Perceptually Motivated Method for Image Inpainting Comparison

I.A. Molodetskikh[1], M.V. Erofeev[1], D.S. Vatolin[1]
ivan.molodetskikh@graphics.cs.msu.ru|merofeev@graphics.cs.msu.ru|dmitriy@graphics.cs.msu.ru
[1]Lomonosov Moscow State University, Moscow, Russia

The field of automatic image inpainting has progressed rapidly in recent years, but no one has yet proposed a standard method of evaluating algorithms. This absence is due to the problem's challenging nature: image-inpainting algorithms strive for realism in the resulting images, but realism is a subjective concept intrinsic to human perception. Existing objective image-quality metrics provide a poor approximation of what humans consider more or less realistic.

To improve the situation and to better organize both prior and future research in this field, we conducted a subjective comparison of nine state-of-the-art inpainting algorithms and propose objective quality metrics that exhibit high correlation with the results of our comparison.

Keywords: *image inpainting, objective quality metric, quality perception, subjective evaluation, deep learning.*

1. Introduction

Image inpainting, or hole filling, is the task of filling in missing parts of an image. Given an incomplete image and a hole mask, an inpainting algorithm must generate the missing parts so that the result looks realistic. Inpainting is a widely researched topic. Many classical algorithms have been proposed [5, 26], but over the past few years most research has focused on using deep neural networks to solve this problem [12, 16, 17, 19, 23, 31, 32].

Because of the many avenues of research in this field, the need to evaluate algorithms emerges. The goal of an inpainting algorithm is to make the final image as realistic as possible, but image realism is a concept intrinsic to humans. Therefore, the most accurate way to evaluate an algorithm's performance is a subjective experiment where many participants compare the outcomes of different algorithms and choose the one they consider the most realistic.

Unfortunately, conducting a subjective experiment involves considerable time and resources, so many authors resort to evaluating their proposed methods using traditional objective image-similarity metrics such as PSNR, SSIM and mean l_2 loss relative to the ground-truth image. This strategy, however, is inadequate. One reason is that evaluation by measuring similarity to the ground-truth image assumes that only a single, best inpainting result exists—a false assumption in most cases.

Thus, a perceptually motivated objective metric for inpainting-quality assessment is desirable. The objective metric should approximate the notion of image realism and yield results similar to those of a subjective study when comparing outputs from different algorithms.

We conducted a subjective evaluation of nine state-of-the-art classical and deep-learning-based approaches to image inpainting. Using the results, we examine different methods of objective inpainting-quality evaluation, including both full-reference methods (taking both the resulting image and the ground-truth image as an input) and no-reference methods (taking the resulting image as an input).

2. Related work

Little work has been done on objective image inpainting-quality evaluation or on inpainting detection in general. The somewhat related field of manipulated-image detection has seen moderate research, including both classical and deep-learning-based approaches. This field focuses on detecting altered image regions, usually involving a set of common manipulations: copy-move (copying an image fragment and pasting it elsewhere in the same image), splicing (pasting a fragment from another image), fragment removal (deleting an image fragment and then performing either a copy-move or inpainting to fill in the missing area), various effects such as Gaussian blur, and recompression. Among these manipulations, the most interesting for this work is fragment removal with inpainting.

The approaches to image-manipulation detection can be divided into classical [13, 20], and deep-learning-based approaches [2, 21, 34, 35]. These algorithms aim to locate the manipulated image regions by outputting a mask or a set of bounding boxes enclosing suspicious regions. Unfortunately, they are not directly applicable to inpainting-quality estimation because they have a different goal: whereas an objective quality-estimation metric should strive to accurately compare realistically inpainted images similar to the originals, a forgery-detection algorithm should strive to accurately tell one apart from the other.

3. Inpainting subjective evaluation

The gold standard for evaluating image-inpainting algorithms is human perception, since each algorithm strives to produce images that look the most realistic to humans. Thus, to obtain a baseline for creating an objective inpainting-quality metric, we conducted a subjective evaluation of multiple state-of-the-art algorithms, including both classical and deep-learning-based ones. To assess the overall quality and applicability of the current approaches and to see how they compare with manual photo editing, we also asked professional photo editors to fill in missing regions of the test photos.

3.1 Test data set

Since human photo editors were to perform inpainting, our data set could not include publicly available images. We therefore created our own private set of test images by taking photographs of various outdoor scenes, which are the most likely target for inpainting.

Fig. 1. Images for the subjective inpainting comparison. The black square in the center is the area to be inpainted.

