Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images

ROBERTO AMOROSO*, University of Modena and Reggio Emilia, Italy
DAVIDE MORELLI*, University of Modena and Reggio Emilia, Italy and University of Pisa, Italy
MARCELLA CORNIA, University of Modena and Reggio Emilia, Italy
LORENZO BARALDI, University of Modena and Reggio Emilia, Italy
ALBERTO DEL BIMBO, University of Florence, Italy
RITA CUCCHIARA, University of Modena and Reggio Emilia, Italy and IIT-CNR, Italy

Recent advancements in diffusion models have enabled the generation of realistic deepfakes by writing textual prompts in natural language. While these models have numerous benefits across various sectors, they have also raised concerns about the potential misuse of fake images and cast new pressures on fake image detection. In this work, we pioneer a systematic study of the authenticity of fake images generated by state-of-the-art diffusion models. Firstly, we conduct a comprehensive study on the performance of contrastive and classification-based visual features. Our analysis demonstrates that fake images share common low-level cues, which render them easily recognizable. Further, we devise a multimodal setting wherein fake images are synthesized by different textual captions, which are used as seeds for a generator. Under this setting, we quantify the performance of fake detection strategies and introduce a contrastive-based disentangling strategy which let us analyze the role of the semantics of textual descriptions and low-level perceptual cues. Finally, we release a new dataset, called COCOFake, containing about 600k images generated from original COCO images.

CCS Concepts: • **Computing methodologies → Image representations**; **Matching**; *Computer vision tasks.*

Additional Key Words and Phrases: multimodal deepfakes, vision-and-language, generative models

ACM Reference Format:
Roberto Amoroso, Davide Morelli, Marcella Cornia, Lorenzo Baraldi, Alberto Del Bimbo, and Rita Cucchiara. 2023. Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images.

1 INTRODUCTION

Machine-generated images have gained extensive popularity in the digital world due to the spread of GANs [22, 30, 31, 41] and diffusion models [13, 44, 47, 50]. While image generation tools can be employed for lawful goals, such as assisting content creators, generating simulated datasets, or enabling multimodal interactive applications, they have raised concerns regarding their potential for illegal and malicious purposes [1, 6, 25]. These include the forgery of natural images and the generation of images in support of fake news. In this context, assessing the authenticity of images becomes a fundamental goal for security and for guaranteeing the trustworthiness of AI algorithms.

Most of the past approaches for deepfake detection have employed perceptual cues [17, 19, 64], including frequency analysis, the detection of artifacts, or pixel discontinuities. Furthermore, a

*Both authors contributed equally to this research.

Authors' addresses: Roberto Amoroso, University of Modena and Reggio Emilia, Modena, Italy, roberto.amoroso@unimore.it; Davide Morelli, University of Modena and Reggio Emilia, Modena, Italy and University of Pisa, Pisa, Italy, davide.morelli@unimore.it; Marcella Cornia, University of Modena and Reggio Emilia, Modena, Italy, marcella.cornia@unimore.it; Lorenzo Baraldi, University of Modena and Reggio Emilia, Modena, Italy, lorenzo.baraldi@unimore.it; Alberto Del Bimbo, University of Florence, Florence, Italy, alberto.delbimbo@unifi.it; Rita Cucchiara, University of Modena and Reggio Emilia, Modena, Italy and IIT-CNR, Pisa, Italy, rita.cucchiara@unimore.it.

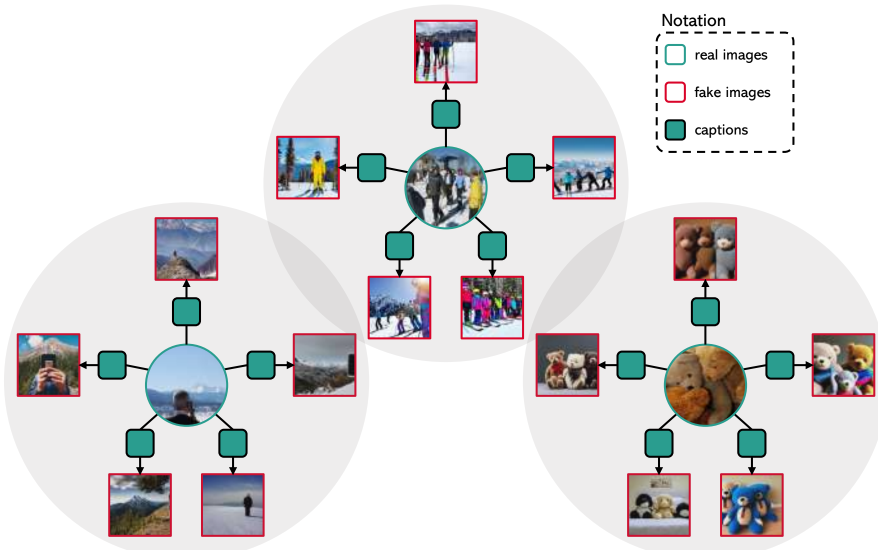

Fig. 1. Overview of our multimodal deepfake detection setting, in which five subsets of the semantics contained in a given image are employed to generate as many fake images.

significant portion of the early studies has focused exclusively on fake faces [35, 36, 48]. Today's generators [14, 20, 44, 45, 47, 50] are general-purpose, text-driven, and exhibit higher generation quality. If we look at images generated by Stable Diffusion [47] (see Fig. 3), we might notice that some of them appear hyper-realistic and, thus, easily recognizable, while other contains semantic anomalies. However, most of them are realistically plausible.

In this paper, we aim at developing a systematic study on deepfake detection, in an era when generated content is becoming increasingly realistic and text-driven. We do this in a multimodal setting that enables us to examine deepfake detection from both a perceptual and a semantic perspective. Specifically, given an image, we consider different textual descriptions and fake images generated by using each of the descriptions as a prompt (Fig. 1). In this manner, we build clusters sharing similar semantics, containing one real image and multiple fake images. Under this setting, we first investigate the effectiveness of contrastive and classification-based visual features in training a classifier to recognize deepfakes. Surprisingly, we find out that high-level contrastive-based features learned on image and text pairs are very effective in discriminating between real and generated images. We hypothesize that low-level perceptual features also percolate into such descriptors, even though they are trained at a semantic level.

While these findings might be effective in defending us from current generators, we can expect that tomorrow's generators will increase their quality and become less detectable via low-level features. Thus, we devise a contrastive-based disentanglement strategy that enables us to remove the contribution of low-level features. This approach establishes a fair setting in which generated images cannot be distinguished at a perceptual level. Under this setting, we propose and discuss a general procedure for discriminating between fake and real images based on semantic information. To evaluate the effectiveness of the proposed method, we introduce a new dataset, namely COCOFake, which comprises approximately 600k images generated from original COCO images.

