Attention-based Adversarial Training for Seamless Nudity Censorship

Conference Paper · March 2019

CITATIONS
0

READS
665

3 authors:

Gabriel Simões
Pontifícia Universidade Católica do Rio Grande do Sul
16 PUBLICATIONS **48** CITATIONS
SEE PROFILE

Jônatas Wehrmann
Pontifícia Universidade Católica do Rio Grande do Sul
22 PUBLICATIONS **145** CITATIONS
SEE PROFILE

Rodrigo C. Barros
Pontifícia Universidade Católica do Rio Grande do Sul
93 PUBLICATIONS **1,094** CITATIONS
SEE PROFILE

Attention-based Adversarial Training for Seamless Nudity Censorship

Gabriel S. Simões, Jônatas Wehrmann and Rodrigo C. Barros
Machine Intelligence and Robotics Research Group
School of Technology, Pontifícia Universidade Católica do Rio Grande do Sul
Av. Ipiranga, 6681, 90619-900, Porto Alegre, RS, Brazil
Email: {gabriel.simoes.001, jonatas.wehrmann}@acad.pucrs.br, rodrigo.barros@pucrs.br

***Abstract*—The amount of digital pornographic content over the Internet grows daily and accessing such a content has become increasingly easier. Hence, there is a real need for mechanisms that can protect particularly-vulnerable audiences (e.g., children) from browsing the web. Recently, object detection methods based on deep neural networks such as CNNs have improved the effectiveness and efficiency of identifying and blocking pornographic content. Even though improvements in detecting intimate parts have been significant, the occlusion of the content is still primarily done by either blurring or removing regions of the image in an intrusive fashion. A recent study has addressed the problem of censoring the pornographic content in a non-intrusive way by generating the so-called seamless censorship via cycle-consistent generative adversarial networks. Such an approach has managed to automatically add bikinis to naked women without explicit supervision or paired training data. In this paper, we extend that method by designing a novel cycle-consistency framework that leverages sensitive information from an attention-based multi-label convolutional neural network. We evaluate the quality of our novel generative model by conducting a web survey with over 1000 opinions regarding the resulting images from our method and from baseline approaches. Results of the survey show that our method considerably improves the state-of-the-art on the seamless censorship task.**

***Index Terms*—adversarial training, attention, convolutional neural networks, deep learning, GANs, pornography censorship.**

I. Introduction

The amount of adult content available on the internet grows daily. Cooper [1] associates the growing of that material to three main factors: (i) *accessibility*, since it is easy to access pornographic content online; (ii) *affordability*, given that the adult content is available with low monetary cost; (iii) *anonymity*, which protects users and encourages access.

Considering mostly the accessibility aspect from [1], there is a clear need for automatic approaches capable of identifying adult content and censoring it when accessed by vulnerable audiences (e.g., children and specific religious groups). For instance, an automatic system could be used in live broadcasts to protect audiences from explicit body-part exposition. This important task unfortunately has not received enough attention from the scientific community in order to allow the development of automatic methods for censoring explicit content. There is also the need for widely-spread benchmarks to properly evaluate novel data-driven models.

Seminal work for pornography censorship based on deep learning has mostly focused on class-based predictions, hence images or frames from a video containing pornographic content have to be fully removed [2]. An alternative for the classification approach is to generate bounding boxes surrounding the intimate parts, so the images can be partly censored by either blurring the enclosed regions or adding black boxes. However, even the bounding box approach does not hide the fact that the image is originally pornographic, bearing in mind the intrusiveness of such a method.

In an attempt to develop a non-intrusive approach for pornography censorship, More et al. [3] have addressed the problem as an image-to-image translation task, where images from domain A (naked women) are converted to another domain B (women wearing bikinis). Such a method has the advantage of translating images with no explicit supervision (bounding boxes or segmentation masks) and does not require paired training examples (e.g., the same person with and without a bikini). That method addresses the lack of instance level supervision by using two domain sets, each one representing the concepts of A and B. Thus, one needs to train a generator to map $G : A \rightarrow B$, which will then be capable of transforming naked-women images into their counterparts (women in bikinis). Another contribution of [3] is the construction of a novel unaligned dataset containing either nude women or women wearing bikini.

