17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Content Filtering in Streaming Video Using Domain Adaptation

Utsav Shah* Muhammad Aqmar Mitsuru Nakazawa* Björn Stenger*
Rakuten Institute of Technology, Rakuten Group, Inc.

Abstract

This paper addresses the problem of content filtering in live streaming video. We consider the case where positive data, content to be filtered, is not readily available on the target platform. We therefore use positive data from other sources and apply domain adaptation to classify new data on the target platform. In order to map features of source and target domains into a common feature space, we optimize a Wasserstein distance (WD) loss and binary cross entropy loss, such that class distributions remain separated in the new feature space. Our baseline model achieves state-of-the-art results on the public NPDI dataset, and we show that WD-based domain adaptation improves the accuracy in the absence of positive samples in the target domain.

1 Introduction

Video streaming services employ content moderation to provide a family-friendly environment to customers. In this paper, we consider the task of adult content filtering. This task includes classifying video content based on appearance and scene context [1]. One major challenge is the collection of a representative dataset. This has been approached by collecting data from multiple domains, *i.e.*, selecting positive data from various media platforms and negative data from different sources [2, 3, 4, 5]. This strategy allows for collecting large datasets, but leads to a mismatch of training and test distributions owing to differences in scene environments and camera views [6, 7]. Prior work in cross-dataset settings assumes that the data has a similar distribution during training and testing [3, 8]. However, this assumption does not always hold in real situations, especially when no positive samples in the target domain are available.

In this paper, we formulate the task of adult content filtering in video as a domain adaptation problem. We use positive data from other sources with different distributions to that of the target platform. To reduce domain discrepancy between source and target, we apply domain adaptation based on Wasserstein distance minimization [9]. This allows us to learn a domain-invariant feature representation between all source data and negative-only target data distributions, see Figure 1. In order to filter content, we sample frames from the video stream and classifiying them using a CNN (ResNet-101). In order to create a strong baseline model, we include task-specific data augmentation as well as an attention mechanism. To validate the design choices, we carry out experiments on the public NPDI dataset [2] as well as data from a video streaming service. We evaluate the method in three settings: within-domain, cross-domain, and adapted-domain. To the best of our knowledge, this is the first work to apply feature distribution-based domain adaptation to the task of adult content detection.

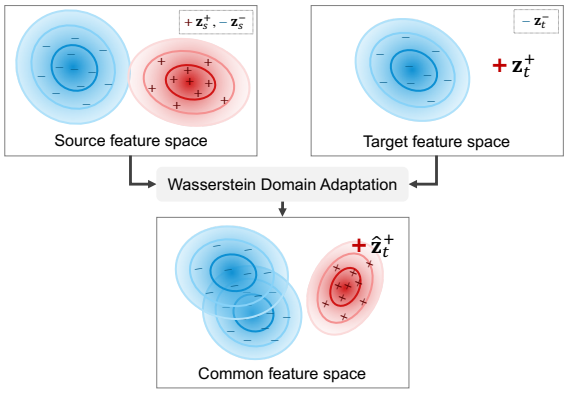

Figure 1: Domain adaptation. *Wasserstein domain adaptation maps source and target features ($\mathbf{z}_s$ and $\mathbf{z}_t$, respectively) into a common space with reduced between-domain discrepancy and preserved between-class discrimination.*

2 Prior work

Adult Scene Detection. A common approach to adult content detection is skin color detection in video frames [10, 11, 12, 13]. Additional cues such as shape and appearance have been used for increased robustness [14, 4, 15, 16]. More recently, deep learning-based approaches improved classification performance over hand-crafted features [17]. To focus on specific local image regions, Zhang *et al.* [18] proposed visual attention for a bags-of-visual-words (BoVW) model, which achieved better performance than a model without attention [19]. Wang *et al.* [20] adopted an attention-gated mechanism combined with a deep network, and showed that this mechanism helps improve performance. To focus on local image regions, Jin *et al.* [3] modeled an input image as a bag of regions and applied weighted multiple instance learning. Several works proposed deep learning architectures where local and global context are jointly taken into consideration [21, 22]. To detect the concept of provocative intent, Ganguly *et al.* [23] proposed a hierarchical model that includes multiple factors, such as exposed skin area, body pose, gestures, facial expressions and scene context.