Contributions In summary, the main contributions of this work are as follows:

- We develop a framework that utilizes machine-generated variants of natural images to investigate the detectability of diffusion model-generated images at the semantic and perceptual levels. By filtering the semantic content of natural images through natural language descriptions, we create a dataset of machine-generated images that can be used to investigate the performance of fake detection against modern diffusion models.
- We demonstrate that contrastive-based features can be effectively employed for fake detection against modern diffusion models, with high recognition rates.
- We propose a contrastive-based disentanglement approach to distinguish between low-level and semantic features in modern visual extractors. This allows us to distinguish between natural images and the generated ones using only semantic cues while neglecting the perceptual ones. This is important for the future development of more realistic generators.
- We generate and release the COCOFake dataset, which contains over 600k fake images linked to natural images through captions. This dataset can be used to test and evaluate the performance of fake detection algorithms against diffusion model-generated images.

2 RELATED WORK

General deepfake detection. In recent years, with the growth and diffusion of generative models, several research efforts [12, 59] have been made to effectively detect synthetic images generated by GANs [22, 30, 31, 41, 66] and other deep learning-based architectures [34, 56]. While initial works did not concentrate on the generalization capabilities of deepfake detectors [40, 48], subsequent approaches [4, 10, 21, 23, 39, 60] focused instead on the development of generic detectors that can be applied to different generators, thus avoiding the need to have a specific detector for each generative model. On the same line, different solutions [17, 19, 64] proposed to detect deepfakes based on the spectrum of GAN-generated images. In fact, CNN-based generative models usually leave a distinguishable fingerprint over generated images, due to transposed convolutions [17, 64], up-sampling operations [5, 19], and the spectral bias of convolution layers [18, 32]. Some works in similar directions also focused on associating fake images to the corresponding generator among several known GANs [28, 63] or extending deepfake detection to the video domain [11, 24].

Detection of deepfakes generated with diffusion models. While all aforementioned methods are tailored for detecting deepfakes generated by GANs, a few works extended the analysis to deepfake images coming from diffusion models [13, 42, 44, 47, 50]. Among them, Wolter et al. [61] proposed to detect fake images based on their wavelet-packet representations taking into account features from the pixel and frequency space. Ricker et al. [46] also tackled the frequency domain, analyzing different factors that influence the spectral properties of these images. Similarly, Corvi et al. [9] introduced an analysis of the forensics traces left by common diffusion models and investigated whether deepfake detectors tailored for GANs can also distinguish images generated by diffusion models. Finally, Sha et al. [55] analyzed and compared deepfakes generated by different text-to-image diffusion models, investigating the possibility of correctly attributing deepfake images to the diffusion model that generated them. Overall, these studies highlight the need for developing detection methods that can effectively detect deepfakes generated by various types of generative models, including diffusion models.

Datasets for deepfake detection. The availability of large datasets has played a crucial role in the development of deepfake detection techniques. One of the most widely used datasets is FaceForensics++ [48], which contains videos of real and fake faces generated using several generative models. The dataset provides both raw and manipulated videos with different compression rates and resolutions, allowing the evaluation of deepfake detection methods under different scenarios. Another popular dataset is Celeb-DF [36], which contains videos of celebrities manipulated using different techniques including GANs and face swapping. Celeb-DF also provides several levels of difficulty, ranging from low-quality to high-quality forgeries, making it suitable for evaluating both traditional and advanced deepfake detection methods. Other datasets have been proposed, such as DeeperForensics-1.0 [27], which contains manipulated videos generated using multiple GAN-based models, and DFDC [15], composed of thousands of videos of real and fake faces.

Despite the availability of these datasets, there is still a need for more diverse and challenging datasets that reflect the increasing sophistication of deepfake generation methods. In particular, while current datasets mainly focus on faces, there is a lack of datasets for detecting deepfakes in other types of images, such as natural scenes. The proposed COCOFake dataset aims to address this limitation by providing a large-scale dataset of natural images and their corresponding synthetic images generated by diffusion models, along with natural language captions linking them. This allows for the evaluation of deepfake detection methods in a more complex and diverse context and also enables the development of methods that can identify semantic inconsistencies between natural and synthetic images.

3 PROPOSED METHOD

3.1 Notation and Preliminaries

We propose a framework for studying and detecting multimodal generated fake images, which encompasses the identification and separation of their perceptual and semantic components. In the rest of the paper, we will employ the following notation: I_R will indicate a natural (real) image, C a textual description (*i.e.*, a caption), and I_F will indicate a fake image produced by a generator. Under this setting, a *parent* real image I_R can be the seed for N different *children* fake images $I_{F,i}$ given a set of textual descriptions $\{C_i\}$ of I_R, with $i = 1, ..., N$, by using each of the descriptions as prompt for the generator.

Semantic and style components of an image. The information content of an image can be credited to many factors. For simplicity, we assume that an image I, regardless of its authenticity, embodies two information contributions, namely a *semantic component* $\mathcal{H}_{sem}(I)$ and a perceptual or *style component* $\mathcal{H}_{sty}(I)$. The former represents the content that could be expressed in a textual sentence, while the latter describes the image appearance, encompassing elements such as colors, textures, brightness, and low-level visual cues. Given a real image I_R, we can therefore express its total information $\mathcal{H}$ as a function of its semantic and style components, as follows:

$$\mathcal{H}(I_R) = f(\mathcal{H}_{sem}(I_R), \mathcal{H}_{sty}(I_R)). \tag{1}$$

However, when an image is described through a natural language sentence, only a portion of its semantics is actually conveyed inside the caption. In other words, natural language descriptions act as a filter for the semantic content of the image. Hence, we introduce $\Delta\mathcal{H}_{sem}(I, C)$ to represent the portion of semantic information described by a caption C. By analogy, we could say that the textual descriptions of an image act as DNA fragments that can be utilized to generate an offspring of images.

Generating offspring with natural language utterances. From an input image I_R we can, therefore, extract N semantic information subsets $\Delta\mathcal{H}^i_{sem}(I_R, \cdot)$ and feed them to a generator obtaining N different fake images $I_{F,i}$, with $i = 1, ..., N$. We define *semantic cluster* the ensemble of the starting real image I_R and the offspring of N fake images $I_{F,i}$ generated from it. For instance, given a real image dataset such as COCO [37], containing K images, each represented by $N = 5$ captions, we could create K clusters of $N + 1$ images with one parent and N children.

3.2 Learning to Discriminate Real and Fake images

Once a dataset in the aforementioned form has been built, we first measure to what extent real and generated images can be discriminated independently from their membership to a semantic cluster. Instead of doing this by learning ad-hoc visual features, we investigate the usage of state-of-the-art pre-trained visual models. In other words, given a dataset containing both natural and generated images, we investigate the development of a model which can identify real images by relying on the full set of their visual features, as extracted with a pre-trained backbone. Regarding the generation of the images, in the following we will employ Stable Diffusion [47], which is freely available and represents a state-of-the-art approach. Nevertheless, the approach could be easily extended to other generators.

