Maintaining Natural Image Statistics with the Contextual Loss - Supplementary

Roey Mechrez*, Itamar Talmi*, Firas Shama, Lihi Zelnik-Manor

The Technion - Israel

1 Implementation details

1.1 Super Resolution

We adopt the SRGAN architecture [1][1] and train it on only 800 images from the DIV2K dataset [2], for 1500 epochs with $h = 0.1$ (for the contextual loss). Our network is initialized by first training using only the $L2$ loss for 100 epochs. We used TensorFlow [3] and Adam optimizer [4] with the default parameters ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-08$).

In the total objective we set: $\lambda_{\text{CX}} = 0.1$, $\lambda_{\text{GAN}} = 1e-3$, and $\lambda_{L2} = 10$. The images $G(s)^{\text{LF}}, y^{\text{LF}}$ are low-frequencies obtained by convolution with a Gaussian kernel of width 21×21 and $\sigma = 3$. For the Contextual loss feature extraction we used layer $conv3\_4$ of VGG19 [5]

1.2 Normal Estimation

We chose as architecture the Cascaded Refinement Network (CRN) [6][2] originally suggested for label-to-image and was shown to yield great results in a variety of other tasks [7]. For the contextual loss we took as features 5×5 patches of the normal map (extracted with stride 2) and layers $conv1\_2, conv2\_2$ of VGG19. In our implementation we reduced memory consumption by random sampling of all three layers into 65×65 features.

In the total objective we set: $\lambda_{\text{CX}} = 1$, and $\lambda_{L2} = 0.1$. The normal-maps $G(s)^{\text{LF}}, y^{\text{LF}}$ are low-frequencies obtained by convolution with a Gaussian kernel of width 21×21 and $\sigma = 3$. We tested with both $\lambda_{L1} = 1$ and $\lambda_{L1} = 0$, which removes the third term.

We will release the code upon acceptance.

* indicate authors contributed equally

[1] We used the implementation in https://github.com/tensorlayer/SRGAN

[2] Authors release https://github.com/CQFIO/PhotographicImageSynthesis

2 Additional Experiments

A qualitative comparison to pixel-to-pixel losses The popular trend in training generator networks is to use pixel-to-pixel loss functions such as $L1$ or $L2$ since these directly minimize the PSNR. However, the resulting images are often considered to be non-realistic by human raters [1,8,9]. To show the benefits of using a statistical objective, such as the Contextual loss, we designed a simple experiment on a super-resolution task. The goal of the experiment is to compare a network trained with the Contextual loss with a network trained with $L2$ (or $L1$) - i.e., one that aims at optimal PSNR.

To do this we chose a simplified super-resolution setup, where we train an image-specific network, (i.e., training on a single image), to increase the resolution of that specific image. This simplified setup essentially tries to overfit the network to the specific image, in order to reveal the network's ability to reconstruct the image under least challenging conditions.

As architecture we adopted SRResNet [1] and trained it with either $L1$, or $L2$, or the Contextual loss as the objective. The features fed to the Contextual loss were vectorized RGB patches of size 5×5 (stride 2). At each iteration we compared between the high resolution (HR) target image y and the output of the network $G(s)$, where s is the low resolution (LR) image. The training data consisted of random crops (of size 384×384) extracted from a single image (of size 1072×712). The crops were $\times4$ down-sampled (to size 96×96) yielding pairs of LR-HR images. Training lasted 10K iterations.

Results are presented in Figure 1. As can be seen, optimizing the network with $L2$ produces blurred high-resolution images. This is in spite of the simplified problem setup and the loss being a direct match for PSNR. Results with $L1$ were very similar and hence were excluded from the figure. Training with the Contextual loss, on the other hand, resulted in high quality reconstruction. The network intelligently hallucinated many of the fine details, that were missing in the low-resolution. This outcome is well aligned with the observations in [10], where it was shown that such detail generation is impossible when using pixel-to-pixel loss functions like $L2$ and $L1$.

