Seamless Nudity Censorship: an Image-to-Image Translation Approach based on Adversarial Training

Martin D. More, Douglas M. Souza, Jônatas Wehrmann, 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: {martin.more, douglas.souza.002}@acad.pucrs.br, rodrigo.barros@pucrs.br

***Abstract*—The easy access and widespread of the Internet makes it easier than ever to reach content of any kind at any moment, and while that poses several advantages, there is also the fact that sensitive audiences may be inadvertently exposed to nudity content they did not ask for. Virtually every work on nudity and pornography censorship focus solely on performing binary classification, where the result is used to decide whether to completely ignore the accessed content or not. Such an approach may compromise user experience because the entire content, either images or frames of a video, has to be removed/blocked. In this paper, we propose a paradigmatic shift in the literature of adult censorship: instead of detecting and excluding the identified content, we propose to automatically *filter out* only the sensitive regions of an image. For that, we have developed an image-to-image translation approach based on adversarial training that implicitly locates sensitive regions in images and covers them whilst preserving its semantics, i.e., putting appropriate clothing. We test this novel approach on images of nude women, in which we are capable of automatically generating bikinis that cover the sensitive parts without the additional effort of previously annotating body parts. Our results are visually impressive, proving that it is possible to perform seamless nudity censorship with small effort of data collection and annotation.**

I. Introduction

The easy access and widespread of the Internet especially through mobile phones makes it easier than ever to reach content of any kind at any moment. This convenience, however, often comes at a price: in many cases, people are exposed to content they did not ask for. An example is when people inadvertently access nudity content. Studies[1] show that eight out of ten 18-year-olds think it is too easy for young people to accidentally see nudity or even pornography online. This is a concerning issue in many levels, especially in an era where people are joining the Internet at early ages. Statistics[2] show that 93.2% of boys and 62.1% of girls have seen online pornography before the age of 18. This scenario motivates the development of computational approaches that are capable of monitoring and automatically detecting pornography, with the final goal of protecting sensitive populations.

Earlier work on nudity and pornography identification focused on identifying body parts that could be used to detect sensitive media, such as faces, human skin, and nipples [1]–[4], while recent studies address the problem using state-of-the-art representation learning approaches to automatically learn features capable of distinguishing between sensitive and non-sensitive content [5]–[7]. Those studies focus solely on performing binary classification, using the result to decide whether to completely ignore the accessed content or not. Note that, in practice, such an approach may compromise user experience because the entire content, being images or frames of a video, would have to be removed/blocked.

A more reasonable approach would be to manually identify and filter only the sensitive regions of an image or video using some form of content blocking mechanism, as depicted in Figures (1a)-(1c). To the best of our knowledge, this type of approach for automatic sensitive content filtering has not yet been explored in the specialized literature. With that in mind, we propose in this work a novel type of adult content filtering task where the goal is to automatically *filter out* only the sensitive regions of an image. The motivation behind this task is to avoid ruining user experience while consuming content that may occasionally contain explicit content.

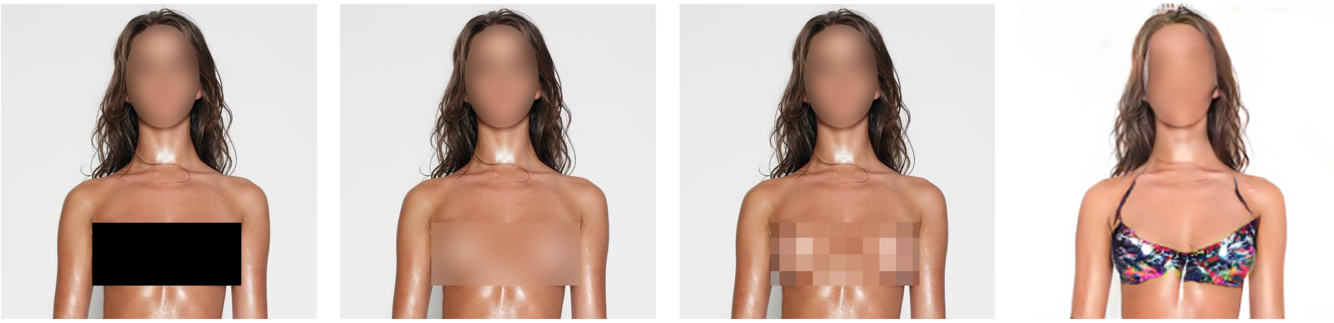
The task of filtering out sensitive regions in nudity content could be addressed in different ways. A possibility that stands out is to cast such a problem as an object detection task, where the objects at hand would be the sensitive body parts in nudity content. Hence, state-of-the-art object detection approaches such as YOLO [8], Faster-RCNN [9], or RetinaNet [10] could be used to locate the sensitive regions and then one could apply an image filter within the detected bounding boxes (as shown in Figure 1a). Such a strategy would definitely be an improvement over binary classification. However, it presents two major problems: (i) the censorship would still be perceived by the person consuming the content; and (ii) training an object detector requires a large dataset with per-object boxes manually annotated, which is both time-consuming and tedious.