In order to evaluate the discriminative power of current pre-trained visual features, we model the discriminator as a two-class linear classifier, so that input visual features are only linearly projected before taking the final decision on their realism. Formally, given a real image $I_R \in \mathbb{R}^{3 \times H \times W}$ and an image encoder $E_I : \mathbb{R}^{3 \times H \times W} \rightarrow \mathbb{R}^D$, we extract a vectorial image feature F_I as

$$F_I = E_I(I_R). \tag{2}$$

The features F_I are then fed into a linear layer $L : \mathbb{R}^D \rightarrow \mathbb{R}$, whose output is thresholded to classify between *real* (*i.e.*, 0) and *fake* (*i.e.*, 1) images. As it will be discussed in the experimental section, our findings indicate that this is (still) a relatively simple task even when employing a state-of-the-art generator. This is, most likely, due to the fact that fake images are sightly different in terms of low-level cues with respect to real images.

3.3 Semantic Preservation Analysis

As a second analysis, we investigate the preservation of semantic information among real and generated images. To do so, we consider a multimodal embedding space, in which both images and texts can be projected [3, 43, 51]. Specifically, we verify if, starting from a generated image, we can retrieve the particular caption used as prompt during its generation. In other words, we test if the subset of the real semantic information $\Delta\mathcal{H}_{sem}(I_F, C)$ associated with a caption C is still recognizable in the visual features extracted from the generated image.

Formally, given a caption C describing a real image I_R, and a textual encoder E_T, we tokenize and extract the textual features F_T as:

$$F_T = E_T(C). \tag{3}$$

For each visual feature of a given fake image I_F, we verify the ability to retrieve the corresponding textual feature used to create I_F through the generator model.

As it will be shown in the experimental section, we find out that (a) the alteration of low-level cues induced by the generator does not affect the semantic contribution coming from the original image, and that (b) the semantic contribution of the generator does not obfuscate the original semantic content.

3.4 Disentangling Semantics and Style

As the detection of fake images is likely promoted by the difference in low-level cues between generated and real images, we finally investigate a more challenging setting in which the style component induced by the generator is disentangled and removed. To do so, we learn a model which identifies the style component of the generator which is common to all generated images. We then measure whether, after eliminating such a component, the remaining semantic information is sufficient to discriminate between real and fake images. Noticeably, this corresponds to a more challenging setting where all the common low-level traits left by the generator are removed and not

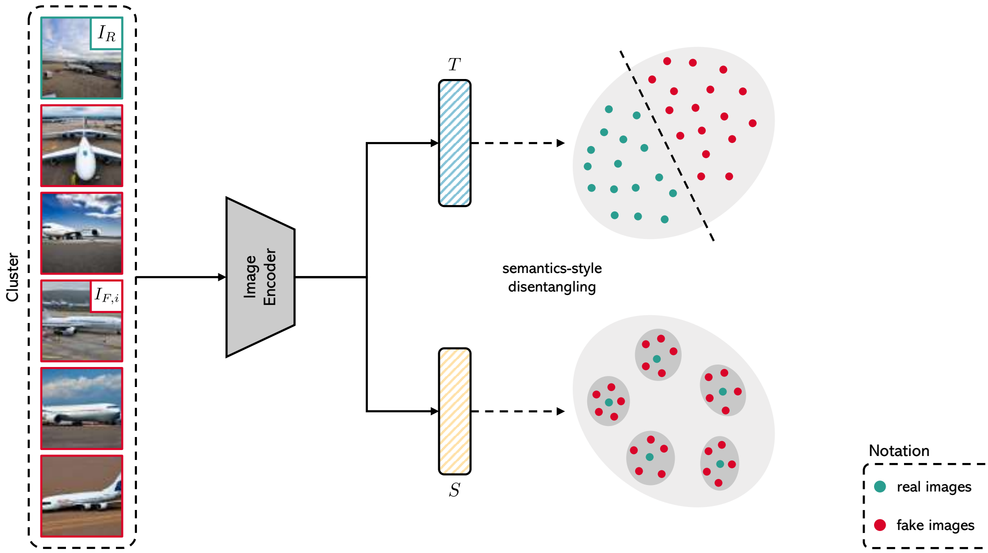

Fig. 2. Schema of our approach for disentangling semantics and style for deepfake detection.

employed to perform deepfake classification. In other words, this also corresponds to recognizing fakes generated by an "ideal" generator which does not leave common low-level traits.

In order to perform this analysis, we propose a new contrastive-based learning model which can project images in a semantic space and in a style space (Fig. 2). For a good style-semantic disentanglement we expect that, in the style embedding latent space, the feature vectors of real images should be separated from features of fake images in a cluster-agnostic way, while in the semantic embedding latent space the cluster compactness should be preserved. Specifically, we train two separate linear projections T and S, where T focuses on style while S on semantics. For the T layer we aim at increasing the distance between fake and real elements, regardless of their membership in a specific cluster. For the S layer, instead, we want to create compact clusters of elements sharing the same semantic content, while increasing the distance among two fake elements or two real elements.

We express these requirements through two loss components $\mathcal{L}_c$ and $\mathcal{L}_{fr}$. The former attracts elements of the same cluster, while the latter attracts elements having the same label (*i.e.*, real and fake). From here, we can define the losses needed to train T and S, respectively, as follows:

$$\begin{aligned} \mathcal{L}_T &= \mathcal{L}_{fr} - \mathcal{L}_c, \\ \mathcal{L}_S &= \mathcal{L}_c - \mathcal{L}_{fr}. \end{aligned} \tag{4}$$

To implement both $\mathcal{L}_c$ and $\mathcal{L}_{fr}$, we leverage a Supervised Contrastive Loss [33], defined as follows:

$$\mathcal{L}_{SupCon} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathcal{F}_i \mathcal{F}_p^\intercal / \tau)}{\sum_{a \in A(i)} \exp(\mathcal{F}_i \mathcal{F}_a^\intercal / \tau)}, \tag{5}$$

where $i \in I \equiv \{1, ..., N+1\}$ represents the index of an arbitrary sample, $\mathcal{F}$ are ℓ_2-normalized input features of a given image, τ is a temperature parameter, and $A(i) \equiv I / \{i\}$. $P(i)$ is the set of indices of all items sharing the same label of i, and $|P(i)|$ is its cardinality.

Depending on the nature of the labels used in the training of the supervised contrastive loss, we can implement repulsive and attractive forces in the form of the loss components $\mathcal{L}_c$ and $\mathcal{L}_{fr}$. In

$\mathcal{L}_c$, in particular, we assign the same label to elements belonging to the same cluster, while in $\mathcal{L}_{fr}$ we assign the same label to all real samples, and the same label to all fake images. The objective of $\mathcal{L}_c$ is to attract elements of the same cluster, while $\mathcal{L}_{fr}$ pushes real and fake images.

3.5 The COCOFake Dataset for Multimodal Deepfake Recognition

In literature, to the best of our knowledge, there are no multimodal datasets containing texts, real and fake images which are compatible with our multimodal setting. Thus, we generate and release the COCOFake dataset, an extension of COCO [37]. Each real image in COCOFake is paired with five fake images that are conditionally generated based on each of the captions associated with the same image. We employ the Stable Diffusion model [47] as our generator. Specifically, we employ version 1.4 of Stable Diffusion, which has been pre-trained on the English image-text pairs of the LAION-5B dataset [53] and finetuned at a resolution equal to 512×512 on the LAION-Aesthetics subset[1]. During image generation, we employ the safety checker module to reduce the probability of explicit images and disable the invisible watermarking of the outputs to prevent easy identification of the images as machine-generated.

