Learning to Detect Genuine versus Posed Pain from Facial Expressions using Residual Generative Adversarial Networks

Learning to Detect Genuine versus Posed Pain from Facial Expressions using Residual Generative Adversarial Networks

Mohammad Tavakolian[1], Carlos Guillermo Bermudez Cruces[1,2], and Abdenour Hadid[1]
[1] Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland
[2] School of Telecommunications Engineering, Technical University of Madrid, Spain

Abstract— **We present a novel approach based on Residual Generative Adversarial Network (R-GAN) to discriminate genuine pain expression from posed pain expression by magnifying the subtle changes in the face. In addition to the adversarial task, the discriminator network in R-GAN estimates the intensity level of the pain. Moreover, we propose a novel Weighted Spatiotemporal Pooling (WSP) to capture and encode the appearance and dynamic of a given video sequence into an image map. In this way, we are able to transform any video into an image map embedding subtle variations in the facial appearance and dynamics. This allows using any pre-trained model on still images for video analysis. Our extensive experiments show that our proposed framework achieves promising results compared to state-of-the-art approaches on three benchmark databases, i.e., UNBC-McMaster Shoulder Pain, BioVid Head Pain, and STOIC.**

I. INTRODUCTION

Discriminating spontaneous facial expressions from posed ones is a challenging problem, due to the high visual similarity between the two phenomena. In both situations, a subject's face shows similar features, which are difficult to discriminate. Fortunately, spontaneous expressions generally occur in less controlled situations with some involuntary subtle movements in the subject's facial muscles in a response to a stimuli. Capturing these subtle facial dynamics is however non-trivial. Moreover, head movements further complicate the detection of spontaneous expressions. On the other hand, it is well-substantiated that posed facial expressions are usually exaggerated and have higher occurrence frequency than spontaneous ones [4]. For instance, analyzing facial action units reveals that deliberate smiles are more intense than spontaneous smiles [5], since the subjects have more control over their facial movements to express an exaggerated smile.

Studies on Facial Action Coding System (FACS) [6] have showed that facial Action Units (AUs) react differently when someone deliberately shows facial expressions [7]. Based on this finding, some works have been conducted on detecting and analyzing AUs that are more associated with categorical facial expression, i.e., disgust, anger, sadness, happiness, surprise, and contempt. By extracting and combining appearance features representing facial textures and geometric features, most of previous works have strived to obtain a discriminative representation of spontaneous facial expressions. There is still a correlation among geometric and appearance features that prevents to capture subtle dynamic of spontaneous facial expressions. The recent great advancements in face analysis are stimulated by approaches that learn deep representations from the region of interest and perform classification or regression on the top of the learned representations [8]. Although these methods perform well and learn discriminative representations, they seem to fail to detect subtle changes in the face. It is indeed challenging to devise a rich representation encoding the subtle facial changes.

One basic approach is to improve the resolution of small regions of interest by enlarging the scale of input images and generate high-resolution feature maps [9]. On the other hand, some other methods produce multi-scale representations using hierarchical techniques [10]. However, scaling the input dimension usually increases the computational cost for training and testing. In addition, multi-scale representations do not capture dynamics of the face that have discriminative information for facial analysis. We argue that detecting correlations between subtle and large-scale facial movements is important. Once these correlations are detected, we can transform the representation of subtle movements using learned large-scale information maintaining the dynamic of the face.

In this paper, we propose a Residual Adversarial Generative Network (R-GAN) to generate a magnified representation of subtle facial movements. We encode the small motion of facial components by capturing and encoding the appearance and dynamic of video sequences. Our proposed method further improves the summarized representations by magnifying the small spatiotemporal variations alike macro-expressions. Figure 1 shows the summarized representations of given videos and detected subtle facial movements. Similar to the classical GAN [11], R-GAN has two stacked networks, i.e., a generator and a discriminator network. In order to intensify small dynamics and make their representation more discriminative, we use fine-grained details from early layers of the generator by adding a residual connection. Using this technique, the generator network learns magnified representation of the input. Using the intermediate representations, the discriminator network is trained to detect posed and genuine pain expressions. We also add a classification network to the discriminator to classify the detected facial expression. Using this strategy, we further validate the accuracy of the whole model. Moreover, the classification network supervises the adversarial learning of the model.

Among the salient contributions of this paper, we can cite:

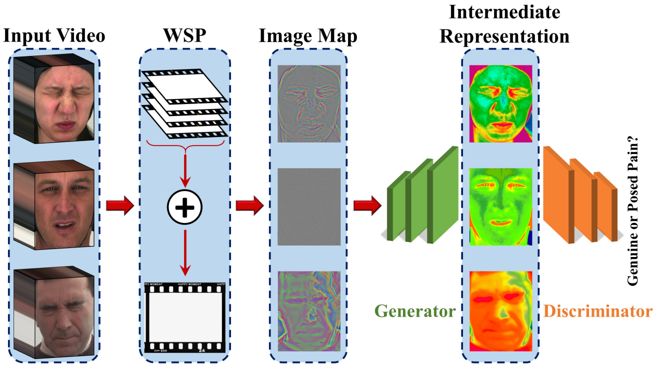

Fig. 1: Our proposed framework: by capturing the appearance and dynamic changes, an input video is summarized into a single image map that is fed into a residual generative adversarial network for discriminating posed from genuine pain.

(1) We present a Residual Generative Adversarial Network (R-GAN) for discriminating genuine pain expressions from posed ones. In contrast to the classical GAN, we include a residual representation learning network to the generator for magnifying the subtle dynamics of the face; (2) we encode the spatiotemporal variations of the video by summarizing the entire video frames into one image map preserving the video dynamics. This approach significantly improves the training process and relaxes the need for a large amount of data (videos) for training the networks; (3) we conduct extensive experiments on three benchmark and publicly available datasets to evaluate the performance of the proposed method.