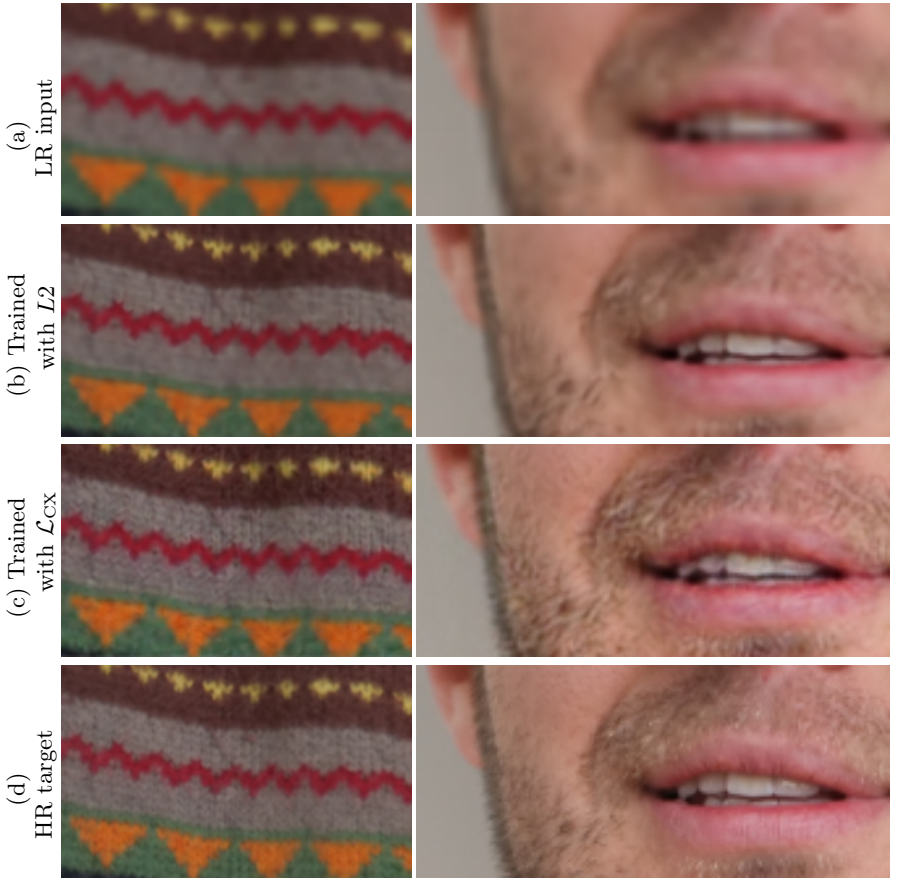
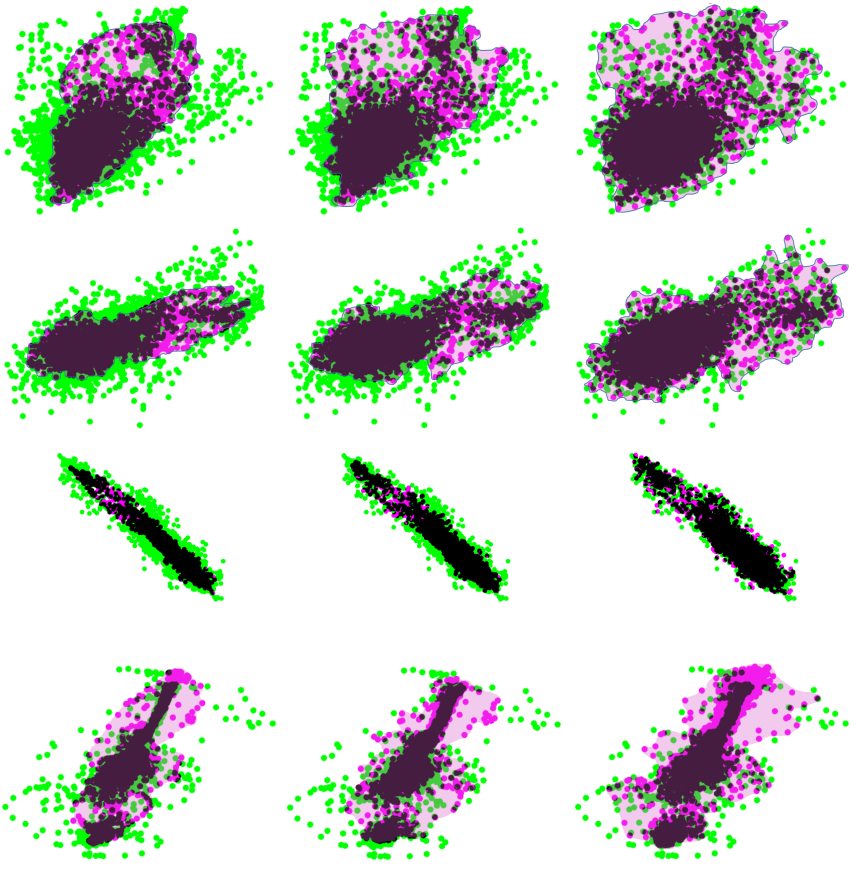


Fig. 1. Maintaining natural image statistics in super-resolution: (b),(c) High-resolution images produced by SRResNet when trained with either $L2$ or $\mathcal{L}_{\text{CX}}$, respectively. Training was done on a single image, see text for details. It can be seen that using the Contextual loss allows hallucination of fine details on the sweater texture, the cracks on the lips, and the facial hair. Conversely, with $L2$, the generated images are blurred (results with $L1$ are highly similar, hence, excluded).