The main motivation of the work in More et al. [3] is to avoid ruining the user experience while consuming content that may occasionally contain nudity. Their solution workflow was inspired by CycleGan [4], though the authors had to remove the background of the input images to bring the generator focus to the specific subject in order to achieve better-looking images. Such a strategy, however, has the disadvantage of losing an important component of the original image, which is the background. Concurrently, Mo et al. [5] propose a method that incorporates the instance information of multiple target objects in the framework of generative adversarial networks (GAN), called instance-aware GAN (InstaGAN), which translates both an image and the corresponding set of instance attributes while maintaining the permutation invariance property. The method uses object segmentation masks for instance information, which is a good representation for instance shapes since it

contains object boundaries, while ignoring other details such as color and background. However, their method depends on semantic segmentation labels (i.e., pixel-wise annotation) for model training, constraining the adaptation for new problems where pixel-level annotation is not available.

In this work, we address several previous limitations of the seamless censorship approach to improve the overall quality of the generated images. We aim to preserve peripheral parts of the image (e.g., background and faces) while maintaining our focus on covering the body parts. Our solution comprises a multi-label convolutional network that is trained to identify 5 classes (body parts): i) *butt*, ii) *breasts*, iii) *penis*, iv) *vagina*, and v) *no-nudity*, where *no-nudity* means the absence of sensitive parts. The architecture of the network implements a Scaled Dot-Product Attention branch [6], which generates attention masks that focus on the main subject of an input image. To improve the standard method, we merge the attention mask to the input volume before and then inside the generator of a cycle-consistent framework. The intuition is that the attention mask will be capable of highlighting the target areas of the image, contributing to shift the focus of the generator to act only over the sensitive areas. To improve peripheral (and thus overall) image quality, we merge the generator output with the original input image.

To evaluate the results of our method, we have conducted a web survey that have collected more than 1000 opinions. The survey results demonstrate that the human-perceived quality of our generated images is significantly superior than previous methods for the seamless censorship task.

II. Related Work

In this section, we discuss work that address the two main concepts related to this paper: generative networks in the context of image-to-image translation; and studies that provide datasets and methods for identifying/classifying adult content in both images and videos.

A. Image-to-Image Translation

Generative Adversarial Networks (GANs) [7] is a framework that trains two networks simultaneously in a zero-sum game. During training, a generator G produces synthetic images while the discriminator D learns to identify whether the input was drawn from a real dataset or was produced by the generator. The generator thus learns to produce realistic images that can trick the discriminator into producing false responses. The game is defined as follows:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))], \quad (1)$$

where $\mathbf{z}$ is a low-dimensional latent vector drawn from a known distribution (such as uniform or Gaussian) and is fed as input noise to G. In traditional GANs, the final G model can generate multiple images by varying the sample from the latent vector $\mathbf{z}$.

Whereas in traditional GANs the generated images are unconditioned, Conditional Generative Adversarial Networks (CGANs) can generate images based on a certain input [8]. This type of framework pave the way for tasks that focus on changing specific characteristics of the image, such as super-resolution [9] and image inpainting [10]. Generally speaking, CGANs are a way to solve image-to-image translation tasks, where images from a certain domain A can be mapped to their corresponding image in domain B .

Unless the experiment design comprises a task in which the image can be collected in both domains (such as images during day and night or winter and summer), finding paired examples from both domains to train the model becomes cumbersome and sometimes demand expensive expert help. To overcome this limitation, Zhu et al. [4] proposed CycleGAN, an unpaired image-to-image translation approach. More recently, More et al. [3] extended CycleGAN to seamlessly cover intimate body parts. To enhance the generated image quality, the method comprises a step that detaches the background of the input images to shift the focus of the generator to the specific subject at hand. This solution contributes for better covering results, but at the expense of losing peripheral characteristics of the original image, especially the background. More et al. [3] also contributed by publicly providing a novel unaligned dataset containing images from both domains, i.e., naked women vs. women wearing bikinis.

B. Adult Content Filtering

Avila et al. [11] introduced one of the first datasets for adult content detection, namely NPDI. It comprises nearly 80 hours from 802 videos downloaded from the internet. NPDI is divided into two disjoint classes: adult and non-adult videos. The non-adult class is further sub-divided in 200 easy-to-classify videos and 200 hard-to-classify videos. The latter includes videos with scenes of people in beaches, wrestling, and swimming.

A novel dataset for adult content classification, namely *DataSex*, was introduced by Simões et al. [12]. The authors provided the largest dataset for binary classification of pornographic images. It contains a collection of $\approx 300,000$ images that are equally distributed into adult and benign classes. They also provide splits for training and validation purposes. *DataSex* was built by crawling around $300,000$ publicly available images from adult websites. Simões et al. [12] reports classification results of $\approx 95\%$ accuracy in DataSex's test set, achieved by fine-tuning a pre-trained GoogleNet [13].