Overall, referring to the splits commonly used in captioning literature [7, 8, 29, 52], the COCO dataset comprises 113,287 training images, 5,000 validation, and 5,000 test images. Preserving the same splits, COCOFake is composed of 679,722 training images, 30,000 validation, and 30,000 test images. Sample real-generated image clusters from the COCOFake dataset are shown in Fig. 3. For each example, we present the real image alongside the five fake images generated from each of the five captions from the original COCO dataset. As can be seen, the generated images are generally coherent with the corresponding caption. However, in some cases, the generated images are overly realistic with brighter colors and a more professional photographic style than the real counterpart. This can be attributed to the dataset employed in the finetuning phase (*i.e.* the LAION-Aesthetics subset) of the Stable Diffusion model [47], used to generate fake images. In Fig. 4 we report less realistic examples from the COCOFake dataset, again showing the original image and the five fake images with the corresponding captions. Failure cases include hallucinating the semantic content of the caption (first two rows), incorrect understanding of the caption (third row), abstract rendering of objects (traffic lights in the third row), and unrealistic rendering of human poses (last row).

To assess the robustness of our analysis, we further generate the test and training set using a different version of Stable Diffusion (*i.e.*, v2.1). This version exploits OpenCLIP [62] instead of CLIP [43] and is trained on a subset of LAION-5B [53] filtered for explicit pornographic material. The updated version of Stable Diffusion provides superior results compared to version 1.4 due to the use of a larger and more diverse training dataset. Under this setting, we compare the performance of our method on images generated by different versions of Stable Diffusion, providing insights into the impact of the generative model on the deepfake detection performance.

4 EXPERIMENTAL EVALUATION

4.1 Implementation Details

Image encoders. We test two families of backbones: the first are trained for classification on ImageNet [49], while the second are trained on a cross-modal setting on large-scale datasets using the contrastive loss. Due to the nature of the task these networks were trained for, only the latter family provides also text encoders E_T.

Specifically, we employ a ResNet [26] model with 48 convolutional layers and a Vision Transformer (ViT) [16] architecture in its B/32 configuration. The ViT encoder takes as input squared patches extracted from the input image and consists of a sequence of multi-head self-attention

[1]https://laion.ai/blog/laion-aesthetics/

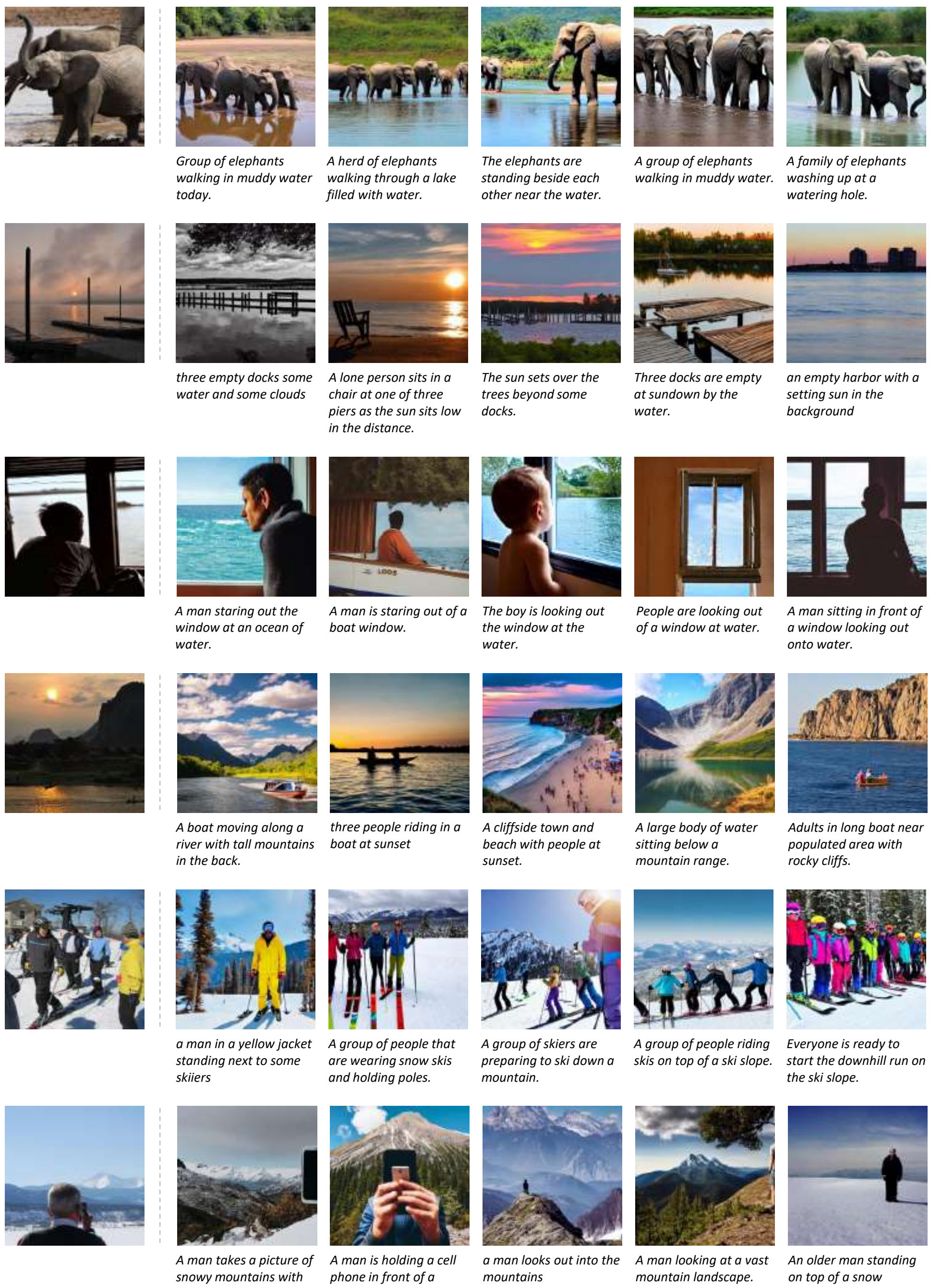

Fig. 3. Sample images from the COCOFake dataset. The leftmost column shows the original (real) image, while the remaining columns show fake images generated from each of the five COCO captions.