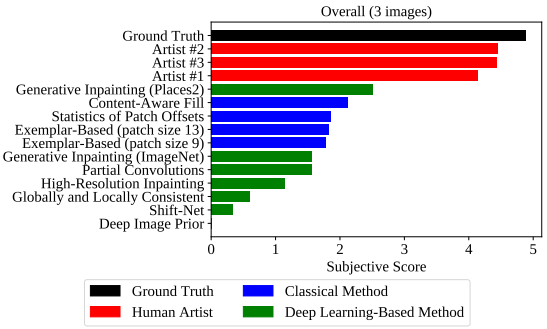

Fig. 2. Subjective-comparison results across three images inpainted by human artists.

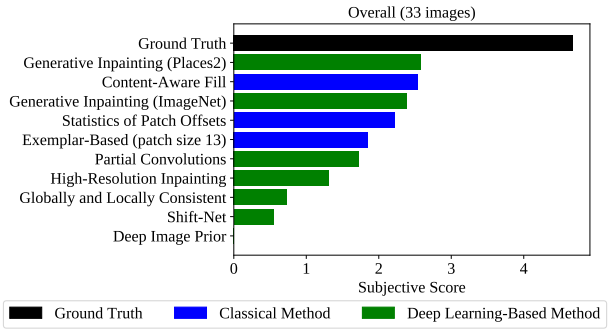

Fig. 3. Subjective-comparison results for 33 images inpainted using automatic methods.

Each test image was 512×512 pixels with a square hole in the middle measuring 180×180 pixels. We chose a square instead of a free-form shape because one algorithm in our comparison [30] lacks the ability to fill in free-form holes. The data set comprised 33 images in total. Fig. 1 shows examples.

3.2 Inpainting methods

We evaluated three classical [1, 5, 7] and six deep-learning-based approaches [10, 16, 27, 29, 30, 32]. Additionally, we hired three professional photo-restoration and photo-retouching artists to manually inpaint three randomly selected images from our test data set.

3.3 Test method

The subjective evaluation took place through the http://subjectify.us platform. Human observers were shown pairs of images and asked to pick from each pair the one they found most realistic. Each pair consisted of two different inpainting results for the same picture (the set also contained the original image). In total, 6945 valid pairwise judgements were collected from 215 participants.

The judgements were then used to fit a Bradley-Terry model [3]. The resulting subjective scores maximize likelihood given the pairwise judgements.

3.4 Results of the subjective comparison

Fig. 2 shows the results for the three images inpainted by the human artists. The artists outperformed all

Fig. 4. Comparison of inpainting results from Artist #1 and statistics of patch offsets [7] (preferred in the subjective comparison).

automatic algorithms, and out of the deep-learning-based methods, only generative image inpainting [32] outperformed the classical inpainting methods.

The individual results for each of these three images appear in Fig. 5. In only one case did an algorithm beat an artist: statistics of patch offsets [7] scored higher than one artist on the "Urban Flowers" photo. Fig. 4 shows the respective results. Additionally, for the "Splashing Sea" photo, two artists actually "outperformed" the original image: their results turned out to be more realistic.

We additionally performed a subjective comparison of various inpainting algorithms among the entire 33-image test set, collecting 3969 valid pairwise judgements across 147 participants. The overall results appear in Fig. 3.

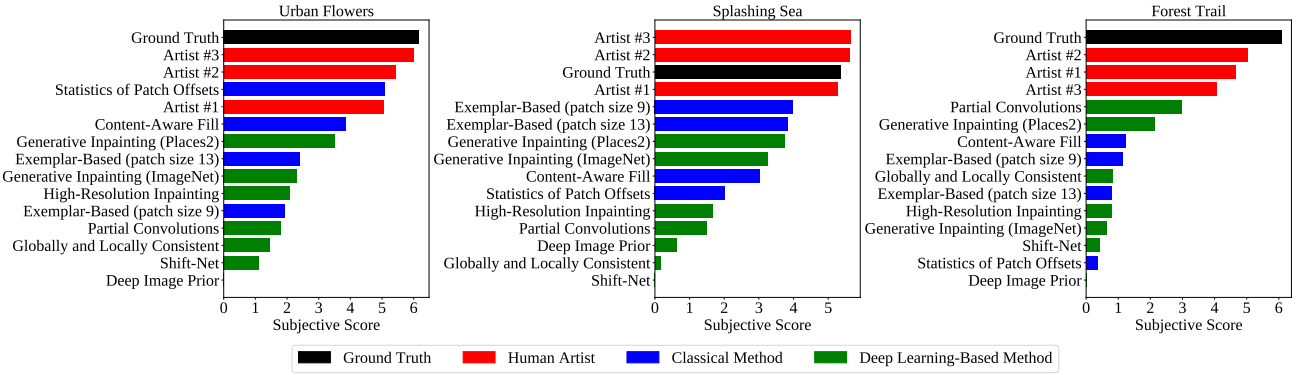

Fig. 5. Results of the subjective study comparing images inpainted by human artists with images inpainted by conventional and deep-learning-based methods.