The work described in [14] is the first to use deep neural networks for pornography classification in videos. It proposes a method that requires fine-tuning two distinct ConvNets, namely *AlexNet* [15] and *GoogLeNet* [13]. Next, the pre-trained models are fine-tuned in each fold of the NPDI dataset. Note that such an approach requires training 10 distinct models: one model per training fold (5) and per network (2). In order to avoid overfitting, the authors apply strong dropout rates and data augmentation with randomly selected image-crops in the training phase.

Recently, Wehrmann et al. [16] presented ACORDE (Adult Content Recognition with Deep Neural Networks). The proposed approach uses a convolutional architecture as a feature extractor and a Long Short-Term Memory network (LSTM) [17] to perform video classification. The method extracts feature vectors from the keyframes of NPDI to construct video semantic descriptors that feed an LSTM responsible for analyzing the video. The entire pipeline works in an end-to-end fashion, eliminating the fine-tuning phase and the ConvNet re-training. ACORDE establishes itself as the current state-of-the-art for adult video detection in NPDI.

III. Method

In this paper, we propose a novel method for seamless nudity censorship in images, namely *AttGAN+*. For such, we make use of an adversarial-training image-to-image translation approach that draws bikinis over nude female bodies, preserving peripheral parts of the image such as the background and people's faces. Our solution encloses an attention convolutional network which is trained to identify sensitive (intimate) body parts to create attention maps that will be used to help guiding the generators within the image-to-image translation framework. We embed the attention masks as additional information to reinforce the need of focusing on the intimate body parts that should be transformed in the output image.

A. Attention Network

Our main contribution is regarding the use of an additional attention network responsible for recognizing images that contain explicit content, so the generative network can focus on those regions in order to generate state-of-the-art seamless censorship. For training such a network, we need a dataset for body-part recognition. Given that the ones available for censorship or nudity detection do not contain labels for each body-part, we introduced a novel dataset, namely *Dataset for Pornography Censorship* (*DPC*). This dataset has been manually annotated in a per-body-part granularity (see Section IV for more details).

In order to learn the attention maps, we use a pre-trained ResNet-152 [18] removing the last two layers (last fully-connected and global pooling layers). By removing those layers, the network outputs a tensor $\mathbb{R}^{f \times w \times h}$, where f is the number of filters, w is the width, and h is the height of the feature map. Since we need an attention map for each class, we add a convolutional layer with c filters, where c is the number of classes, generating a $c \times w \times h$ feature tensor. For generating class scores, we apply an average global pooling so the spatial dimensions are summarized into a c-dimensional vector. This vector is then activated with the logistic sigmoid function σ, generating the final model predictions denoted by $\hat{\mathbf{Y}}$. Note that by averaging all the spatial dimensions directly onto the class space, we enforce the image regions related to each class to present scores large enough to outperform those related to classes absent from the image.

The attention network is trained in a multi-label fashion given that a single image may contain several explicit body parts altogether. Therefore, similarly to [19]–[21], we optimize a per-neuron binary cross-entropy loss function, as follows:

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\Big[\mathbf{Y}_{ij} \times log(\hat{\mathbf{Y}}_{ij}) + (1-\mathbf{Y}_{ij}) \times log(1-\hat{\mathbf{Y}}_{ij}) \quad (2)$$

where N is the number of instances within a mini-batch, C is the number of classes, and $\mathbf{Y}_{ij}$ is the j^{th} ground truth label for the i^{th} instance.

After training the attention network using *DPC*, we remove the final average global pooling function so the spatial dimensions are preserved. Each spatial position is then normalized within the range $[0, 1]$ by applying a 2D-SOFTMAX function on the class dimension of the output. Ultimately, we aggregate all the maps from classes that represent any explicit body part through a max-pooling operator. Such an aggregation results in an attention tensor of size $2 \times w \times h$, in which the first $w \times h$ map represents all the regions responsible for the explicit body parts, while the second depicts regions without any sensitive content. Note that, in theory, by introducing the first activation map to the generative network within the image-to-image translation framework, it should be able to only modify regions of the original image that contain explicit content, while keeping unchanged the *safe* regions of the image. Formally, the forward pass in the attention network that is used to generate the attention map is denoted by $AN(I) = M$ where M is the $2 \times w \times h$ attention mask generated from input image I.