Ideally, it would be desirable to have a method that could censor sensitive content in a totally non-intrusive manner so that the user would not notice the nudity, but also avoiding the need of manually annotating a large amount of body parts for

[1] http://www.burnet.edu.au/news/435_burnet_studies_shed_light_on_sexual_behaviour_of_teenagers

[2] http://enough.org/stats-youth-and-porn

978-1-5090-6014-6/18/$31.00 ©2018 IEEE



(a) Censor Box (b) Blur (c) Pixellation (d) Seamless (ours)

Fig. 1. Techniques for censoring sensitive regions of an image. *(a)-(c)*: **manual** strategies commonly used for localized censorship. So far, no studies have addressed this problem with an automatic approach. *(d)*: result of our fully-automatic seamless censoring approach using unpaired image-to-image translation.

data-driven approaches. Figure 1d presents this ideal scenario, which is precisely the approach proposed in this paper. In a nutshell, we have developed an image-to-image translation approach based on adversarial training that implicitly locates sensitive regions in images and covers them whilst preserving the semantics of the image (i.e., putting appropriate clothing). Formally, we translate an image $\mathbf{x}$ from the sensitive content domain X to an image $\mathbf{y}$ of the non-sensitive content domain Y where sensitive parts are covered preserving the semantics of the source domain. Note that the data needed for this task are images from domains X and Y, which are easy to acquire and no special annotation is required.

A recurrent issue in image-to-image approaches is that it is necessary to have *aligned* pairs of samples $\{\mathbf{x}_i, \mathbf{y}_i\}$ to train models that map from domain X to domain Y. Our method is based on state-of-the-art image-to-image translation techniques that allow us to learn a model that maps from an unsafe image domain (nude women) to a safe image domain (women wearing bikinis) using *unpaired* training samples, avoiding the need (and the cost) of obtaining aligned pairs of samples. We show several impressive results regarding the automatic generation of bikinis in images of nude women, proving that it is possible to perform seamless nudity censorship with small effort of data collection and annotation.

II. Background

In this section, we briefly review the state-of-the-art in adversarial training via Generative Adversarial Networks (GANs) as well as the most important image-to-image translation approaches to date.

A. Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) [11] is a framework for estimating generative models via an adversarial training process. It consists of two models that are based on neural networks: a generator, G, and a discriminator, D. The discriminator attempts to distinguish between real images sampled from the training data ($\mathbf{x} \sim p_{\text{data}}$) and synthetic images generated by the generator model ($\hat{\mathbf{x}} \sim p_g$). The generator, on the other hand, tries to produce realistic-looking samples to fool the discriminator. Formally, G and D are set to play the following minimax game:

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

where $\mathbf{z}$ is a low-dimensional latent vector drawn from a simpler known distribution (such as uniform or Gaussian) and is fed as input noise to G. After training the models, the generator network is capable of producing a wide variety of images, depending on the values of the latent vector $\mathbf{z}$.

Ideally, since we are optimizing the mapping between latent space and complex images, small modifications in $\mathbf{z}$ should map to small modifications in image space. Thus, this latent space representation can often be used to traverse the natural image manifold. This is useful, for instance, to create natural interpolation between results, or even performing image editing tasks [12]. One should not expect, though, the latent space to be semantically organized, i.e., the particular dimensions of $\mathbf{z}$ may not correspond to semantically-coherent attributes (even though approaches such as *InfoGAN* [13] attempt to perform this mapping).

The traditional GANs framework is *unconditioned*, which means that there is no control over the modality of the data being generated. It is possible, however, to create Conditional Generative Adversarial Networks (CGANs) [14] by feeding the models with additional information $\mathbf{c}$, such as class labels or data from other modalities. For doing so, we modify the original GANs value function (Equation 1) to:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x},\mathbf{c}}[\log D(\mathbf{x},\mathbf{c})] + \mathbb{E}_{\mathbf{z},\mathbf{c}}[\log(1 - D(G(\mathbf{z},\mathbf{c}),\mathbf{c}))], \tag{2}$$

where $\mathbf{x} \sim p_{\text{data}}$, $\mathbf{c} \sim p_{\text{data}}$, and $\mathbf{z} \sim p_z$. Architecture-wise, the conditional information can be added to the models in numerous ways, the most common being a simple concatenation between $\mathbf{z}$ and $\mathbf{c}$ for the generator, and a concatenation operation on the feature map axis of the input for the discriminator. Simply adding conditional input to the models does not necessarily entail any changes to the adversarial loss; however, several studies [13], [15], [16] have modified training procedures to take advantage of this extra input. GANs have already been conditioned on data with several different modalities, such as class labels [14], attribute vectors [17], text [18], images [19], and videos [20]. It has been shown [21] that models typically benefit from using a conditional input; one possible explanation being that this helps the model to better navigate the space of possible outputs.

One of the main drawbacks of GANs is that the training process is very sensitive to the choice of hyperparameters. Several recent studies propose constraints and improvements to the framework to improve training stability and synthesis results [15], [22]–[25]. GANs have shown remarkable results for tasks such as image synthesis [26], [27], image super-resolution [28], image inpainting [29], image editing [30], text-to-image synthesis [18], image-to-image translation [19], [31], and many more.

