Person Inpainting with Generative Adversarial Networks

Gizem Esra Ünlü

1 Introduction

Image inpainting is a long-standing problem in computer vision where the goal is to recover the original image from the corrupted image. Filling the missing pixels so that an observer who does not know the original image cannot detect the changes is challenging since the inpainted regions must be realistic-looking and semantically plausible. The completed regions should be consistent with the rest of the image for pleasing results. This problem can be applied to a wide variety of problems, e.g. photo editing to remove unwanted objects or 3D object generation from occluded 2D images.

While recent inpainting methods have proven to work well on various texture, object, face and street-view databases, no work exclusively targets the in-the-wild human body inpainting task, specifically in the generative neural networks domain. For the case where missing human body parts are to be recovered, the challenge arises from the complexity of the data itself: masked areas of humans in various actions such as daily activities or sports are hard to predict since joints can be in numerous positions which can only inferred from the semantics of the image. Also, the existing generative methods are unable to learn the human figure which is apparent from the fact that for fully masked cases e.g. a fully masked hand, the system erases the hand and fills masked area with the background rather that recovering the body part. The inability to infer that there is a human in the background inhibits these methods from recovering the missing body parts. Another challenge is that masks can be anywhere on the image: background pixels as well as human areas should be inpainted successfully in equal measure (Fig. 1).

G. E. Ünlü (✉)
Bogazici University, Istanbul, Turkey
e-mail: gizem.unlu1@boun.edu.tr

S. Escalera et al. (eds.), *Inpainting and Denoising Challenges*,
The Springer Series on Challenges in Machine Learning,
https://doi.org/10.1007/978-3-030-25614-2_9

Fig. 1 Given an image with arbitrary masks (**a**) which can be on the body parts or the background, inpainting (**b**) is applied to recover the missing parts. Using this result with the recovered body parts, pose information is extracted from a complete human figure. (**c**) OpenPose [6] library was used for pose estimation

Previous inpainting approaches that are not deep learning based such as [5, 7], calculate randomized patch correspondences and select the closest one to the masked area in an iterative scheme. The downside of such methods is that it assumes the the missing patch is present in the background and fills the hole accordingly, unable to generate or hallucinate original objects which is necessary for the case of fully masked.

More recent deep learning based approaches use generative models for the inpainting task, with adversarial training [4] as the focus. As the pioneering work in this area, [8] proposed Context Encoder, an encoder-decoder based network which has an adversarial loss in addition to the regular L2-loss inherent to autoencoder type networks with success. In [9], a second discriminator was introduced to the previous work, to stabilize the training and produce more-realistic results by assessing the produced inpainted result both locally and globally. Inspired by these, [10] implemented an attention module, to find and borrow features to fill the holes from similar patches in the background (Fig. 2).

In this work, we apply Generative Adversarial Networks to the complex problem of human body part inpainting with a two stage coarse-to-fine generator/completion network and two critics: local and global. Experiments are done within Chalearn's Image Inpainting Challenge and we showcase the performance of GLS-GAN [16] loss in the inpainting domain. Example results are shown in Fig. 3.

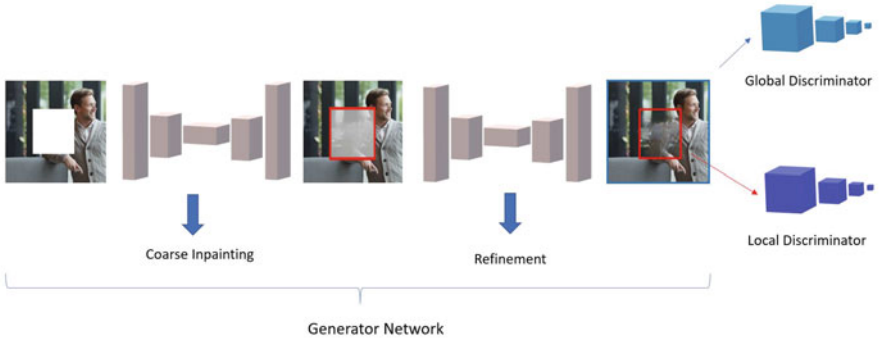

Fig. 2 Network architecture used for inpainting which was proposed in [10]. A masked image is given as input, an initial inpainting is done by the coarse network. The coarse result is then fed to refinement network to produce the final inpainting result. Two critics, global and local discriminators, assess the inpainting performance