II. RELATED WORK

The facial expression analysis methods can be classified into two broad categories, i.e., static and dynamic. Static methods analyze the facial structure in a single frame without considering the contextual information of consecutive frames in a video sequence [12], [13]. On the other hand, dynamic approaches [14], [15] take the advantage of temporal relationship of adjacent frames to enhance the performance. The existing methods mainly focus on classification of different facial expressions. However, few works have investigated the problem of distinguishing posed facial expressions from spontaneous ones. Valstar *et al.* [16] revealed that the temporal modeling of brow movements is crucial for detecting spontaneous facial expressions. Cohen and Schmidt [7] showed the importance of the amplitude and duration of facial movements. Park *et al.* [17] used face region specific weight factors to achieve a new representation of the subtle expressions.

Although being successful, the aforementioned studies mainly aim at differentiating posed and spontaneous facial expressions. To be specific, distinguishing genuine pain from posed one is important in some medical and criminal applications [18]. According to FACS, analyzing facial expression of pain provide information for detecting genuine pain [19]. Hill *et al.* [19] showed that the facial actions presented during deliberately dissimulated pain expressions exhibit significantly different temporal behaviour than genuine pain expressions. Littlewort *et al.* [20] used Support Vector Machines (SVM) to determine posed pain by extracting Gabor wavelet features from the faces.

Due to the subtlety of facial movements, an effective magnified representation of facial dynamic is required. Huang *et al.* [21] proposed a super-resolution technique based on recurrent convolutional neural network to capture the long-term dynamics in the face videos. Kappeler *et al.* [22] applied super-resolution convolutional neural network on a group of motion compensated frames with a fixed temporal scale. Yang *et al.* [23] used a joint spatiotemporal residual network for intensifying the dynamics in the videos. Leding *et al.* [24] applied Generative Adversarial Network (GAN) to image super-resolution. However, these methods seem to suffer from efficiently capturing the complex underlying subtle movements.

III. POSED VERSUS GENUINE PAIN

In this present work, we propose a novel framework to address the challenging task of discriminating genuine pain expressions from posed ones. First, we introduce a video summarization method, which is based on spatiotemporal pooling, to encode the dynamic and appearance of a video sequence into an image map. Then, we present a GAN alike model for differentiating genuine pain expressions from posed ones.

A. *Weighted Spatiotemporal Pooling*

It is well known that the performance of deep models is highly dependent on the size and quality of the training data. For instance, when dealing with videos, they are usually divided into segments (mainly to increase the number of samples) before feeding them to a CNN. However, complex deep models dealing with video sequences, tend to involve a lot of parameters for tuning. Although some methods treat videos as sequences of temporally stacked frames, these methods are usually not able to efficiently extract the dynamic of the videos. In particular, representing the appearance and dynamic of the face is essential for the facial expression analysis due to the correlation between different facial expressions. This motivates us to propose a novel spatiotemporal pooling method for capturing and encoding the appearance and dynamics of the whole video sequence into a single image map that is used as the input of our deep model.

In order to encode the appearance and the dynamics in the videos, one should capture the optimal spatiotemporal information. To encode and capture such information, we rely on the fact that visual attention is usually given to the regions that have more descriptive information [25]. Inspired by information theory, the local information of an image can be quantified in terms of sequences of bits. We extend this theory to the context of video sequences, where a sequence of frames exhibits spatiotemporal characteristics. Capturing spatiotemporal information correlates with a powerful statistical video model. To reduce the high dimensionality of the

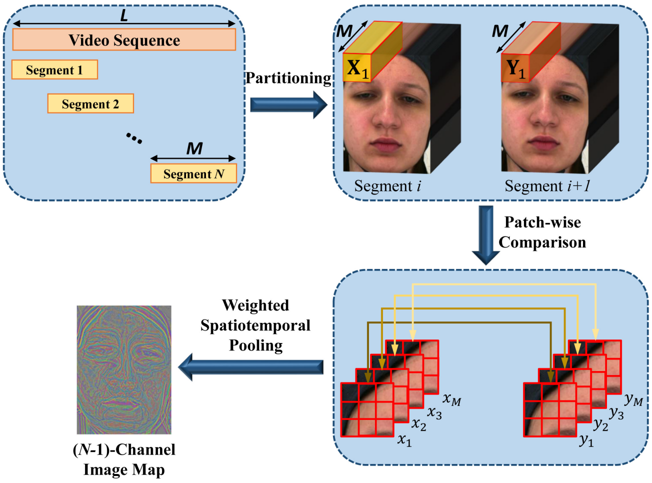

Fig. 2: An overview of the proposed WSP. The video is divided into overlapping segments of the same length. A path-wise comparison is carried out between two consecutive segments to build a summarized image representing the appearance and dynamic changes within two segments of the video.

data, we follow a Markov assumption that the probability density of a point is estimated by its neighboring points. Therefore, the task is to obtain a statistical model of groups of neighboring points. The Gaussian Scale Mixture (GSM) is a powerful tool for this purpose. Based on GSM, the neighborhood is typically built on the top of a group of neighboring coefficients in a multi-resolution image transform domain.

Let $\mathbf{x} \in \mathbb{R}^K$ be a column vector encompassing neighboring points. Based on GSM, $\mathbf{x}$ can be modelled as a product of two independent components, i.e., $\mathbf{x} = \alpha U$, where U is a zero-mean Gaussian vector with convariance matrix $\mathbf{C}_U$ and α is a mixing multiplier. The value of α varies over space and time. Hence, $\mathbf{x}$ is a zero-mean Gaussian vector with covariance $\mathbf{C}_\mathbf{x} = \alpha^2 \mathbf{C}_U$. The mutual information between frames of two time instances provides useful information about the appearance and dynamic of a video sequence. We propose a model to capture spatiotemporal information by comparing frames of two overlapping consecutive segments of the video and computing a weighted score of variations between them. In this way, we can summarize two segments of the video into one channel image. By repeating this process for all segments, we obtain a multichannel image that summarizes the appearance and dynamic of the whole video sequence. Figure 2 illustrates our proposed Weighted Spatiotemporal Pooling method (WSP).

