Title: METHOD FOR VISUAL PARAPHRASE ATTACK SAFE AND DISTORTION FREE IMAGE WATERMARKING FOR AI- GENERATED IMAGES

[001] The present invention relates to a watermarking method for an image, particularly to the AI-generated image. More particularly, the present invention relates to the distortion-free watermarking method and prevents visual paraphrasing.

BACKGROUND OF THE INVENTION

[002] Image watermarking is a technique that protects digital images from unauthorized access, usage and distribution. Typically, image watermarking involves an embedding of a visible and invisible mark onto the image to protect the image from attack. A report by the European Union Law Enforcement Agency predicts that by 2026, up to 90% of online content could be synthetically generated which raises concerns among policymakers, who cautioned that ”Generative AI could act as a force multiplier for political disinformation which leads to mandate the watermarking of AI-generated images, videos, and audio.

[003] However, concerns remain regarding the vulnerability of invisible watermarking techniques to tampering and the potential for malicious actors to bypass the watermarking entirely. Further, the AI-powered de-watermarking attacks, especially the newly introduced visual paraphrase attack have shown an ability to fully remove watermarks, resulting in a paraphrase of the original image. In visual paraphrase attacks, an image is altered while preserving its core semantic regions, termed as Non-Melting Points (NMPs).

[004] Also, the available watermarking techniques have many gaps and limitations such as watermarks may be easily removed or altered by image editing tools, invisible watermarks are mostly susceptible to compression, cropping, or resizing, which may degrade or destroy the watermark and quality degradation of the watermark. Also, in case someone manages to remove or modify the watermark, the protection is lost. In addition, unauthorized distribution may still occur through methods like screen capture or video recording. Once digital content is distributed widely, tracing and protecting the content using watermarks becomes difficult, especially if the watermarks are not robust enough to survive in different environments or after multiple generations of copying.

Scaling and Distribution Issues: Watermarking becomes more complicated when dealing with the mass distribution of content (e.g., movies or songs), where tracking all instances of piracy may be unfeasible.

[005] A patent application CN117615075A titled "Watermark adding and watermark identifying method, device, equipment and readable storage medium" discloses about a watermark adding and watermark identifying method, device, equipment and readable storage medium; the method comprises the following steps: acquiring information to be processed; inputting information to be processed into a first generation model to obtain first multimedia content; the first multimedia content comprises preset watermark information; wherein the first generation model and the identification model are obtained by combined training; the authentication model is used to authenticate whether the multimedia content is generated for the first generation model.

[006] A non-patent literature titled "The Brittleness of AI-Generated Image Watermarking Techniques: Examining Their Robustness Against Visual Paraphrasing Attacks" discloses about rapid advancement of text-to-image generation systems. However, in this paper, we argue that current image watermarking methods are fragile and susceptible to being circumvented through visual paraphrase attacks. The proposed visual paraphraser operates in two steps. First, it generates a caption for the given image using KOSMOS-2, one of the latest state-of-the-art image captioning systems. Second, it passes both the original image and the generated caption to an image-to-image diffusion system. During the denoising step of the diffusion pipeline, the system generates a visually similar image that is guided by the text caption. The resulting image is a visual paraphrase and is free of any watermarks. Further, the disclosed subject matter provides a critical assessment, empirically revealing the vulnerability of existing watermarking techniques to visual paraphrase attacks.

[007] Therefore, the available techniques for the visual paraphrasing are not accurate and fails to prevent attacks on images. Also, there is a need of a method for visual paraphrase attack safe and distortion free image watermarking for AI- generated images.

OBJECTIVE OF THE INVENTION

[008] The primary objective of the present invention is to provide a method for visual paraphrase attack safe and distortion free image watermarking for AI- generated images.

[009] The primary objective of the present invention is to provide the method for visual paraphrase attack safe and distortion-free image watermarking for AI- generated images to prevent invisible watermarking.

[0010] The primary objective of the present invention is to provide the method for visual paraphrase attack safe and distortion-free image watermarking for AI- generated images for preventing de-watermarking attacks, especially visual paraphrase attack.

[0011] The primary objective of the present invention is to provide the method for visual paraphrase attack safe and distortion-free image watermarking for AI- generated images for fully removing watermarks which caused a paraphrase of the original image.

[0012] Another objective of the present invention is to provide the method for visual paraphrase attack safe and distortion-free image watermarking for AI- generated images preventing counter reverse engineering efforts aimed at locating NMPs to disrupt the embedded watermark.

[0013] Yet another objective of the present invention is to provide the method for visual paraphrase attack safe and distortion-free image watermarking for AI- generated images which is accurate.

[0014] Other objectives and advantages of the present invention will become apparent from the following description taken in connection with the accompanying drawings, wherein, by way of illustration and example, the aspects of the present invention are disclosed.

SUMMARY OF THE INVENTION

[0015] The present invention relates to a method for visual paraphrase attack safe and distortion-free image watermarking for AI- generated images comprising steps of identification of one or more stable regions in an image unaffected by paraphrasing. At least five paraphrased images are generated such as paraphrase 1, paraphrase 2, paraphrase 3, paraphrase 4 and paraphrase 5. Intersection over union (IoU) is applied, across variations, to locate non-melting points. Further, the potential non-melting points (NMPs) are identified by using non-maximum suppression (NMS). Further, the non-maximum suppression (NMS) merges highly overlapping boxes, ensuring that only the most representative areas are retained, resulting in a clean set of stable

regions across the images. The one or more watermark signals are embedded with the identified non-melting points (NMPs) using a multi-channel strength watermarking approach. Further, resilience of the NMP-embedded watermarks are assessed against paraphrasing. The NMP-embedded watermarks are paraphrased by using one or more techniques and adaptive image enhancement is applied on the image. The adaptive image enhancement is applied to improve watermarked image quality by blending it with the original for balancing quality and watermark strength. The robustness is analyzed of the watermark by evaluating one or more parameters.