2 Related Work

Natural image generation using generative networks has shown promising results. Since the release of Goodfellow et al.'s paper [4] *Generative Adversarial Networks* (GAN), numerous works on GAN stability have been proposed for the notoriously unstable adversarial training, resulting in an abundance of network architectures, losses, and regularization/normalization techniques as well as significant amount of 'tricks'[11]. One of the leading architectures is Deep Convolutional Generative Adversarial Networks (DCGAN)[12], where the GAN idea was first combined with convolutional layers with architectural constraints that are said to provide a more stable training setting. Some loss functions that are said to stabilize training are Wasserstein GANs (WGAN) [13], WGAN-GP[14] and least squares GAN (LSGAN) [15]. A more recent work in the in this area is Generalized Loss Sensitive GAN (GLSGAN [16]) which is a regularized model that can produce better samples from a probability distribution and is shown to be a generalized family of functions with WGAN and LSGAN as its special cases.

For the image inpainting task, older non-learning based methods try to recover the missing information from the neighbouring areas of the mask via a distance field [1, 2], but they fall short in performance when the masked area is relatively big and texture variance is high. Patch-matching algorithms such as [5, 7], iteratively search for the best corresponding patches in the foreground pixels to fill the corrupted regions without producing semantically-accurate inpainting with high computational costs. Recently, deep learning based methods which use convolutional layers have shown superior performance [3, 8, 10]. In [3], an encoder-decoder based network was developed for filling irregular sized holes in images, as opposed to rectangular masks seen in other works, e.g., [8–10]. This work also introduces a novel partial convolution operation for the image completion where the information

