16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

# UMGAN: Generative adversarial network for image unmosaicing using perceptual loss

Kamran Javed Nizam Ud Din Seho Bae Rahul S. Maharjan Donghwan Seo
Juneho Yi
College of Information and Communication Engineering,
Sungkyunkwan University, Suwon 16419, Korea
{kamran, nizam, bseho, rmsingh, dong88.seo, jhyi}@skku.edu

## Abstract

*Image mosaicing conceals sensitive parts of an image. The objective of this work is to recover hidden semantic structure under mosaiced parts, especially focusing on facial images. While recent image completion methods based on deep learning have shown promising results on recovering damaged parts in an image, they have not addressed the problem of image unmosaicing. We present a Generative Adversarial Network (GAN) approach to image unmosaicing called UMGAN, which is an image-to-image translation method. We have found that exploiting perceptual loss together with low level $l_1$ loss and high level Structural SIMilarity (SSIM) loss is quite effective to attain visually plausible and semantically consistent results. We have evaluated our method on the CelebA and MIT-CBCL image datasets and achieved better perceptual results than state-of-the-art image completion methods.*

## 1 Introduction

Tremendous amount of media contents are broadcast daily by electronic and press media which harm the privacy of depicted persons. It is necessary to obscure the privacy against inference attacks [1]. For privacy reasons, sensitive information is obscured by image obfuscation techniques. Among these techniques are blurring, masking, head inpainting, and mosaicing (pixelation).

*Mosaicing* conceals identity by tessellating sensitive parts of an input image into tiles. Each tile is filled with the average value of the pixels, belonging to that tile and then spurious noise is added. Thus, recognition or recovery of the mosaiced entity (*e. g.,* face) is challenging. Mosaicing is mostly used for obfuscation of sensitive objects on television broadcasting, press, and social media to obscure entities. On the other hand, revealing a visually plausible and semantically consistent image particularly at obfuscated parts is required to meet public interests.

The goal of this research is to uncover mosaiced parts of an image, focusing on facial area. This problem is called as *image unmosaicing*. Given a mosaiced image as input, the aim is to generate a complete image with plausible and natural looking results particularly at the mosaiced parts.

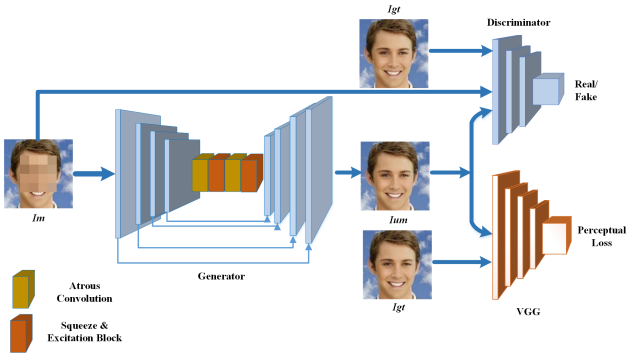



Figure 1: *The UMGAN model architecture*

We formulate image unmosaicing as an image-to-image translation problem and employ Generative Adversarial Network (GAN) for generating unmosaiced images conditioned on mosaiced images. To this end, we present a GAN approach as shown in Fig. 1. GAN generates unmosaiced image with fine details.

The main contributions of this research are summarized as follows:

- We propose *UMGAN*, a novel method for image unmosaicing using perceptual loss as feature level penalty.
- Our GAN network uses a combination of low level, $l_1$ loss, high level, *SSIM* loss and perceptual loss as reconstruction loss along with adversarial loss, which produce unmosaiced images that are both visibly plausible and semantically consistent.
- Extensive experiments show better performance of our method than state-of-the-art image completion techniques.

## 2 Related Work

**Image completion:** Given an image with some parts missing or masked, intent of image completion is to fill in the missing or corrupted parts with appropriate contents. There are many ways to do image

completion. Traditional approaches propagates image appearance information from neighboring pixels to fill holes [2] [3]. However, these approaches can only fill narrow target holes, where the texture is stationary. This scheme may result in artifacts for data where texture and color variance is large. Patch-based approaches work well for non-stationary data by pasting relevant patches from source image into target image [4]. However, these approaches work in iterative manner which is computationally expensive. Hence, they are ineffective for real-time applications.