B. AttGAN+

Figure 1 presents our novel framework for seamless nudity censorship, *AttGAN+*. Our solution is inspired on [4], preserving the original architectures of generators G and discriminators D. We add an additional convolutional network to the flow which comprises the following 7 steps: i) the attention mask generation when input A flows within the attention CNN, resulting in an attention map (upsampled according to the image input size); the final attention mask is a $256 \times 256 \times 1$ tensor originated from the max of the first 4 channels of the attention output. ii) we fuse the attention mask as additional information to the input image creating a new channel, transforming the original image into a volume of dimension $W \times H \times 4$; this volume flows through the generator network up to the second convolutional layer; iii) we sum the attention map across all output channels of the first convolutional layer activations produced by the first convolutional layer of G_{AB}; iv) we invert the original attention mask; v) we concatenate the inverted mask to the generated image (Fake B) to build the volume prior to the reconstruction; vi) we sum the inverted attention mask across all output channels of the first convolutional layer activations inside G_{BA}; and vii) we match the reconstructed image with the original input. The generator and discriminator architectures

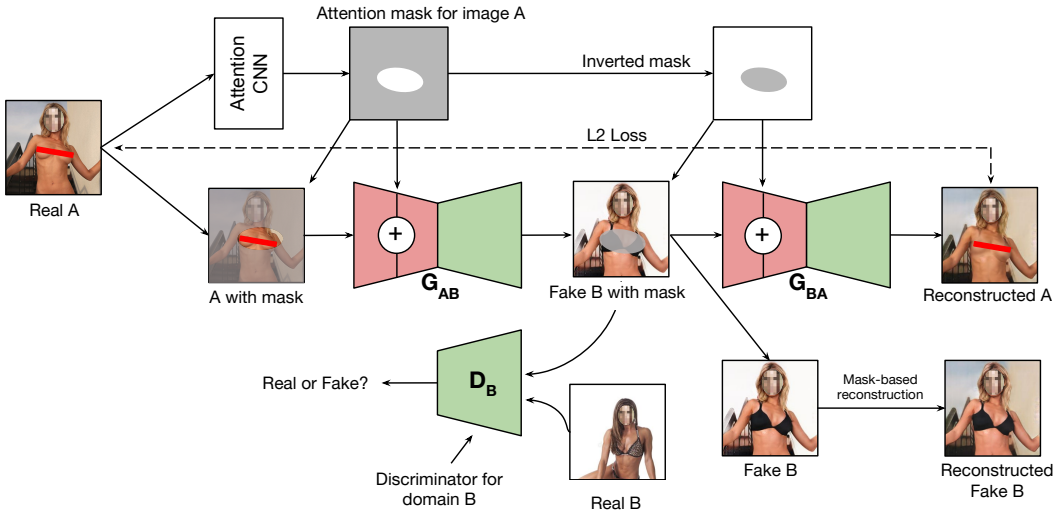

Fig. 1: *AttGAN+* framework. Real A is an image sampled from domain A, i.e., the set of images with naked women. Analogously, Real B is an image sampled from domain B, which is the set of women wearing bikini. Generator G_{AB} transforms images from domain A into B, and G_{BA} maps an image that was generated as belonging to B (Fake B) back to domain A. Both G_{AB} and G_{BA} are guided by the attention masks of each domain. This is done by concatenating the mask at the input, and also through a summation operation for every channel at the output of the first convolutional layer. The inverted mask serves the same purpose, but highlighting regions without sensitive content. For privacy reasons we have manually blurred all faces and covered the intimate parts with red tags.

were based on the work by More et al. [3], which uses a 9-Blocks ResNet Generator – an autoencoder that interposes residual connections and bottleneck layers inspired on [22].

We test with other two variations of *AttGAN+*: *AttGAN*and *AttGAN++*, as detailed next.

1) AttGAN: The first approach that we have developed to combine the attention mask with the cycle-consistent image-to-image translation framework is called *AttGAN*. In *AttGAN*, we use the attention data generated by AN as additional information to the framework by only concatenating it at the generators G_{AB} and G_{BA} input, which means *AttGAN* comprises only steps i), ii), iv), and v) of *AttGAN+*.

2) AttGAN++: Our second variation, namely *AttGAN++*, enhances *AttGAN+* by merging the G_{AB} output with the original input image. We use the attention maps produced by AN to make a guided merging, where sensitive nude parts are covered by fake bikinis, generated by G_{AB}. The idea of *AttGAN++* is to maintain the improvements of *AttGAN+* while keeping peripheral regions of the image entirely preserved.

IV. *DPC*: *Dataset for Pornography Censorship*

DPC is a dataset that contains $3,000$ images for detection of intimate parts. Each image contains at least 1 object from any of the following classes: *butt*, *breast*, penis (*frontalm*), or vagina (*frontalf*). *DPC* is divided into training ($2,100$ images), test (600), and validation (300) sets. It comprises images crawled from the *wild* presenting large variability in terms of:

- scale/size, where the image dimensions range from 170 to $3,000$ pixels;
- lighting conditions;
- scene composition; and
- ethnicity of the components.