We assume that the neighborhood $\mathbf{x}$ undergoes a series of spatiotemporal variations after passing time t, resulting in a deformed neighborhood $\mathbf{y}$.

$$\mathbf{y} = g\mathbf{x} + V = g\alpha U + V \tag{1}$$

where the spatiotemporal deformation is modeled using a gain factor g followed by an additive independent Gaussian noise V with covariance $\mathbf{C}_V = \sigma_v^2 \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. One limitation of the above model is that, it does not account for noise and only consider the spatiotemporal variations in the ideal case. To address this problem, we add the noise element to the reference and deformed neighborhood models.

$$\mathbf{p} = \mathbf{x} + N_1 = sU + N_1 \tag{2}$$

$$\mathbf{q} = \mathbf{y} + N_2 = gsU + V + N_2 \tag{3}$$

where N_1 and N_2 are independent Gaussian noise with covariance matrices $\mathbf{C}_{N_1} = \mathbf{C}_{N_2} = \sigma_n^2 \mathbf{I}$. The parameter σ_n^2 is the uncertainty of noisy observations. So, we can derive the covariance matrices of $\mathbf{y}$, $\mathbf{p}$, and $\mathbf{q}$ as:

$$\mathbf{C}_\mathbf{y} = g^2 \alpha^2 \mathbf{C}_U + \sigma_v^2 \mathbf{I} \tag{4}$$

$$\mathbf{C}_\mathbf{p} = \alpha^2 \mathbf{C}_U + \sigma_n^2 \mathbf{I} \tag{5}$$

$$\mathbf{C}_\mathbf{q} = g^2 \alpha^2 \mathbf{C}_U + \sigma_v^2 \mathbf{I} + \sigma_n^2 \mathbf{I} \tag{6}$$

At each point, the information of the reference and deformed frames is obtained by the mutual information $I(\mathbf{x}|\mathbf{p})$ and $I(\mathbf{y}|\mathbf{q})$, respectively. We aim to approximate the perceptual information content from both frames. To be specific, we subtract the common information shared between $\mathbf{p}$ and $\mathbf{q}$ from $I(\mathbf{x}|\mathbf{p})$ and $I(\mathbf{y}|\mathbf{q})$. So, we define a weight based on the mutual information content as:

$$w = I(\mathbf{x}|\mathbf{p}) + I(\mathbf{y}|\mathbf{q}) - I(\mathbf{p}|\mathbf{q}) \tag{7}$$

To solve Equation (7), it should be noted that $\mathbf{x}$, $\mathbf{y}$, $\mathbf{p}$, and $\mathbf{q}$ are all Gaussian for a given fixed α. Therefore, the mutual information approximation can be computed using the determinants of the covariances.

$$I(\mathbf{x}|\mathbf{p}) = \frac{1}{2} \log \left[\frac{|\mathbf{C}_\mathbf{x}| \times |\mathbf{C}_\mathbf{p}|}{|\mathbf{C}_{(\mathbf{x},\mathbf{p})}|} \right] \tag{8}$$

$$I(\mathbf{y}|\mathbf{q}) = \frac{1}{2} \log \left[\frac{|\mathbf{C}_\mathbf{y}| \times |\mathbf{C}_\mathbf{q}|}{|\mathbf{C}_{(\mathbf{y},\mathbf{q})}|} \right] \tag{9}$$

$$I(\mathbf{p}|\mathbf{q}) = \frac{1}{2} \log \left[\frac{|\mathbf{C}_\mathbf{p}| \times |\mathbf{C}_\mathbf{q}|}{|\mathbf{C}_{(\mathbf{p},\mathbf{q})}|} \right] \tag{10}$$

where

$$\mathbf{C}_{(\mathbf{x},\mathbf{p})} = \begin{bmatrix} \mathbf{C}_\mathbf{x} & \mathbf{C}_\mathbf{xp} \\ \mathbf{C}_\mathbf{px} & \mathbf{C}_\mathbf{p} \end{bmatrix} \tag{11}$$

$$\mathbf{C}_{(\mathbf{y},\mathbf{q})} = \begin{bmatrix} \mathbf{C}_\mathbf{y} & \mathbf{C}_\mathbf{yq} \\ \mathbf{C}_\mathbf{qy} & \mathbf{C}_\mathbf{q} \end{bmatrix} \tag{12}$$

$$\mathbf{C}_{(\mathbf{p},\mathbf{q})} = \begin{bmatrix} \mathbf{C}_\mathbf{p} & \mathbf{C}_\mathbf{pq} \\ \mathbf{C}_\mathbf{qp} & \mathbf{C}_\mathbf{q} \end{bmatrix} \tag{13}$$

According to the fact that U and N_1 are independent, Equation (11) can be simplified using:

$$\mathbf{C}_\mathbf{xp} = \mathbf{C}_\mathbf{px} = \mathbb{E}\left[\mathbf{x}\mathbf{p}^T\right] = \alpha^2 \mathbf{C}_U = \mathbf{C}_\mathbf{x} \tag{14}$$

This results in:

$$|\mathbf{C}_{(\mathbf{x},\mathbf{p})}| = \left| \begin{bmatrix} \mathbf{C}_\mathbf{x} & \mathbf{C}_\mathbf{x} \\ \mathbf{C}_\mathbf{x} & \mathbf{C}_\mathbf{p} \end{bmatrix} \right| = |\sigma_n^2 \mathbf{C}_\mathbf{x}| \tag{15}$$

Similarly, we can derive:

$$\mathbf{C_{yq}} = \mathbf{C_{qy}} = g^2\alpha^2\mathbf{C}_U + \sigma_v^2\mathbf{I} = \mathbf{C_y} \tag{16}$$

$$\mathbf{C_{pq}} = \mathbf{C_{qp}} = g^2\alpha^2\mathbf{C}_U \tag{17}$$

$$|\mathbf{C}_{(\mathbf{y},\mathbf{q})}| = \left|\begin{bmatrix}\mathbf{C_y} & \mathbf{C_y}\\ \mathbf{C_y} & \mathbf{C_q}\end{bmatrix}\right| = |\sigma_n^2\mathbf{C_q}| \tag{18}$$