B. Image-to-image Translation

Many problems in computer graphics and computer vision can be reduced to performing image-to-image translation, which is the task of transforming an image $\mathbf{x}$ from domain X to an image $\mathbf{y}$ in a different domain Y. Some concrete examples of such tasks are: (i) image colorization; (ii) edge detection; (iii) generating photographs from sketches; and (iv) image prediction from a normal map. Initially, solutions to such tasks relied on hand-crafted mapping functions, but a recent trend is to create methods using automatically-learned features, which are commonly extracted via Convolutional Neural Networks (CNNs) [32], [33].

Formally, given a labeled dataset of *paired* images $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$, *paired* image-to-image translation can be treated as a supervised learning problem where a model learns the mapping $f_\theta : \mathbf{x} \mapsto \mathbf{y}$ by minimizing a loss function, which is sometimes formulated around a per-pixel classification or regression task [32]–[35]. However, minimizing element-wise distances, (e.g., L1 and L2) tends to produce blurry results [29], [33], [36] and still requires the loss function to be tuned for the task at hand. A more recent trend is to leverage adversarial learning for image-to-image translation [37]–[40]. One extremely successful example is "*pix2pix*" [19], which provides a data-agnostic framework for paired image-to-image translation based on CGANs. The original framework produces images with a resolution of 256×256, but a recent study [41] has improved on the architecture of *pix2pix* to support generating high-resolution (4096×2048) images.

We are continuously experiencing an increase in dataset sizes and diversity. However, it is still uncommon to find datasets containing large quantities of *paired* image samples $\{\mathbf{x}, \mathbf{y}\}$ to support image-to-image translation tasks. Although paired images can be easily collected for some tasks, such as image colorization, it can be extremely difficult and expensive to obtain paired data for tasks where the desired output is highly complex or not even well defined, like artistic stylization or object transfiguration. For this reason, researchers have developed many approaches [42]–[44] for *unpaired* image-to-image translation, where the goal is to learn mapping functions to somehow relate separate domains X and Y, given training samples $\{\mathbf{x}_i\}_{i=1}^{N}$ where $\mathbf{x}_i \in X$ and $\{\mathbf{y}_i\}_{j=1}^{M}$ where $\mathbf{y}_i \in Y$. One of the most recent approaches, called *cycleGAN* [31], adapts the *pix2pix* framework to the unpaired setting, achieving good results in tasks such as style transfer, object transfiguration, season transfer, and photo enhancement.

III. Proposed Approach

We propose an image-to-image translation approach based on adversarial training that implicitly locates sensitive regions in images that contain nudity and covers them whilst preserving the semantics of the image, i.e., automatically generating clothing to cover the nudity. We translate an image $\mathbf{x}$ from the sensitive content domain X (pool of images containing nude women) to an image $\mathbf{y}$ of the non-sensitive content domain Y (pool of images containing women in bikinis) where sensitive parts are covered by bikinis though preserving the semantics of the original image. The data needed for this task are images from domains X and Y, which are easy to acquire and no special annotation is required, as presented in Section III-A.

Our proposed approach follows the architecture presented in Figure 2. We draw inspiration from [31], where paired data is not required for performing image-to-image translation. The key idea is to perform adversarial training to learn realistic mappings between domains. Specifically, the framework consists of two mapping generators, $G : X \mapsto Y$ and $F : Y \mapsto X$, and two discriminators, D_X and D_Y. D_X distinguishes between real images $\{\mathbf{x}\}$ and translated images $\{F(\mathbf{y})\} = \{\hat{\mathbf{x}}\}$, while D_Y discriminates between real images $\{\mathbf{y}\}$ and translated images $\{G(\mathbf{x})\} = \{\hat{\mathbf{y}}\}$. The loss function optimized by our approach is presented in Section III-B, whereas the network architectures for the generators and discriminators are described in Section III-C.

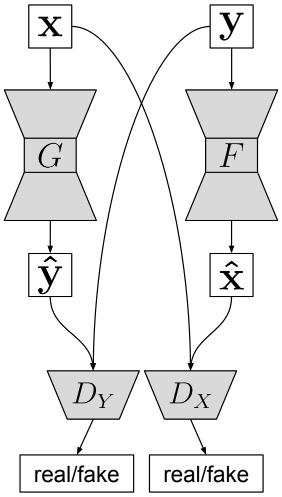


Fig. 2. Proposed image-to-image translation approach for seamless censoring of nudity content via adversarial training.

A. Dataset

Previous studies on adult content detection conducted experiments either by using pre-established datasets [45]–[48] of images and videos containing adult and regular content, or by creating a custom dataset that better suits their needs by scraping the Internet. Since the task that the existing datasets aim to support is binary classification of content, some images and videos are totally unrelated to adult content (e.g., cartoons, videos of animals, landscapes), which do not help us with the task of translating between domains in a constrained, seamless fashion. Given that the contents of existing datasets are suboptimal for our image-to-image model, we opted for collecting our own dataset from scratch.

