Open Set Classification of GAN-based Image Manipulations via a ViT-based Hybrid Architecture

Jun Wang Omran Alamayreh Benedetta Tondi Mauro Barni
Department of Information Engineering and Mathematics, University of Siena
j.wang@student.unisi.it, omran@diism.unisi.it benedetta.tondi@unisi.it, barni@dii.unisi.it

Abstract

Classification of AI-manipulated content is receiving great attention, for distinguishing different types of manipulations. Most of the methods developed so far fail in the open-set scenario, that is when the algorithm used for the manipulation is not represented by the training set. In this paper, we focus on the classification of synthetic face generation and manipulation in open-set scenarios, and propose a method for classification with a rejection option. The proposed method combines the use of Vision Transformers (ViT) with a hybrid approach for simultaneous classification and localization. Feature map correlation is exploited by the ViT module, while a localization branch is employed as an attention mechanism to force the model to learn per-class discriminative features associated with the forgery when the manipulation is performed locally in the image. Rejection is performed by considering several strategies and analyzing the model output layers. The effectiveness of the proposed method is assessed for the task of classification of facial attribute editing and GAN attribution.

1. Introduction

Synthetic manipulation of face images has become ubiquitous and is being increasingly used in a wide variety of applications [15], thus posing a serious threat to public trust. Many detectors have been proposed to classify images forged by generative models as fakes/synthetic [14]. There are cases where just knowing that the image is fake is not enough and more information is required on the synthetic manipulation undergone by the image. This is the case, for instance, when the synthetic manipulation[1] consists of local attribute editing, rather than in the generation of a synthetic image from from scratch [26, 30], in which case it is preferable to also provide evidence to support the judgment that the image is fake, rather than simply saying that the image has been manipulated. Several methods addressed this problem via binary detectors, judging the image as real or fake, which also have the ability to localize the manipulation, e.g. outputting binary localization masks or attention maps [18, 23, 44], or via multi-class classifiers [34], that classify the type of facial attribute editing performed by generators. In yet some other cases there is interest in knowing the specific type of architecture used to generate the manipulation (synthetic image attribution), Methods have been proposed that perform attribution via multi-class classifiers by relying on artifacts or signatures (fingerprints) left by the models in the generated images [22, 39].

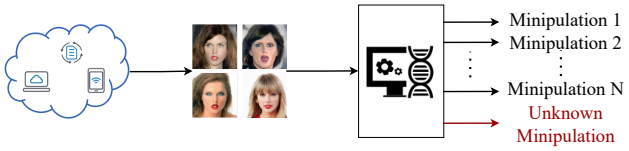

Figure 1. Open set scenario of synthetic manipulation classification (classification with rejection option) considered in this paper.

A common drawback with all the above binary and multi-class classification approaches is that their application is limited to closed-set scenarios. For instance, Generative Adversarial Network (GAN) attribution methods can correctly attribute the image only if it comes from a GAN architecture among those seen during training, and they are incapable of identifying or revealing unseen GAN types. This seriously limits the applicability of these methods in real-world settings, where the images seen during operation time may be edited in different ways or generated by architectures not seen during training, with the consequence that the predictions made by methods are not trustable.

In this paper, we address the problem of classification of synthetic facial manipulation in open set scenarios, proposing a method for classification with rejection option. To build the closed-set classifier, the method combines a Vision Transformer (ViT) with a hybrid approach for simultaneous classification and localization, inspired to [33, 34]. Then,

[1] In the following, we generically refer to the case of local GAN manipulation and generation from scratch with the term synthetic manipulation.

a dedicated module performs rejection/acceptance of the sample by analyzing the model output layers. Rejection is performed by considering the maximum logit score (MLS), maximum softmax probability (MSP), and the OpenMax approach. More specifically, in our architecture, the input sequence to the ViT is formed from feature maps extracted from a CNN ([7, 36]), and then the ViT module is used to exploit feature maps correlation, via the self-attention mechanism. In the general case of local manipulation, the same features shared by Vit-based classification heads are also utilized by a localization branch via a fully convolution network. The goal of the localization branch is to force the network to focus on the most significant parts of the image (attention mechanism) [33].The overall architecture, which includes the feature extraction network, localization branch and ViT module, is then trained in an end-to-end fashion. Our method follows some recent works in machine learning, showing that ViT allows achieving improved performance for out-of-distribution detection [9] and open set recognition [1], compared to standard CNN architectures.

Experiments are carried out considering the classification of facial attribute editing and GAN attribution. For the classification of synthetic facial attributes, we considered 19 editing types, with manipulations performed by InterfaceGAN [30] and StyleCLIP [26]. For GAN attribution, the performance is assessed considering facial images generated by several modern generative models, namely, LGSM [31], StyleGAN2 [20], StyleGAN3 [19], Taming transformer [8] and Latent Diffusion [28]. For both tasks, experiments were performed considering different combinations of in-set and out-of-set manipulations. The results confirm that the proposed architecture, and in particular the use of ViT, is beneficial and allows to significantly improve open set performance without impairing the accuracy on closed-set samples, outperforming state-of-the-art methods in the literature for open set recognition (OSR).