Putting all the above equations together, we can simplify the mutual information weight function in Equation (7) as:

$$w = \frac{1}{2}\log\left[\frac{|\mathbf{C}_{(\mathbf{p},\mathbf{q})}|}{\sigma_n^{4K}}\right] \tag{19}$$

where

$$|\mathbf{C}_{(\mathbf{p},\mathbf{q})}| = |\left(\left(\sigma_v^2+\sigma_n^2\right)\alpha^2+\sigma_n^2g^2\alpha^2\right)\mathbf{C}_U + \sigma_n^2\left(\sigma_v^2+\sigma_n^2\right)\mathbf{I}| \tag{20}$$

Applying an eigenvalue decomposition to the covariance matrix $\mathbf{C}_U = \mathbf{O}\Lambda\mathbf{O}^T$, where $\mathbf{O}$ is an orthogonal matrix and Λ is a diagonal matrix with eigenvalues λ_k for $k = 1,\ldots,K$ along its diagonal entries, we can compute $|\mathbf{C}_{(\mathbf{p},\mathbf{q})}|$.

$$|\mathbf{C}_{(\mathbf{p},\mathbf{q})}| = |\mathbf{O}\{\left(\sigma_v^2+(1+g^2)\sigma_n^2\right)\alpha^2\Lambda + \sigma_n^2(\sigma_v^2+\sigma_n^2)\mathbf{I}\}\mathbf{O}^T| \tag{21}$$

Due to the orthogonal property of $\mathbf{O}$ and the expression between two $\mathbf{O}$, $|\mathbf{C}_{(\mathbf{p},\mathbf{q})}|$ is obtained as a closed-form equation.

$$|\mathbf{C}_{(\mathbf{p},\mathbf{q})}| = \prod_{k=1}^{K}\{\left(\sigma_v^2+(1+g^2)\sigma_n^2\right)\alpha^2\lambda_k + \sigma_n^2(\sigma_v^2+\sigma_n^2)\} \tag{22}$$

Hence, Equation (19) can be expressed as

$$w = \frac{1}{2}\sum_{k=1}^{K}\log\left(1+\frac{\sigma_v^2}{\sigma_n^2}+\left(\frac{\sigma_v^2}{\sigma_n^4}+\frac{1+g^2}{\sigma_n^2}\right)\alpha^2\lambda_k\right) \tag{23}$$

The obtained weight function shows an interesting connection with the local deformation withing frames of video. According to the deformation model in Equation (1), the variations from $\mathbf{x}$ to $\mathbf{y}$ are characterized by the gain factor g and the random deformation σ_v^2. As g is a scale factor along the signal direction, it does not cause any changes in the structure of the image. Thus, the structural deformations are captured by σ_v^2. Our weight function increases monotonically with σ_v^2. This demonstrates that more weights are cast to the areas that have larger variations.

We still need to approximate a set of parameters, i.e., $\mathbf{C}_U$, α^2, g, and σ_v^2, to use the weight function of Equation (23). We estimate $\mathbf{C}_U$ as

$$\hat{\mathbf{C}}_U = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T \tag{24}$$

where N is the number of evaluation widows and $\mathbf{x}_i$ is the i-th neighborhood vector. The multiplier α is spatially varying and can be approximated using a maximum likelihood estimator.

$$\hat{\alpha}^2 = \frac{1}{K}\mathbf{x}^T\mathbf{C}_U^{-1}\mathbf{x} \tag{25}$$

We can also obtain the deformation parameters g and σ_v^2 by optimizing the following least square regression problem.

$$\hat{g} = \arg\min_{g}\|\mathbf{y}-g\mathbf{x}\|_2^2 \tag{26}$$

By taking the first-order derivative from Equation (26), we have:

$$\hat{g} = \frac{\mathbf{x}^T\mathbf{y}}{\mathbf{x}^T\mathbf{x}} \tag{27}$$

Putting this into Equation (1), we can compute σ_v^2 using V^TV/K, which results in:

$$\hat{\sigma}_v^2 = \frac{1}{K}\left(\mathbf{y}^T\mathbf{y}-\hat{g}\mathbf{x}^T\mathbf{y}\right) \tag{28}$$

We compute the information content weight by moving a sliding window across each frame of two consecutive frames (see Fig. 2), where the window covers a $H\times W$ spatial neighborhood at each location. This process results in an information content weight map for each two overlapping segments of the video. Let x_i and y_i be the i-th points in the reference frame $\mathbf{X}$ and the deformed frame $\mathbf{Y}$, respectively. The Mean Square Error (MSE) between two frames is given by

$$\text{MSE} = \frac{1}{P}\sum_{i=1}^{P}(x_i-y_i)^2 \tag{29}$$

where P is the total number of points in the frame. We define an information content weighted MSE for the corresponding location of the central point in the spatial neighborhood using Equation (23). Assuming $x_{j,i}$ and $y_{j,i}$ are the i-th point at the j-th frame and $w_{j,i}$ be the information content weight computed at the corresponding location, we derive Weighted Spatiotemporal Pooling (WSP) as:

$$\text{WSP}(\mathbf{x},\mathbf{y}) = \prod_{j=1}^{M}\left(\frac{\sum_i w_{j,i}(x_{j,i}-y_{j,i})^2}{\sum_i w_{j,i}}\right) \tag{30}$$

where M is the length of each segment of the video. Repeating this process for all two consecutive segments, we obtain $N-1$ single images, which encode the appearance and dynamic variations within the whole video. By stacking all the obtained images together, we build a $(N-1)$-channel image map that will be used as an input for facial expression analysis.

B. *Residual Generative Adversarial Network*

In the previous section, we performed a Weighted Spatiotemporal Pooling (WSP) to exploit the appearance and the dynamic variations of a video sequence by summarizing the whole length of the video into an image. Here, we introduce a new Residual Generative Adversarial Network (R-GAN) to differentiate genuine pain expression from posed one. We present a different architecture for the generator network that produces magnified representations for subtle changes in the facial structures. In addition, the discriminator network supervises the generative process by computing an adversarial loss and a classification loss. Figure 3 shows an overview of our proposed R-GAN architecture.