*Email:{utsav.shah, mitsuru.nakazawa, bjorn.stenger}@rakuten.com

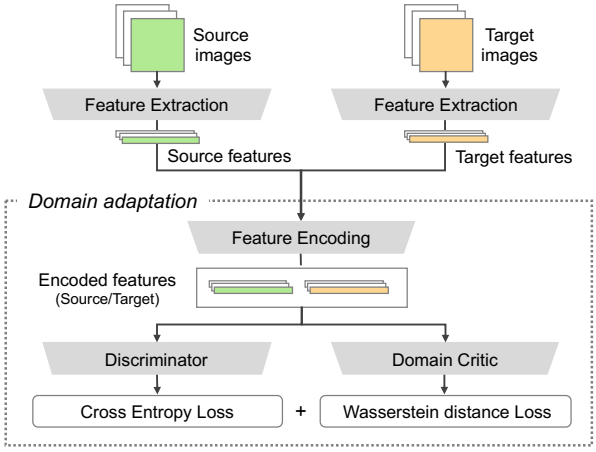

Figure 2: Proposed System. *We apply domain adaptation in the feature space by optimizing a cross entropy loss to preserve discrimination and a Wasserstein distance loss to reduce the distribution shift between source and target domains.*

Domain Adaptation. Domain shift is a well-known problem in visual recognition tasks when there is a distribution change between training and test data, namely, source and target domain [6, 7]. In the adult scene detection task, domain shift is caused by large differences in camera views and lighting conditions. In order to maintain detection performance, domain adaptation is an effective approach to minimize the distribution mismatch between two domains. Various domain adaptation approaches have been developed for visual recognition tasks. The review by Patel *et al*. [6] categorizes these into the following categories: feature augmentation, feature transformation, model parameter adaptation, dictionary learning, domain re-sampling, and multiple intermediate representations. Wang *et al*. [7] reviews approaches for visual domain adaptation developed for deep neural networks.

3 Method

Our method is composed of feature extraction and domain adaptation stages, see Figure 2. We assume that source images include both positive and negative data, whereas target images consist of negative data only. In conjunction with encoding extracted features in a common feature space, we reduce domain shift and preserve discrimination between positive and negative classes.

3.1 Feature extraction

We extract features using a ResNet-101 [24] and add three modules to improve feature extraction: data augmentation, an attention mechanism, and false positive filtering. For data augmentation we use RandAugment [25] for regularization during training. Additionaly, we synthesize images that are closer to target distribution. Specifically, we segment humans from images using DeepLabV3 [26] and superimpose them on natural or artificial backgrounds in representative sizes. To include an attention mechanism, we adopt the Attention Branch Network (ABN) architecture [27]. ABN optimizes an image-level attention map, which is combined with image feature maps to generate probability scores. Even though our method is not explicitly based on skin detection, skin areas are inherently learned as useful features for adult scene detection and can result in false positives, especially during close-up scenes of faces. We additionally run a face-detector to reduce the number of false positives.

3.2 Domain Adaptation

To reduce the distribution mismatch between source and target feature distributions we apply domain adaptation [28]. More specifically, given a feature vector $\mathbf{z}_s$ in the source domain S and $\mathbf{z}_t$ in the target domain T, the goal is to learn a transformation $W : S \rightarrow T$. This task can be approached as an optimal transport problem, where we seek to minimize the cost to move points from one distribution to another, measured by the Wasserstein distance, also known as Earth Mover's distance. Shen *et al*. [9] proposed a representation learning method guided by the Wasserstein distance to reduce domain discrepancy. Compared to prior domain adaptation methods [29, 30, 31], Wasserstein distance-based representation learning [9] has been shown to have high discriminative power and a more stable gradient when minimizing domain discrepancy [32]. The main difference to [9] is that we use the deep network described in Section 3.1 as initial feature extraction module and pass the features into an encoder layer, in which the domain-invariant feature space parameters are learned.