To build *DPC*, we randomly selected image samples from DataSex, the pornographic classification dataset presented in [12]. Those images were annotated for object detection following two steps: i) initial annotation, and ii) reviewing step. In the first step, all images were divided into four groups, each of which assigned to a human annotator. In the second step, annotators reviewed each other's work. The complete work pipeline took about a month to be completed.

Given the body parts subject, the total annotated area is relatively small. For instance, PASCAL VOC [23] presents an average annotated area of 20.8% across all images, while in *DPC* such an area consists of $\approx 11.8\%$. A total of $6,500$ objects were manually annotated, implying an average of 3.4 objects per image (1 being the minimum number of annotations and 11 the maximum). Table I shows the bounding box (annotations) distribution through classes and images.

TABLE I: Body-part distribution in *DPC*.

	butt	*breast*	*penis*	*vagina*	total
bounding boxes	$1,200$	$2,693$	$1,265$	$1,383$	$6,541$
images	$1,122$	$1,537$	$1,134$	$1,335$	$3,000$

V. Experimental Setup

In this section, we introduce the datasets that were used for training both the multi-label classification model (AN) and the generative models. We also briefly describe how we evaluate

Fig. 2: The same sample from domain A after translation within 5 different training epochs. Results from method *AttGAN++*. For privacy reasons we have manually blurred all faces.

the quality of the generated images and compare them with baseline methods.

A. Datasets

We train our attention network AN in a multi-label fashion to predict 5 classes: *no-nudity*, *butt*, *breast*, *penis* and *vagina*. In this context, *no-nudity* means the absence of intimate parts and the other classes represent content that must be covered. To train the network we use *DPC*, a dataset originally built for addressing the problem of pornography censorship as an object detection task. We first adapt the dataset for a multi-label task by using the available bounding box labels. The task becomes the prediction of whether an object from a class appears in the image, and we no longer worry in detecting its (x, y) position or finding multiple occurrences. *DPC* is a subset of DataSex [12], a binary pornographic dataset for image classification.

1) Bikini dataset: The Bikini dataset that we adopt to validate our method and run our experiments was presented by More et al. [3]. To build the dataset, the authors scrapped images from the Internet for nude women and women wearing bikinis. They keep only one single person per image. The dataset was divided into training and test sets. For nude women (domain A) the final image count was 921 for training and 103 for test, and for women wearing bikinis (domain B) the final image count was 1044 training images and 117 test images.

B. Hyperparameters

For training our approach, we start from scratch and we keep all the hyperparameters the same as those used in the work by More et. al. [3]. The generator and discriminator loss functions were the same as [4], however we train our models for 500 epochs with an initial learning rate of 1×10^{-4}, which we keep by 100 epochs, and we decay it linearly to zero for 400 epochs. We conducted experiments for 3 versions of the method: *AttGAN*, *AttGAN+*, and *AttGAN++*.

C. Evaluation

To qualitatively validate the generated results, we have distributed a web form composed of 50 images from our test set. For each input image, we ask people to compare the *baseline* (the work by More et al. [3]) with *AttGAN* and *AttGAN+* specifically for the seamless censorship task. When none of the images seem to present adequate results, users were instructed to check option D, which always represented *poorly-translated* images. Options A, B and C presented images generated by the *baseline*, by *AttGAN*, and by *AttGAN+*. Those options were shuffled not to bias users into checking the same alternative. Note that *AttGAN++* is not included in the survey since it was created after we had evaluated the survey results.

VI. Experiments

In this section, we detail experiments that show the performance of 3 variations of our method, namely *AttGAN*, *AttGAN+*, and *AttGAN++*. For each variation, we keep optimization hyperparameters and loss functions equivalent to [3], which is the work we use as baseline method (hereby called simply *baseline*). We use input dimensions equal to 256×256 for the attention network and for the generators G_{AB} and G_{BA}. The attention network output has dimensions 8×8. We apply bilinear upscale algorithm to make the attention mask size compatible with the generator input. Figure 2 depicts the training evolution for 5 different training epochs from *AttGAN++*, for on of the samples of our test set. It is clear that the optimization process is generating better and better results as training advances.