Context Encoders (*CE*) was a pioneering work that used Convolution Neural Networks (CNN) for image completion problem [5]. *CE* recovers large missing part of an image conditioned on its surroundings. They achieved this by using a combination of pixel-wise $l_2$ reconstruction loss and adversarial loss. However, it produces undesirable artifacts and deficiency in high frequency details. Recently, *CE* based structural inpainting produced better results to complete complex structures by using additional perceptual reconstruction losses [6]. For semantically consistent image completion, Liu *et al.* proposed a method to employ supervision under multiple level of loss functions for image completion [7]. Motivated by [5, 6, 7], we make use of two different levels of reconstruction losses, low-level ($l_1$) and high-level (*SSIM*) along with adversarial loss.

Iizuka *et al.* proposed a network model that can complete arbitrary missing regions in an image by introducing a Globally and Locally consistent Image completion (*GLI*) approach [8]. Often times, its output has noticeable noise and artifacts especially when holes are at margins. Recently, Yu *et al.* made several improvements to *GLI* by introducing a two stage coarse-to-fine generative image completion model with a novel Contextual Attention (*CA*) layer [9]. It learns novel image structure by explicitly considering on related feature patches from relevant surrounding regions. A joint CNN optimization framework was introduced by Yang *et al.* to hallucinate missing image regions by modeling global content constraint and local texture constraint [10]. A multi-scale neural patch synthesis algorithm was used for high resolution image inpainting.

**Generative Adversarial Network (GAN):** GAN has shown promising results in image generation tasks since it was invented [11]. It consists of two models: a generator $\boldsymbol{G}$ that captures the data distribution, and a discriminator $\boldsymbol{D}$ that estimates the probability that a sample came from the training data rather than $\boldsymbol{G}$. GAN utilizes adversarial training to learn the generator and discriminator alternatively and has shown powerful ability to generate natural images [12, 13]. Due to GAN's ability to generate images, it has widely been used for research problems such as super resolution [14, 15], texture synthesis [16, 17], domain translation [18] [19] and image completion [5, 6, 7, 8, 9, 10, 20].

We opt to use an UNET [21] like architecture with Pix2Pix [18] to learn adapted loss and capture details particularly at unmosaiced parts. UNET performs well on bio-medical segmentation applications due to its skip connections at multiple scales of convolution. Pix2Pix effectively learns the mapping and loss function from one domain to another domain.

## 3 UMGAN

We present a GAN architecture called *UMGAN* which processes a mosaiced image and generates its unmosaiced image. GAN predicts unmosaic images with fine details. Fig. 1 shows the overall architecture of *UMGAN*.

We employ an UNET-like architecture with skip connections between the mirror layers of the encoder and decoder at three different scales in the generator. Specifically, $i$-th layer in the encoder is concatenated with ($N$ - $i$)-th layer in the decoder, where $N$ represents the total number of layers in the generator. Skip connections prevent information loss at bottleneck due to small feature map size. Different from the UNET, two layers of atrous convolution (*rate*: 2,4) and two squeeze and excitation (SE) blocks [22] are alternatively used in the middle of the generator network. The atrous convolution not only captures larger field of view but also decreases the number of parameters. A larger field of view helps yield more semantically coherent results. SE block improves the representational power of a network by enabling it to perform dynamic channel-wise feature recalibration. It learns the weights for each channel in the feature space. The encoder consists of five convolution layers. The decoder is the same as the encoder except convolution is replaced by deconvolution (transpose convolution). Here, each convolution and deconvolution layer means *relu+conv+instant_norm*. We used Patch-GAN discriminator which penalizes dissimilar structure at the scale of patches [18]. The last layer of generator and discriminator have *tanh* and *sigmoid* activation function respectively.

Additionally, we have used feature level reconstruction error from pre-trained VGG-19 [23] as perceptual loss. Specifically, feature maps output of layer 3, 4 and 5 of VGG-19 are used to calculate perceptual loss because middle layers have both structural and low level information.