The rest of the paper is organized as follows. Section 2 introduces the related work on the generation and detection/classification of synthetic faces, and the most relevant methods for open-set classification in machine learning. The proposed architecture is presented in Section 3. Section 4 describes the experimental methodology and setting. The results and the comparisons with the state-of-the-art are finally reported and discussed in Section 5. Finally, we draw conclusions in section 6.

2. Related work

2.1. AI-synthesized faces and their detection

Artificial Intelligence (AI)-synthesized faces can be either fully synthetic when the faces are generated from scratch using generative models, or locally manipulated, e.g. when a single facial attribute or multiple attributes are modified by the model while the other attributes remain unchanged.

A wide variety of generative models, notably GANs [19, 20] and diffusion models [28, 31], are nowadays able to generate high-resolution images from scratch with an unprecedented level of realism. Inspired by the superior performance of the StyleGAN series [20] in synthesizing high-resolution and high-quality images, StyleGAN architectures have been adopted for image editing, achieving high-quality edited images. Among them, we mention InterFaceGAN [30] and StyleCLIP [27]. InterFaceGAN proposes a framework to interpret the disentangled face representation learned by the StyleGAN model and studies the properties of the facial semantics encoded in the latent space, showing that it is possible to edit the semantic attribute through linear subspace projection. StyleCLIP is a text-based interface for StyleGAN-based image manipulation. StyleCLIP mainly uses the Contrastive Language-Image Pre-training (CLIP) model to edit the latent code through the user input language description, so as to achieve the purpose of editing the image. With regard to the defences, several methods have been developed in the last years to discriminate between fake and real [35–37]. Attempts have also been made to attribute the GAN model and/or the type of architecture generating the image. In many cases, model-level attribution is performed by relying on the estimation of a fingerprint characterizing the GAN model, e.g. [22, 42, 43]. Other works address the task of architecture attribution, attributing the fake image to the source architecture instead of the specific model, see for instance [39], where a multi-class classifier is proposed to discriminate among different architectures in a closed-set setting, even when the images are generated by using different initialization, loss and dataset, hence possessing distinct model-level fingerprints. Recently, a method for the classification of the synthetic face editing performed by the GAN has also been proposed in [34]. The method relies on a patch-driven hybrid classification network with localization supervision, that classifies the editing among a pool of possible manipulations (closed-set setting), with good robustness against post-processing. As a drawback of this approach, the pre-training of the patch-based models is time-consuming.

2.2. Open Set Recognition

Open set recognition (OSR), first formalized in [29] for classical machine learning, addresses the problem of determining whether an input belongs to one of the classes used to train a network. Such a problem has received increasing attention, especially in the last years [11]. A method to address OSR with deep neural network-based approaches, named OpenMax, was presented in [2]. An extra class is added for the prediction, to model the unknown class case. OpenMax adapts meta-recognition concepts to the activa-

tion patterns in the penultimate layer of the network for unknown modelling. The Extreme Value Theory (EVT) is used to estimate the probability of the input being an outlier. Several works have shown that in many cases easy strategies that look at the softmax probability or the logits can also effectively judge if the sample comes from an unknown class [10], e.g. exploiting the fact that the maximum output score tends to be smaller for inputs from unknown classes (out-of-set) [6, 32], or that the energy of the logit vector tends to be lower for out-of-set samples [21].

Other approaches explored reconstruction errors obtained via autoencoders for open set rejection [24, 25, 40]. In [5], Yang et al. designed a suitable embedding space for open set recognition using convolutional prototype learning, that abandons softmax, and implements classification by finding the nearest prototype in the Euclidean norm in the feature space (GCPL). In [38], a novel learning framework for OSR, called reciprocal point learning (RPL), is proposed. The method is extended in [4] (ARPL) via an adversarial mechanism that generates confusing training samples, to enhance the distinguishability of known and unknown classes. Recently, in [17], a method that combines autoencoders with, respectively, prototype learning (PCSSR) and reciprocal point learning (RCSSR) has been proposed.

To the best of our knowledge, in the literature pertaining to synthetic manipulation detection, all the methods proposed so far are limited to the closed-set scenario. An exception is represented by [12], where an algorithm to discover new GANs from a given unlabeled set and cluster them is proposed. More specifically, in the method in [12], unseen classes are considered during network training as a unique -unlabeled - class (discovery set). The images belonging to this class are then clustered using the learned features, attributing them to new labels.

2.2.1 Vision Transformers for OSR

Transformers were originally proposed for natural language processing (NLP). Recently, they have been successfully exploited in computer vision tasks with great results [16]. The extension of transformers to the image domain, namely, Vision Transformers (ViT) [7], are self-attention architectures that process the image as a sequence of image patches, that are treated the same way as tokens (words) in the NLP case. Following the paradigm of the ViT architecture in [7], a series of variants of the original structures have been proposed to further improve the performance on image tasks.