We detail our adaptation approach in the following. Given feature vectors $\mathbf{z}_s^+$ and $\mathbf{z}_s^- \in \mathbb{R}^d$ from positive and negative classes, respectively, in the source domain, and features from the negative class in the target domain, $\mathbf{z}_t^- \in \mathbb{R}^d$, we apply these steps:

1. We pass feature vectors $\mathbf{z} \in \mathbb{R}^d$ into another fully connected (FC) layer as a learnable mapping function G to output $\hat{\mathbf{z}} \in \mathbb{R}^m$ where $m < d$. We call this layer feature encoder with parameter ϑ_G.

2. The encoded feature vectors $\hat{\mathbf{z}}_s^-$, $\hat{\mathbf{z}}_s^+$, and $\hat{\mathbf{z}}_t^-$ are processed in two separate FC layers: a domain critic layer and a discriminator layer. The domain critic layer learns a function F to transform the encoded feature representation $\hat{\mathbf{z}} \in \mathbb{R}^m$ to a real number $z \in \mathbb{R}$. More specifically, the domain critic layer estimates the Wasserstein distance in order to reduce source and target domain discrepancy in the encoded feature space. Assuming we only have the negative data distribution in the target domain, the domain critic loss $\mathcal{L}_w$ between the representation of source and target distribution is calculated as:

$$\mathcal{L}_w = \frac{1}{n_s} \sum_{\mathbf{z}_s \in S} F\left(\hat{\mathbf{z}}_s\right) - \frac{1}{n_t^-} \sum_{\mathbf{z}_t^- \in T} F\left(\hat{\mathbf{z}}_t^-\right), \quad (1)$$

where n_s is the number of samples in the source domain, and n_t^- the number of negative samples in the target domain. The discriminator layer C maps $\mathbb{R}^m$ to $\mathbb{R}^c$, where c is the number of classes, here $c = 2$. The discriminator layer preserves the discrimination capability of the encoded features with the binary cross-entropy loss $\mathcal{L}_c$:

$$\mathcal{L}_c = -\frac{1}{n_s} \sum_{i=1}^{n_s} \sum_{j=0}^{1} \mathbb{1}\left(y_i^s = j\right) \log\left(C\left(\hat{\mathbf{z}}_s\right)_j\right), \quad (2)$$

where y_i^s is a binary label for the i-th source sample and $\mathbb{1}$ is the indicator function.

3. To train a domain invariant embedding, we optimize the following total loss function in an adversarial manner:

$$\min_{\vartheta_G, \vartheta_c} \left(\mathcal{L}_c + \alpha \max_{\vartheta_w} \left(\mathcal{L}_w - \beta \mathcal{L}_{\text{grad}}\right)\right), \quad (3)$$

where α is a balancing coefficient between the discrimination and domain invariance losses, and β is a coefficient to control the contribution of $\mathcal{L}_{\text{grad}} = (||\nabla_{\hat{\mathbf{z}}} G||_2 - 1)^2$, a gradient penalty to enforce the Lipschitz constraint [32, 33]. Additional feature representations $\hat{\mathbf{z}}$ are sampled at random points between source and target domains to calculate $\mathcal{L}_{\text{grad}}$. We optimize ϑ_G, ϑ_c, and ϑ_w via standard back-propagation. The between-domain discrepancy of the feature representations is iteratively reduced by minimizing the Wasserstein distance.

4. Test data $\mathbf{z}_t$ in the target domain is classified by mapping it into the domain-invariant feature space $\hat{\mathbf{z}}_t = G(\mathbf{z}_t)$ followed by applying a softmax classifier to the output of the discriminator $C(\hat{\mathbf{z}}_t)$.