[0016] The method of identifying the non-melting points in the image as claimed in claim 1, wherein the method comprises step of detecting salient region in the image based on a number of features. The potential non-melting point is identified by using non-maximum suppression (NMS). Further, the non-maximum suppression (NMS) merges highly overlapping boxes, ensuring that only the most representative areas are retained, resulting in a clean set of stable regions across the images. A stability score, based on the frequency of the boxes across one or more images, to each of the boxes, with lower scores indicating higher occurrences across multiple paraphrased images.

[0017] Further, the non-melting points (NMS) merges highly overlapping boxes ensuring that only the most representative areas are retained, resulting in a clean set of stable regions across the one or more images. Further, the adding adversarial noise to the watermarked image, which disrupts the detection of the one or more stable regions. Further, the one or more techniques include but not limited to a random patching and a noisy burnishing. Further, the random patching is used to embed additional watermarks. Further, the noisy burnishing is used to prevent NMP detection. Furher, the location of the non-melting points (NMS) and the non-melting points (NMS) bounding boxes are encrypted through a security key, for ensuring security and practicality, decrypted by an authenticate user. Further, the one or more parameters including but not limited to distortion and detectability.

[0018] A system for implementing the method for visual paraphrase attack safe and distortion-free image watermarking for AI- generated images comprising a detecting unit, an image processing unit, an embedding unit, a processing unit and an analyzing unit. Further, the detecting unit is installed to detect one or more stable regions in an image unaffected by paraphrasing. Further, the image processing unit is installed to process the images for identifying the non-melting point

(NMP). Further, the embedding unit is installed for merging the one or more watermark signals with the identified non-melting points (NMPs). Further, the processing unit is installed to process the images to ensure the prevention of the paraphrasing of the NMP-embedded watermark and ensure watermarked image quality. Further, the analyzing unit is installed for evaluating the robustness of the watermark.

BRIEF DESCRIPTION OF DRAWINGS

[0019] The present invention will be better understood after reading the following detailed description of the presently preferred aspects thereof with reference to the appended drawings, in which the features, other aspects and advantages of certain exemplary embodiments of the invention will be more apparent from the accompanying drawing in which:

[0020] Figure 1(a) illustrates a flow chart of a method of visual paraphrase attack safe and distortion-free image watermarking technique for AI-generated images;

[0021] Figure 1(b) illustrates a pictorial representation of the method for image watermarking ;

[0022] Figure 2(a) illustrates a pictorial representation of comparison of various methods of the saliency detection of the image;

[0023] Figure 2(b) illustrates a pictorial representation of the non-melting points (NMPs) in the paraphrased images;

[0024] Figure 3 illustrates progression of multichannel watermarking at different diffusion steps;

[0025] Figure 4 illustrates a pictorial representation of noisy burnishing disrupts saliency detection in watermarked images hindering attackers from locating (NMPs) or altering watermarked areas;

[0026] Figure 5 illustrates a pictorial representation of a method of embedding encrypted metadata with NMP locations in images.

[0027] Figure 6 illustrates a pictorial representation of a comparison of two sets of images before and after adaptive enhancement;

[0028] Figure 7 illustrates a pictorial representation of the impact of successive paraphrasing attacks on PECCAVI (green) and ZoDiac (black) watermarked images with detection scores; and

[0029] Figure 8 illustrates a flow chart for a method of watermark detection.

DETAILED DESCRIPTION OF INVENTION

[0030] The following detailed description and embodiments set forth herein below are merely exemplary out of the wide variety and arrangement of instructions which can be employed with the present invention. The present invention may be embodied in other specific forms without departing from the spirit or essential characteristics thereof. All the features disclosed in this specification may be replaced by similar other or alternative features performing similar or same or equivalent purposes. Thus, unless expressly stated otherwise, they all are within the scope of the present invention.

[0031] Accordingly, those of ordinary skill in the art will recognize that various changes and modifications of the embodiments described herein can be made without departing from the scope of the invention. In addition, descriptions of well-known functions and constructions are omitted for clarity and conciseness.

[0032] The terms and words used in the following description and claims are not limited to the bibliographical meanings but are merely used to enable a clear and consistent understanding of the invention. Accordingly, it should be apparent to those skilled in the art that the following description of exemplary embodiments of the present invention are provided for illustration purpose only and not for the purpose of limiting the invention.

[0033] It is to be understood that the singular forms “a,” “an,” and “the” include plural referents unless the context clearly dictates otherwise.

[0034] It should be emphasized that the term “comprises/comprising” when used in this specification is taken to specify the presence of stated features, integers, steps, or components but

does not preclude the presence or addition of one or more other features, integers, steps, components, or groups thereof.

[0035] The present invention relates to a method of visual paraphrase attack safe and distortion free image watermarking technique for AI-generated images. The method strategically embeds watermarks within stable regions known as a non-melting point (NMPs). The method also incorporates noisy burnishing to counter reverse engineering efforts aimed at locating the NMPs to disrupt the embedded watermark, thereby enhancing durability.

[0036] This method introduces PECCAVI, the first visual paraphrase attack safe and distortion free image watermarking technique. In visual paraphrase attacks, an image is altered while preserving its core semantic regions, termed Non-Melting Points (NMPs). The PECCAVI strategically embeds watermarks within the NMPs and employs multi-channel frequency domain watermarking. The method also incorporates noisy burnishing to counter reverse engineering efforts aimed at locating the NMPs to disrupt the embedded watermark, thereby enhancing durability. The concept of visual paraphrasing attack refers to generating variations of an image that retain the same semantic content while altering the visual presentation. Further, the method includes an image-to-image paraphrased approach that ensures reliable adherence to the original image's appearance and meaning, even within this variable parameter space, delivering a consistent and structurally faithful visual paraphrase.