Recent literature on machine learning has shown that ViT can achieve very good performance for out-of-distribution detection in image classification tasks [9], outperforming standard CNNs. Among the works exploiting ViT for open set recognition, we mention [1,3]. The method in [1], in particular, combines ViT with energy-based rejection for open set scene classification in remote sensing imagery. Following the above literature, in this paper, we propose to exploit ViT for open-set classification of synthetic image manipulation. To the best of our knowledge, this is the first attempt in this direction.

3. Proposed method

The general problem of open set classification of synthetic manipulation addressed in this paper is illustrated in Figure 1. A forensic classifier with rejection option classifies the type of synthetic manipulation, among those in a kwown set, at the same time being capable to reject unknown samples, namely, samples that were subject to a different manipulation or generation procedure with respect to those in the known set.

As we said, we focus on the case of facial manipulations. Formally, given a synthesized face image $x \in \mathbb{R}^{H\times W\times 3}$ (height = H, weight = W), the system assigns to x a label y and a mask M associated to the manipulation. Given the set with the N known manipulations considered during training, the predicted label may take $N+1$ values, where the $N+1$-th value identifies the rejection class (unknown manipulation). If we let $\hat{y}$ denote the output (predicted label) of the N-class classification network, the final classification function $\phi(x)$ takes the following expression: $\phi(x) = \hat{y}$ if x is accepted as an in-set sample, or $\phi(x) = N+1$ otherwise. In the following, we denote with $p \in \mathbb{R}^N$ is the probability vector (after softmax) of the network associated with the N classes in the closed-set. The localization mask M is used to indicate the pixels where the image has been manipulated (pixels for which $M = 1$ indicate the manipulated areas). More in general, a localization mask may just highlight regions of interest (like an attention mask), without necessarily corresponding to a manipulation mask.

3.1. Proposed ViT-based Hybrid Architecture

The general scheme of the proposed method is shown in Figure 2. The network is composed of two branches for classification and localization, respectively. A ResNet50 network is used as the backbone for feature extraction. Following [13,35], a modification of the original ResNet architecture is considered, where we remove the sampling operation in the first convolutional layer of the network, setting the stride parameter to 1, with the kernel size fixed to 3. The features are then input to a transformer-based module performing the N-class classification and to a fully convolutional network (FCN) head for the localization, as detailed below. Hence, in our scheme, the input sequence to the ViT is formed from feature maps of the CNN, as an alternative to raw image patches [7].

ViT-based classification module. Let f_r denote the vector of extracted features. We indicate with (H_f, W_f) the

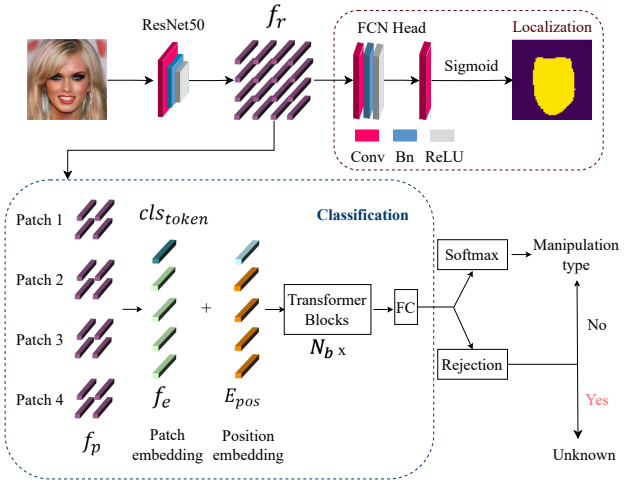

Figure 2. Overall architecture of the proposed method.

size of the feature maps, and with D_f the number of the channels/maps. Then, $f_r \in \mathbb{R}^{H_f \times W_f \times D_f}$. The following preprocessing is applied before feeding the ViT module. For a given patch size P, f_r is first reshaped into a sequence of $P \times P \times D_f$ patches, in number $N_p = H_f W_f / P^2$. The special case $P = 1$ corresponds to the case when the input sequence is obtained by simply flattening the spatial dimensions of the feature map and projecting to the transformer dimension. The input sequence obtained after this flattening layers is $f_p \in \mathbb{R}^{N_p \times \left(P^2 D_f\right)}$. Following the general procedure with ViT, patch embedding is performed via mapping to D_p dimensions via linear projection. $f_e = f_p \cdot E_p$ is the output, of shape of $N_p \times D_p$, obtained after the patch embedding operation, where E_P denotes the embedding matrix, $E_p \in \mathbb{R}^{\left(P^2 D_f\right) \times D_p}$. A placeholder data structure cls_{token}, used to store information that is extracted from other tokens in the sequence f_e, is prepended to the beginning of input sequence f_e (randomly initialized). Position embeddings $E_{pos} \in \mathbb{R}^{(N_p+1) \times D_p}$ are added to the patch embeddings to retain positional information, thus getting the sequence of vectors $\{cls_{token}, f_e\} + E_{pos}$, that is then fed to a standard transformation encoder, like those used in NLP. The transformer encoder is composed of N_b identical transformer blocks, each one constructed of alternating layers of multi-headed self-attention (MHA) and multi-layer perception (MLP) blocks, with a layernorm (LN) applied before every block, followed by residual connections after every block, see [7] for more details. Finally, a fully connected layer is attached to the transformer encoder and outputs the predicted probability vector p for the N enclosed classes.

