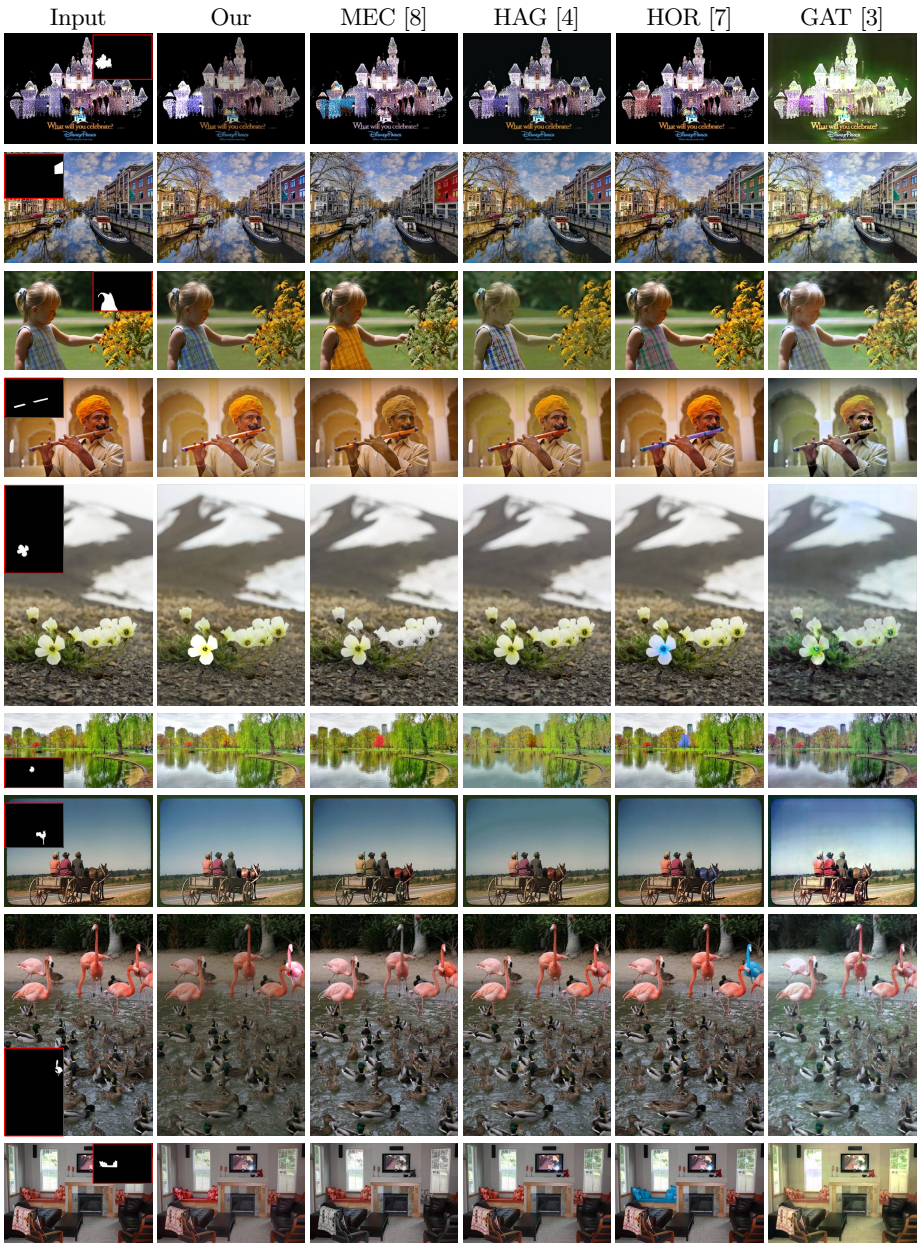

Fig. 7: More model comparisons on the Mechrez dataset.