[0037] This method introduces PECCAVI is the first visual paraphrase attack-safe, distortion-free image watermarking technique. With the rise of AI-generated misinformation, the PECCAVI may contribute significantly to the greater social good. The method surpasses existing watermarking existing techniques which it requires substantial computational resources.

[0038] The method (PECCAVI), for visual paraphrase attack safe and distortion-free image watermarking for AI-generated images, is focused on the placement of the watermark, method of watermarking, the need for a more sophisticated detection mechanism, assessment method resistance to visual paraphrase attacks, and whether the watermarking process distorts the 1original concept excessively.

[0039] Figure 1(a) in combination with figure 1(b) illustrates the method of visual paraphrase attack safe and distortion-free image watermarking technique for AI-generated images. Figure 1(b) illustrates steps of the method for image watermarking encompasses NMP detection, multi-channel watermark embedding, noisy burnishing, and encrypted metadata addition. These components collectively ensure robust, low-distortion watermarks that resist paraphrase attacks, safeguarding AI-generated images from unauthorized alterations.

[0040] Further, the method comprises steps of identification of one or more stable regions that are unaffected by paraphrasing, at step 102. Further, at least five paraphrases are generated, at step 104. Further, the non-melting points are located, across variations, by applying intersection over union (IoU), at step 106. Further, the potential non-melting point (NMP) is identified by using non-maximum suppression (NMS). Further, the non-maximum suppression (NMS) merges highly overlapping boxes, ensuring that only the most representative areas are retained, resulting in a clean set of stable regions across the images, at step 108. Further, the non-melting points (NMPs) are embedded with one or more watermark signals, at step 110. Further, the NMP-embedded watermarks are subjected to resilience assessment against paraphrasing, at Step 112. Further, the NMP-embedded watermarks are prevented from paraphrasing by using one or more techniques, Step 114. Further, the watermarked image quality is improved by blending it with the original image for balancing quality and watermark strength by applying adaptive image enhancement, at step 116.

[0041] Further, the method of identifying the non-melting points in the image comprises the detection of salient regions in the image based on a number of features of the images. Further, the features may include but not limited to color contrast, texture, or edges. Further, the potential non-melting point is identified by using non-maximum suppression (NMS). Further, the non-maximum suppression (NMS) merges highly overlapping boxes, ensuring that only the most representative areas are retained, resulting in a clean set of stable regions across the images. Further, a stability score is assigned, based on the frequency of the boxes across one or more images, to each of the boxes, with lower scores indicating higher occurrences across multiple paraphrased images.

[0042] In the preferred embodiment, at first, the method for visual paraphrasing attacks safe and distortion-free image watermarking for AI-generated images comprises steps of, an identification of one or more stable regions unaffected by paraphrasing. The stable regions are ideal for embedding watermark signals, as the stable regions are less likely to be altered. Also, the stable regions are referred as non-melting points or non-melting regions.

[0043] Further, the non-melting points (NMPs) detection involves two main steps including salient region detection and non-melting points (NMPs) location. Further, the salient region detection in image processing identifies the most "salient" or visually prominent areas within the image, based on unique features like color contrast, texture, or edges. Further, the saliency detection is performed using XRAI method, as shown in Figure 2.

[0044] In an another embodiment, the salient region detection is performed through a number of methods including Vanilla Gradient, Smooth Grad, Vanilla Integrated (Vanilla I Grad), Smooth Grad Integrated (Smooth I Grad), XRAI, Grad CAM, Smooth Grad Grad CAM, and MSI-Net, as shown in figure 2(a).

[0045] Figure 2(b) NMPs across various paraphrased versions are displayed. Each row corresponds to a different paraphrase, with four columns: the original image, saliency map, top salient regions highlighted in red, least in blue, and refined bounding boxes obtained after applying NMS and IoU. Further, the non-melting points (NMPs) are located by identifying the most stable areas in the image.

[0046] Further, at-least five automatic paraphrases are generated. After identifying key regions in each paraphrased image, applying Intersection over Union (IoU) to locate the most stable areas across variations, referred to as the non-melting points (NMPs). The intersection over union (IoU) quantifies the overlap between boxes across images, highlighting regions that consistently appear in similar locations. Once potential NMPs are identified which is selected using non-maximum suppression (NMS). The non-maximum suppression (NMS) merges highly overlapping boxes, ensuring that only the most representative areas are retained, resulting in a clean set of stable regions across images. Each box is then assigned a stability score, based on the frequency of the

box, across images, with lower scores indicating higher occurrences across the paraphrased images. This score reflects each region's robustness as the non-melting points (NMPs). If no stable regions are found, a default box is added to ensure robustness.

[0047] Figure 3 illustrates progression of multichannel watermarking at different diffusion steps. Each step shows how the watermark integrates into the image through the multi-channel frequency domain. Further, the one or more watermark signals are embedded with the identified non-melting points (NMPs) using a multi-channel strength watermarking approach to enhance robustness against de-watermarking attacks. Further, the watermark strength is determined before embedding the watermark with the non-melting points (NMPs). As stronger paraphrasing removes watermarks more effectively. Therefore, using a higher watermark strength in such non-melting points (NMPs) to make the method more resilient. Watermark strength is determined by the distance between rings within the watermark, with smaller distances indicating greater strength. For example, Channel 4 shows a smaller ring distance (0.5), while Channel 3 reflects a larger distance (0.75), as shown in figure 3. Strength values range from 0 to 1.0, depending on the number of paraphrases containing the NMP: $Ws = \max(0.1, 1 - 0.25 \cdot (n - 1))$, $n \in \{1, 2, 3, 4, 5\}$. Here n represents the number of regions that an NMP appears in out of the at least five paraphrases. Further, the watermark is embedded across multiple channels, which provides higher detection scores due to overlapping bounding box watermarks.