FCN localization module. The extracted features f_r are also input to an FCN computing the estimated manipulation mask M. Such FCN consists of two convolutional layers, a batch normalization layer, a ReLU layer and finally a sigmoid layer to map the values in the $[0, 1]$ range. As we mentioned, the main reason for the introduction of the localization branch is to guide the classification and help force the network to focus on the most significant parts of the image, in the case of local manipulation, that has been shown to have a beneficial effect on the classification accuracy and generalization capabilities [33].

The overall N-class classification architecture is trained end-to-end by minimizing a combination of the cross-entropy (CE) loss, associated with the classification task, and the mean squared error (MSE) of localization, respectively. Formally, $loss_{hyb} = \lambda_{cls} \cdot \text{CE}(y, p) + \lambda_{loc} \cdot \text{MSE}(G, M)$, where G denotes the ground truth localization mask and λ_{cls} and λ_{loc} balance the trade-off between localization and classification tasks.

The impact of each part of the proposed architecture, and in particular, the localization branch and the ViT module, is assessed in the experiments. For the ViT, we considered $N_b = 4$ transformer blocks. Different patch sizes P of the ViT were considered in our experiments.

3.2. Out-of-Set Rejection

In order to detect samples whose manipulations do not belong to the knowm set, we considered three rejection strategies, two of which perform the rejection by analyzing the model output after or before the softmax activation layer, namely maximum softmax probability (MSP) [6] and maximum logits score (MLS) [32], respectively, and OpenMax [2].

When MSP and MLS approaches are adopted, lower scores associated with the predicted class reflect the uncertainty of the network prediction, providing evidence of unknown classes (out-of-set). Then, the final output of our open set classifier is obtained as follows

$$\phi(x) = \begin{cases} \hat{y}, & \text{if } \arg\max(h) > th \\ N+1, & \text{otherwise} \end{cases} \quad (1)$$

where h is the model output (the softmax probability in the MSP, the logit scores in the MLS), and th is a predefined threshold. When the OpenMax is adopted, then the output of the closed-set classifier is accepted ($\phi(x) = \hat{y}$) if $p_o < th'$, where p_o is the probability of the sample being an outlier, estimated by the method, and th' is the decision threshold. Otherwise, it is rejected ($\phi(x) = N + 1$).

4. Experimental Setup

4.1. Datasets

GAN editing dataset. To build the dataset, we first use the PTI inversion method to reconstruct the images and extract the latent code. Image attributes are manipulated by InterfaceGAN [30] and StyleCLIP [26]. We selected 5,992

Table 1. Summary of the 19 editing classes (18 + 'None').

Editing tools	Edit types
PTI	T0: None (Reconstructed)
InterfaceGAN	**Expression** (T1-T2): Smile, Not smile, **Aging** (T3,T4): Old, Young
StyleCLIP	**Expression** (T5, T6): Angry, Surprised **Hairstyle** (T7-T12): Afro, Purple_hair, Curly_hair, Mohawk, Bobcut, Bowlcut **Identity change** (T13-T18): Taylor_swift, Beyonce, Hilary_clinton, Trump, Zuckerberg, Depp

images from CelebAHQ dataset and each image is edited with 18 edit types: 4 facial attributes are edited with InterfaceGAN, and 14 facial attributes with StyleCLIP. The 'None' type corresponds to the case of the image reconstructed with no editing (obtained via the PTI inversion method). An overview of our dataset is provided in Table 1. We exploited a pre-trained face parsing model [41] to group the various edited attributes into four categories: expression, aging, hairstyle, identity change. We rely on these categories to construct the localization masks used for training. Figure 3 shows an example of manipulated face image for each edited attribute.

GAN attribution dataset. To build the GAN attribution dataset used in our experiments we considered five GAN architectures: StyleGAN2 [20], StyleGAN3 [19], Taming Transformer [8], Latent Diffusion [28] and LSGM [31]. For each architecture, we considered 50k images. The models were trained on the FFHQ/FFHQU dataset. In all the cases, we used pre-trained models released by the authors. For LSGM, Taming transformers and Latent diffusion models, the resolutions of the images are 256×256, while for StyleGAN models the images are generated with both 256×256 and 1024×1024 resolution.

4.2. Experimental setting

To train our model for GAN face editing classification, the dataset of real images is split as follows: 4400 images are used to generate the editing used for training, 1592 for those used for testing, for a total of 83600 (4400×11) images for training and 30248 (1592×19) for testing. Cross-validation is implemented during training by randomly splitting the training set in 4000×11 images used for training and 400×11 images used for validation, every 10 epochs. Training is performed via Adam optimizer with learning rate $lr = 10^{-5}$ and batch size $bs = 32$ for 100 epochs. The input size is set to $256 \times 256 \times 3$. We ran comparison with the state-of-the-art methods in the field of OSR, i.e., GCPL [5], RPL [38], ARPL [4], CAC [24], PC-SSR and RCSSR [17], mentioned in Section 2.2. All these methods are trained using the code released by the authors

Figure 3. Examples of edited images by InterfaceGAN [30] (first row) and StyleCLIP [26] (second and third rows).