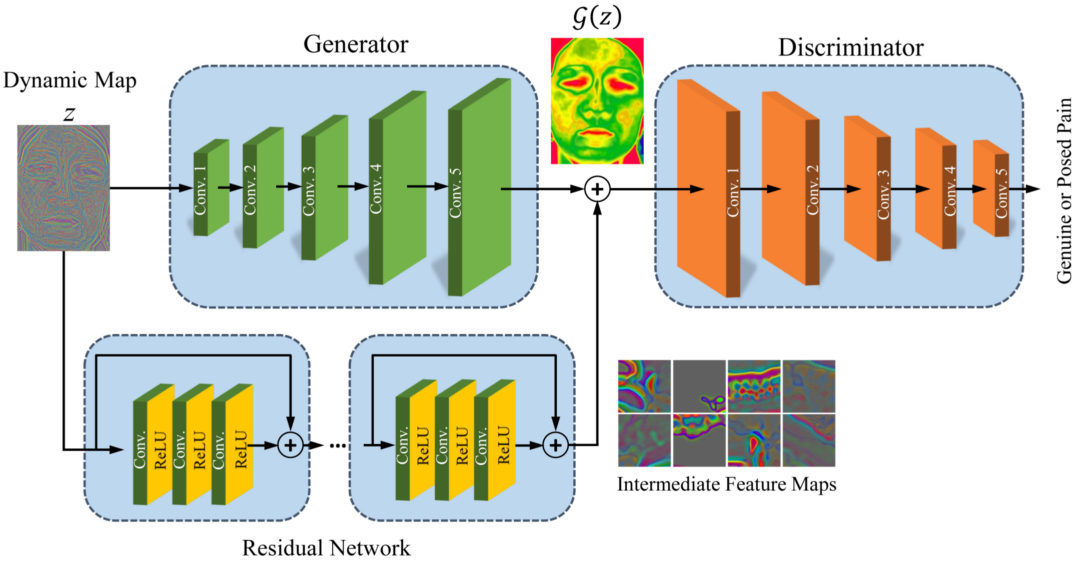

Fig. 3: The proposed Residual Generative Adversarial Network (R-GAN). The generator is coupled with a deep residual network that takes data with fine-grained details from the input. The residual representations are used to improve the original generator output by magnifying the areas with high variations. The discriminator has two parts, i.e., the adversarial part and the classification part. The adversarial network decides whether the input is genuine or not, while the classification network estimates the intensity of the pain level.

The objective function of the basic GAN [20] is based on a minimax optimization problem, which is expressed as:

$$\min_{G}\max_{D} L(G,D) = \mathbb{E}_{q\sim p_{data(q)}} \log D(q) + \mathbb{E}_{z\sim p_z(z)} \log(1-D(G(z))) \quad (31)$$

where G stands for the generator that maps data z from the input distribution $p_z(z)$ to the distribution $p_{data(q)}$ over data q. D represents the discriminator that estimate the probability of a sample that comes from the distribution $p_{data(q)}$ rather than G. In this paper, q and z are the representations for magnified and subtle spatiotemporal variations of the face, i.e., F_m and F_s, respectively. Our goal is to train the generator so that it maps the representations of subtle changes F_s to the magnified ones $G(F_s)$ that can be used for discrimination process. Due to subtlety of changes in the face, learning the representation $G(F_s)$ that matches F_m is a non-trivial task. Therefore, we presents a new residual generator network that has a residual connection in addition to the vanilla generator model. In this way, the model provides low-level features of the subtle variations f from which the generator learns to produce the residual representation between the representations of magnified and subtle changes through a residual learning process. Hence, we reformulate the objective function of GAN as:

$$\min_{G}\max_{D} L(G,D) = \mathbb{E}_{F_m\sim p_{data(F_m)}} \log D(F_m) + \mathbb{E}_{F_s\sim p_{F_s}(F_s|f)} \log(1-D(F_s+G(F_s|f))) \quad (32)$$

As in Figure 3, the generator produces a latent representation showing the variations on the face. The discriminator network has two tasks, i.e., the adversarial task for differentiating between the posed pain expression and the genuine ones and the classification task for estimating the intensity of the facial pain expressions. We learn the parameters of R-GAN in an alternative way to optimize the minimax problem.

Let G_{Θ_g} denotes the generator network with parameters Θ_g. We compute Θ_g by optimizing the loss function L_{dis}, which is a weighted summation of the adversarial loss L_{dis_a} and the classification loss L_{dis_c}.

$$\Theta_g = \arg\min_{\Theta_g} L_{dis}\left(G_{\Theta_g}(F_s)\right) \quad (33)$$

The adversarial part of the discriminator is trained such that it maximizes the probability of assigning the correct label to both the generated magnified representation $G_{\Theta_g}(F_s)$ and the input with noticeable variations F_m. We denote the parameters of the adversarial part of the discriminator D_{Θ_a} as Θ_a and estimate it by solving the following optimization problem:

$$\Theta_a = \arg\min_{\Theta_a} \left\{-\left[\log D_{\Theta_a}(F_m) + \log\left(1-D_{\Theta_a}\left(G_{\Theta_g}(F_s)\right)\right)\right]\right\} \quad (34)$$

Solving the above optimization problem allows the discriminator to distinguish the difference between the generated magnified representation for the subtle variations and the real ones from the genuine pain expressions. To justify the detection performance from the generated magnified representations, the classification part of the discriminator should be first trained using the features of genuine pain to achieve reasonable classification accuracy. Denoting D_{Θ_c} as the classification part of the discriminator, we calculate its parameters Θ_c by optimizing the following loss function:

$$\Theta_c = \arg\min_{\Theta_c} L_{dis_c}(F_m) \quad (35)$$

Taking the generated intermediate representation as input, the discriminator deals with it in two parts, i.e., the adversarial and the classification parts. The adversarial part comprises two fully-connected layers followed by a sigmoid layer that yields in the adversarial loss. The classification part has two fully-connected layers followed by a softmax layer that is used for computation of the classification loss. Suppose the adversarial loss is L_{dis_a} and the classification loss is L_{dis_c}, the discriminator loss is calculated as the sum of both losses, i.e., $L_{dis} = L_{dis_a} + L_{dis_c}$.