[0048] Further, the NMP-embedded watermarks are assessed against paraphrasing. Further, the NMP-based watermarking faces two key challenges including assessing the resilience of NMP-embedded watermarks against further paraphrasing and anticipating potential countermeasures from attackers who may reverse engineer methods to detect and distort NMPs, reducing watermark detectability. Two strategies are implemented to ensure robustness of the NMP-embedded watermarks including random patching to embed additional watermarks, and noisy burnishing to prevent NMP detection.

[0049] Further, the random patching is implemented to enhance the security of the non-melting points (NMPs). Once all NMPs are detected and saved, the smallest NMPs are identified. Further, an additional NMP is generated, of the same shape as the smallest NMP, at a random, non-overlapping location. The selection may be randomized using a vendor-specific pseudo-random

algorithm. Watermarks are then embedded in these randomly placed NMPs, similar to the original NMPs, using either single-channel or multi-channel approaches.

[0050] Figure 4 illustrates noisy burnishing disrupts saliency detection in watermarked images hindering attackers from locating (NMPs) or altering watermarked areas. This technique preserves the frequency-based watermark, ensuring high detectability while enhancing security against tampering. Attackers may attempt to identify salient regions of the image to remove the watermark. This may be countered by adding adversarial noise to the watermarked image, which disrupts the detection of these salient regions. Noisy burnishing disrupts saliency detection in watermarked images hindering attackers from locating NMPs or altering watermarked areas. The method preserves the frequency-based watermark, ensuring high detectability while enhancing security against tampering.

[0051] Figure 5 illustrates a pictorial representation of embedding encrypted metadata with NMP locations in images, ensuring secure watermark verification while meeting regulatory standards and resisting reverse engineering by hiding precise watermark coordinates. A significant challenge is enabling vendors or text to image system providers to verify that the image is PECCAVI-watermarked without knowing the exact locations of random patches or NMPs. Further, the method embeds encrypted metadata containing this information—random patch locations and NMP bounding boxes. Using a secure key, vendors may decrypt the metadata to verify the watermark, ensuring both security and practicality.

[0052] Further, the method aligns with California's mandate for accessible AI detection tools and robust watermarking for AI-generated content, as shown in figure 5. While noisy burnishing locations are stored, one might question storing random patch data. If vendors know the pseudo-random patch algorithm, they could compute these locations, but storing the computed locations reduces computational overhead and allows flexibility in case the algorithm changes.

[0053] Further, adaptive image enhancement is applied to improve the watermarked image quality by blending it with the original image for balancing quality and watermark strength. Adaptive image enhancement is applied to improve watermarked image quality by blending it with the original $\bar{x}_0 = \hat{x}_0 + \gamma(x_0 - \hat{x}_0)$ where $\gamma \in [0, 1]$ balances quality and watermark strength. The

goal is to find the smallest γ such that similarity $S(\bar{x}_0, \hat{x}_0) \geq s^*$, typically using SSIM, as shown in Figure 6.

[0054] Further, the robustness of the method is analyzed by evaluating one or more parameters. Further, the one or more parameters including but not limited to distortion and detectability. For analyzing image quality distortion, metrics such as PSNR, SSIM, FID and CMMD is used which are further used to measure the watermark's impact on visual fidelity, both perceptually and structurally. For watermark detectability, the method that is PECCAVI's resilience is analyzed against classical attacks like brightness adjustments, Gaussian noise, JPEG compression, and varying paraphrasing strengths using the average WDP. Together, the metrics provide a balanced view of both pixel-level (PSNR, SSIM) and feature-level (FID, CMMD) distortion, helping us assess overall image quality. For distortion metric results, as shown in Table 1.

[0055] A summary of the metrics, presented in Table 1, highlights PECCAVI's effectiveness in preserving high image quality while ensuring robust watermark retention under diverse attack scenarios. Meta's Watermark Anything Model (WAM) enables imperceptible, localized image watermarking, embedding, locating, and decoding multiple watermarks in small regions of high-resolution images. Our evaluation shows that PECCAVI outperforms WAM in resisting visual paraphrasing attacks. For image quality distortion, we use metrics such as PSNR, SSIM, FID.

[0056] Watermarked image quality is compared in terms of PSNR, SSIM, FID, and CMMD scores. Watermark robustness is compared on the bases of average WDP before and post attacks on MS-COCO dataset. We evaluate on four basic attacks and on diffusion based paraphrase attack with different strength values. For any given watermarked image, its paraphrases are prepared using its captions from the dataset for varying strength values and a fixed guidance scale value of 7.5. For an image, the watermark detection probability is then calculated by averaging the probabilities over the paraphrases, and the final value is obtained by averaging over the whole data subset. λ refers to the percentile of salient regions being considered when using the PECCAVI scheme of watermarking.

[0057] In an another embodiment, the efficacy of PECCAVI is evaluated across diverse text to image (T2I) models, including Stable Diffusion, Stable Diffusion XL (SDXL), Stable Diffusion 2.1 (SD 2.1), DALL-E 3, and Midjourney 6. Further, the process produced a dataset known as MS

COCOAI, where captions and images from the original MS COCO dataset fed into the text to image models to generate and store corresponding images and results are presented in Table 1.