**Integrated Objective Functions:** Most of the image completion methods uses following loss categories: *Naturalness* loss (*e. g.,* Adversarial loss [11], Least squares GAN loss [24] or Wasserstein GAN loss [25]) and *reconstruction* loss. We opted to use reconstruction loss along with adversarial loss.

$$\mathcal{L}_{total} = \mathcal{L}_{adv} + \lambda.\mathcal{L}_{recon}, \tag{1}$$

$\lambda$ is constant to adjust the weight between adversarial loss and reconstruction loss. Reconstruction loss as expressed in equations 2,

$$\mathcal{L}_{recon} = \mathcal{L}_{l_1} + \mathcal{L}_{ssim} + \mathcal{L}_{perc}, \tag{2}$$

comprises of pixel-wise penalty $l_1$, structure level *SSIM* loss and VGG feature level penalty $\mathcal{L}_{perc}$ from pre-trained VGG-19 [23] . In addition to $\mathcal{L}_{recon}$, adversarial GAN loss $\mathcal{L}_{adv}$ [11] is used as expressed in equation 3.

$$\mathcal{L}_{adv} = min_G max_D \mathbb{E}_{I_m, I_{gt}}[\log(D(I_m, I_{gt})) + \mathbb{E}_{I_m} \log(1 - D(I_m, G(I_m)))] \quad (3)$$

Here **G** tries to minimize this objective against an adversarial **D** that tries to maximize it. Generator's output is an unmosaiced image $I_{um}$, with semantically correct structure with realistic result.

## 4 Experimental Setup

We have evaluated *UMGAN* on the publicly available, CelebA Face dataset [26]. It contains face images of various celebrities with wild backgrounds. We have aligned the faces using OpenFace dlib [27] and created mosaic dataset with mosaic size 128x128 out of 256x256 image. In order to compare the performance of our method with that of representative image completion methods, *CE* [5], *GLI* [8], and *CA* [9], we have trained these methods on the mosaic dataset.

For training, a pair of mosaiced image and ground truth are fed into the proposed network. The proposed network is trained using Adam optimizer [28] with momentum 0.5 and learning rate $2 \times 10^{-4}$. We used batch size 10 with random flip augmentation and trained the network for 500 epochs. We trained the network for different values of $\lambda$ and found $\lambda = 75$ shows better results in term of texture and naturalness. *UMGAN* is implemented using tensorflow and trained for 72 hours on NVIDIA GeForce 1080Ti GPU.

## 5 Results and Comparisons

Perceptual loss is feature level penalty to the network for achieving perceptually consistent unmosaiced images. To investigate the effectiveness of perceptual loss, we train the model with and without using perceptual loss. Fig. 2 shows the results of ablation study. *UMGAN* without perceptual loss is unable to recover eye color consistent with the ground truth. Whereas, by adding perceptual penalty, not only eye color but hair texture is accurately recovered under the mosaiced part. Perceptual loss provides more freedom to the regressor while focusing on meaningful image properties under the mosaiced part.

We provide the qualitative and quantitative evaluations of our model and compare it with other representative image completion methods, *CE* [5], *GLI* [8] and *CA* [9] by training them for unmosaicing problem.

**Qualitatively**: Fig. 3 reports a visual comparison. It can be seen that *CE*, *GLI* and *CA* failed to *a)* reconstruct plausible face semantics *b)* recovered part is not consistent with surroundings. We think that

(a) (b) (c) (d)