Adversarial Loss. Trying to deceive the discriminator with the generated representations, an adversarial loss is used to enforce the generator to produce the magnified representation for the faces with subtle facial variations similar as that of the large variations. The adversarial loss L_{dis_a} is defined as:

$$L_{dis_a} = -\log D_{\Theta_a}\left(G_{\Theta_g}\left(F_s\right)\right) \tag{36}$$

Classification Loss. Receiving the generated representation as input, the classification part of the discriminator network computes the probabilities for different pain intensity levels. The classification loss L_{dis_c} is simply equivalent to the softmax cross entropy loss of the output comparing with the ground-truth.

IV. EXPERIMENTS

To evaluate the performance of the proposed framework, we performed comprehensive experiments on three benchmark and publicly available databases, i.e., the UNBC-McMaster Shoulder Pain Expression Archive [26], the BioVid Heat Pain [27], and the STOIC [28]. First, we analyze the differences between facial expressions of the posed and genuine pain. Then, we extensively evaluate the performance of R-GAN and the video summarization method. Finally, we show that our experimental results are competitive compared to the state-of-the-art.

A. Experimental Data

UNBC-McMaster. The UNBC-McMaster Shoulder Pain Expression Archive [26] is widely used for pain expression analysis. This database contains videos of spontaneous facial expressions of subjects performing a series of active and passive range-of-motion tests to their either affected or unaffected limbs in two sessions. Each video sequence is annotated in a frame-level fashion by FACS, resulting in 16 discrete pain intensity levels based on action units. In our experiments, we used the active test set that has 200 face videos of 25 subjects with 48,398 frames.

BioVid. The BioVid Heat Pain database [27] has been collected from 90 participants from three age groups. Four distinct pain levels have been induced in the right arm of each subject. Bio-physiological signals such as the Skin Conductance Level (SCL), the electrocardiogram (ECG), the electromyography (EMG), and the electroencephalogram (EEG) have been recorded. In our experiments, we only use Parts A and D of this database, which contain spontaneous and posed facial expressions, respectively. Part A includes 8,700 videos of 87 subjects which are labeled with respect to pain stimulus intensity. Part D contains 630 videos of 90 subjects, who show posed facial expressions.

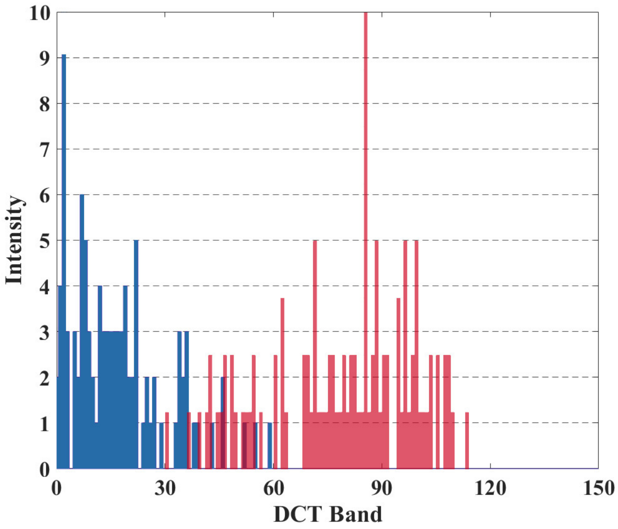

Fig. 4: The frequency analysis of the facial structure in presence of genuine pain (blue) and posed pain (red).

STOIC. The STOIC database [28] includes posed facial expressions of 34 actors from 20 to 45 years old. The database has 80 video sequences showing the peak of facial expressions and the length of the videos have been reduced to 500 ms.

B. Posed Pain vs. Genuine Pain

We analyzed the reaction of the facial structure to the pain stimulus. From the labeled samples, we took discrete cosine transform (DCT) and measured the intensity of each individual frequency band. Figure 4 shows the results of our observations. As can be seen, genuine pain expression mainly has low-frequency components (see the blue histogram in Figure 4). On the other hand, posed pain expression mainly has medium to high-frequency component (see the red histogram in Figure 4). These results validates our initial hypothesis that the facial expression of the posed pain is more exaggerated, as the subject tries to show high intensity of the pain by shrinking their facial components. These changes in the face usually occur quick that increase the intensity of higher DCT bands.

C. Parametric Analysis of Weighted Spatiotemporal Pooling

As described in Section III-A, the proposed WSP divides the video sequence into the fixed-size overlapping segments. The volumetric patch-wise comparison between two consecutive segments allows WSP to perform a spatiotemporal pooling by capturing the appearance and dynamic variation within the video. Using this technique, we can summarize the entire length of the video into a single image with a fixed number of channels. The performance of the WSP depends on two parameters, i.e., the length of segments and the spatial size of patches. In Table I, we evaluated the performance of the proposed method in terms of adversarial accuracy (detecting posed and genuine pain) and the classification accuracy (estimating pain intensity level) by changing the

TABLE I: Comparison of the adversarial accuracy (%) and the classification accuracy (%) of the proposed method versus different length of segments on the BioVid database [27].

	Length of Segment					
	5	10	15	20	25	30
Adversarial	82.53	83.54	85.05	84.33	83.69	82.71
Classification	85.34	90.08	92.51	92.43	91.59	90.34

TABLE II: Comparison of the adversarial accuracy (%) and the classification accuracy (%) of the proposed method versus different spatial size of patches on the BioVid database [27].

	Spatial Size of the Patch				
	3 × 3	5 × 5	7 × 7	9 × 9	11 × 11
Adversarial	82.70	85.05	83.91	79.36	77.21
Classification	90.33	92.51	91.67	88.49	85.90

length of segments. As can be seen, the accuracy improves by increasing the segment's length to 15 frames. However, the performance drops as we increase the length of segments from 15 to 30 frames. These results show that longer segments contain more redundant spatiotemporal information so that WSP cannot capture all the subtle variations in the whole video.