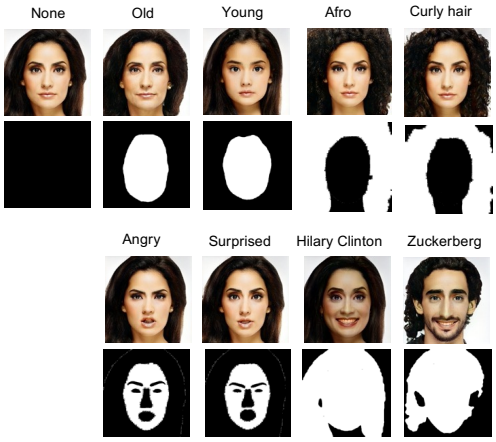

Figure 4. Examples of images and masks obtained with different editing from each category. From left to right: 'None', aging (2), hairstyle (2), expression (2), identity change (2).

on our dataset with default setting and input size 224×224.

As for GAN attribution, for each architecture, the images are split into 35000:5000:10000 for training, validation and testing, respectively. Training is carried out using the same optimizer, learning rate and batch size as above, for 50 epochs. Performance in the closed set is evaluated by measuring the classification accuracy, while the AUC of the ROC curve obtained by varying the thresholds th (th' for OpenMax) is measured to evaluate the rejection performance in open set.

In the case of the classification of GAN face editing, the manipulation is performed locally and the localization branch is employed to guide the training. The ground truth localization masks highlighting the regions of interest in the images are used to train the model and are obtained as detailed in the following. We decided to use different masks for every category (expression, aging, hairstyle, identity change). We focus on the whole face area for aging editing while we consider the hair region for hairstyle editing. For identity editing, the focus area covers the whole face and hair since both of them are relevant in the characterization

Table 2. Splitting of editing types considered in the experiments.

Groups	In-set	Out-of-set
G0	T0, T2, T3, T5, T6, T7 T8, T9, T13, T14, T15	T1, T4, T10, T11 T12, T16, T17, T18
G1	T0, T1, T2, T5, T6, T13 T14, T15, T16, T17, T18	T4, T3, T7, T8 T9, T10, T11, T12
G2	T0, T1, T2, T5, T6, T7 T8, T9, T10, T11, T12	T4, T3, T13, T14 T15, T16, T17, T18
G3	T0, T1, T2, T3, T4, T11 T12, T13, T14, T15, T18	T5, T6, T7, T8 T9, T10, T16, T17
G4	T0, T1, T3, T4, T6, T10 T12, T15, T16, T17, T18	T2, T5, T7, T8 T9, T11, T13, T14

Table 3. Performance in closed-set and open-set, using different rejection strategies, for different configurations (G0-G9). The AUC is reported in the open-set.

Config		G0	G1	G2	G3	G4
Closed-set	Accuracy	88.99	94.68	87.03	94.34	95.25
Open-set	MSP	79.35	79.63	71.49	84.54	83.97
	OpenMax	81.83	81.89	**81.39**	74.86	81.34
	MLS	**85.34**	**91.36**	78.34	**91.98**	**89.75**
Config		G5	G6	G7	G8	G9
Closed-set	Accuracy	92.65	95.51	89.24	94.94	95.94
Open-set	MSP	82.29	87.29	75.50	84.49	83.30
	OpenMax	78.62	86.20	**83.72**	85.00	83.73
	MLS	**88.05**	**95.23**	82.43	**93.13**	**91.77**

of identity. Finally, for expression editing, the profiles of the mouth, eyes, eyebrows and nose are enhanced in the masks by removing the corresponding segmented regions, being highly related to expressions. Some examples of masks are shown in Figure 4.

5. Results

In this section, we report and discuss the results we got for the tasks of classification of GAN face editing and GAN attribution. Most of the experiments, in particular, the comparison with general state-of-the-art methods for OSR in machine learning, as well as an ablation study on the impact of the various elements of the proposed architecture and parameters, are reported for the former case. This is the case, in fact, where all the components of the proposed ViT-based hybrid network are considered, including the localization branch.

5.1. GAN face editing classification

Experiments were carried out considering 10 different configurations of in-set and out-of-set editing types, referred to as G0-G9. In each case, 11 editing types are considered as in-set classes, while the remaining 8 are taken out-of-set. Table 2 reports G0-G4 configurations, with the 'None' class always included as in-set. Configurations G5-G9 are obtained from G0-G4 by switching the first in-set and out-of-set type, hence with the 'None' class in the out-of-set.

Table 3 reports the closed-set accuracy and the open-set performance achieved with the 3 rejection strategies, for the various configurations. The average accuracy of the classification on the $N = 11$ closed-set classes in the closed-set is 92.86. Regarding the open-set performance, the MLS is the strategy that gives the best results. In particular, with MLS we got AUC = 88.74 on average, in contrast to 81.19 and 81.86 for MSP and OpenMax respectively. Notably, the configurations for which the best closed-set performance is achieved correspond to those, that perform better in the open-set scenario. Therefore, in the following, results are reported for the MLS strategy, unless stated otherwise.