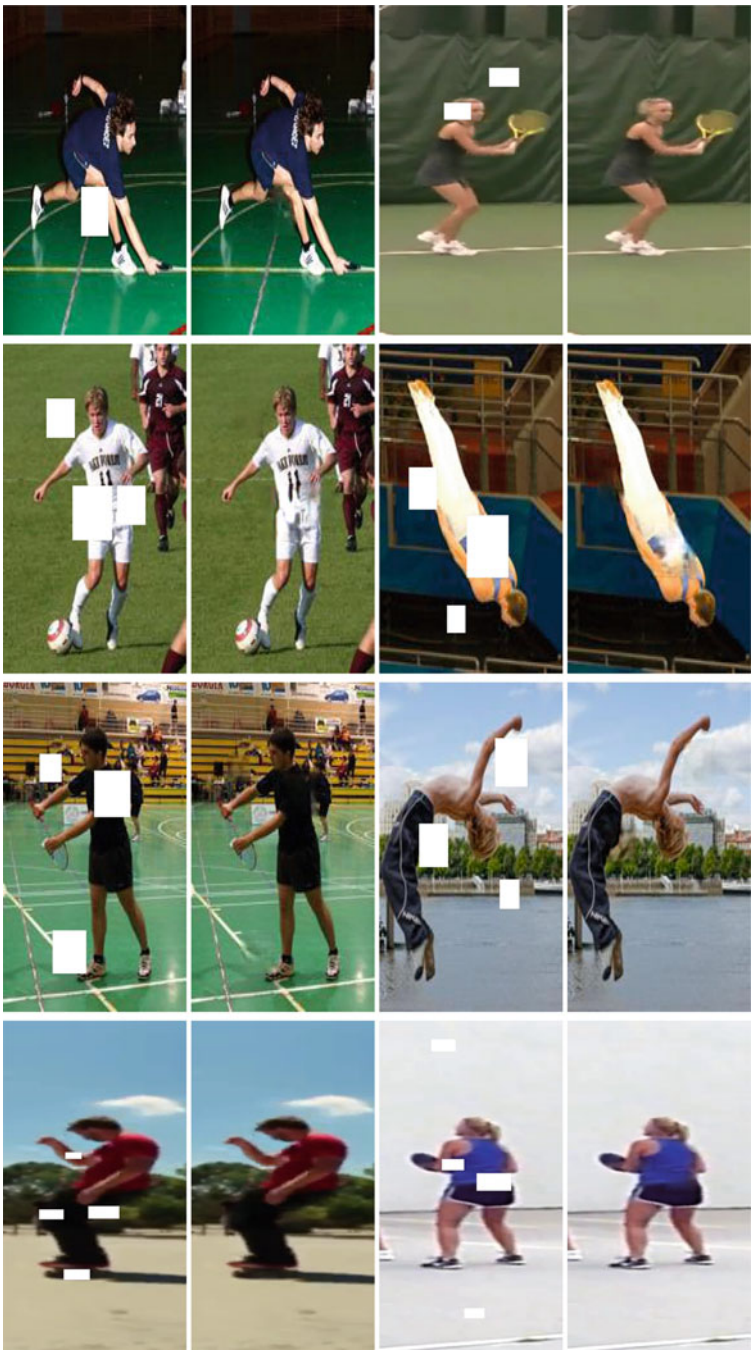

Fig. 3 Some results of the model on the test images are shown here. Masked images and their corresponding completions are given. The inpainted areas consistent with rest of the image with the body parts recovered adequately

for the masked areas are propagated only from existing pixels and excluding the empty pixels, which enable the network to condition the output on "valid inputs".

Generative networks, a sub-class of deep neural networks, are used quite often in the image inpainting task [8–10]. We build upon the recent work [10] and apply it to the human body part recovery task with modifications to the loss function to observe the effects on the problem.

3 Method

For the human-body inpainting task, we use a deep convolutional neural network based architecture shown in Fig. 2. The inpainting network fills the missing pixels in the input image, while two discriminator networks assess this output to determine how consistent the completion is both locally and globally. Training the completion network against the discriminators adversarially yields the final inpainting network that is able to produce realistic results.

3.1 Network Architecture

Our approach based on the inpainting architecture from the paper [10]. The inpainting network consists of two sub-networks: coarse inpainting network and refinement network. The coarse network takes a masked image to produce an initial completion which is then fed to the refinement network to obtain the final inpainting results. Dilated convolutions are used to incorporate surrounding context by expanding the receptive field of kernels which increases inpainting success. A fully convolutional attention module is integrated to the refinement network which learns to match the most relevant background patches to the patches in the masked area. Both discriminators, global and local, are fully convolutional with Leaky RELU as the activation function.

3.2 Loss Function

The network is trained with a combination of reconstruction losses and adversarial loss which was used in previous inpainting works [8–10]. The coarse network in the two-stage completion network is trained with L1 reconstruction loss only, whereas the refinement network is trained with both an L1-loss as well as an adversarial loss. This mixture of loss was well studied in recent inpainting literature and allows for a stable the training process [9].

Unlike the previous approaches where either a DCGAN based loss [8, 9] or a Wasserstein Loss (WGAN-GP) [10], we use the loss proposed in [16] for Loss-

sensitive GANs which was shown to outperform the original GAN[4] formulation and exhibits comparable results compared to other existing GAN models in terms of their ability to generalize to the underlying data distribution.