Another important parameter, which affects the performance of WSP, is the spatial size of the patches. Table II compares the performance of the proposed framework by changing the spatial size of patches. Using small windows, the receptive field of WSP is also small. As a result, the spatiotemporal pooling is performed in a tiny region of the video, which ignores most of important relationships between neighboring points. However, choosing a large size for patches' window leads to low accuracy. We argue that this drop in the performance is due to capturing too much information and increasing the complexity of the obtained summarized image. Based on our experiments, the optimal spatial size of patches is 5×5. Moreover, more accurate observation of the both Tables I and II shows that changes in the classification accuracy is lower than the changes in the adversarial accuracy. These results demonstrate that the classification part of the discriminator network is robust to the small changes of the input sample. However, detection of genuine pain expressions is highly correlated to the subtle variations in the face. Therefore, an accurate representation of the appearance and dynamic of the face is crucial for discrimination of the genuine and posed facial expressions.

D. Comparative Analysis

To evaluate to performance of the proposed framework, we compared our proposed method with the state-of-the-art approaches. Table III draws comparisons between our proposed framework with Variational Auto-Encoder (VAE) [29], Fast-RCNN [8], Faster-RCNN [30], GAN [11], and Conditional GAN (CGAN) [31] on three benchmark databases, i.e., UNBC [26], BioVid [27], and STOIC [28]. We conducted two series of experiments: (i) without using WSP, and (ii) using WSP, to analyze the importance of capturing subtle facial variations for differentiating posed pain expression from genuine one. As can be seen from Table III, our proposed R-GAN consistently outperforms other methods by a noticeable margin. However, the results are worse when we did not apply WSP. Although the proposed method achieved a very high accuracy in STOIC database, we assert that it is due to the small size of the database. To the contrary, R-GAN is able to discriminate the posed and genuine pain on larger databases such as UNBC and BioVid.

TABLE III: Comparison of the accuracy (%) of the proposed method and the state-of-the-art approaches on discrimination between the posed and genuine pain expressions.

Method		UNBC	BioVid	STOIC
No WSP	VAE [29]	59.33	48.63	72.35
	Fast-RCNN [8]	66.92	52.97	78.66
	Faster-RCNN [30]	68.42	54.77	80.25
	GAN [11]	76.25	61.25	84.79
	CGAN [31]	79.91	65.79	86.03
	R-GAN	**86.40**	**76.23**	**89.64**
WSP	VAE [29]	68.35	54.75	84.33
	Fast-RCNN [8]	76.95	61.74	86.43
	Faster-RCNN [30]	79.33	62.91	87.95
	GAN [11]	84.69	74.85	90.82
	CGAN [31]	87.51	79.42	92.33
	R-GAN	**91.34**	**85.05**	**96.52**

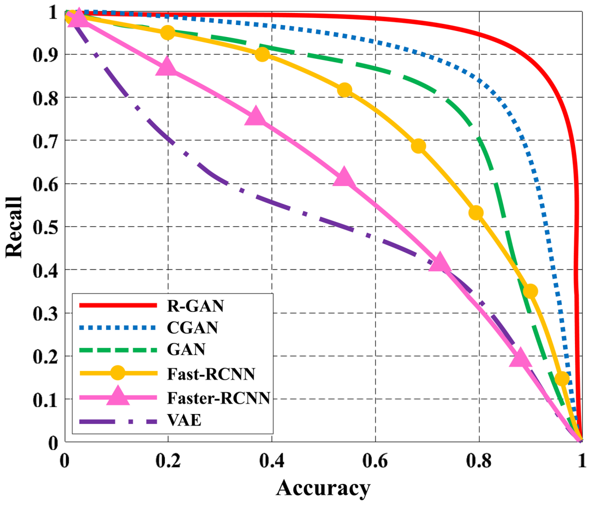

Fig. 5: Comparisons of the overall detection performance on BioVid database [27].

To provide more insights into the performance of the proposed method, Figure 5 depicts the recall-accuracy curve for the aforementioned methods by applying WSP on the BioVid database [27]. This curve further demonstrates the effectiveness of the proposed R-GAN learning capability.

Figure 6 shows some mis-detection examples of the proposed method. Some subjects who experience real pain are classified as actors exhibiting posed pain, while some posed pain expression samples are detected as genuine pain expression. We argue that these mis-detections are due to the failure of the proposed method in modeling inter-person reactions to the pain. As can be seen in Figure 6, some subjects skillfully pretend a painful experience, while some people, who suffer from the real pain, either suppress their feeling or react differently, e.g., by smiling.

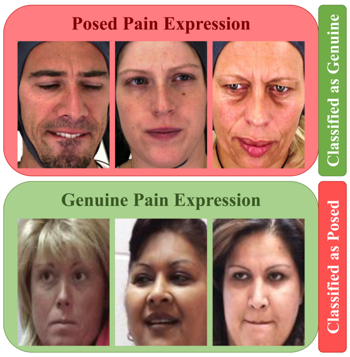

Fig. 6: Examples of mis-classified pain expressions. **Top:** Posed pain expressions classified as genuine ones, **Bottom:** Genuine pain expressions classified as posed ones.

V. CONCLUSION

We presented a novel framework for distinguishing genuine pain from posed pain. We proposed a Weighted Spatiotemporal Pooling (WSP) to capture and summarize the appearance and dynamic of the face in the whole length of the video into one single image map. This strategy allows to use pre-trained models on images for video analysis applications. The WSP captures subtle variations of the facial structure, which are crucial for facial expression analysis. In addition, we introduced Residual Generative Adversarial Network (R-GAN) for discriminating genuine pain from posed one. A residual network is connected to the generator network to enhance the magnification of the subtle facial variations. The discriminator has two parts namely adversarial and classification part. The adversarial network makes a decision between genuine and posed pain expression, while the classification network estimates the intensity of pain level. We showed the effectiveness of our proposed framework on three benchmark, publicly available databases and achieved state-of-the-art performance.

ACKNOWLEDGMENT

The financial support of the Academy of Finland, Infotech Oulu, Nokia Foundation, and Tauno Tönning Foundation is acknowledged.

REFERENCES

[1] M. Hoque, L. P., Morency, and R. W. Picard, *Are You Friendly or Just Polite? Analysis of Smiles in Spontaneous Face-to-Face Interactions*, in Proc. Affect. Comput. and Intell. Interact., pp. 135-144, 2011.