In Figure 5, we report an example of predicted masks in the various cases, for the G0 configuration, for visual assessment. Although the localization has been considered only to supervise the training, like an attention mechanism, and not for localization purposes, by looking at the figure, we can observe that in many cases the method is able to produce similar masks, namely masks with a similar white region (focus area), in both closed-set and open-set images, for editing types belonging to the same category (see Table 1). This indicates that the network tends to look at areas of the image that are most relevant for the discrimination of the manipulation, and also for open-set inputs.

The comparison of the proposed method with state-of-the-art algorithms is reported in Table 4, for the configurations G0, G3 and G4. We see that the proposed method achieves the best results in all the cases in both closed and open-set. In particular, our ViT-based hybrid algorithm gets an AUC of 85.34, 91.98 and 89.75 in G0, G3 and G4, respectively, getting an improvement with respect to the best-performing method from the state-of-the-art always larger than 4% in both Accuracy and AUC. It is worth observing that all these methods have been proposed to address general problems of OSR in deep learning, and adopted for standard image classification tasks and object recognition, e.g. MNIST or CIFAR classification. Hence, they are not designed for forensic problems and in particular manipulation classification tasks, where the classification often relies on the analysis of subtle traces, and the goal in the open set scenario is being able to reveal unseen alterations of similar content or the presence of different fingerprints.

5.1.1 Ablation Study

We conducted an ablation study to investigate the effects of the patch size P used in the ViT module and to validate the

Table 4. Comparison with state-of-the-art method. Results are reported for the G0, G3 and G4 configurations.

Methods	G0		G3		G4	
	Closed-set (Accuracy)	Open-set (AUC)	Closed-set (Accuracy)	Open-set (AUC)	Closed-set (Accuracy)	Open-set (AUC)
GCPL [5]	73.72	73.25	40.93	69.46	43.16	65.48
RPL [38]	74.43	76.21	70.19	81.46	65.76	71.18
ARPL [4]	82.64	81.73	87.80	84.93	90.7	79.89
CAC [24]	77.86	74.95	83.33	78.57	85.09	77.63
PCSSR [17]	84.10	74.49	90.79	85.42	92.25	83.63
RCSSR [17]	83.70	72.95	90.60	86.87	91.67	85.32
Ours	**88.99**	**85.34**	**94.34**	**91.98**	**95.25**	**89.75**

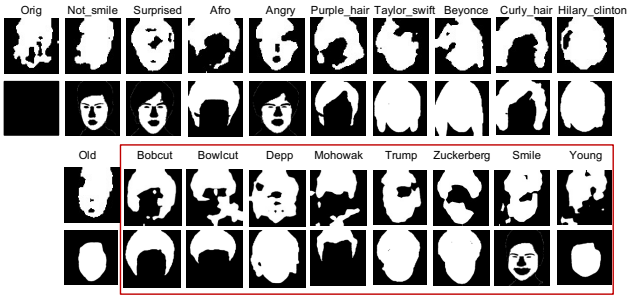

Figure 5. Example of localization masks for the 18 editing types. Predicted (top) and ground truth (bottom) masks are visualized. The masks in the red box refer to the out-of-set editing types.

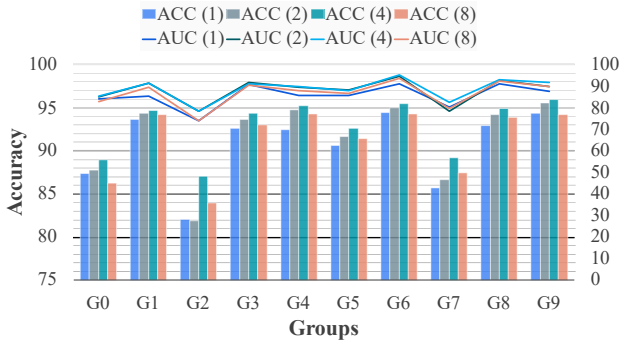

Figure 6. Ablation study on the impact of patch size P of ViT under the various configurations. Vertical bars show closed-set Accuracy, while the line plots show the AUC for open-set.

effectiveness of each component of the proposed architecture.

Impact of different patch sizes. Figure 6 shows the results using different patch sizes P in the ViT module, namely $P = 1$, 2, 4 and 8 (the legends reports the P setting among brackets). We see that increasing the patch size, up to $P = 4$, is beneficial for both closed-set and open-set performance. However, when the patch size increases further, namely, above 4, results do not improve, and actually a performance drop is observed (of around 1.6% in Accuracy and 2% in AUC on the average). Then, from our experiments, with P =4 the ViT achieves the best trade-off between the exploitation of the spatial and of the feature maps correlation.