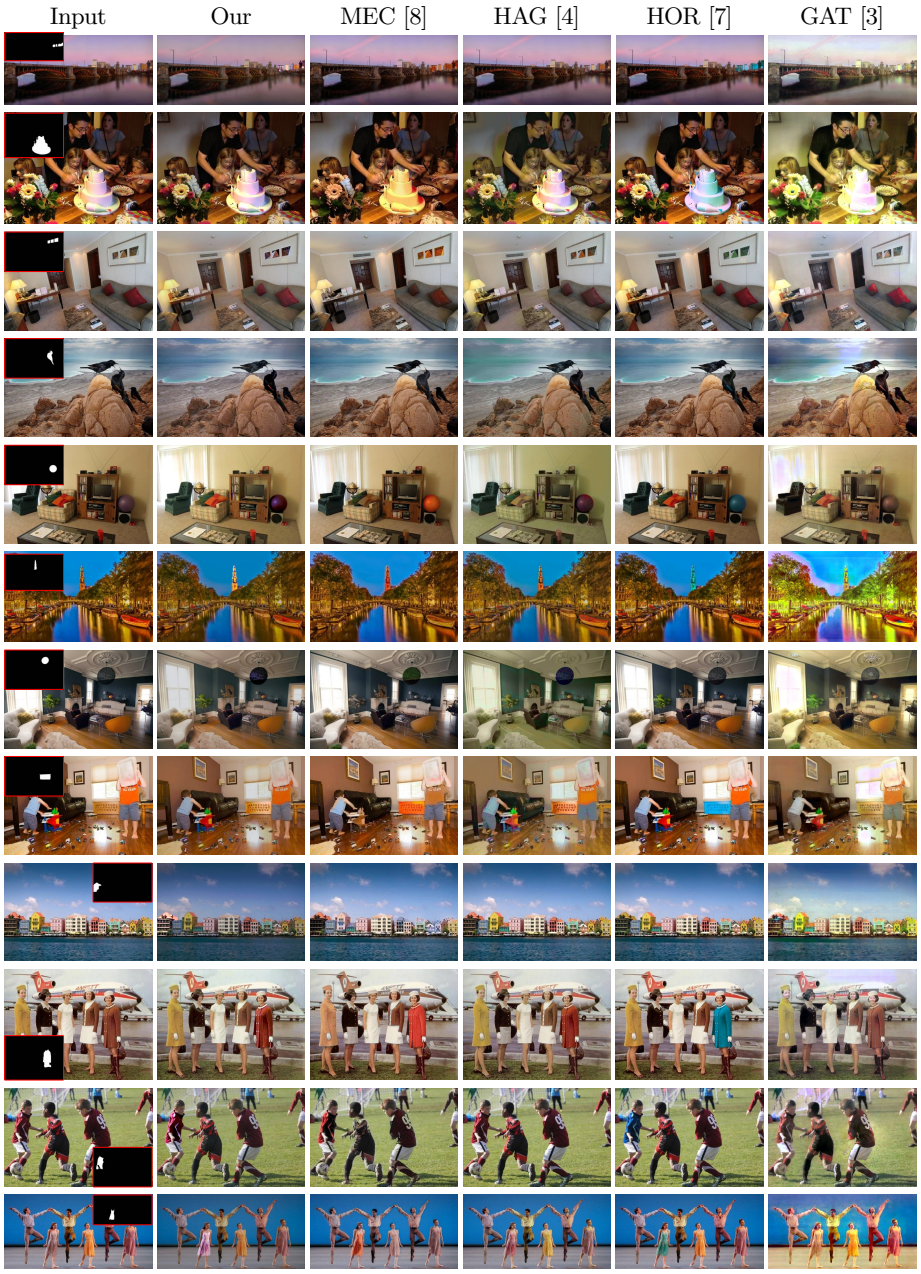

Fig. 8: More model comparisons on the Mechrez dataset.

Fig. 9: More results on CoCoClutter dataset.