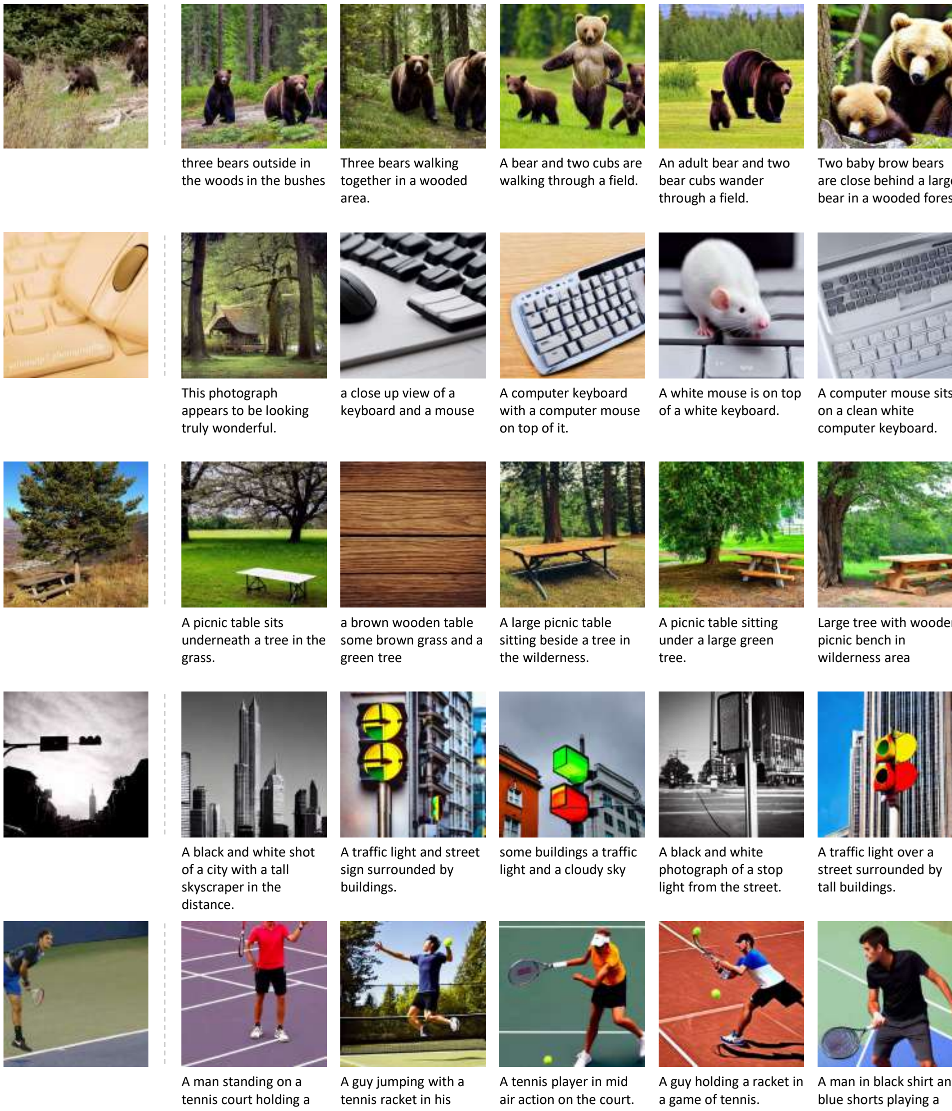

Fig. 4. Less realistic images from the COCOFake dataset. The leftmost column shows the original (real) image, while the remaining columns show fake images generated from each of the five COCO captions.

layers [58]. Both these architectures are trained on the ImageNet dataset [49] that contains around 1.3M images.

As cross-modal architectures, we use two models coming from CLIP [43]. In particular, we employ CLIP RN50 and CLIP ViT-B/32 models, both pre-trained on the OpenAI WebImageText (WIT) dataset, composed of 400 million image-text pairs collected from the web. Moreover, we employ the open source implementation of CLIP (*i.e.*, OpenCLIP [62]), trained with a post-ensemble method for improving robustness to out-of-distribution samples. In our experiments, we consider two versions of the OpenCLIP ViT-B/32 model: one trained on the LAION-400M dataset [54] that contains 400 million CLIP-filtered image-text pairs crawled from the web, the other trained on the larger LAION-2B composed of 2 billion image-text pairs [53].

Linear probing. In our experiments, we also conduct linear probes. In this case, we follow the approach of [43] and employ the features extracted from the backbones to train a logistic regression model with ℓ_2 penalty and LBFGS solver [2, 65]. To balance the training samples, we employ one randomly extracted fake image for each cluster.

Disentanglement architecture and training details. When disentangling semantics and styles, we train the two linear layers S and T, which perform a linear projection to the same dimensionality of the backbone visual features. To train these layers, we employ AdamW [38] as optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use a batch size of 1,024 and a learning rate of 0.001, training all models for 25 epochs.

4.2 Metrics

To assess the performance of our proposed methodology and evaluate spatial relationships between elements in the embedding spaces, we define six different metrics. These aim to quantify the capability to discriminate between real and fake images and to quantify disentanglement.

Overall and Full Cluster Accuracy. These two metrics measure the real/fake classification accuracy both over the entire dataset and inside each cluster. The former metric is cluster-independent and is computed using all the elements of a dataset split (*i.e.*, validation, test). The latter, instead, is a cluster-based metric that scores if all elements of a cluster are correctly classified as real or fake, and the metric is then averaged across all clusters.

Min and Max Intra-Cluster Distance Accuracy. These two metrics are employed to evaluate the relative spatial positions of the elements inside a cluster. In particular, for each cluster, we measure the distances between the real image and each of the fake images belonging to the cluster. We then check how many times the real image is the item having the minimum or maximum distance with respect to all the others in the cluster. In other words, for each cluster, the min distance accuracy scores if the real image feature is on average the nearest to all the fake image features, while the max distance accuracy scores if it is the most distant one.

Exact Pair and Intra-Cluster Retrieval. These metrics are used to evaluate the goodness of the retrieval task (see Sec. 3.3), in which given a generated image we seek to retrieve its parent caption. The former metric is a recall@k computed considering as ground-truth, for each fake image, the caption used for generating it. The latter, instead, is a recall@k that measures for a given fake image if the retrieved caption matches one of the five captions of the cluster the image belongs to.

4.3 Performance of Visual Features

Unsupervised classification. We start by assessing the capabilities of existing image features to discriminate between real and generated images, in an unsupervised setting. We employ the min and max distance accuracy metrics defined above and check the presence of spatial relationships between real and generated images inside each cluster.

Results are reported in Table 1 on the test and validation sets of both Stable Diffusion v1.4 and v2.1. We employ six different visual backbones, namely two ResNet-50 pre-trained on ImageNet and OpenAI WIT and four ViT-B/32 pre-trained on ImageNet, OpenAI WIT, LAION-400M, and LAION-2B. As it can be seen, according to the features extracted from the aforementioned backbones, the real image of each cluster tends to be the one with maximum distance with respect to all the other elements. This suggests that these features are discriminative for the task of deepfake classification and that they percolate low-level features which allow for distinction between real and generated items inside of each semantic cluster. Noticeably, the maximum distance accuracy increases when

Table 1. Minimum and maximum distance accuracy on validation and test sets of COCOFake, using different visual backbones.