They confirm our observations from the first comparison: among the deep-learning-based approaches we evaluated, generative image inpainting [32] seems to be the only one that can outperform the classical methods.

4. Objective inpainting-quality estimation

Using the results we obtained from the subjective comparison, we evaluated several approaches to objective inpainting-quality estimation. In particular, we used these objective metrics to estimate the inpainting quality of the images from our test set and then compared them with the subjective results. For each of the 33 images, we applied every tested metric to every inpainting result (as well as to the ground-truth image) and computed the Pearson and Spearman correlation coefficients with the subjective result. The final value was an average of the correlations over all 33 test images.

4.1 Full-reference metrics

To construct a full-reference metric that encourages semantic similarity rather than per-pixel similarity, as in [11], we evaluated metrics that compute the difference between the ground-truth and inpainted-image feature maps produced by an image-classification neural network. We selected five of the most popular architectures: VGG [22] (16- and 19-layer deep variants), ResNet-V1-50 [8], Inception-V3 [25], Inception-ResNet-V2 [24] and Xception [4]. We used the models pretrained on the ImageNet [6] data set. The mean squared error between the feature maps was the metric result.

We additionally included the structural-similarity (SSIM) index [28] as a full-reference metric. SSIM is widely used to compare image quality, but it falls short when applied to inpainting-quality estimation.

4.2 No-reference metrics

We picked several popular image-classification neural-network architectures and trained them to differentiate images without any inpainting from partially inpainted images. The architectures included VGG [22] (16- and 19-layer deep), ResNet-V1-50 [8], ResNet-V2-50 [9], Inception-V3 [25], Inception-V4 [24] and PNASNet-Large [15].

For training, we used clean and inpainted images based on the COCO [14] data set. To create the inpainted images, we used five inpainting algorithms [5, 7, 10, 29, 32] in eight total configurations.

The network architectures take a square image as an input and output the score—a single number where 0 means the image contains inpainted regions and 1 means the image is "clean." The loss function was mean squared error. Some network architectures were additionally trained to output the predicted class using one-hot encoding (similar to binary classification); the loss function for this case was softmax cross-entropy.

The network architectures were identical to the ones used for image classification, with one difference: we altered the number of outputs from the last fully connected layer. This change allowed us to initialize the weights of all previous layers from the models pretrained on ImageNet, greatly improving the results compared with training from random initialization.

For some experiments we tried using the RGB noise features [34] and the spectral weight normalization [18].

In addition to the typical validation on part of the data set, we also monitored correlation of network predictions with the subjective scores collected in Section 3. We used the networks to estimate the inpainting quality of the 33-image test set, then computed correlations with subjective results in the same way as the final comparison. The training of each network was stopped once the correlation of the network predictions with the subjective scores peaked and started to decrease (possibly because the networks were overfitting to the inpainting results of the algorithms we used to create the training data set).

4.3 Results

Fig. 6 shows the overall results. The no-reference methods achieve slightly weaker correlation with the subjective-evaluation responses than do the best full-reference methods. But the results of most no-reference methods are still considerably better than those of the