[0058] In an another embodiment, paraphrased images are generated at different strengths, with lower s-values keeping more original details and higher values allowing greater alteration. WDP assesses watermark retention, while SSIM measures similarity to the original image. PECCAVI shows high WDP at lower strengths, retaining watermark integrity even through moderate paraphrasing.

[0059] In an another embodiment, the method is compared with a Meta's Watermark Anything Model (WAM) which enables imperceptible, localized image watermarking, embedding, locating, and decoding multiple watermarks in small regions of high-resolution images. And, evaluation shows that PECCAVI outperforms WAM in resisting visual paraphrasing attacks, as shown in figure 7.

[0060] Figure 7 illustrates the cumulative impact of successive paraphrasing attacks on PECCAVI (green) and ZoDiac (black) watermarked images is depicted, with detection scores. PECCAVI shows superior resilience, maintaining stable scores under high-strength paraphrasing attacks, demonstrating its durability over ZoDiac.

[0061] In an another embodiment, the performance of PECCAVI watermarking scheme is comapred with various post-process image watermarking methods. The methods are compared under a number of attack schemes such as Brightness Enhancement with a factor of 0.5, Gaussian Noise with a std of 0.05, JPEG compression with a quality factor of 50, and Visual Paraphrasing, using stable-diffusion-xl-base-1.0 with image captions and paraphrase strengths of 0.1 and 0.2. Further test the method on VAE-based image compression model having a quality setting of stable diffusion-based image regeneration model with denoising steps using stable-diffusion-2-1-base. The results are produced on 500 images randomly sampled from the COCO Dataset, as shown in Table1.

[0062] Table 1- Watermarked image quality is compared in terms of PSNR, SSIM, FID, and CMMD scores.

		Image Quality				Avg. Watermark Detection Probability (WDP)					
						Pre-Attack	Post-Attack				
Method	λ	PSNR	SSIM	FID	CMMD		Brightness	Gaussian Noise	JPEG	Paraphrase (s=0.1)	Paraphrase (s=0.2)
DwtDctSVD	-	41.04	0.988	1.447	8.88e-06	0.98	0.01	0.14	0.65	0.00	0.00
Stable Signature	-	42.91	0.98	0.34	3.41e-06	0.99	0.75	0.73	0.65	0.59	0.51
WAM	-	46.05	1.00	0.33	0.01	1.00	0.62	0.61	0.58	0.63	0.56
ZoDiac	-	28.47	0.92	124.76	1.14e-3	1.00	0.92	0.90	0.89	0.81	0.70
PECCAVI with different saliency methods											
PECCAVI (Vanilla Integrated)	Top 30	31.50	0.95	17.61	2.48e-5	0.96	0.94	0.93	0.93	0.68	0.64
	Top 40	31.26	0.94	25.82	2.97e-5	0.94	0.94	0.94	0.95	0.72	0.68
	Top 50	31.31	0.95	22.86	4.48e-5	0.96	0.94	0.95	0.95	0.71	0.67
PECCAVI (MSI Net)	Top 30	30.64	0.94	20.81	3.1e-5	0.98	0.94	0.95	0.95	0.83	0.77
	Top 40	30.57	0.94	21.96	4.7e-5	0.99	0.98	0.98	0.97	0.83	0.78
	Top 50	30.71	0.94	23.39	3.9e-5	0.99	0.97	0.98	0.97	0.85	0.81
PECCAVI (XRAI)	Top 30	29.56	0.93	55.60	5.7e-5	0.99	0.98	0.98	0.98	0.91	0.85
	Top 40	29.87	0.93	34.12	4.7e-5	0.99	0.99	0.99	0.98	0.89	0.84
	Top 50	29.84	0.93	31.61	3.6e-5	0.99	0.98	0.98	0.99	0.89	0.84

[0063] Table 1 shows Watermarked image quality is compared in terms of PSNR, SSIM, FID, and CMMD scores. Watermark robustness is compared on the bases of Average WDP before and post attacks on MS-COCO dataset. We evaluate on 4 basic attacks and on diffusion based paraphrase attack with different strength values. For any given watermarked image, its paraphrases are prepared using its captions from the dataset for varying strength values and a fixed guidance scale value of 7.5. For an image, the watermark detection probability is then calculated by averaging the probabilites over the paraphrases, and final value is obtained by averaging over the whole data subset. λ refers the percentile of salient regions being considered when using PECCAVI scheme of watermarking.

[0064] PECCAVI - Watermark Detection The PECCAVI watermark detection process, illustrated in figure 8, employs reverse-engineering to retrace the likely steps used to embed the watermark. The method begins by retrieving and decrypting metadata containing NMP and random patch locations. Further, the method involves scanning of multiple channels for these watermark signals and ultimately provides a watermark detection probability for the image.

[0065] A system for implementing the method for visual paraphrase attack safe and distortion-free image watermarking for AI-generated images comprising a detecting unit, an image processing unit, an embedding unit, a processing unit and an analyzing unit. Further, the detecting unit detects one or more stable regions in an image unaffected by paraphrasing. Further, the image processing unit processes the images to identify the non-melting point (NMP). Further, the embedding unit for merging one or more watermark signals with the identified non-melting points (NMPs). Further, the processing unit to process the images to ensure the prevention of the

paraphrasing of the NMP-embedded watermark and ensure watermarked image quality. Further, analyzing unit for evaluating the robustness of the watermark.

[0066] While this invention has been described in connection with what is presently considered to be the most practical and preferred embodiment, it is to be understood that the invention is not limited to the disclosed embodiments, but, on the contrary, is intended to cover various modifications and equivalent arrangements included within the scope of the appended claims.