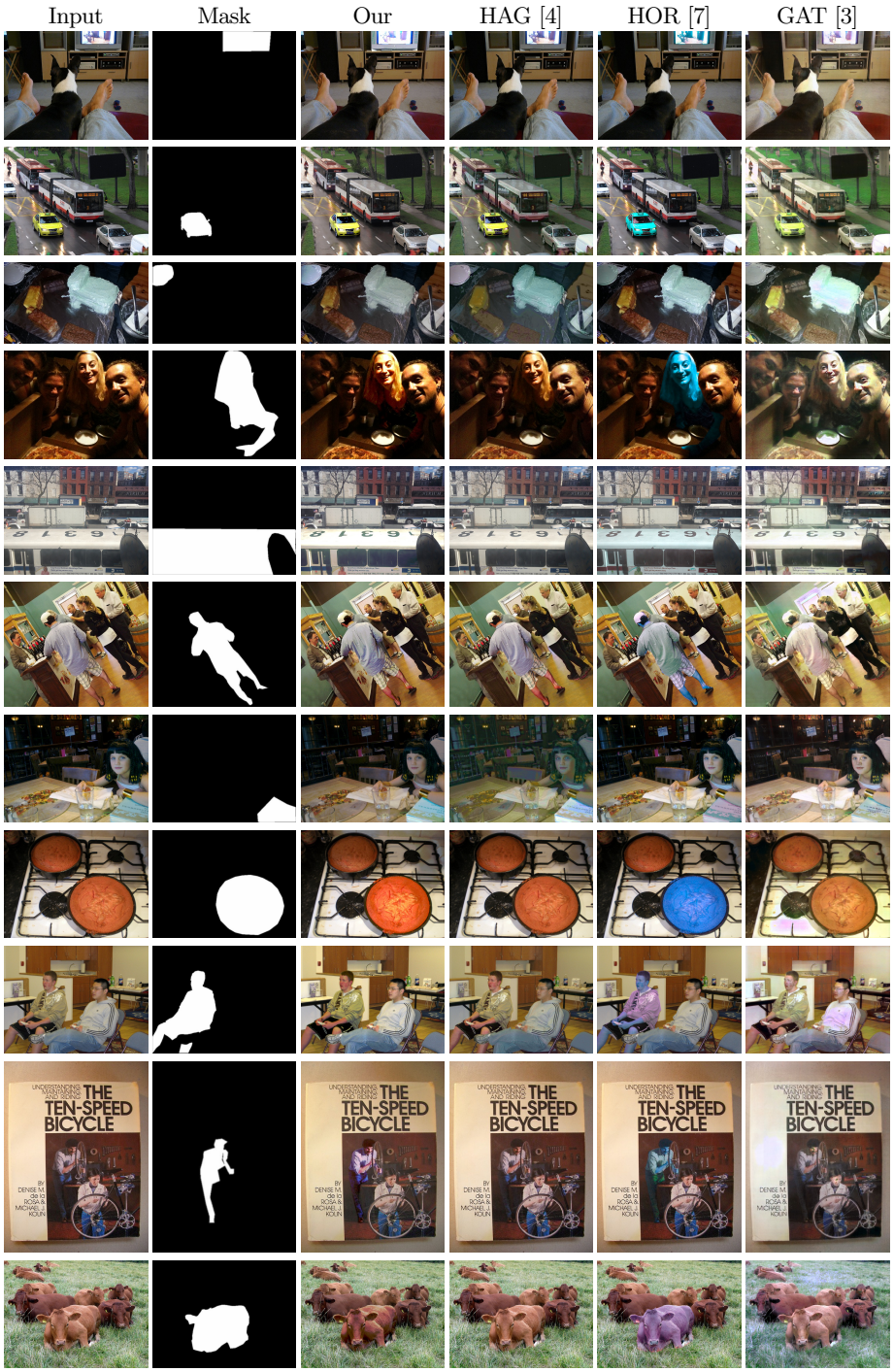

Fig. 10: More results on CoCoClutter dataset.