We scraped images from the Internet for both domains: nude women and women wearing bikinis. We filtered results, keeping only images where a single person appears. The dataset was further split into training and test sets. For women wearing bikinis (X), the final image count was $1,044$ training images and 117 test images; for nude women (Y), the final image count was 921 for training and 103 for test. We make the dataset public for research purposes[3].

B. Loss Function

The natural choice for adversarial loss is the classic GANs loss (Equation 1). However, we adopt the LSGANs [49] loss as it has shown to be stable and produce good results. The adversarial loss of G and D_Y is given by:

$$\begin{aligned}\mathcal{L}_{\text{Adv}}(G, D_Y, X, Y) = \; & \mathbb{E}_{\mathbf{y}\sim p_{\text{data}}(\mathbf{y})}[(D_Y(\mathbf{y}) - 1)^2] + \\ & \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[D_Y(G(\mathbf{x}))^2].\end{aligned} \tag{3}$$

The loss for the mapping function for $F : Y \mapsto X$ is similarly given by $\mathcal{L}_{\text{Adv}}(F, D_X, Y, X)$. In theory, this adversarial loss can be used to force the generators to produce realistic-looking samples, but it does not introduce enough constraints to guarantee similarity between images across domains. Hence, to force similarity between translations and reduce the search space for mapping functions, we explore the property that a translation should be "cycle-consistent" [31]. Mathematically, cycle consistency implies that G and F should be inverses of each other and both mappings should be bijections, which implies that an image translation cycle should recreate the original image. A *forward cycle consistency* is expressed as $\mathbf{x} \to G(\mathbf{x}) \to F(G(\mathbf{x})) \approx \mathbf{x}$ and a *backward cycle consistency* corresponds to $\mathbf{y} \to F(\mathbf{y}) \to G(F(\mathbf{y})) \approx \mathbf{y}$. This is used to formulate a *cycle-consistency loss*:

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G, F) = \; & \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\|F(G(\mathbf{x})) - \mathbf{x}\|_1] + \\ & \mathbb{E}_{\mathbf{y}\sim p_{\text{data}}(\mathbf{y})}[\|G(F(\mathbf{y})) - \mathbf{y}\|_1].\end{aligned} \tag{4}$$

Combining the adversarial losses and the cycle-consistency loss, we can formulate the full objective, where λ_{cyc} controls the relative importance between objectives:

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) = & \mathcal{L}_{\text{Adv}}(G, D_Y, X, Y) + \\ & \mathcal{L}_{\text{Adv}}(F, D_X, Y, X) + \\ & \lambda_{cyc}\mathcal{L}_{\text{cyc}}(G, F).\end{aligned} \tag{5}$$

C. Network Architecture

We test two popular architectures for the generators and a standard architecture for the discriminators.

N-Layers Discriminator. We use a simple discriminator architecture with a growing number of convolutional filters towards the end. The discriminators use Leaky ReLU as activation function and apply instance normalization [50] after every convolutional layer, except the first and the last.

9-Blocks ResNet Generator. We experiment with the 9-Blocks ResNet generator inspired by [51]. This architecture consists of an autoencoder that applies residual connections between bottleneck layers. Additionally, it applies ReLU as activation and instance normalization after the convolutions.

U-Net 256 Generator. We also experiment with a popular architecture for the generator. The U-Net [52] consists of an autoencoder with residual connections between layers that operate at the same spatial dimensions. Our implementation also applies instance normalization, Leaky ReLU activation for the encoder, and ReLU for the decoder. The U-Net has shown good results for image segmentation as well as for image-to-image translation tasks.

IV. Experiments

We train models that operate at the resolution of 256×256 pixels. The generators and discriminators are trained using simultaneous gradient descent, where at each training step we update the weights of D_Y, D_X, G, and F, respectively. We use Adam optimizer with a learning rate of 2×10^{-4}, $\beta_1 = 0.5$ and $\beta_2 = 0.99$ for all networks. The weight for the cycle consistency term λ_{cyc} is set to 10. We train the networks for 400 epochs using batches of size 1. We decay the learning rate linearly as training progresses. Aditionally, we translate images every 100 seconds for visual inspection reasons. Training took 2 days on a single NVIDIA® 1080 Ti GPU employing the PyTorch framework.

A. Results on the Original Dataset

Figure 3 depicts results achieved by using the original dataset for training our networks. Note that the ResNet generator (second row) consistently produced better results when compared to the U-Net approach (third row). Even though the latter learned to remove some sensitive parts (e.g., the nipples were often erased), it presented difficulties in properly positioning the bikini and often distorted the original image. On the other hand, the ResNet generator have shown better quality in both generating good-looking bikinis and covering explicit body parts.

When training our models, we have noticed that the backgrounds seem to have significant influence in the speed of the learning process as well as in the quality of the results.

[3] Link omitted due to double-blind review.