4 Experiments

To validate the proposed method, we conducted the following three experiments. First, to confirm that the extracted features are suitable for the task of adult scene detection, we evaluate them on a standard within-domain classification task, *i.e.*, using the same dataset as the source and target domains. Next, for cross-domain evaluation, we train the model on the source distribution and evaluate its performance on different target distribution, comparing it with a number of other publicly available methods for adult content detection. Finally, we evaluate the proposed domain adaptation method, comparing it with two other adaptation methods on the same task.

4.1 Datasets

Live-streaming dataset (Live). To prepare a dataset containing representative image data, we crawled 214 archived videos from a live-streaming service, mainly containing footage of one person talking, singing or dancing, wearing casual clothes. It also contains many face close-ups as well as images without any people. We extracted 120,000 images by randomly sampling 1% of the frames. All images in this dataset belongs to the negative class.

Web-crawled dataset (Web). To construct a dataset containing both positive and negative classes, we crawled 230,000 images from the web. These images include scenes of people wearing diverse set of clothing as well as close-up face images and images with no humans, as in the Live dataset. Images of people wearing no top or no clothes are labeled as positive, all other images are labeled as negative.

NPDI Dataset. The public NPDI dataset has been widely used as a benchmark for adult content detection [2]. It contains 16,727 images, selected from 800 hours of video data. We only consider 'porn' and 'non-porn (easy)' categories of this dataset as these directly translate to our definition of positive and negative classes, respectively. Following Wang *et al*. [7], we confirm that 895 of 6,387 positive images are incorrectly labeled and exclude these in the experiments. In total, we use 5,692 positive and 6,784 negative images from the NPDI dataset. We follow the 5-fold cross validation protocol for the evaluation.

Table 1: Within-domain evaluation *on the combined Web and Live dataset. R: ResNet-101, GA: augmentation by generated images, RA: RandAugment, Att: ABN attention mechanism.*

	Acc	P@99% R	F1	mAP
R	98.86	87.72	97.16	86.58
R+GA	98.89	86.65	97.23	88.85
R+GA+RA	98.89	**93.22**	97.27	87.60
R+GA+RA+Att	**99.02**	92.99	**98.65**	**95.76**

4.2 Results

To measure the performance, we used four metrics: binary accuracy (Acc), precision@99%recall (P@99%R), F1-score (F1), and mean Average Precision (mAP).

4.2.1 Within-domain evaluation

In this experiment we combine the Live and Web datasets and split the merged dataset randomly into training and test sets with a 80:20 ratio. Table 1 shows the results from this within-domain evaluation, confirming that our model performs well across several metrics. We use data augmentation with images generated by overlaying segmented human foreground images onto background images (GA) as well as RandAugment [25] (RA). We observe an improvement for three of the metrics by introducing the

Table 2: Cross-domain evaluation on NPDI. *Comparison of Nude.js [34, 35], PornDetector [36], and NudeNet [37, 38] with our approach. The training datasets do not contain any NPDI images.*

	Acc	P@99% R	F1	mAP
Nude.js [34, 35]	60.29	34.49	59.55	41.84
PornDetector [36]	78.53	34.25	77.51	39.80
NudeNet [37, 38]	83.50	34.57	81.42	29.11
Ours (R+M+RA+At)	**89.14**	**34.67**	**87.51**	**46.30**

Table 3: Domain adaptation results on NPDI *(Target) based on model trained on other (Source) dataset. △: Trained only on Web+Live dataset. □: Trained on Web+Live+(NPDI negative) dataset.*

	Acc	P@99% R	F1	mAP
Ours (w/o adaptation)$^{\triangle}$	89.14	34.67	87.51	46.30
Ours (w/o adaptation)$^{\square}$	93.10	34.58	92.36	30.89
Euclidean-based adaptation	93.23	34.58	92.50	20.14
GMM-based adaptation	93.19	35.37	92.14	78.45
WD-based adaptation	**93.70**	**56.01**	**92.97**	**96.92**

ABN attention mechanism (Att). We use this model trained with data augmentation and attention as baseline model in the following experiments.