Impact of different architectures Figure 7 compares the results achieved by the proposed architecture including the ViT module for the classification and the localization branch (FCN), with those achieved by the same method by removing the FCN, and those of the baseline ResNet50, where the standard ResNet50 is used for the multi-class classification. In this case, the rejection is performed in a similar way, by analyzing the output layer of the last FC of ResNet50, before the softmax (MLS). A significant performance gain is obtained with the proposed method in all the configurations. In particular, combining the use of ViT for processing the feature maps with the hybrid approach we got a gain in performance of up to 10% in Accuracy and 9% in AUC.

5.2. GAN attribution

In this section, we report the results we got for open-set GAN attribution. By focusing on fully synthetic images, we do not include the localization branch in the proposed architecture, but only the ViT module. Experiments are carried out considering 4 different splittings of the 5 architectures. The in-set and out-of-set architectures for each configuration are detailed in the following: S1) in-set: LSGM, StyleGAN2 and Taming transformer; out-of-set: StyleGAN3 and Latent diffusion; S2) in-set: StyleGAN2, StyleGAN3 and Latent diffusion; out-of-set: LSGM and Taming transformer; S3) in-set: LSGM, StyleGAN2 and StyleGAN3; out-of-set: Taming transformer and Latent diffusion; S4) in-set: LSGM, StyleGAN2 and Latent diffusion; out-of-set: StyleGAN3 and Taming transformer.

Table 5 shows the closed-set and open-set performance achieved by the proposed architecture (*ResNet50+ViT*) in all the configurations. The results of the baseline are also reported (*ResNet50*). We see that the advantage we got with

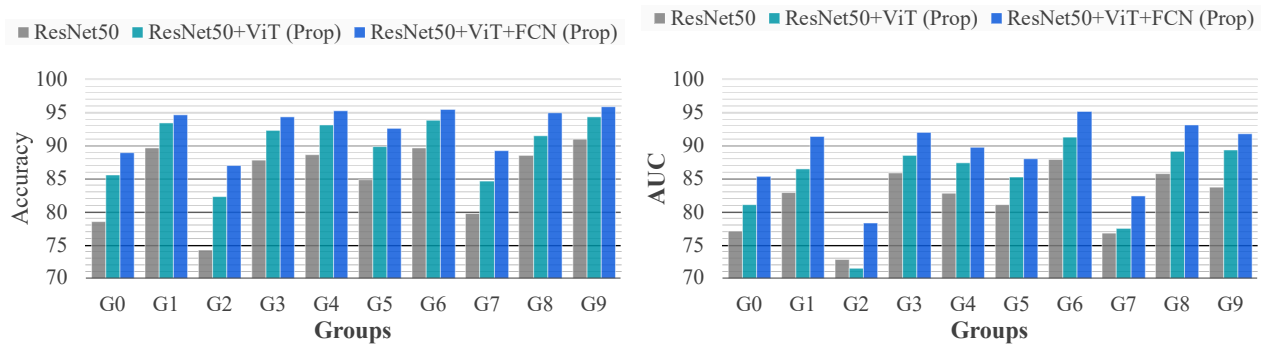

Figure 7. Performance in closed-set (left) and open-set (right) for different configurations (G0-G9).

Table 5. Results on GAN attribution task.

Config	Method	Closed-set (Accuracy)	Open-set (AUC)		
			MSP	OpenMax	MLS
S1	ResNet50	97.76	77.23	64.60	76.32
	ResNet50+ViT (prop)	**99.86**	**92.73**	**92.70**	**92.72**
S2	ResNet50	78.26	39.90	33.40	43.40
	ResNet50+ViT (prop)	**82.38**	**72.49**	**65.02**	**70.39**
S3	ResNet50	92.80	78.78	57.83	69.82
	ResNet50+ViT (prop)	**98.56**	**82.74**	**78.13**	**83.31**
S4	ResNet50	81.82	67.43	61.66	69.98
	ResNet50+ViT (prop)	**94.61**	**90.31**	**93.56**	**93.60**

respect to the baseline is even bigger in this case than that in the previous case of face editing classification. In particular, when the rejection strategies are mounted on top of the baseline architecture, that is, by considering the features extracted by a standard ResNet50 classifier for the analysis, the rejection performance is very poor, with an AUC lower than 70% in most cases. Our method instead can achieve a much higher AUC, that goes above 90 % for S1 and S4. Under the S1 and S2 configurations, the results are worse. We observe that these configurations include both StyleGAN2 and 3 in the training set, hence resulting in a lower diversity of the in-set dataset, that might be the reason for the worse generalization capability to open-set scenarios. Finally, we observe that, as before, the MLS is the strategy that gives the best performance on average, even if in this case the 3 rejection strategies work very similarly. These results confirm that the features extracted with our architecture are representative and allow a good characterization of the various architectures, yielding good discrimination also in the open-set scenario.

6. Conclusion

We have presented a method to address the problem of open-set classification of synthetic manipulations. A multi-class classifier with rejection option is implemented, that classifies the manipulation, at the same time being capable to reject an unknown (unseen) manipulation. To address this task, we resort to a ViT-based hybrid architecture that explores global attention from patches while being guided by manipulation localization. Rejection is performed via several approaches, that rely on the analysis of the output logits and scores, and on outlier probability estimation. Experiments demonstrate the effectiveness of the proposed method, also compared to other state-of-the-art methods for open-set classification, for the task of classification of GAN face editing and GAN attribution.