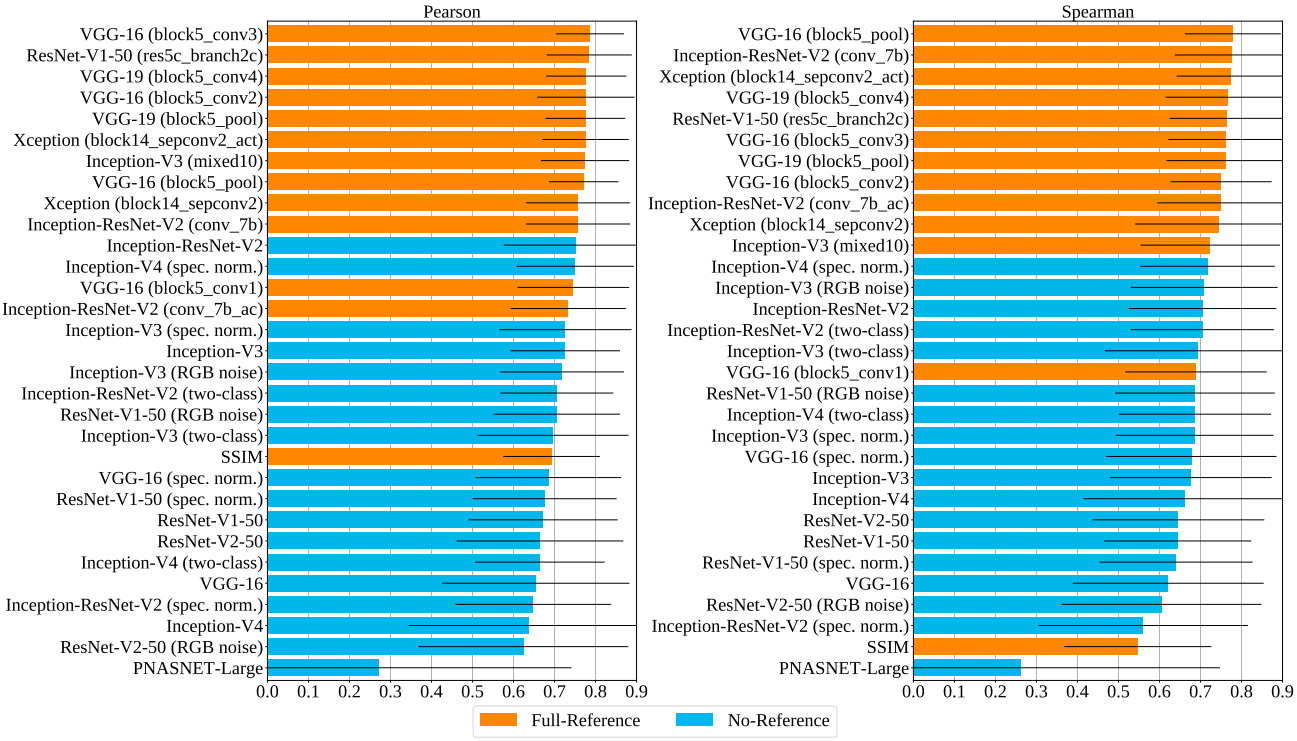

Fig. 6. Mean Pearson and Spearman correlations between objective inpainting-quality metrics and subjective human comparisons. The error bars show the standard deviations.

full-reference SSIM. The best correlation among the no-reference methods came from the Inception-V4 model with spectral weight normalization.

It is important to emphasize that we did *not* train the networks to maximize correlation with human responses. We trained them to distinguish "clean" images from inpainted images, yet their output showed good correlation with human responses. This confirms the observations made in [33] that deep features are good for modelling human perception.

5. Conclusion

We have proposed a number of perceptually motivated no-reference and full-reference objective metrics for image-inpainting quality. We evaluated the metrics by correlating them with human responses from a subjective comparison of state-of-the-art image-inpainting algorithms.

The results of the subjective comparison indicate that although a deep-learning-based approach to image inpainting holds the lead, classical algorithms remain among the best in the field.

We achieved good correlation with the subjective-comparison results without specifically training our proposed objective quality-evaluation metrics on the subjective-comparison response data set.

6. Acknowledgement

This work was partially supported by Russian Foundation for Basic Research under Grant 190100785 a.

7. References

[1] https://research.adobe.com/project/content-aware-fill/.

[2] J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj, and B. S. Manjunath. Exploiting spatial structure for localizing manipulated image regions. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[3] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[4] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, 2004.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009.

[7] K. He and J. Sun. Statistics of patch offsets for image completion. In *European Conference on Computer Vision*, pages 16–29. Springer, 2012.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645, 2016.

[10] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):107, 2017.

[11] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[12] H. Li, G. Li, L. Lin, H. Yu, and Y. Yu. Context-aware semantic inpainting. *IEEE Transactions on Cybernetics*, 2018.

[13] H. Li, W. Luo, X. Qiu, and J. Huang. Image forgery localization via integrating tampering possibility maps. *IEEE Transactions on Information Forensics and Security*, 12(5):1240–1252, 2017.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755, 2014.

[15] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy. Progressive neural architecture search. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[16] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. In *The European Conference on Computer Vision (ECCV)*, 2018.

[17] P. Liu, X. Qi, P. He, Y. Li, M. R. Lyu, and I. King. Semantically consistent image completion with fine-grained details. *arXiv preprint arXiv:1711.09345*, 2017.

[18] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[19] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[20] C.-M. Pun, X.-C. Yuan, and X.-L. Bi. Image forgery detection using adaptive oversegmentation and feature point matching. *IEEE Transactions on Information Forensics and Security*, 10(8):1705–1716, 2015.

[21] R. Salloum, Y. Ren, and C.-C. J. Kuo. Image splicing localization using a multi-task fully convolutional network (mfcn). *Journal of Visual Communication and Image Representation*, 51:201–209, 2018.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] Y. Song, C. Yang, Z. Lin, H. Li, Q. Huang, and C. J. Kuo. Image inpainting using multi-scale feature image translation. *arXiv preprint arXiv:1711.08590*, 2, 2017.

[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[26] A. Telea. An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, 9(1):23–34, 2004.

[27] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[29] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan. Shift-net: Image inpainting via deep feature rearrangement. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[30] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[31] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018.

[32] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[34] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis. Learning rich features for image manipulation detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[35] X. Zhu, Y. Qian, X. Zhao, B. Sun, and Y. Sun. A deep learning approach to patch-based image inpainting forensics. *Signal Processing: Image Communication*, 67:90–99, 2018.