1. A method for visual paraphrase attack safe and distortion-free image watermarking for AI-generated images comprising;

Step 1: identifying one or more stable regions in an image unaffected by paraphrasing;

Step 2: generating at least five paraphrased image;

Step 3: applying intersection over union (IoU), across variations, to locate non-melting points;

Step 4: identifying the potential non-melting point (NMP) by using non-maximum suppression (NMS), wherein the non-maximum suppression (NMS) merges highly overlapping boxes, ensuring that only the most representative areas are retained, resulting in a clean set of stable regions across the images;

Step 5: embedding one or more watermark signals with the identified non-melting points (NMPs) using a multi-channel strength watermarking approach;

Step 6: assessing resilience of the NMP-embedded watermarks against paraphrasing;

Step 7: preventing paraphrasing of the NMP-embedded watermarks by using one or more techniques;

Step 8: applying adaptive image enhancement to improve watermarked image quality by blending it with the original for balancing quality and watermark strength; and

Step 9: analyzing the robustness of the watermark by evaluating one or more parameters.

2. A method of identifying the non-melting points in the image as claimed in claim 1, wherein the method comprises:

Step 1: detecting salient region in the image based on a number of features;

Step 2: identifying the potential non-melting point by using non-maximum suppression (NMS), wherein the non-maximum suppression (NMS) merges highly overlapping boxes, ensuring that only the most representative areas are retained, resulting in a clean set of stable regions across the images; and

Step 3: assigning a stability score, based on the frequency of the boxes across one or more images, to each of the boxes, with lower scores indicating higher occurrences across multiple paraphrased images.

3. The method (100) as claimed in claim 1, wherein non-melting points (NMS) merges highly overlapping boxes ensuring that only the most representative areas are retained, resulting in a clean set of stable regions across the one or more images.

4. The device (100) as claimed in claim 1, wherein the adding adversarial noise to the watermarked image, which disrupts the detection of the one or more stable regions.

5. The device (100) as claimed in claim 1, wherein the one or more techniques include but not limited to a random patching and a noisy burnishing.

6. The device (100) as claimed in claim 1, wherein the random patching is used to embed additional watermarks.

7. The device (100) as claimed in claim 1, wherein the noisy burnishing is used to prevent NMP detection.

8. The device (100) as claimed in claim 1, wherein the location of the non-melting points (NMS) and the non-melting points (NMS) bounding boxes are encrypted through a security key, for ensuring security and practicality, decrypted by an authenticate user.

9. The device (100) as claimed in claim 1, wherein the one or more parameters including but not limited to distortion and detectability.

10. A method for watermark detection in the image as claimed in claim 1, wherein the method, comprising steps of -

i. retrieving and decrypting metadata containing non melting points (NMPs) and random patch locations;

ii. scanning multiple channels for the watermark signals present in the image; and

iii. providing a watermark detection probability for the image.

11. A system for implementing the method for visual paraphrase attack safe and distortion-free image watermarking for AI- generated images comprising:

(a) a detecting unit to detect one or more stable regions in an image unaffected by paraphrasing;

(b) an image processing unit to process the images for identifying the non-melting point (NMP);

(c) an embedding unit for merging the one or more watermark signals with the identified non-melting points (NMPs);

(d) a processing unit to process the images to ensure the prevention of the paraphrasing of the NMP-embedded watermark and ensure watermarked image quality; and

(e) an analyzing unit for evaluating the robustness of the watermark.

ABSTRACT

METHOD FOR VISUAL PARAPHRASE ATTACK SAFE AND DISTORTION FREE IMAGE WATERMARKING FOR AI- GENERATED IMAGES

The present invention relates to a method for visual paraphrase attack safe and distortion-free image watermarking for AI- generated images comprising steps of identification of one or more stable regions in an image unaffected by paraphrasing. At least five paraphrased images generated. Intersection over union (IoU) applied, across variations, to locate non-melting points. The potential non-melting point (NMP) is identified by using non-maximum suppression (NMS). Further, the non-maximum suppression (NMS) merges highly overlapping boxes, ensuring that only the most representative areas are retained, resulting in a clean set of stable regions across the images. One or more watermark signals is embedded with the identified non-melting points (NMPs) using a multi-channel strength watermarking approach. The NMP-embedded watermarks is assessed against paraphrasing. The NMP-embedded watermarks prevented from paraphrasing. Adaptive image enhancement is applied to improve watermarked image quality by blending it with the original for balancing quality and watermark strength.

Figure 1(a-b)

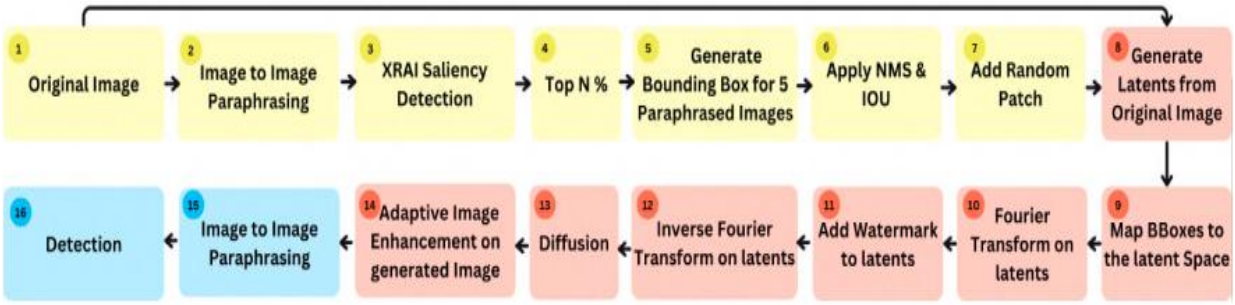

Figure 1(a)