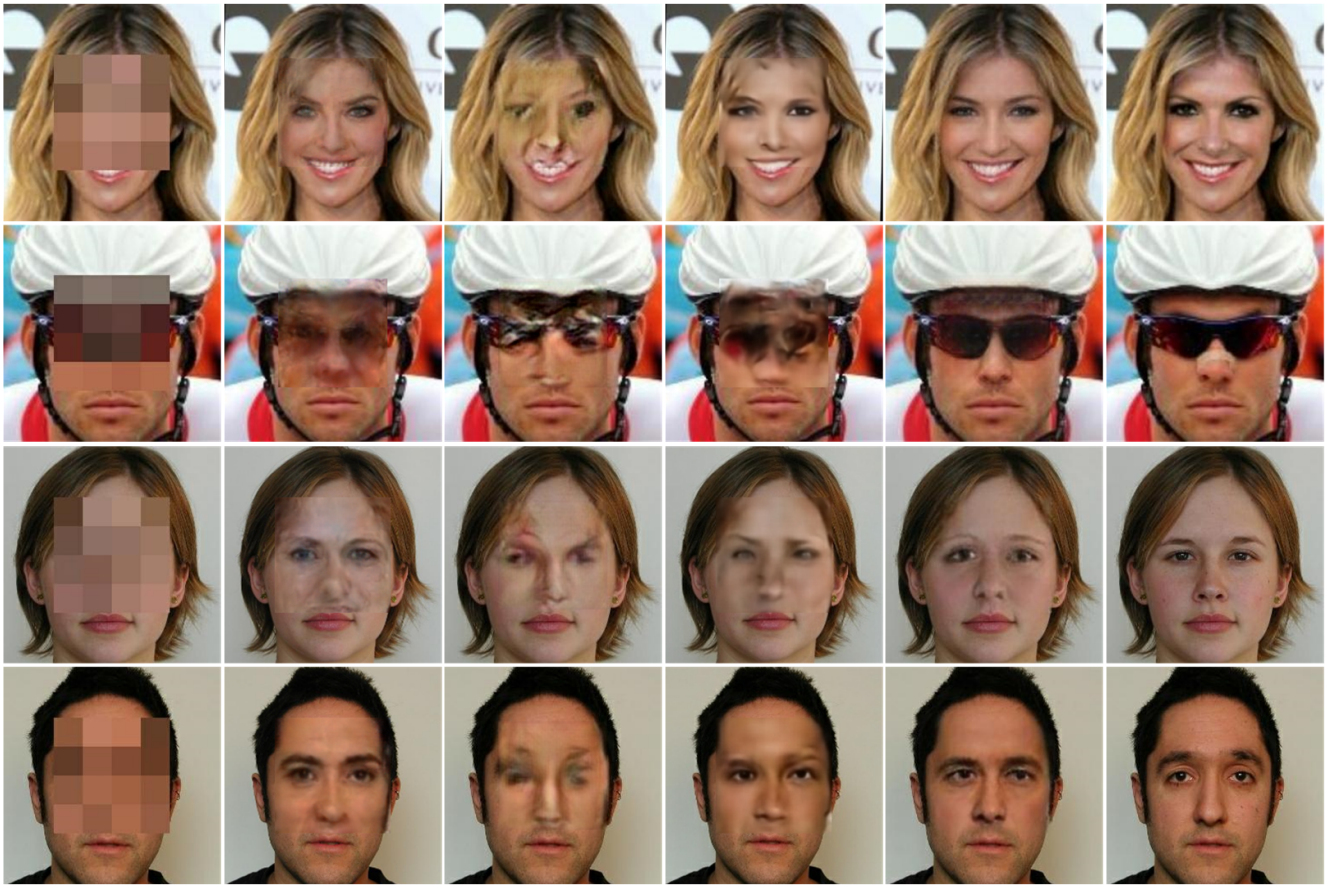
Figure 2: *Visual comparison of UMGAN with and without using perceptual loss.* ***(a)*** *Mosaiced images $I_m$,* ***(b)*** *Without using perceptual loss,* ***(c)*** *Using perceptual loss,* ***(d)*** *Ground truth images $I_o$. Note that perceptual loss successfully recovers correct eye color and hair texture.*

it is because their methods try to replace un-corrupted area adjacent to mosaiced parts with the ground truth, which may not be consistent with the recovered part. In contrast, the unmosaicing results of *UMGAN* are not only natural and visually plausible but also recover consistent structure throughout the image.

**Quantitatively**: Quantitative evaluation is challenging to judge whether an unmosaicing result contains visually plausible image structures and textures as there are many possible solutions different from the ground truth face. Nevertheless, we report the quantitative performance comparison in term of *MSE*, *SSIM*, and *PSNR*. As Table 1 shows that quantitatively, *UMGAN* performs better than *CE*, *GLI* and *CA* for the

Table 1: Performance comparison of different methods.

| Methods | MSE | SSIM | PSNR |
|---|---|---|---|
| *CE* [5] | 3036 | 0.483 | 18.52dB |
| *GLI* [8] | 3395 | 0.465 | 18.02dB |
| *CA* [9] | 1636 | 0.842 | 21.86dB |
| *UMGAN* | **481** | **0.853** | **26.68dB** |



(a) (b) (c) (d) (e) (f)

Figure 3: *Visual comparison with representative image completion methods* ***(a)*** *Mosaiced images* $I_m$, ***(b)*** *CE [5],* ***(c)*** *CA [9],* ***(d)*** *GLI [8],* ***(e)*** *UMGAN,* ***(f)*** *Ground truth images* $I_o$. *The first two rows show the results for CelelbA dataset and the next couple of rows for MIT-CBCL images*

image unmosaicing problem. By carefully adjusting loss function, it allows to generate unmosaiced images more close to the original images with better semantics.