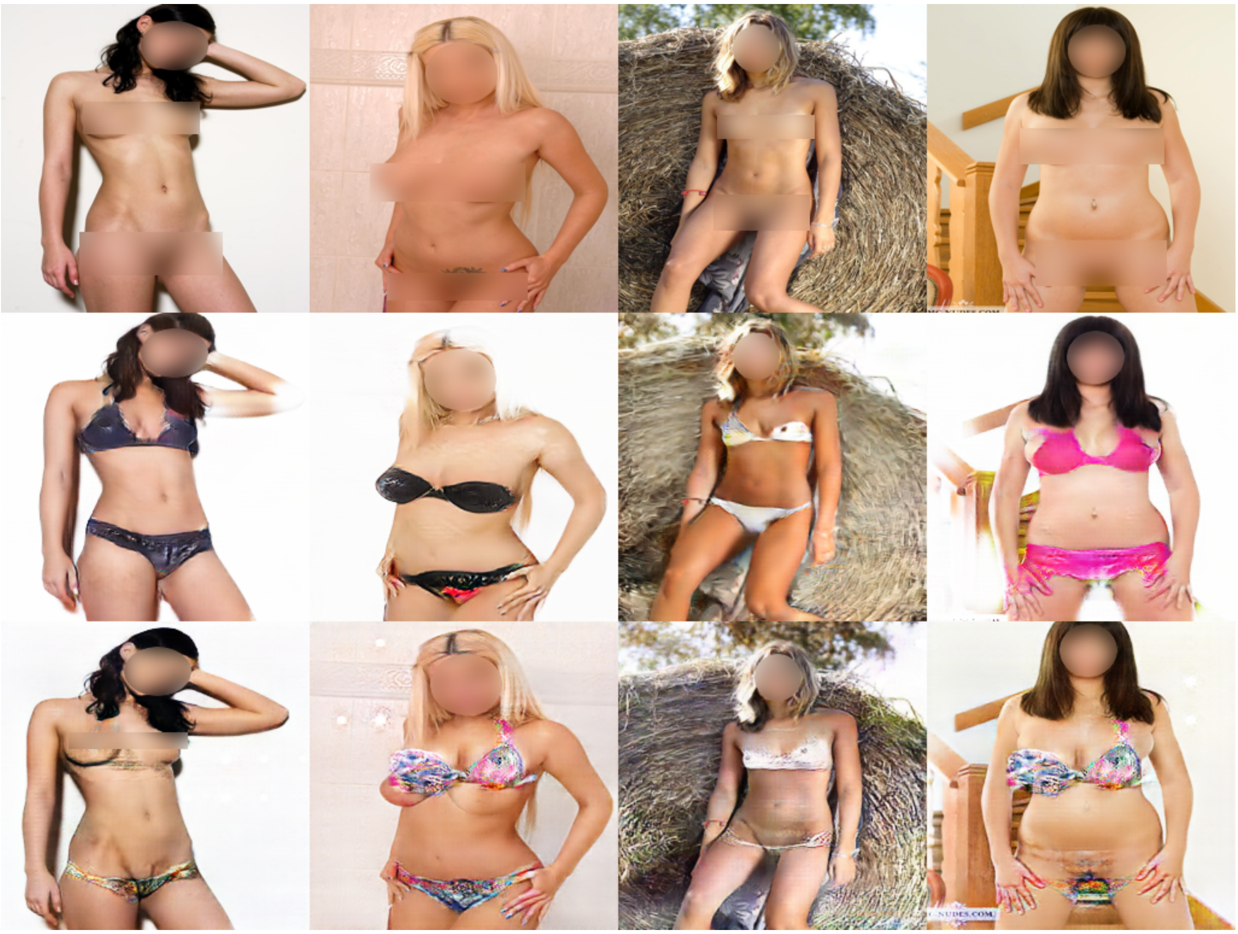


Fig. 3. Results after training on the original dataset. Top row: real images (manually censored for protecting the reader). Middle row: results using 9-Blocks ResNet generator. Bottom row: results using a U-Net 256 generator (blurring applied to unsatisfactory results).

Hence, we have modified our original dataset by removing the background of each image, as detailed next.

B. Results after Removing Backgrounds

A problem we observed during training is that the networks tried to understand the relationship between the image background and the task of censoring nude body parts on the women present in the foreground, even though such a relation does not exist in real life. As a countermeasure, we decided to perform experiments using a clean version of the dataset that comprises only pictures of people in front of a white background. Theoretically, by using "background-free" images, the networks can properly focus on learning the task at hand, rather than approximating irrelevant noisy background variations between images.

We built such a dataset version by segmenting the people in all images with the aid of Mask R-CNN [53], the state-of-the-art approach for semantic and instance segmentation. In a nutshell, Mask R-CNN's basic structure is quite similar to Faster R-CNN, the difference being that it predicts binary masks for each RoI (Region of Interest) to allow pixel-level segmentation. In most cases, this background removal strategy successfully removed the backgrounds of the images in our dataset. However, we noticed some error cases in which Mask R-CNN was unable to find any person, or performed incorrect segmentation. Given that such miss-segmented instances introduce a controlled amount of noise for both image classes, we decided to keep those imperfect images in this dataset.

Figure 4 shows images generated by our approach trained over the no-background version of the dataset. Note that these results are arguably more consistent than those provided by models trained with the original dataset in Figure 3. Once again, one can observe that the ResNet-based model outperformed U-Net one, by generating images with the sensitive parts properly covered with real-looking bikinis. In addition, it introduced much less distortion than its competitor.