4.2.2 Cross-domain evaluation

To evaluate generalization to a different dataset, we use the model trained on our Live + Web dataset and evaluate it on NPDI. We split the NPDI dataset based on the standard fold data list. Table 2 shows the results for this cross-domain experiment. We compare the performance of our method with publicly available adult content detectors: Nude.js [34, 35], a method based on skin detection, and PornDetector [36] and NudeNet [37, 38], which are based on deep learning. The proposed method consistently shows better results on NPDI across all metrics.

4.2.3 Domain adaptation

For this experiment we consider images from the merged Web and Live dataset as coming from the source domain. The negative class of the NPDI dataset forms the negative target distribution, ignoring the positive NPDI samples in this experiment.

The first row in Table 3 shows the results of our baseline model. The accuracy increases by including the negative data of NPDI (row 2). Starting with this model, we compare alternative domain adaptation methods based on distribution shift. For this, we translate the target features by the difference vector measured from the target negative

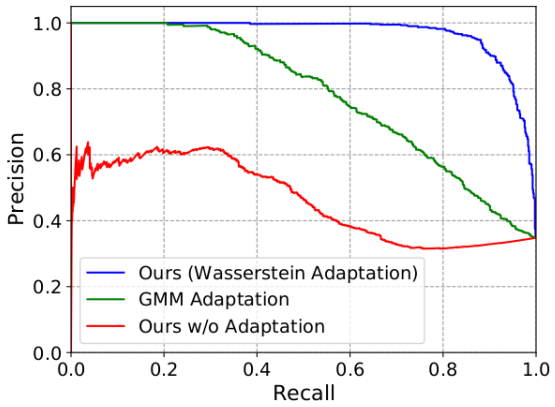

Figure 3: Precision-recall curve *for domain adaptation comparison on the NPDI evaluation set (target domain). The Wasserstein distance-based adaptation approach significantly outperforms the GMM-based approach.*

cluster mean to the one in the source domain. We then classify the features by either minimum Euclidean distance (Euclidean-based adaptation) or maximum log-likelihood using the trained Gaussian Mixture Model (GMM) with 4 mixture components (GMM-based adaptation) [39].

For our Wasserstein distance (WD) based method we empirically set hyper-parameters in (3) to $\alpha = 0.1$ and $\beta = 10.0$. The number of nodes in the feature encoder's input, hidden, and output layers are 2048, 1000, and 500, respectively. The number of nodes in the domain critic's input, hidden, and output layers are 500, 100, and 1, respectively. The discriminator input and output FC layers contain 500 and 2 nodes, respectively. An Adam optimizer is used with learning rates of 10^{-3} for the domain critic and 10^{-4} for the discriminator, respectively.

Table 3 and the precision-recall curve in Figure 3 show the effectiveness of applying domain adaptation to our adult content filtering model. Wasserstein domain adaptation shows the best performance across all metrics, maintaining discriminative power between classes in the target domain. The high precision at high recall values is especially important for our application.

5 Conclusion

In this paper we proposed domain adaptation based on Wasserstein distance minimization to improve cross-domain recognition performance of adult content detection. We designed our content filtering model using task-specific image data augmentation, an attention mechanism, and false-positive filtering. Through experiments, we confirmed that domain adaptation effectively mitigates the domain-shift problem between source and target and improves content filtering performance in a cross-dataset setting even without positive training samples from the target domain.

References

[1] Christian X. Ries and Rainer Lienhart. "A survey on visual adult image recognition". In: *Multimedia Tools and Applications* 69 (2012), pp. 661–688.