[2] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, *DISFA: A Spontaneous Facial Action Intensity Database*, IEEE Trans. Affect. Comput., vol. 4, no. 2, pp. 151-160, 2013.

[3] A. Gutirrez-Garca and M. G. Calvo, *Discrimination Thresholds for Smiles in Genuine Versus Blended Facial Expressions*, Cogent Psychology, vol. 2, no. 1, 2015.

[4] E. G. Krumhuber and A. S. R. Manstead, *Can Duchenne Smiles be Feigned? New Evidence on Felt and False Smiles* Emotion, vol. 9, no.6, pp. 807-820, 2009. 1

[5] K. L. Schmidt, Z. Ambadar, J. F. Cohn, and L. I. Reed, *Movement Differences Between Deliberate and Spontaneous Facial Expressions: Zygomaticus Major Action in Smiling*, Jour. of Nonverbal Behavior, vol. 30, no. 1, pp. 37-52, 2006. 1

[6] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, 1978. 1

[7] J. F. Cohn and K. L. Schmidt, *The Timing of Facial Motion in Posed and Spontaneous Smiles*, Int. Jour. of Wavelets, Multiresolution and Inf. Proces., vol. 2, no. 2, pp. 121-132, 2004. 1, 2

[8] R. Girshick, *Fast R-CNN*, in Proc. IEEE ICCV, pp. 1440-1448, 2015. 1, 7

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. *SSD: Single Shot mMultibox Detector*, arXiv preprint arXiv:1512.02325, 2015. 1

[10] F. Yang, W. Choi, and Y. Lin, *Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers*, in Proc. IEEE CVPR, pp. 2129-2137, 2016. 1

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative Adversarial Nets*, in Proc. NIPS, pp. 2672-2680, 2014. 1, 7

[12] I. Kotsia, S. Zafeiriou, and I. Pitas, *Texture and Shape Information Fusion for Facial Expression and Facial Action Unit Recognition*, PR, vol. 41, no. 3, pp. 833-851, 2008. 2

[13] Y. Liu, X. Yuan, X. Gong, Z. Xie, F. Fang, and Z. Luo, *Conditional Convolution Neural Network Enhanced Random Forest for Facial Expression Recognition*, PR, vol. 84, pp. 251-261, 2018. 2

[14] M. Demirkus, D. Precup, J. J. Clark, and T. Arbel, *Hierarchical Spatio-Temporal Probabilistic Graphical Model with Multiple Feature Fusion for Binary Facial Attribute Classification in Real-World Face Videos*, IEEE Trans. PAMI, vol. 38, no. 6, pp. 1185-1203, 2016. 2

[15] J. Chen, Z. Chen, Z. Chi, and H. Fu, *Facial Expression Recognition in Video with Multiple Feature Fusion*, IEEE Trans. Affect. Comput., vol. 9, no. 1, pp. 38-50, 2018. 2

[16] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn, *Spontaneous vs. Posed Facial Behavior: Automatic Analysis of Brow Actions*, in Proc. ICMI, pp. 162-170, 2006. 2

[17] S. Park and D. Kim, *Spontaneous Facial Expression Classification with Facial Motion Vectors*, in Pro. IEEE FG, pp. 1-6, 2008. 2

[18] C. F. Bond and B. M. DePaulo, *Accuracy of Deception Judgments*, Personality and Social Psychology Review, vol. 10, no. 3, pp. 214-234, 2006. 2

[19] M. L. Hill and K. D. Craig, *Detecting Deception in Pain Expressions: The Structure of Genuine and Deceptive Facial Displays*, Pain, vol. 98, no. 1, pp. 135-144, 2002. 2

[20] G. C. Littlewort, M. S. Bartlett, and K. Lee, *Automatic Coding of Facial Expressions Displayed During Posed and Genuine Pain*, IVU, vol. 27, no. 12, pp. 1797-1803, 2009. 2, 5

[21] Y. Huang, W. Wang, and L. Wang, *Bidirectional Recurrent Convolutional Networks for Multi-frame Super-resolution*, in Proc. NIPS, pp. 235243, 2015. 2

[22] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, *Video Super-Resolution With Convolutional Neural Networks*, IEEE Trans. Comput. Imag., vol. 2, no. 2, pp. 109-122, 2016. 2

[23] W. Yang, J. Feng, G. Xie, J. Liu, Z. Guo, and S. Yan, ”Video Super-Resolution Based on Spatial-Temporal Recurrent Residual Networks,” CVIU, vol. 168, pp. 79-92, 2018. 2

[24] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, *Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network*, in Proc. IEEE CVPR, pp. 105-114, 2017. 2

[25] S. H. Khatoonabadi, N. Vasconcelos, I. V. Baji, and Y. Shan, *How Many Bits Does It Take for a Stimulus to be Salient?*, in Proc. IEEE CVPR, pp. 5501-5510, 2015. 2

[26] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, *Painful Data: The UNBC-McMaster Shoulder Pain Expression Archive Database*, in Proc. IEEE FG, pp. 57-64, 2011. 6, 7

[27] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. C. Traue, S. Crawcour, P. Werner, A. Al-Hamadi, and A. O. Andrade, *The BioVid Heat Pain Database Data for The Advancement and Systematic Validation of An Automated Pain Recognition System*, in Proc. IEEE Int. Conf. Cybern., pp. 128-131, 2013. 6, 7

[28] S. L. Roy, and C. Roy, C. Éthier-Majcher, I. Fortin, P. Belin, and F. Gosselin, *STOIC : A Database of Dynamic and Static Faces Expressing Highly Recognizable Emotions*, 2009. 6, 7

[29] D. P. Kingma and M. Welling, *Auto-Encoding Variational Bayes*, In Proc. ICLR, 2014. 7

[30] S. Ren, K. He, R. Girshick, and J. Sun, *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*, IEEE Trans. PAMI vol. 39, no. 6, pp. 1137-1149, 2017. 7

[31] M. Mirza and S. Osindero, *Conditional Generative Adversarial Nets*, CoRR, abs/1411.1784, 2014. 7