C. Byproduct Results

The task we propose to address in this paper is to develop a data-oriented approach for censoring sensitive regions in images. We opted for using an *unpaired* image-to-image approach since creating a dataset of annotated image pairs of women wearing bikinis (X) and nude women (Y) that was large enough to train our models would be insurmountable work. Fortunately, as we have shown in Section IV-A, this approach was capable of producing impressive results as it

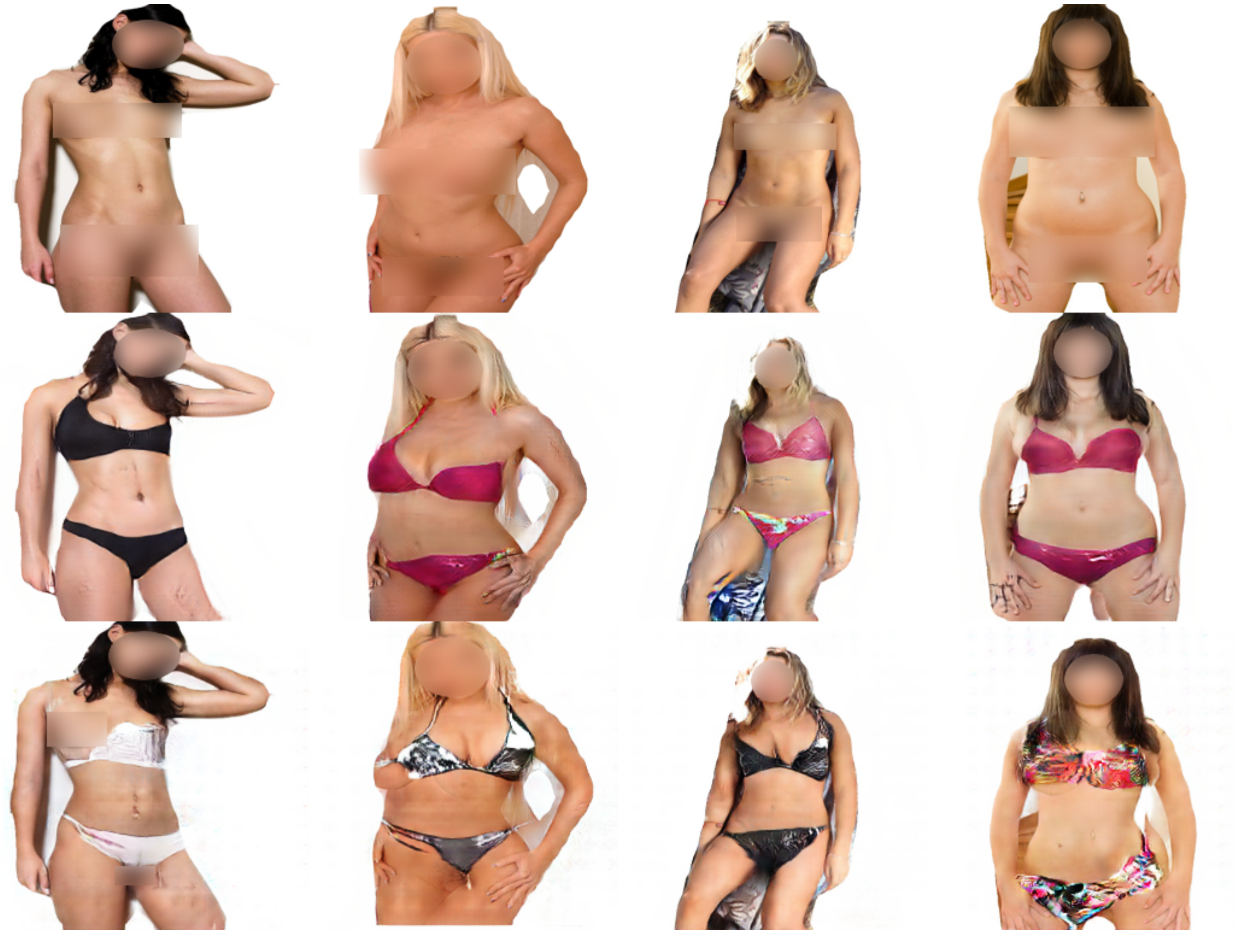


Fig. 4. Results after training on the no-background dataset. Top row: real images (manually censored for protecting the reader). Middle row: results using 9-Blocks ResNet generator. Bottom row: results using a U-Net 256 generator (blurring applied to unsatisfactory results).

stands. However, note that – as depicted in Figure 2 – we train two distinct generators: (i) $F : Y \mapsto X$, which maps from nude women to women wearing bikinis; and (ii) $G : X \mapsto Y$, which translates from women wearing bikinis to nude women. Even though G is part of the architecture, the task performed by the model is not in the scope of this work. Nevertheless, we present results obtained when using generator G in Figure 5, showing that the generator learned to perform its task successfully.