As oppose to the original GAN formulation where the goal of the discriminator is to determine if the sample is real or fake, GLS-GAN learns a loss function L_θ, parametrized by θ, with the objective of measuring how different a fake sample is from a real one. For the inpainting case, let x be the original image, $\tilde{x}$ be the corrupted/masked image and G_ϕ, parametrized by ϕ is the generator/completion network. Then, it is assumed that the loss for a real sample x, $L_\theta(x)$, is lower than the loss for a completed image $G_\phi(\tilde{x})$ by a margin:

$$L_\theta(x) \leq L_\theta(G_\phi(\tilde{x})) - \triangle(x, G_\phi(\tilde{x}))$$

where $\triangle(x, G_\phi(\tilde{x}))$ measures the difference between x and $G_\phi(\tilde{x})$, calculated simply as L1-distance.

After stating the main idea of GLS-GAN, the actual objectives functions of the critic L_θ and generator G_ϕ which are trained alternately, are written below. Following the notation from [16], for a fixed generator G_{ϕ_*}, the objective function for the discriminator is given as follows: let P_{data} and $P_{completed}$ be the real and completed image distributions. Then,

$$J(\theta, \phi_*) = \underbrace{\mathbb{E}}_{x \sim P_{data}(x)} L_\theta(x) + \lambda \underbrace{\mathbb{E}}_{\substack{x \sim P_{data}(x) \\ \tilde{x} \sim P_{completed}(\tilde{x})}} C\Big(\triangle(x, G_{\phi_*}(\tilde{x})) + L_\theta(x) - L_\theta(G_{\phi_*}(\tilde{x}))\Big)$$

in which the cost function C chosen as a leaky rectified linear function such that $C_v(a) = max(a, va)$ with slope $v \in (-\inf, 1]$ and λ is a positive balancing parameter. The generator's objective is minimized as shown below: for a fixed discriminator L_{θ_*},

$$H(\theta_*, \phi) = \underbrace{\mathbb{E}}_{\tilde{x} \sim P_{completed}(\tilde{x})} L_{\theta_*}(G_\phi(\tilde{x}))$$

4 Experiments

The performance of the inpainting network was evaluated on a dataset which was released within "Chalearn Looking at People Satellite Workshop ECCV'18"[17] which contains humans in arbitrary poses. The quantitative and qualitative results are given in the following sections.

4.1 Datasets

Chalearn's human inpainting dataset is created by collecting images from multiple human activity and pose datasets and consists of 41,076 images which are aligned to have a human at the center. This complex dataset contain humans in diverse poses and environments, such as various sports activities both indoors and outdoors. We use the given training/validation/test splits with percentages 70, 15, 15, respectively. This dataset does not have a fixed data size, therefore, all images are resized to 256×256 for training purposes. No data augmentation was used.

Implementation Details In all our experiments, the pre-trained model of [10] on Places2 [18] dataset is used and we finetune with Chalearn's dataset. For training, we use Adam optimizer[19] with $\beta_1 = 0.5$ and $\beta_2 = 0.9$ and learning rate is 1e−4. All models are with a batch size of 32 for 10K iterations (Fig. 4).

4.2 Qualitative Results

Since the challenge test dataset consists of images various resolutions as small as 71×154 and as big as 1819×1080 and different masks sizes, the qualitative results of the model should be examined in three cases: small images, big images, and mask size and placement. The results for the first case can be seen in Fig. 3 which show adequate inpainting performance in recovering missing body-parts as well as background patches. For the second case of big images, the performance degrades visibly and the model generates unrealistic hallucinations which is to be expected because the training of the inpainting network was done with images of resolution 256×256. An example inpainting for an image of resolution 1152×720 is shown in Fig. 5 first row where the model hallucinates unrealistic patches. The last case for the qualitative results explains the model's behaviour for masks that cover a body part fully, i.e. the head is completely masked. In this case, the model acts as an eraser, fills the body part with background which is explained by the fact that training is done without a body-part oriented approach. The model's further performance on test set images with various sizes can be seen in Fig. 4.