		Validation Set		Test Set	
Backbone	**Dataset**	Min Dist. Accuracy	Max Dist. Accuracy	Min Dist. Accuracy	Max Dist. Accuracy
Stable Diffusion v1.4:					
RN50	ImageNet	8.50	23.58	8.82	24.82
ViT-B/32	ImageNet	6.84	23.12	6.88	23.88
CLIP RN50	OpenAI WIT	3.72	38.48	3.60	41.24
CLIP ViT-B/32	OpenAI WIT	3.30	38.88	3.24	40.10
OpenCLIP ViT-B/32	LAION-400M	5.28	31.94	5.00	32.02
OpenCLIP ViT-B/32	LAION-2B	1.40	42.80	1.72	44.00
Stable Diffusion v2.1:					
RN50	ImageNet	5.98	29.62	6.62	30.16
ViT-B/32	ImageNet	5.12	29.18	4.92	30.00
CLIP RN50	OpenAI WIT	2.40	46.72	2.20	48.28
CLIP ViT-B/32	OpenAI WIT	2.92	42.08	2.98	44.18
OpenCLIP ViT-B/32	LAION-400M	4.58	34.06	4.62	36.02
OpenCLIP ViT-B/32	LAION-2B	1.88	42.64	1.78	43.80

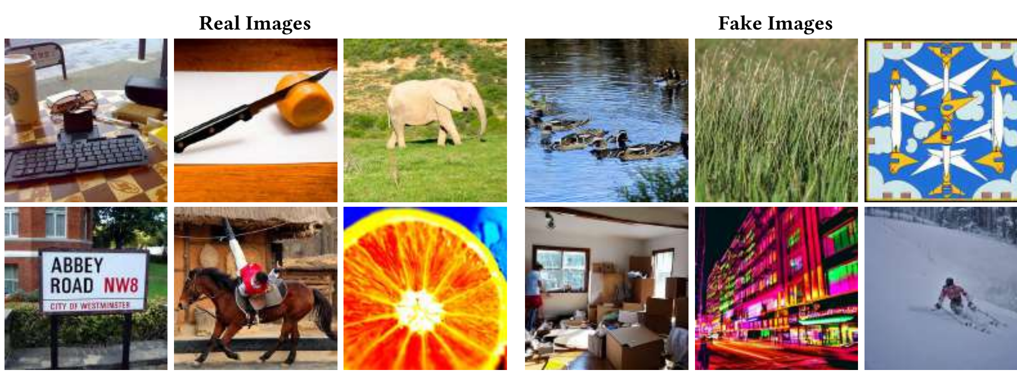

Fig. 5. Sample misclassification errors on both real (left) and fake (right) images, using OpenCLIP ViT-B/32 trained on LAION-2B as the visual encoder.

considering backbones trained on multimodal datasets with respect to backbones trained on classification, suggesting that image-text matching promotes the percolation of perceptual features.

Comparing the results Stable Diffusion v2.1 exhibits an improvement over v1.4 as evidenced by an increase in the maximum distance metric and a decrease in the minimum distance metric. This suggests that the features extracted from v2.1 are better separable and hence the generated images are more easily detected.

Linear probing. Following the approach popularized by [43], we train a linear projection through logistic regression on top of the features extracted from the aforementioned backbones.

We notice that all the selected visual features exhibit a significant capability in linearly discriminating real and fake images, in both the validation and test set of the COCOFake dataset. In continuity with the previous experiment, in Table 2 we observe that contrastive-based visual

Table 2. Overall and full cluster accuracy results on the validation and test sets, using linear probing and features of different backbones trained on the COCOFake training set.

		Validation Set		Test Set	
Backbone	**Dataset**	Overall Accuracy	Full Cluster Accuracy	Overall Accuracy	Full Cluster Accuracy
Stable Diffusion v1.4:					
RN50	ImageNet	90.31	57.56	90.62	57.94
ViT-B/32	ImageNet	87.64	47.62	87.16	47.32
CLIP RN50	OpenAI WIT	99.07	94.60	99.17	95.30
CLIP ViT-B/32	OpenAI WIT	99.11	94.84	98.97	94.24
OpenCLIP ViT-B/32	LAION-400M	97.88	88.18	97.83	87.80
OpenCLIP ViT-B/32	LAION-2B	99.68	98.01	99.64	97.84
Stable Diffusion v2.1:					
RN50	ImageNet	81.71	34.94	82.31	35.84
ViT-B/32	ImageNet	76.71	24.68	77.31	26.92
CLIP RN50	OpenAI WIT	93.54	69.08	93.74	69.64
CLIP ViT-B/32	OpenAI WIT	94.41	72.30	94.72	73.62
OpenCLIP ViT-B/32	LAION-400M	83.30	38.48	84.32	40.74
OpenCLIP ViT-B/32	LAION-2B	98.88	93.68	98.96	94.08

Table 3. Exact pair and intra-cluster retrieval results.

		Validation Set						Test Set					
		Exact Pair			Intra-Cluster			Exact Pair			Intra-Cluster		
Backbone	**Dataset**	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5
CLIP RN50	OpenAI WIT	31.33	49.05	56.93	41.91	58.46	66.01	30.98	48.38	56.42	42.09	58.35	65.93
CLIP ViT-B/32	OpenAI WIT	32.12	50.43	58.36	43.34	60.15	67.42	31.96	49.67	57.51	43.24	59.3	66.78
OpenCLIP ViT-B/32	LAION-400M	36.48	55.36	63.28	47.17	63.62	70.73	35.53	54.49	62.56	46.72	62.92	70.22
OpenCLIP ViT-B/32	LAION-2B	40.34	59.44	67.18	50.78	66.64	73.58	39.57	58.78	66.18	50.46	66.34	73.03

backbones showcase significantly higher accuracy levels, up to 98.01% of full cluster accuracy on the validation set, and up to 99.68% overall accuracy on the same split. This further confirms the observation that contrastive-based backbones extract and project into their embedding space, low-level and perceptual features that allow discriminating current deepfakes. To assess the robustness of the method we further test the trained classifiers on the data generated using Stable Diffusion v2.1. As can be observed in the lower part of Table 2, the trained classifier performs comparably also in this setting. Overall, these two experiments show that the pre-trained visual backbones exhibit high discrimination power when identifying deep fakes.

In light of the high accuracy levels of the aforementioned experiment, in Fig. 5 we report sample misclassified images. It can be noted, in particular, that fake images incorrectly classified as authentic (right side of the figure) depict close-ups and artistic drawings, whose authenticity is visually harder to guarantee.

Semantic Preservation. We then conduct the retrieval-based analysis anticipated in Sec. 3.3, in which we look for the original caption used to generate a particular image inside of a multimodal embedding space. The objective of this experiment is to assess whether the semantic information contained in the caption is preserved after the generation and to what extent the generation process alters semantic features.

Table 4. Overall and full cluster accuracy results on the semantic space S and minimum/maximum distance accuracy results on the style space T. These results are obtained by training on the COCOFake training set under the disentanglement setting and evaluating on validation and test sets of the COCOFake dataset, using data extracted from both Stable Diffusion v1.4 and v2.1.