D. Robustness Analysis

Unlike previous studies in adult content detection, our image-to-image approach is not concerned with detecting and classifying content as sensitive material or not. However, the task we aim to solve requires, even if implicitly, for our model to be capable of discerning between these two types of content, considering that domains X and Y contain different sample distributions. If our model cannot implicitly capture the difference between domains, this would mean that our generators, G and F, would not be capable of translating samples for their respective domains, and our discriminators, D_X and D_Y, would not be able to discern between real and fake images. Therefore, to test the robustness of our

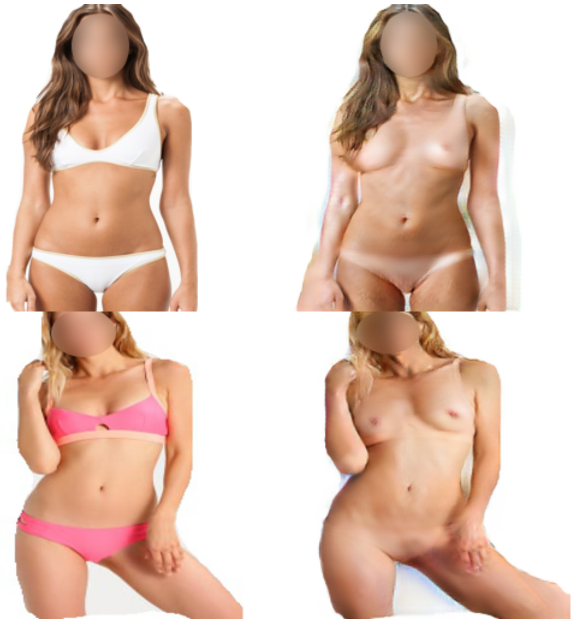


Fig. 5. Results of generator G that maps women in bikini to nude women. Left: real images. Right: results using 9-Blocks ResNet generator.

model in the task at hand (applying appropriate clothing to women), we present generator F with samples of women already wearing bikinis. We expect the model to perceive that the type of content already matches the output domain and perform no modification whatsoever. As presented in Figure 6, our model passes the proposed robustness check, since it does not significantly modify the original images.

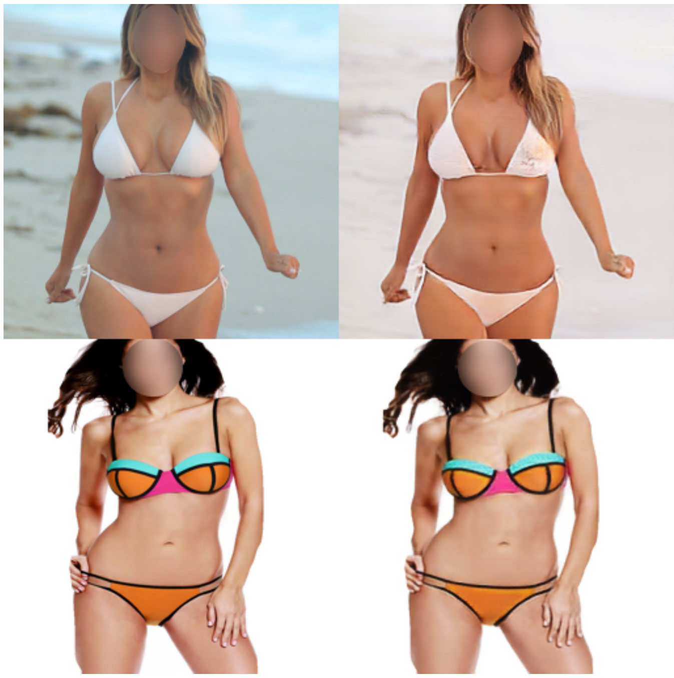


Fig. 6. Robustness analysis. Left: real images. Right: images with minimal modifications created by generator F.

V. Related Work

In this section, we discuss related work that provide datasets and methods for identifying/classifying adult content in both images and videos. Note that no work so far attempts at automatically censoring nude content. They simply indicate whether an image/frame or video contain adult content.

Avila et al. [47] introduced one of the first datasets for adult content detection, namely NPDI. Such dataset 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. [48]. The authors provide the largest dataset for binary classification of pornographic images. It comprises a collection of $\approx 300,000$ images that are equally distributed in adult and benign classes. They also already provide splits for training and validation purposes. DataSex was built by crawling around $300,000$ publicly available images from adult websites. Simões et al. [48] report classification results of $\approx 95\%$ accuracy in DataSex's test set by fine-tuning a pre-trained GoogleNet.

The work described in [5] is the first to use deep neural networks for pornography classification in videos. That work proposes a method that requires fine-tuning two distinct ConvNets, namely *AlexNet* [54] and *GoogLeNet* [55]. 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. [7] presented ACORDE (Adult Content Recognition with Deep Neural Networks), which is a method that uses a convolutional architecture as a feature extractor and a Long Short-Term Memory network (LSTM) [56] to perform video classification. ACORDE 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.

VI. Conclusion

We proposed in this paper an image-to-image translation approach based on adversarial training that implicitly locates sensitive regions in images that contain nudity and covers them whilst preserving the semantics of the image. We translate an image $\mathbf{x}$ from the sensitive content domain X (nude women) to an image $\mathbf{y}$ of the non-sensitive content domain Y (women in bikinis), where the sensitive parts are automatically covered by bikinis. Our approach does not require paired training samples and produces impressive highly-realistic results, paving the way for solving the novel task of seamless nudity censorship. For future work, we intend to analyze the impact of different architectural choices and loss functions on the generated images, and also to embed our approach in a browser application to protect audiences from accessing undesired content.