A. Attention Network Results

Table II shows results for four versions of AN, ResNet-[34, 50, 101, 152], with an adapted final convolutional layer. The last layer can also be a transposed convolutional layer (denoted by *T*), making the output volume slightly larger since we use kernel size of 3. All models were trained and evaluated on the multi-label dataset *DPC*(validation set). In addition, we also computed AN accuracy for nudity detection in the Bikini training set. Results show that for deeper networks, the use of transposed convolutions seem helpful for achieving better predictive performance across distinct datasets. In addition, one can see that the best performing model is the ResNet-152T, which outperformed all other networks in both evaluation sets. Hence, this model was used for the attention map extraction in all versions of *AttGAN+*.

B. Generation Results

Figure 3 shows 8 test-set samples translated by the baseline [3] and by the 3 variations of our method. Figure 3b

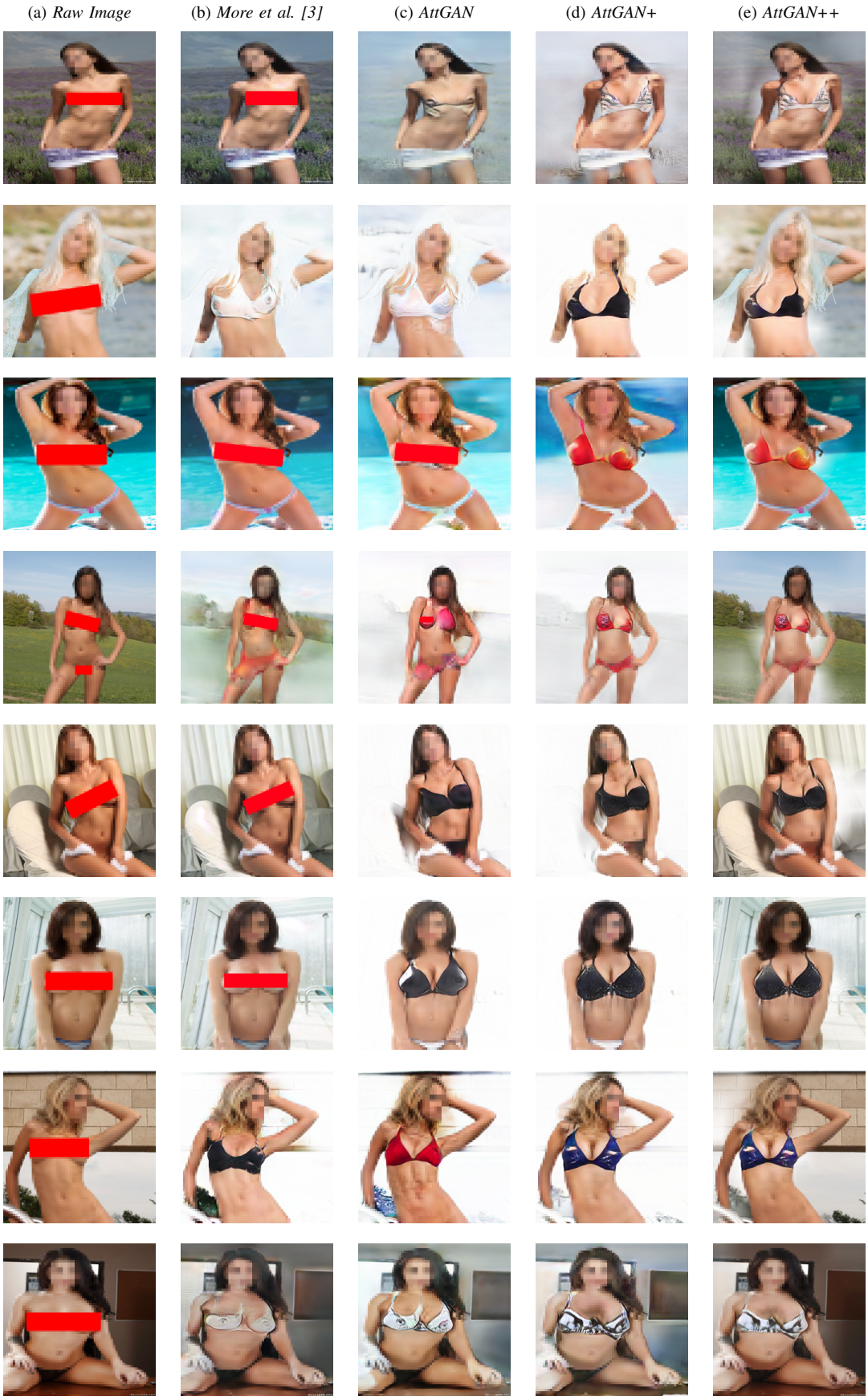

Fig. 3: Results after 500 epochs. a) original image. b) baseline by More et al. [3]. c) *AttGAN*. d) *AttGAN+*. e) *AttGAN++*. For privacy reasons we have manually blurred all faces and covered the intimate parts with red tags.