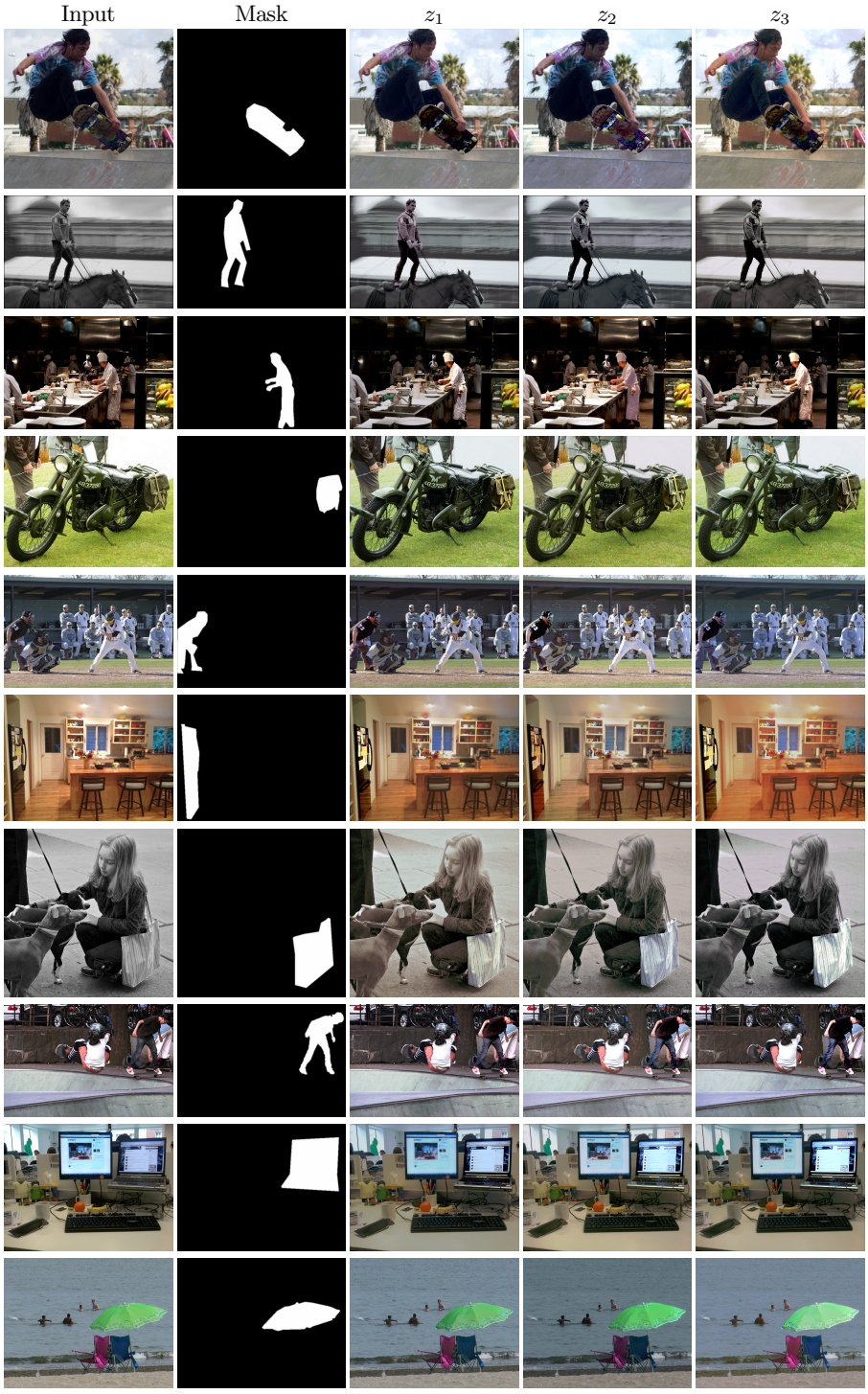

Fig. 11: More results on stochastic image generation.

Fig. 12: Additional results for decreasing human attention.

References

1. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: SAM: Pushing the Limits of Saliency Prediction Models. In: CVPR Workshops (2018)
2. Fosco, C., Newman, A., Sukhum, P., Zhang, Y.B., Zhao, N., Oliva, A., Bylinskii, Z.: How much time do you have? modeling multi-duration saliency. In: CVPR (2020)
3. Gatys, L.A., Kümmerer, M., Wallis, T.S., Bethge, M.: Guiding human gaze with convolutional neural networks. arXiv preprint arXiv:1712.06492 (2017)
4. Hagiwara, A., Sugimoto, A., Kawamoto, K.: Saliency-based image editing for guiding visual attention. In: Proceedings of the 1st international workshop on pervasive eye tracking & mobile eye-based interaction (2011)
5. Hu, Y., He, H., Xu, C., Wang, B., Lin, S.: Exposure: A white-box photo post-processing framework. SIGGRAPH (2018)
6. Kümmerer, M., Wallis, T.S., Bethge, M.: Deepgaze ii: Reading fixations from deep features trained on object recognition. arXiv preprint arXiv:1610.01563 (2016)
7. Mateescu, V.A., Bajić, I.V.: Attention retargeting by color manipulation in images. In: Proceedings of the 1st International Workshop on Perception Inspired Video Processing (2014)
8. Mechrez, R., Shechtman, E., Zelnik-Manor, L.: Saliency driven image manipulation. WACV (2019)
9. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR (2018)