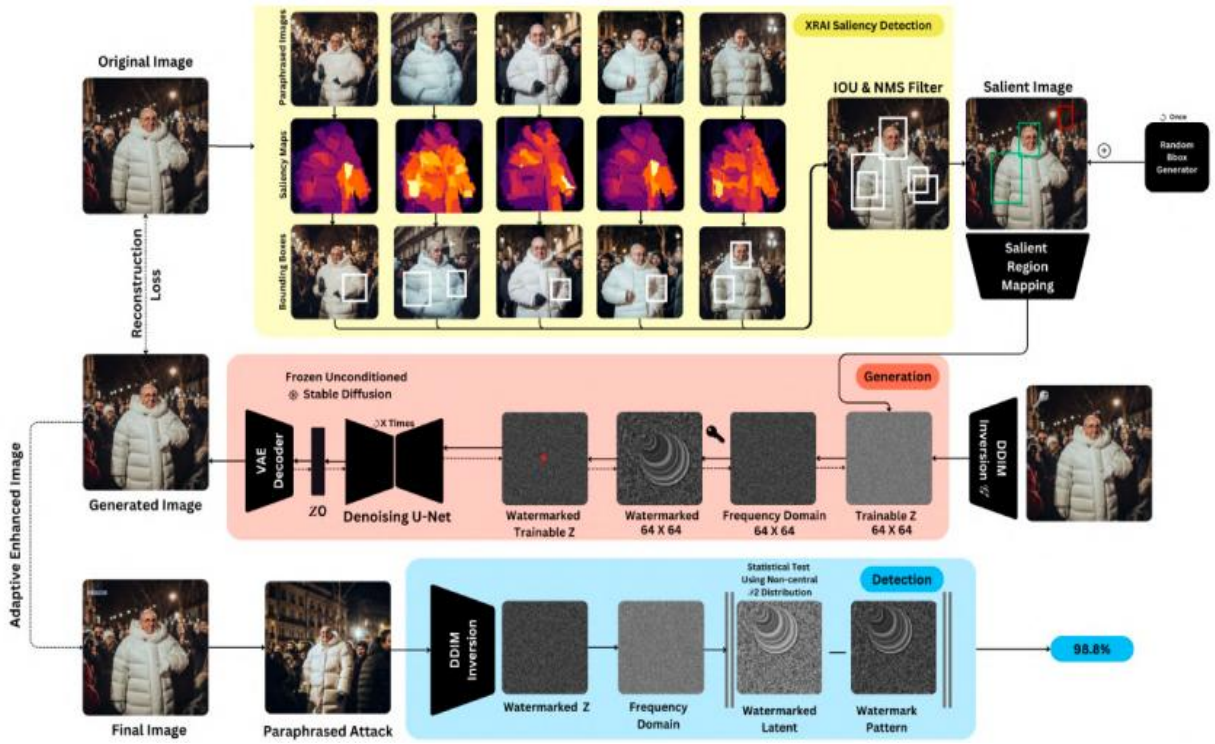

Figure 1(b)

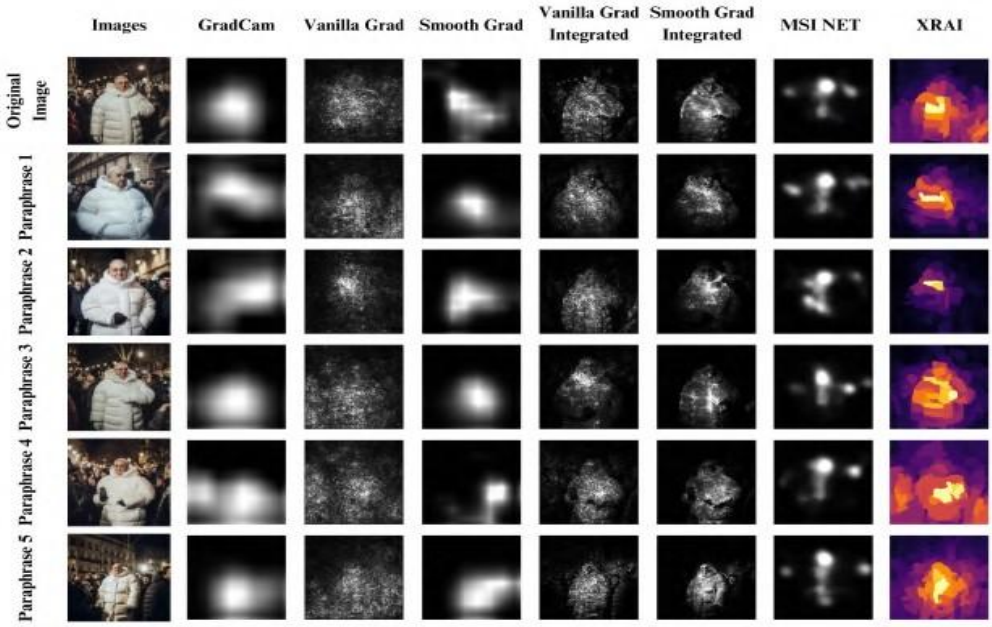

Figure 2 (a)