TABLE II: Attention network results.

Model	Validation	Training (Bikini)
ResNet-34	97.83%	98.42%
ResNet-34T	96.17%	96.92%
ResNet-50	97.33%	96.17%
ResNet-50T	97.33%	97.67%
ResNet-101	97.33%	97.82%
ResNet-101T	97.17%	98.57%
ResNet-152	97.67%	97.90%
ResNet-152T	**97.90%**	**98.65%**

illustrates the output of the baseline method trained by 500 epochs, while Figures 3c and 3d show outputs for *AttGAN* and *AttGAN+* also after 500 epochs. Observing the outputs, we clearly identify that *AttGAN+* has the best covering capability and the most coherent bikini shapes when compared to the baseline and to *AttGAN*. In Figure 3e we depict *AttGAN++*, which merges the raw input image with the method output. We use the attention mask to guide the merging process. This approach seems to be the best option since it carries all advantages from *AttGAN+* while being better at preserving the peripheral areas such as faces and background.

We perform a web survey to evaluate our method. The survey ask users to identify the best method among the baseline, *AttGAN*, and *AttGAN+*. Table III compile results for 2 particular cases: (i) considering option D (none of the previous methods are good enough), and (ii) considering only the responses among options A (baseline), B (*AttGAN*), and C (*AttGAN+*). The survey was answered by 21 participants resulting in a total of 1050 responses. For the first scenario, we observe that 49.4% of the generated images were perceived by the respondents as poor results, while 35.2% choose *AttGAN+*, 10.5% choose *AttGAN*, and only 4.9% prefer the baseline approach.

TABLE III: Evaluation survey results.

	Baseline (A)	*AttGAN (B)*	*AttGAN+ (C)*	*Poorly (D)*
Case *(i)*	4.9%	10.5%	35.2%	49.4%
Case *(ii)*	9.6%	20.7%	69.7%	-

The second scenario ignores the nonconformity option D, and it clearly depicts the differences between baseline, *AttGAN*, and *AttGAN+*. We have also performed a chi-squared test to check for statistical relevance in the survey's results. The statistical test was conducted only for the second scenario, and it shows the existence of significant differences with a p-value < 0.001.

C. Ablation Study

The source of our intuition to build *AttGAN+* comes from the fact that it is now possible to develop networks that generate attention maps highlighting areas of interest. We assume that additional information, e.g., attention maps that highlight regions according to certain objects, can improve the generative process. We validate our assumption by observing the attention masks produced by AN. Figure 4 illustrates the attention masks generated by the softmax layer of AN for 3 input images from the test set. The intimate regions observed in Figure 4b are coherently aligned with the original image. Figure 4c confirms that assumption when masking the image and fully supports our intuition that by feeding the input with additional information we can improve the generator capabilities and reinforce transformations at sensitive areas.

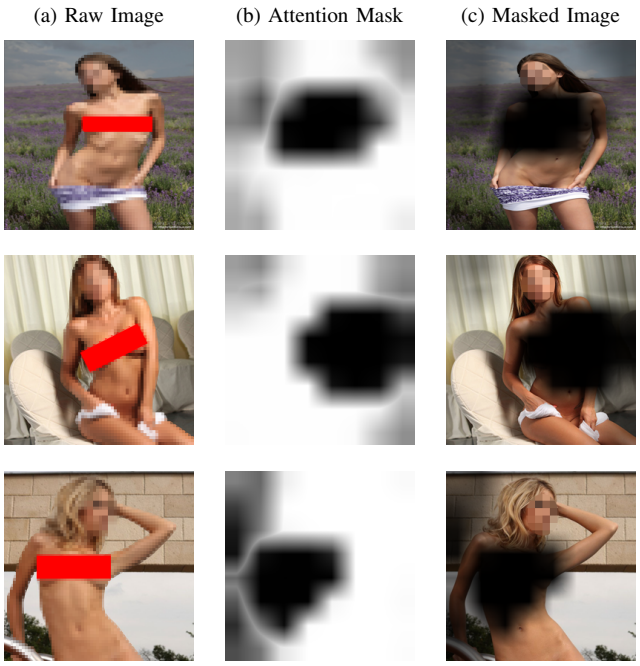

Fig. 4: Attention masks for 3 domain A input images. For privacy reasons we have manually blurred all faces and covered the intimate parts with red tags.