Acknowledgment

We would like to thank the Brazilian research agencies CAPES and FAPERGS for funding this research. We also gratefully acknowledge NVIDIA corporation for donating the GPUs we used for conducting the experiments.

References

[1] H. A. Rowley, Y. Jing, and S. Baluja, “Large scale image-based adult-content filtering.” in *VISAPP (1)*. Citeseer, 2006, pp. 290–296.

[2] J.-S. Lee, Y.-M. Kuo, P.-C. Chung, and E.-L. Chen, “Naked image detection based on adaptive and extensible skin color model,” *Pattern recognition*, vol. 40, no. 8, pp. 2261–2270, 2007.

[3] C. Platzer, M. Stuetz, and M. Lindorfer, “Skin sheriff: a machine learning solution for detecting explicit images,” in *Proceedings of the 2nd international workshop on Security and forensics in communication systems*. ACM, 2014, pp. 45–56.

[4] H. A. Nugroho, D. Hardiyanto, and T. B. Adji, “Nipple detection to identify negative content on digital images,” in *Intelligent Technology and Its Applications (ISITIA), 2016 International Seminar on*. IEEE, 2016, pp. 43–48.

[5] M. Moustafa, “Applying deep learning to classify pornographic images and videos,” *arXiv preprint arXiv:1511.08899*, 2015.

[6] M. Perez, S. Avila, D. Moreira, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Video pornography detection through deep learning techniques and motion information," *Neurocomputing*, vol. 230, pp. 279–293, 2017.

[7] 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.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of IEEE CVPR*, 2016, pp. 779–788.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *arXiv preprint arXiv:1708.02002*, 2017.

[11] 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.

[12] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Neural photo editing with introspective adversarial networks," *arXiv preprint arXiv:1609.07093*, 2016.

[13] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *NIPS*, 2016, pp. 2172–2180.

[14] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[15] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[16] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," *arXiv preprint arXiv:1610.09585*, 2016.

[17] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional gans for image editing," *arXiv preprint arXiv:1611.06355*, 2016.

[18] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.

[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[20] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mse," *arXiv preprint arXiv:1511.05440*, 2015.

[21] J. Gauthier, "Conditional generative adversarial nets for convolutional face generation," *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, vol. 2014, no. 5, p. 2, 2014.

[22] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," *arXiv preprint arXiv:1609.03126*, 2016.

[23] M. Arjovsky and L. Bottou, "Towards principled methods for training gans," *arXiv preprint arXiv:1701.04862*, 2017.

[24] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," *arXiv preprint arXiv:1704.00028*, 2017.

[25] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

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

[27] E. L. Denton, S. Chintala, R. Fergus *et al.*, "Deep generative image models using a laplacian pyramid of adversarial networks," in *Advances in neural information processing systems*, 2015, pp. 1486–1494.

[28] 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.

[29] 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.

[30] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *European Conference on Computer Vision*. Springer, 2016, pp. 597–613.

[31] 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.

[32] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.

[33] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *ECCV*. Springer, 2016, pp. 649–666.

[34] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[35] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *European Conference on Computer Vision*. Springer, 2016, pp. 577–593.

[36] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," *arXiv preprint arXiv:1512.09300*, 2015.

[37] L. Karacan, Z. Akata, A. Erdem, and E. Erdem, "Learning to generate images of outdoor scenes from attributes and semantic layouts," *arXiv preprint arXiv:1612.00215*, 2016.

[38] H. Dong, S. Yu, C. Wu, and Y. Guo, "Semantic image synthesis via adversarial learning," *arXiv preprint arXiv:1707.06873*, 2017.

[39] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2017.

[40] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Advances in Neural Information Processing Systems*, 2017, pp. 465–476.

[41] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," *arXiv preprint arXiv:1711.11585*, 2017.

[42] R. Rosales, K. Achan, and B. J. Frey, "Unsupervised image translation." in *iccv*, 2003, pp. 472–478.

[43] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in neural information processing systems*, 2016, pp. 469–477.

[44] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 700–708.

[45] A. P. Lopes, S. E. de Avila, A. N. Peixoto, R. S. Oliveira, and A. d. A. Araújo, "A bag-of-features approach based on hue-sift descriptor for nude detection," in *Signal Processing Conference, 2009 17th European*. IEEE, 2009, pp. 1552–1556.

[46] A. P. B. Lopes, S. E. de Avila, A. N. Peixoto, R. S. Oliveira, M. d. M. Coelho, and A. d. A. Araújo, "Nude detection in video using bag-of-visual-features," in *Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on*. IEEE, 2009, pp. 224–231.

[47] 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.

[48] 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.

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

[50] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *CoRR*, vol. abs/1607.08022, 2016. [Online]. Available: http://arxiv.org/abs/1607.08022

[51] 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.

[52] 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.

[53] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[55] 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.

[56] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.