Future works will focus on the application of the proposed architecture to different synthetic manipulation classification tasks, considering different image contents, beyond faces. The robustness of the proposed method against post-processing and attacks is also worth investigating. Moreover, the promising results achieved for GAN attribution encourage us to explore further the use of the proposed architecture for this task.

Acknowledgement

This work has been partially supported by the China Scholarship Council (CSC), file No. 202008370186, by the PREMIER project under contract PRIN 2017 2017Z595XS-001, funded by the Italian Ministry of University and Research, and by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under agreement number FA8750-20-2-1004. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

References

[1] Reham Al-Dayil, Yakoub Bazi, and Naif Alajlan. Open-set classification in remote sensing imagery with energy-based vision transformer. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 2211–2214. IEEE, 2022. 2, 3

[2] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016. 2, 4

[3] Feiyang Cai, Zhenkai Zhang, Jie Liu, and Xenofon Koutsoukos. Open set recognition using vision transformer with an additional detection head. *arXiv preprint arXiv:2203.08441*, 2022. 3

[4] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081, 2021. 3, 5, 7

[5] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 507–522. Springer, 2020. 3, 5, 7

[6] C Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970. 3, 4

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*. 2, 3, 4

[8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2, 5

[9] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021. 2, 3

[10] G Gavarini, D Stucchi, A Ruospo, G Boracchi, and E Sanchez. Open-set recognition: an inexpensive strategy to increase dnn reliability. In *2022 IEEE 28th International Symposium on On-Line Testing and Robust System Design (IOLTS)*, pages 1–7. IEEE, 2022. 3

[11] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020. 2

[12] Sharath Girish, Saksham Suri, Sai Saketh Rambhatla, and Abhinav Shrivastava. Towards discovery and attribution of open-world gan generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14094–14103, 2021. 3

[13] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. In *2021 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2021. 3

[14] Luca Guarnera, Oliver Giudice, Francesco Guarnera, Alessandro Ortis, Giovanni Puglisi, Antonino Paratore, Linh MQ Bui, Marco Fontani, Davide Alessandro Coccomini, Roberto Caldelli, et al. The face deepfake detection challenge. *Journal of Imaging*, 8(10):263, 2022. 1

[15] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE transactions on knowledge and data engineering*, 2021. 1

[16] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022. 3

[17] Hongzhi Huang, Yu Wang, Qinghua Hu, and Ming-Ming Cheng. Class-specific semantic reconstruction for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3, 5, 7

[18] Yihao Huang, Felix Juefei-Xu, Qing Guo, Yang Liu, and Geguang Pu. Fakelocator: Robust localization of gan-based face manipulations. *IEEE Transactions on Information Forensics and Security*, 2022. 1

[19] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 2, 5

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2, 5

[21] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020. 3

[22] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 506–511. IEEE, 2019. 1, 2

[23] Ghazal Mazaheri and Amit K Roy-Chowdhury. Detection and localization of facial expression manipulations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1035–1045, 2022. 1

[24] Dimity Miller, Niko Sunderhauf, Michael Milford, and Feras Dayoub. Class anchor clustering: A loss for distance-based open set recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3570–3578, 2021. 3, 5, 7

[25] Poojan Oza and Vishal M Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2307–2316, 2019. 3

[26] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 1, 2, 4, 5

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 5

[29] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013. 2

[30] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020. 1, 2, 4, 5

[31] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021. 2, 5

[32] S Vaze, K Han, A Vedaldi, and A Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *International Conference on Learning Representations (ICLR)*, 2022. 3, 4

[33] Jun Wang, Omran Alamayreh, Benedetta Tondi, and Mauro Barni. An architecture for the detection of gan-generated flood images with localization capabilities. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE, 2022. 1, 2, 4

[34] Jun Wang, Benedetta Tondi, and Mauro Barni. Classification of synthetic facial attributes by means of hybrid classification/localization patch-based analysis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1, 2

[35] J Wang, B Tondi, and M Barni. An eyes-based siamese neural network for the detection of gan-generated face images. *Front. Sig. Proc. 2: 918725. doi: 10.3389/frsip*, 2022. 2, 3

[36] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 615–623, 2022. 2

[37] Xin Wang, Hui Guo, Shu Hu, Ming-Ching Chang, and Siwei Lyu. Gan-generated faces detection: A survey and new perspectives. *arXiv preprint arXiv:2202.07145*, 2022. 2

[38] Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Robust classification with convolutional prototype learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3474–3482, 2018. 3, 5, 7

[39] Tianyun Yang, Ziyao Huang, Juan Cao, Lei Li, and Xirong Li. Deepfake network architecture attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4662–4670, 2022. 1, 2

[40] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4016–4025, 2019. 3

[41] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 5

[42] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7556–7566, 2019. 2

[43] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2019. 2

[44] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. 1