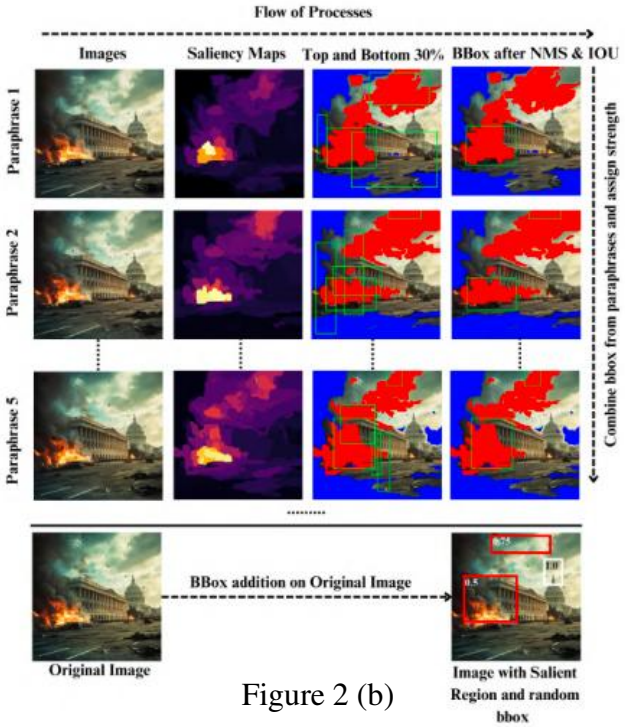

Figure 2 (b)

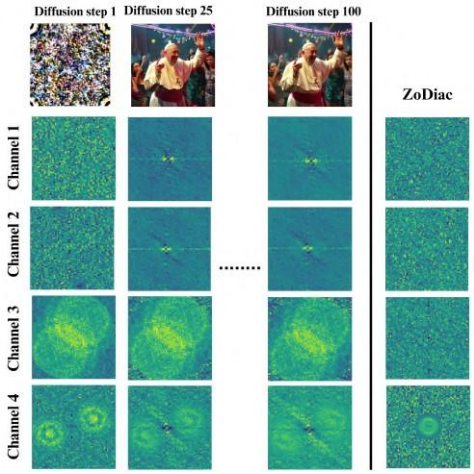

Figure 3

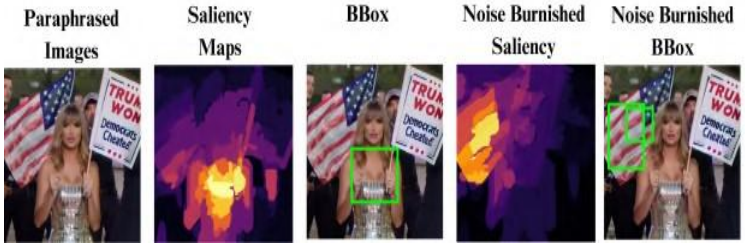

Figure 4

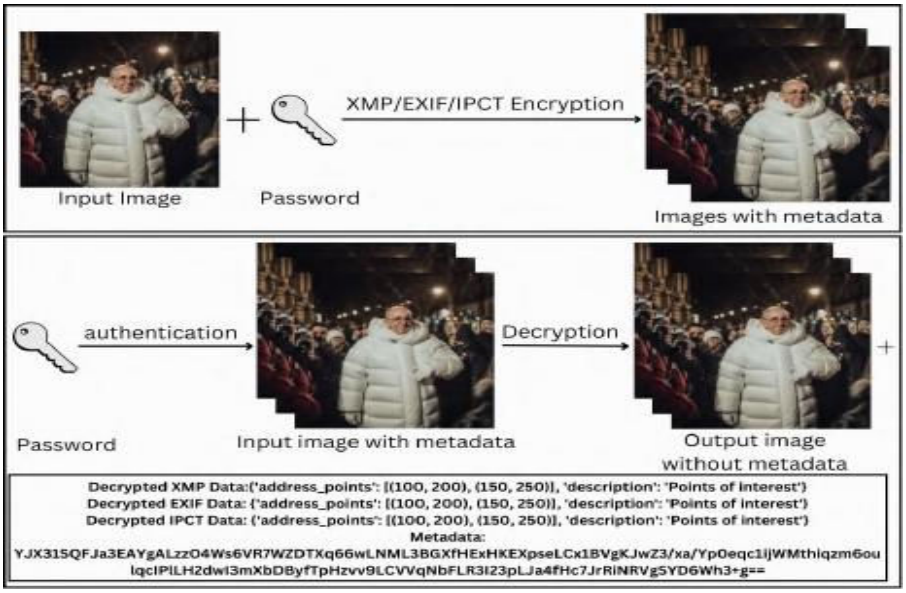

Figure 5

Figure 6

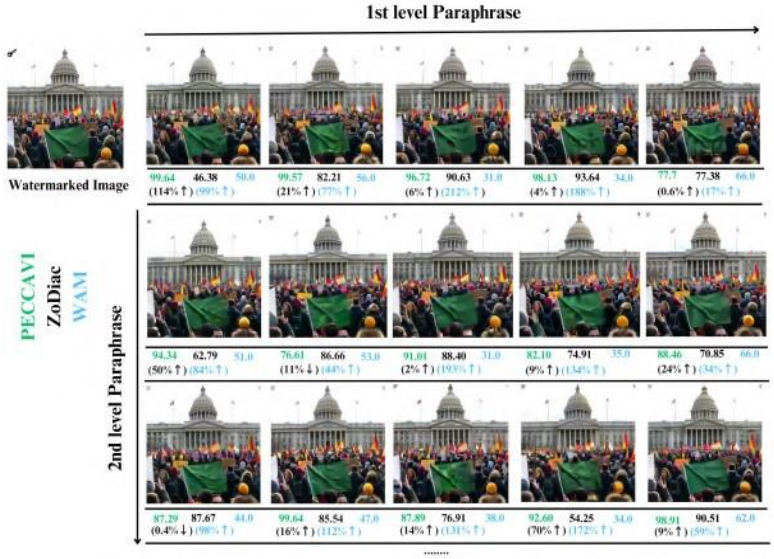

Figure 7


PECCAVI
Watermarked
Image
Decrypt random patches and
NMPs coordinates from metadata
if present
Check Watermark Signals on
Multichannel

Figure 8