[2] Sandra Avila et al. "Pooling in image representation: The visual codeword point of view". In: *Computer Vision and Image Understanding* 117.5 (2013), pp. 453–465. URL: http://dx.doi.org/10.1016/j.cviu.2012.09.007.

[3] Xin Jin, Yuhui Wang, and Xiaoyang Tan. "Pornographic Image Recognition via Weighted Multiple Instance Learning". In: *IEEE Transactions on Cybernetics* (2018), pp. 1–10. arXiv: 1902.03771.

[4] Ana Paula B. Lopes et al. "Nude detection in video using bag-of-visual-features". In: *Proceedings of SIBGRAPI 2009 - 22nd Brazilian Symposium on Computer Graphics and Image Processing* (2009), pp. 224–231.

[5] Martin D. More et al. "Seamless Nudity Censorship: An Image-to-Image Translation Approach based on Adversarial Training". In: *Proceedings of the International Joint Conference on Neural Networks* 2018-July (2018).

[6] Vishal M. Patel et al. "Visual Domain Adaptation: A survey of recent advances". In: *IEEE Signal Processing Magazine* 32.3 (2015), pp. 53–69.

[7] Mei Wang and Weihong Deng. "Deep visual domain adaptation: A survey". In: *Neurocomputing* 312 (2018), pp. 135–153. arXiv: arXiv:1802.03601v4.

[8] Yuhui Wang, Xin Jin, and Xiaoyang Tan. "Pornographic image recognition by strongly-supervised deep multiple instance learning". In: *Proceedings - International Conference on Image Processing, ICIP* 2016-Augus.September (2016), pp. 4418–4422.

[9] Jian Shen et al. "Wasserstein distance guided representation learning for domain adaptation". In: *32nd AAAI Conference on Artificial Intelligence, AAAI 2018* (2018), pp. 4058–4065. arXiv: 1707.01217.

[10] D. A. Forsyth and M. M. Fleck. "Automatic detection of human nudes". In: *International Journal of Computer Vision* 32.1 (1999), pp. 63–77.

[11] Michael Jeffrey Jones, Michael J Jones, and James M Rehg. "Statistical Color Models with Application to Skin Detection Statistical Color Models with Application to Skin Detection". In: *Computer Vision and Pattern Recognition* 46.January (2016), pp. 1–23.

[12] Wayne Kelly, Andrew Donnellan, and Derek Molloy. "Screening for objectionable images: A review of skin detection techniques". In: *Proceedings - IMVIP 2008, 2008 International Machine Vision and Image Processing Conference* October 2008 (2008), pp. 151–158.

[13] Christian Platzer, Martin Stuetz, and Martina Lindorfer. "Skin sheriff: A machine learning solution for detecting explicit images". In: *SFCS 2014 - Proceedings of the 2nd International Workshop on Security and Forensics in Communication Systems* June (2014), pp. 45–55.

[14] Thomas Deselaers, Lexi Pimenidis, and Hermann Ney. "Bag-of-visual-words models for adult image classification and filtering". In: *Proceedings - International Conference on Pattern Recognition* December (2008).

[15] Hyun-Seok Min, W. De Neve, and Yong Man Ro. "Use of semantic features for filtering of malicious content in an IPTV environment". In: *2010 Digest of Technical Papers International Conference on Consumer Electronics (ICCE)*. 2010, pp. 79–80.

[16] Adrian Ulges and Armin Stahl. "Automatic detection of child pornography using color visual words". In: *Proceedings - IEEE International Conference on Multimedia and Expo* (2011), pp. 3–8.

[17] Mohamed Moustafa. "Applying deep learning to classify pornographic images and videos". In: (2015). arXiv: 1511.08899. URL: http://arxiv.org/abs/1511.08899.

[18] Jing Zhang et al. "An approach of bag-of-words based on visual attention model for pornographic images recognition in compressed domain". In: *Neurocomputing* 110 (June 2013), pp. 145–152.