Minimizing the difference between distributions: We next provide another qualitative view on how optimizing the Contextual loss minimizes the difference between the feature distributions of the generated image and target image. This is done by visualizing these distributions, before and after training the network. The goal being to show how after training the distributions become more similar.

We repeat the experiment in the previews section, but this time using straightforward gradient descent, directly updating the image values, instead of a trained CNN. In addition, the source-target pairs were shifted, with respect to each other, by random translation of up to 10 pixels. This was in order to cancel the spatial dependence between the pixels of the source and target.

To visualize the patch distribution of an image we extract from it all 5×5 patches, vectorize them, and apply a random projection onto 2D. In Figure 2 we present such visualizations, where the projections of the patches of the target image and the generated image are overlaid. It can be seen that optimizing with the Contextual loss results in a final generated image with patch distribution highly similar to that of the target. Conversely, when using $L1$ as an objective, the distributions do not converge (and similarly for $L2$, not shown in the figure).



(a) Initial (b) Optimized with $L1$ (c) Optimized with $\mathcal{L}_{\text{CX}}$

Fig. 2. Minimizing the difference between distributions: Random projections of image patches onto 2D illustrate the distribution of patches in an image. The projection of the patches of the **target** are in **green**. The **input** and the final **output** are both in **Magenta**. When two projections agree, they overlap and the color turns to **Black**. We compare between the target image and (a) the input image, (b) the output after optimization with $L1$, and, (c) the output after optimizing with $\mathcal{L}_{\text{CX}}$ (more details in the text). Optimizing with $\mathcal{L}_{\text{CX}}$ reduces the outliers and yields high correlation (more black) with the target, while using $L1$ barely reduces the gap between the densities.

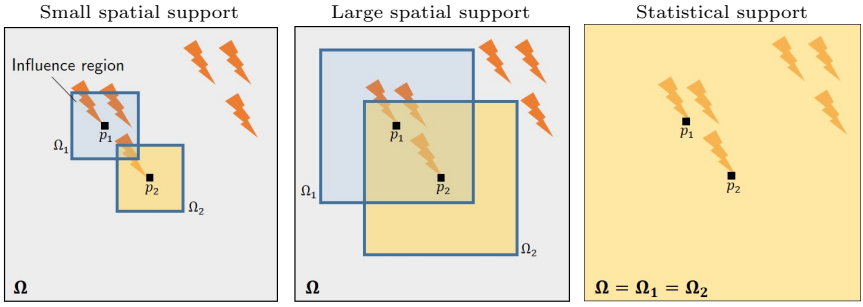


Fig. 3. Importance of spatial support during training: In order to be able to reconstruct specific pattern, illustrate by the lightning, the spatial support should include distant pixels of the same pattern. The pixel p_1 is computed from the influencing region in the spatial support Ω_1, while the pixel p_2 cannot be calculated since the supporting area Ω_2 does not contain the necessary information. This improve when the spatial support is larger. Our solution takes the entire image as the support (context) by utilize the statistical approach of the Contextual loss. Here we illustrate for image SR, with out the loss of generality, same insight is true for all image restoration problems. Similar idea was shown in [13] during test time.

The contribution of a single pixel: In the classical pixel-to-pixel loss function (e.g. $L1$, $L2$, perceptual loss [11]) the contribution of each pixel in the training process is simple, the loss is a sum over all distances. In this manner, each pixel is influence on it self without any relation to the neighborhood. Since the CNN architecture use convolutions and pooling, the spatial support is in practice larger and is term the *receptive field* [12]. A single pixel is computed by a set of mathematical operations taken on its covered region in the input image plane. The receptive field is a function of the network architecture, for example in VGG-16 [5] the receptive field of the last pooling layer is 212.

The support of the contextual loss is non-local in its nature, in practice the contribution of a single pixel is depended in the entire image. This can be simple derives from the contextual loss definition: the contribution of a single pixel, j, is done by taking the max value of A_{ij}. Namely, a single pixel, j, is depended on all $i's$ and vise versa. Conceptually, the contribution of each pixel is calculated by answering the following question: what is the distance of all other pixels $\{x\}_i^N$ to the particular closest y_j? . As a result the receptive field of each pixel without relation to the architecture used is the *entire image*, not directly, rather statistically – thus we name this the statistical receptive field. Inspired from [13] we show in Figure 3 an illustrating of the spatial support importance during SR training.