*CE*, *GLI*, *CA* and *UMGAN* trained on the CelebA dataset are also tested on the MIT-CBCL face dataset [29]. MIT-CBCL face images have plain background whereas, CelebA images are captured in various environments with wild background. The unmosaiced image results of *UMGAN* for MIT-CBCL face images are visually plausible than that of *CE*, *GLI* and *CA* as shown in the last couple of rows of Fig. 3.

Additionally, *CE*, *GLI*, *CA*, and other image completion methods not only need location information of mosaiced part but also cost extra post processing step to remove artifacts along the boundary of the recovered region. Unlike *CE*, *GLI*, and *CA*, *UMGAN* works without location information of mosaiced parts and doesn't require any post processing step. Since our method regenerates a whole image rather than mosaiced part only, we do not have to check on the global and local consistencies separately. Due above mentioned benefits, *UMGAN* can be used where location information of damaged part is not available, for example, in case of recovering mosaiced faces at a designated receiver end. Since *UMGAN* generates perceptually correct and semantically consistent face images, it can help preserving identity in future studies.

## 6 Conclusion

We have presented *UMGAN*, a GAN based method for image unmosaicing. It effectively generates unmosaiced images close to the ground truth image by focusing on the structure recovery and naturalness recovery. Experimental results shows, perceptual loss helps to recover correct eye color and hair texture. Moreover, *UMGAN* is capable of recovering complex face semantic structure.

## 7 Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2016R1D1A1B03930428).

## References

[1] B. Xu, P. Chang, C. L. Welker, N. N. Bazarova, and D. Cosley, "Automatic archiving versus default deletion: What snapchat tells us about ephemerality in design." *In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing. ACM*, pp. 1662–1675, 2016.

[2] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 2000, pp. 417–424.

[3] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE transactions on image processing*, vol. 10, no. 8, pp. 1200–1211, 2001.

[4] V. Kwatra, I. Essa, A. Bobick, and N. Kwatra, "Texture optimization for example-based synthesis," in *ACM Transactions on Graphics (ToG)*, vol. 24, no. 3. ACM, 2005, pp. 795–802.

[5] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE CVPR*, 2016, pp. 2536–2544.

[6] H. V. Vo, N. Q. Duong, and P. Perez, "Structural inpainting," *arXiv preprint arXiv:1803.10348*, 2018.

[7] 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.

[8] S. I. E. S.-S. H. Ishikawa., "Globally and locally consistent image completion." *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, 2017.

[9] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," *Proceedings of the IEEE CVPR*, 2018.

[10] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," *IEEE CVPR*, 2017.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets." *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[12] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv: 1511.06434*, 2015.

[13] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *IEEE Int. Conf. Comput. Vision (ICCV)*, 2017, pp. 5907–5915.

[14] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network." in *CVPR*, vol. 2, no. 3, 2017, p. 4.

[15] B. Wu, H. Duan, Z. Liu, and G. Sun, "Srpgan: Perceptual generative adversarial network for single image super resolution," *arXiv preprint arXiv:1712.05927*, 2017.

[16] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 702–716.

[17] N. Jetchev, U. Bergmann, and R. Vollgraf, "Texture synthesis with spatial generative adversarial networks," *arXiv preprint arXiv:1611.08207*, 2016.

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks." *IEEE CVPR*, pp. 5967–5976, 2017.

[19] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," *arXiv preprint arXiv:1711.09020*, 2017.

[20] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. H. Johnson, and M. N. Do, "Semantic image inpainting with deep generative models." in *CVPR*, vol. 2, no. 3, 2017.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE CVPR*, 2018.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[24] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2813–2821.

[25] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5769–5779.

[26] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE ICCV*, 2015, pp. 3730–3738.

[27] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[28] D. P. Kingma and J. L. Ba, "Adam: Amethod for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2014.

[29] B. Weyrauch, J. Huang, B. Heisele, and V. Blanz, "Component-based face recognition with 3d morphable models," in *Proc. of CVPR Workshop on Face Processing in Video (FPIV'04), Washington DC, 2004*.