Next we investigate the effects of each component of our method presented at Figure 1. Figure 5 illustrates the 3 variations of our method that we progressively add to the baseline. Observe the improvement of adding novel information from the attention masks when concatenated as a new input image channel, and then training by 200 epochs (*AttGAN* in Figure 5a). *AttGAN* keeps improving after 300, 400, and finally 500 epochs (Figure 5b). Now note the improvement when also summing the attention masks at the first convolutional layer output of the generator, which is depicted at Figure 5c. Finally, in Figure 5d we use the attention mask to merge our best method output (*AttGAN+*) with the raw input image, in such a way that peripheral parts such as background and faces are better preserved (the so-called *AttGAN++*).

VII. Conclusions

In this paper we have presented a new method for seamless nudity censorship based on a cycle-consistent image-to-image translation approach enhanced by additional information extracted from an *attention network*. We make use of the attention masks in three distinct strategies to improve the original seminal work by More et al. [3] on seamless nudity censorship,

Fig. 5: Samples from 3 different experiments. a) *AttGAN* with 200 epochs; b) *AttGAN* trained by 500 epochs; c) *AttGAN+* trained by 500 epochs; d)enhanced *AttGAN+*. For privacy reasons we have manually blurred all faces.

which in turn is an evolution of the CycleGAN [4] framework for automatically covering intimate body parts without explicit supervision and/or paired training samples.

We have designed 3 variations of the method and have also conducted a web survey in which 50 test-set images are analyzed by 21 users that need to choose the best approach for automatically generating bikinis in naked women. The survey results indicate a statistically-significant advantage for *AttGAN+*, which was selected as the best approach based on the 1050 collected opinions. Then, we have further evolved *AttGAN+* in an attempt to preserve as best as we could the peripheral parts of the images such as background and faces, resulting in the so-called *AttGAN++*.

For future work, we intend to evaluate the behavior of our method with the generator G_{BA} input also being matched with the original image. In other words, we would like to see if we can successfully include the process performed by *AttGAN++* (Section III-B2) during training. We believe that such improvement will contribute in producing even better reconstruction results. We also intend to evaluate *AttGAN+* in different application domains, especially those used in the literature as benchmarks, e.g., the well-known horses-to-zebras and oranges-to-apples.

ACKNOWLEDGMENT

We would like to thank Motorola Mobility™ and the Brazilian research agencies CAPES, CNPq, and FAPERGS for funding this work.

REFERENCES

[1] A. Cooper, "Sexuality and the internet: Surfing into the new millennium," *CyberPsychology & Behavior*, vol. 1, no. 2, pp. 187–193, 1998.
[2] M. Moustafa, "Applying deep learning to classify pornographic images and videos," *arXiv preprint arXiv:1511.08899*, 2015.
[3] M. D. More, D. M. Souza, J. Wehrmann, and R. C. Barros, "Seamless nudity censorship: an image-to-image translation approach based on adversarial training," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.
[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint arXiv:1703.10593*, 2017.
[5] S. Mo, M. Cho, and J. Shin, "Instagan: Instance-aware image-to-image translation," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=ryxwJhC9YX
[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.
[8] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.
[9] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint arXiv:1609.04802*, 2016.
[10] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.
[11] S. Avila, N. Thome, M. Cord, E. Valle, and A. D. A. AraúJo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 453–465, 2013.
[12] G. Simões, J. Wehrmann, T. Paula, J. Monteiro, and R. C. Barros, "Datasex: um dataset para indução de modelos de classificação para conteúdo adulto," in *KDMiLe*, 2016.
[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE CVPR*, 2015, pp. 1–9.
[14] M. Moustafa, "Applying deep learning to classify pornographic images and videos," *arXiv preprint arXiv:1511.08899*, 2015.
[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.
[16] J. Wehrmann, G. S. Simões, R. C. Barros, and V. F. Cavalcante, "Adult content detection in videos with convolutional and recurrent neural networks," *Neurocomputing*, vol. 272, pp. 432–438, 2018.
[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[19] J. Wehrmann, R. C. Barros, S. N. d. Dôres, and R. Cerri, "Hierarchical multi-label classification with chained neural networks," in *Proceedings of the Symposium on Applied Computing*. ACM, 2017, pp. 790–795.
[20] J. Wehrmann, R. Cerri, and R. Barros, "Hierarchical multi-label classification networks," in *International Conference on Machine Learning*, 2018, pp. 5225–5234.
[21] J. Wehrmann and R. C. Barros, "Convolutions through time for multi-label movie genre classification," in *Proceedings of the Symposium on Applied Computing*. ACM, 2017, pp. 114–119.
[22] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.
[23] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.