4.3 Quantitative Results

Table 1 shows the metrics used to evaluate inpainting performance in Chalearn's Image Inpainting Challenge and the corresponding results of the model we used. The metrics are Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Dissimilarity Index (DSSIM) and Weakly Normalized Joint Distance (WNJD).

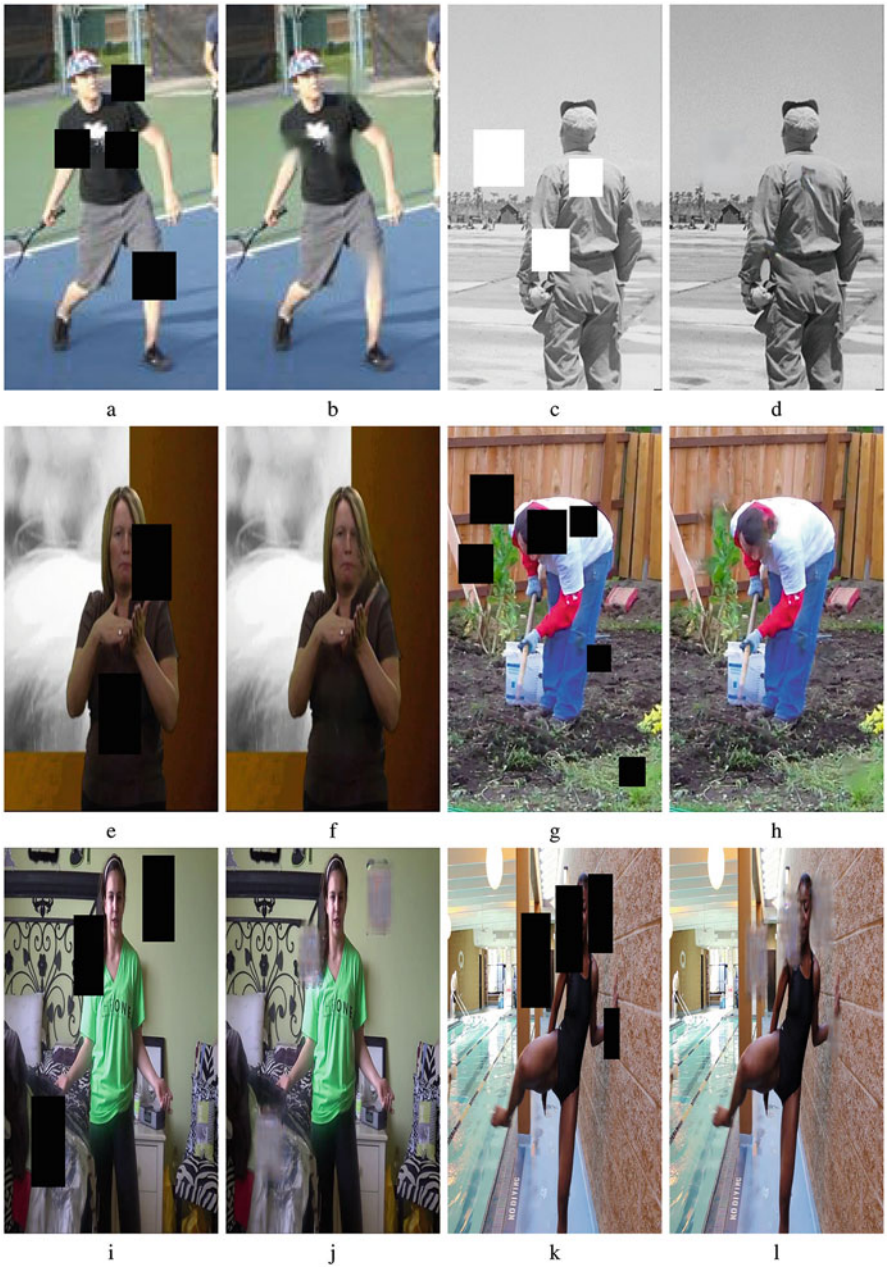

Fig. 4 Comparison of results from the model on the test dataset for different image sizes. The resolutions are as follows: **ab**:102×167, **cd**:258×413, **ef**:322×303, **hg**:448×720, **ij**:1608×1080, **kl**:1920×1080. As can be expected, the inpainting quality degrades as the resolution increases