		Validation Set				Test Set			
Backbone	**Dataset**	Overall Accuracy S	Full Cluster Accuracy S	Min Dist. Accuracy T	Max Dist. Accuracy T	Overall Accuracy S	Full Cluster Accuracy S	Min Dist. Accuracy T	Max Dist. Accuracy T
Stable Diffusion v1.4:									
RN50	ImageNet	62.07	7.74	0.38	88.02	62.96	8.64	0.42	89.08
ViT-B/32	ImageNet	63.78	8.34	1.08	77.70	64.04	8.46	1.30	76.26
CLIP RN50	OpenAI WIT	75.52	21.64	0.00	98.36	74.76	21.40	0.00	98.46
CLIP ViT-B/32	OpenAI WIT	67.33	11.88	0.04	98.36	67.48	12.90	0.20	98.14
OpenCLIP ViT-B/32	LAION-400M	66.47	10.90	0.14	94.82	66.84	10.98	0.10	94.48
OpenCLIP ViT-B/32	LAION-2B	72.40	17.12	0.02	99.30	72.62	17.32	0.06	99.39
Stable Diffusion v2.1:									
RN50	ImageNet	57.77	6.90	0.34	87.46	58.53	6.62	0.52	89.48
ViT-B/32	ImageNet	62.92	8.30	1.92	70.62	63.00	8.86	1.78	72.84
CLIP RN50	OpenAI WIT	65.61	12.66	0.08	96.60	64.77	12.74	0.12	96.42
CLIP ViT-B/32	OpenAI WIT	62.42	9.34	0.14	95.28	62.66	10.06	0.26	94.60
OpenCLIP ViT-B/32	LAION-400M	68.53	12.54	0.66	82.66	68.32	12.02	0.52	83.98
OpenCLIP ViT-B/32	LAION-2B	71.41	16.40	0.08	98.48	71.92	16.98	0.04	98.70

Results are reported in Table 3, using the exact pair and intra-cluster retrieval metrics. Surprisingly, retrieving the exact caption used to generate an image is not always easy, and the process is successful only in 40% of the cases when selecting a proper backbone. Even when considering all captions of the same clusters as positives, moreover, we observe a recall@1 of around 50%, again highlighting the difficulty of the task. This points out that current generators produce images with partially altered semantic features, and is also in line with the previous observation that contrastive-based extractors percolate low-level features.

4.4 Semantic-Style Disentangling Results

We then turn our attention to evaluating the semantic-style disentanglement approach, in which we aim at training two separate embedding spaces, one storing semantic information and the second focusing on style information. We evaluate the semantic projection in terms of full cluster and overall classification accuracy, and the style projection in terms of minimum and maximum distance accuracy. Results are reported in Table 4, on both the validation and test sets of COCOFake and for all the aforementioned backbones. We observe that, in the T space which focuses on style, real and fake images can be properly distinguished, as the real image is always far apart from generated ones. On the contrary, this does not happen in the S space, which focuses on semantics, and in which all elements belonging to the same cluster are pulled together, independently of their authenticity. Still, the identification of deepfakes is feasible, although with lower accuracy, even in this more challenging space, with an accuracy of up to 75% on the validation set. As this corresponds to testing a more challenging generator that leaves fewer lower-level traces, we believe this result might offer interesting insights for future works. Similar but slightly lower results can also be observed when using images generated by Stable Diffusion v2, with an overall accuracy of up to 71%.

The structure of the two spaces can be further visualized in Fig. 6, in which we report 2D t-SNE visualizations [57] of the feature space of the OpenCLIP ViT-B/32 LAION-2B backbone, before and after disentanglement and for both Stable Diffusion v1.4 and Stable Diffusion v2.1. In the original embedding space, as provided by the backbone, real and generated samples appear to be

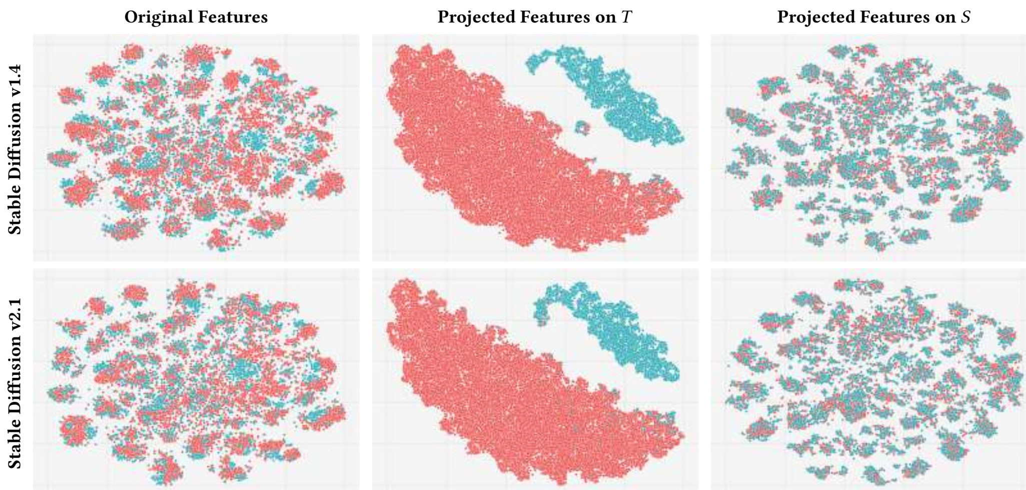

Fig. 6. t-SNE visualizations over the validation set using the original visual features from the OpenCLIP ViT-B/32 LAION-2B backbone (left), the features projected on the T space (style) after disentanglement (middle), and the features projected on the S space (semantics) after disentanglement (right), using Stable Diffusion v1.4 (top) and v2.1 (bottom). Red dots indicate fake images, blue dots indicate real images.

mostly overlapped, even if we do not observe a complete overlap – which is in line with the results presented in Table 1 and Table 2. After the disentanglement, instead, the geometry of the T and S spaces appears completely different: the T space clearly separates real and fake data (with the exception of a few outliers), while in the S space we can observe a complete overlap between real and generated samples and a tendency to group into semantic clusters.

A closer visualization of the original feature space and of the embedding spaces produced by the two projections is reported in Fig. 7. In this case, we report, on each row, the relative positioning of eight sample clusters from the validation set. As it can be seen, the two proposed projections are again effective both in separating real and fake images and in promoting the clustering of images sharing similar semantics regardless of their authenticity.

5 CONCLUSION

This paper proposes a multimodal setting for deepfake detection and analysis, in which real and generated images sharing the same semantics are paired into semantic clusters. In our setting, different semantic projections of a given image, expressed through captions, are employed to generate fake images. Employing the popular Stable Diffusion model as generator, we investigated the performance of contrastive and classification-based visual features, highlighting that diffusion-based deepfakes share common low-level features which make them easily identifiable. Further, we proposed an approach to disentangle semantic and perceptual information, based on supervised contrastive learning. Under this setting, we investigated the classification of authenticity in a semantic space in which low-level cues left by the generator are removed, thus tackling a more challenging scenario. As a complementary contribution, we also collected and released the COCOFake dataset, containing about 600k images generated from COCO using both Stable Diffusion 1.4 and 2.1. We believe that our work can shade further light on the development of deepfake detection strategies, also in consideration of the constant evolution of generator models.