We note that the support discuss above is influencing solaly during training as it reflected in the loss function. This differ from the classical meaning of layer support during test time.

Power-spectrum analysis: Further to Section 5.1 discussing the super resolution trends, we ask to highlight the difference between the PSNR group and the perceptual group in their ability to generate natural looking images. This is done by revisit a decades-old observation, which says that natural images exhibit a typical corresponding power-spectrum [14,15]. We present in Figure 4 the mean power-spectra of images generated by methods from both groups. It can be seen, that the power-spectra corresponding to methods from the second group are by far more similar to that of real images. Furthermore, the power-spectrum corresponding to our method (described next) is the most similar of all to that of the ground-truth.

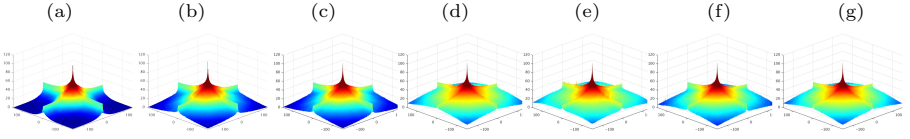
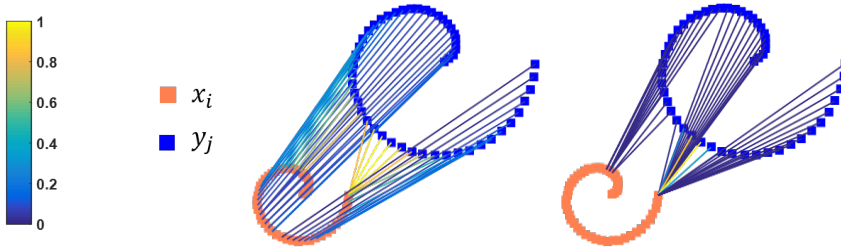


Fig. 4. Natural image statistics: The mean power spectra of images reconstructed by six different super-resolution methods: (a) bicubic (as baseline); (b) LapSRN [16] and (c) SRResNet-MSE [1] are designed for high PSNR and trained with $L2$ as loss; In contrast, (d) SRGAN [1], (e) EnhanceNet [8], and, (f) ours, are all trained with GAN and aim for high perceptual quality. (g) The mean power-spectrum of the ground-truth high-resolution images. It can be seen that methods aimed at perceptual quality (d,e,f) produce images with more natural power-spectrum than methods aimed at high PSNR (a,b,c). The latter lack high-frequencies.

Additional 2D toy examples: In Figure 5 we present an additional 2D example to the one presented in Figure. 3 in the paper.



normalized affinity value (a) The Contextual loss [7] (b) Chamfer distance [17]

Fig. 5. The Contextual loss vs. Chamfer Distance: We demonstrate via a 2D example the difference between two approximations to the KL-divergence (a) The Contextual loss and (b) Chamfer Distance. Point sets Y and X are marked with blue and orange squares respectively. The colored lines connect each y_j with x_i with the largest affinity. The KL approximation of Eq.(8) sums over these affinities. It can be seen that the normalized affinities used by the Contextual loss lead to more diverse and meaningful matches between Y and X.