Fig. 5 Failure cases for our approach. In the first row, the model generates patches that blurry and inconsistent with the rest of the image due to big image size. The second row image contains a mask that fully covers the person's leg which was erased in the completion

Table 1 The results obtained in Chalearn's challenge in image inpainting which earned first place

PSNR	21.8711893588
MSE	0.0158471260207
DSSIM	0.208834181594
WNJD	0.148852195872

5 Conclusion and Future Work

In this work, we have addressed the complex task of human-body inpainting with Generative Adversarial Networks. We have shown that using Generalized Loss Sensitive GAN loss produces good results in the human inpainting problem with respect to several quantitative measures. For future work, we plan to propose human-body specific approaches for person inpainting task which can also work in higher resolution images.

Acknowledgement This work was supported by the Scientific Research Projects Commission within Bogazici University (BAP). Project No: 14504.

References

1. Telea, Alexandru. “An image inpainting technique based on the fast marching method.” Journal of graphics tools 9.1 (2004): 23–34.
2. Bertalmio, Marcelo, et al. “Image inpainting.” Proceedings of the 27th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co., 2000.
3. Liu, Guilin, et al. “Image inpainting for irregular holes using partial convolutions.” Proceedings of the European Conference on Computer Vision (ECCV). 2018.
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y. et al. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672–2680)
5. Barnes, Connelly, et al. “PatchMatch: A randomized correspondence algorithm for structural image editing.” ACM Transactions on Graphics (ToG) 28.3 (2009): 24.
6. Cao, Zhe, et al. “Realtime multi-person 2d pose estimation using part affinity fields.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
7. Barnes, Connelly, et al. “The generalized patchmatch correspondence algorithm.” European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2010.
8. Pathak, Deepak, et al. “Context encoders: Feature learning by inpainting.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
9. Iizuka, Satoshi, Edgar Simo-Serra, and Hiroshi Ishikawa. “Globally and locally consistent image completion.” ACM Transactions on Graphics (TOG) 36.4 (2017): 107.
10. Yu, Jiahui, et al. “Generative image inpainting with contextual attention.” arXiv preprint (2018).
11. Kurach, Karol, et al. “The gan landscape: Losses, architectures, regularization, and normalization.” arXiv preprint arXiv:1807.04720 (2018).
12. Radford, Alec, Luke Metz, and Soumith Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks.” arXiv preprint arXiv:1511.06434 (2015).
13. Arjovsky, Martin, Soumith Chintala, and Léon Bottou. “Wasserstein gan.” arXiv preprint arXiv:1701.07875 (2017).
14. Gulrajani, Ishaan, et al. “Improved training of wasserstein gans.” Advances in Neural Information Processing Systems. 2017.
15. Mao, Xudong, et al. “Least squares generative adversarial networks.” Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017.
16. Qi, Guo-Jun. “Loss-sensitive generative adversarial networks on lipschitz densities.” arXiv preprint arXiv:1701.06264 (2017).
17. http://chalearnlap.cvc.uab.es/dataset/30/description/
18. Zhou, Bolei, et al. “Places: A 10 million image database for scene recognition.” IEEE transactions on pattern analysis and machine intelligence 40.6 (2018): 1452–1464.
19. Kingma, Diederik P., and Jimmy Ba. “Adam: A method for stochastic optimization.” arXiv preprint arXiv:1412.6980 (2014).
    [6] Cao, Zhe, et al. “Realtime multi-person 2d pose estimation using part affinity fields.” arXiv preprint arXiv:1611.08050 (2016).