Fig. 7. t-SNE visualizations on sampled clusters from the Stable Diffusion v1.4 validation set using features extracted from the OpenCLIP ViT-B/32 architecture pre-trained on LAION-2B. We report the original features from the visual backbone (left), the features projected on the T space (style) after disentanglement (middle), and the features projected on the S space (semantics) after disentanglement (right). Dots indicate fake images, triangles indicate real images. Images from the same cluster are shown with the same color.

ACKNOWLEDGMENTS

We thank CINECA for providing computational resources. This work has been supported by the Horizon Europe project "European Lighthouse on Safe and Secure AI (ELSA)" (HORIZON-CL4-2021-HUMAN-01-03), co-funded by the European Union, and by the PNRR project "Future Artificial Intelligence Research (FAIR)", co-funded by the Italian Ministry of University and Research.

REFERENCES

[1] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. 2018. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv preprint arXiv:1802.07228* (2018).

- [2] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* 16, 5 (1995), 1190–1208.
- [3] Angelo Carraggi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2018. Visual-Semantic Alignment Across Domains Using a Semi-Supervised Approach. In *Proceedings of the European Conference on Computer Vision Workshops.*
- [4] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. 2020. What Makes Fake Images Detectable? Understanding Properties that Generalize. In *Proceedings of the European Conference on Computer Vision.*
- [5] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, and Ngai-Man Cheung. 2021. A Closer Look at Fourier Spectrum Discrepancies for CNN-Generated Images Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*
- [6] Robert Chesney and Danielle Citron. 2019. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs* 98 (2019), 147.
- [7] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2020. SMArT: Training Shallow Memory-aware Transformers for Robotic Explainability. In *Proceedings of the IEEE International Conference on Robotics and Automation.*
- [8] Marcella Cornia, Lorenzo Baraldi, Giuseppe Fiameni, and Rita Cucchiara. 2021. Universal Captioner: Inducing Content-Style Separation in Vision-and-Language Model Training. *arXiv preprint arXiv:2111.12727* (2021).
- [9] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2022. On the detection of synthetic images generated by diffusion models. *arXiv preprint arXiv:2211.00680* (2022).
- [10] Davide Cozzolino, Diego Gragnaniello, Giovanni Poggi, and Luisa Verdoliva. 2021. Towards Universal GAN Image Detection. In *Proceedings of the International Conference on Visual Communications and Image Processing.*
- [11] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. 2021. ID-Reveal: Identity-Aware DeepFake Video Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*
- [12] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. 2018. ForensicTransfer: Weakly-supervised Domain Adaptation for Forgery Detection. *arXiv preprint arXiv:1812.02510* (2018).
- [13] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems.*
- [14] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. CogView: Mastering Text-to-Image Generation via Transformers. In *Advances in Neural Information Processing Systems.*
- [15] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The DeepFake Detection Challenge (DFDC) Dataset. *arXiv preprint arXiv:2006.07397* (2020).
- [16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations.*
- [17] Ricard Durall, Margret Keuper, and Janis Keuper. 2020. Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*
- [18] Tarik Dzanic, Karan Shah, and Freddie Witherden. 2020. Fourier Spectrum Discrepancies in Deep Network Generated Images. In *Advances in Neural Information Processing Systems.*
- [19] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. 2020. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *Proceedings of the International Conference on Machine Learning.*
- [20] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors. In *Proceedings of the European Conference on Computer Vision.*
- [21] Sharath Girish, Saksham Suri, Sai Saketh Rambhatla, and Abhinav Shrivastava. 2021. Towards Discovery and Attribution of Open-World GAN Generated Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*
- [22] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems.*
- [23] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. 2021. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In *Proceedings of the IEEE International Conference on Multimedia and Expo.*
- [24] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. 2021. Spatiotemporal Inconsistency Learning for DeepFake Video Detection. In *Proceedings of the ACM International Conference on Multimedia.*
- [25] Douglas Harris. 2018. Deepfakes: False Pornography Is Here and the Law Cannot Protect You. *Duke Law & Technology Review* 17 (2018), 99.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[27] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. 2020. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[28] Matthew Joslin and Shuang Hao. 2020. Attributing and Detecting Fake Images Generated by Known GANs. In *Proceedings of the IEEE Security and Privacy Workshops*.

[29] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[30] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[31] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[32] Mahyar Khayatkhoei and Ahmed Elgammal. 2022. Spatial Frequency Bias in Convolutional Generative Adversarial Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[33] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*.

[34] Durk P Kingma and Prafulla Dhariwal. 2018. Glow: Generative Flow with Invertible 1x1 Convolutions. In *Advances in Neural Information Processing Systems*.

[35] Yuezun Li and Siwei Lyu. 2018. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

[36] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. [Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics]. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision*.

[38] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations*.

[39] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. 2022. Detecting GAN-generated Images by Orthogonal Training of Multiple CNNs. *arXiv preprint arXiv:2203.02246* (2022).

[40] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. 2018. Detection of GAN-Generated Fake Images over Social Networks. In *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval*.

[41] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784* (2014).

[42] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. *arXiv preprint arXiv:2112.10741* (2021).

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*.

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125* (2022).

[45] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the International Conference on Machine Learning*.

[46] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. 2022. Towards the Detection of Diffusion Model Deepfakes. *arXiv preprint arXiv:2210.14571* (2022).

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[48] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Face-Forensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[49] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* (2015).

[50] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487* (2022).

[51] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Positive-Augmented Constrastive Learning for Image and Video Captioning Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[52] Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2022. Retrieval-Augmented Transformer for Image Captioning. In *Proceedings of the International Conference on Content-based Multimedia Indexing.*

[53] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems.*

[54] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *Advances in Neural Information Processing Systems Workshops.*

[55] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. 2022. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Diffusion Models. *arXiv preprint arXiv:2210.06998* (2022).

[56] Arash Vahdat and Jan Kautz. 2020. NVAE: A deep hierarchical variational autoencoder. In *Advances in Neural Information Processing Systems.*

[57] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008), 2579–2605.

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems.*

[59] Luisa Verdoliva. 2020. Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing* 14, 5 (2020), 910–932.

[60] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[61] Moritz Wolter, Felix Blanke, Raoul Heese, and Jochen Garcke. 2022. Wavelet-packets for deepfake image analysis and detection. *Machine Learning* (2022), 1–33.

[62] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. 2022. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[63] Ning Yu, Larry S Davis, and Mario Fritz. 2019. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*

[64] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. 2019. Detecting and Simulating Artifacts in GAN Fake Images. In *Proceedings of the International Workshop on Information Forensics and Security.*

[65] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Software* 23, 4 (1997), 550–560.

[66] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*