3 Additional Results

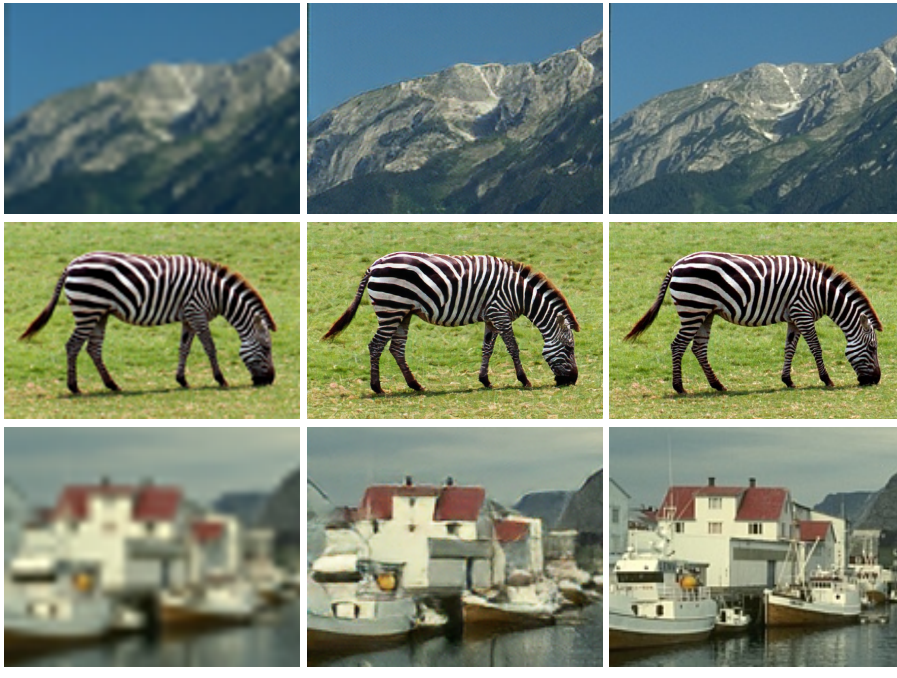
Subtype	Metric	Distortions			Real Algorithms					All
		Trad-itiona	CNN -Based	All	SR	Video Deblur	color-ization	Frame Interp	All	All
Oracle	Human	80.8	84.4	82.6	73.4	67.1	68.8	68.6	69.5	73.9
Low	L2	59.9	77.8	68.9	64.7	58.2	63.5	55.0	60.3	63.2
-level	SSIM	60.3	79.1	69.7	65.1	58.6	58.1	57.7	59.8	63.1
	FSIMc	61.4	78.6	70.0	68.1	59.5	57.3	57.7	60.6	63.8
Net	Squeeze	73.3	82.6	78.0	70.1	60.1	63.6	62.0	64.0	68.6
with	Alex	70.6	83.1	76.8	**71.7**	**60.7**	**65.0**	62.7	**65.0**	68.9
L2	VGG	70.1	81.3	75.7	69.0	59.0	60.2	62.1	62.6	67.0
Net	Squeeze	**75.8**	**83.5**	**79.7**	70.9	60.3	63.2	62.5	64.2	**69.4**
with	Alex	71.4	**83.5**	77.4	71.4	**60.7**	64.6	**62.8**	64.9	69.1
$\mathcal{L}_{\mathbf{CX}}$	VGG	**75.8**	82.1	79.0	70.4	59.3	59.2	62.4	62.8	68.2

Table 1. Results on 2AFC dataset (higher is better) across a spectrum of methods and test sets. Note, that we do not show here the methods that were trained using the 2AFC data for the task of perceptual similarity. The full table, can be found in [18].

How perceptual is the Contextual loss? In table 1 we show full quantitative results across all validation sets and considered metrics, including low-level metrics, supervised networks with L2, and supervised networks with the contextual loss.

Additional super resolution results: Presented in Figures 6 and 7.

Last, the supplementary material include the **first** 10 images (001 to 010) from the DIV2K dataset, these images present the effectiveness of our method on high resolution (2K) images. Note that we attached only 10 images due to space limit of the submission.



Bicubic Ours HR

Fig. 6. Super-resolution zoom examples

Bicubic Ours HR

Fig. 7. Super-resolution zoom examples

References

1. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR. (2017) 1, 2, 7
2. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: CVPR Workshops. (2017) 1
3. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016) 1
4. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 1
5. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 1, 6
6. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: ICCV. (2017) 1
7. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. arXiv preprint arXiv:1803.02077 (2018) 1, 8
8. Sajjadi, M.S., Scholkopf, B., Hirsch, M.: Enhancenet: Single image super-resolution through automated texture synthesis. In: ICCV. (2017) 2, 7
9. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: CVPR. (2018) 2
10. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. arXiv preprint arXiv:1711.10925 (2017) 2
11. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV. (2016) 6
12. Yu, W., Yang, K., Bai, Y., Xiao, T., Yao, H., Rui, Y.: Visualizing and comparing alexnet and vgg using deconvolutional layers. In: ICML. (2016) 6
13. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM TOG (2017) 6
14. Ruderman, D.L.: The statistics of natural images. Network: computation in neural systems (1994) 7
15. Van der Schaaf, v.A., van Hateren, J.v.: Modelling the power spectra of natural images: statistics and information. Vision research (1996) 7
16. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition. (2017) 7
17. Barrow, H.G., Tenenbaum, J.M., Bolles, R.C., Wolf, H.C.: Parametric correspondence and chamfer matching: Two new techniques for image matching. Technical report (1977) 8
18. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. arXiv preprint arXiv:1801.03924 (2018) 9