[19] L. Sui et al. "Research on pornographic images recognition method based on visual words in a compressed domain". In: *IET Image Processing* 6.1 (2012), pp. 87–93.

[20] Liyuan Wang et al. "Porn Streamer Recognition in Live Video Streaming via Attention-Gated Multimodal Deep Features". In: *IEEE Transactions on Circuits and Systems for Video Technology* PP (Dec. 2019), pp. 1–1.

[21] Xinyu Ou et al. “Adult image and video recognition by a deep multicontext network and fine-to-coarse strategy”. In: *ACM Transactions on Intelligent Systems and Technology* 8.5 (2017).

[22] Xizi Wang et al. “Adult Image Classification”. In: *2018 25th IEEE International Conference on Image Processing (ICIP)* (2018), pp. 2989–2993.

[23] Debashis Ganguly, Mohammad H. Mofrad, and Adriana Kovashka. “Detecting sexually provocative images”. In: *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017* c (2017), pp. 660–668.

[24] K. He et al. “Deep Residual Learning for Image Recognition”. In: *CVPR*. 2016, pp. 770–778.

[25] Ekin D Cubuk et al. “Randaugment: Practical automated data augmentation with a reduced search space”. In: *CVPR Workshops*. 2020, pp. 702–703.

[26] Liang-Chieh Chen et al. “Rethinking Atrous Convolution for Semantic Image Segmentation”. In: *CoRR* abs/1706.05587 (2017). arXiv: 1706.05587. URL: http://arxiv.org/abs/1706.05587.

[27] Hiroshi Fukui et al. “Attention Branch Network: Learning of Attention Mechanism for Visual Explanation”. In: *CoRR* abs/1812.10025 (2018). arXiv: 1812.10025. URL: http://arxiv.org/abs/1812.10025.

[28] Kate Saenko et al. “Adapting Visual Category Models to New Domains”. In: *ECCV* (2010), pp. 1–14.

[29] Yaroslav Ganin et al. “Domain-adversarial training of neural networks”. In: *Advances in Computer Vision and Pattern Recognition* 17.9783319583464 (2017), pp. 189–209. arXiv: 1505.07818.

[30] Eric Tzeng et al. “Deep Domain Confusion: Maximizing for Domain Invariance”. In: (2014). arXiv: 1412.3474. URL: http://arxiv.org/abs/1412.3474.

[31] Baochen Sun, Jiashi Feng, and Kate Saenko. “Return of frustratingly easy domain adaptation”. In: *30th AAAI Conference on Artificial Intelligence, AAAI 2016* Figure 1 (2016), pp. 2058–2065. arXiv: 1511.05547.

[32] Martin Arjovsky, Soumith Chintala, and Léon Bottou. “Wasserstein GAN”. In: (2017). arXiv: 1701.07875. URL: http://arxiv.org/abs/1701.07875.

[33] Ishaan Gulrajani et al. “Improved training of wasserstein GANs”. In: *Advances in Neural Information Processing Systems* 2017-December (2017), pp. 5768–5778. arXiv: 1704.00028.

[34] Rigan Ap-Apid. “An algorithm for nudity detection”. In: *5th Philippine Computing Science Congress*. 2005, pp. 201–205.

[35] Nude.py for Nude.js. *Date of Access: February 12, 2021*. URL: https://github.com/hhatto/nude.py.

[36] PornDetector. *Date of Access: February 12, 2021*. URL: https://github.com/bakwc/PornDetector.

[37] NudeNet Blog. *Date of Access: February 4, 2021*. URL: https://praneethbedapudi.medium.com/nudenet-an-ensemble-of-neural-nets-for-nudity-detection-and-censoring-d9f3da721e3.

[38] NudeNet GitHub. *Date of Access: February 12, 2021*. URL: https://github.com/notAI-tech/NudeNet.

[39] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. “Speaker verification using adapted Gaussian mixture models”. In: *Digital Signal Processing: A Review Journal* 10.1 (2000), pp. 19–41.