Global Relational Probing in Vision Transformers: Approach to Detect AI Generated Images

Bijendra Tyagi[1], Ganesh Kumar Yadav[2], Harshvardhan Singh[3],
Anoushka Sinha[4], Rohit Pandey[5], Khushi Gupta[6]
{ bijendra.tyagi@jssaten.ac.in[1], gky14jul@gmail.com [2], 22csaiml062@jssaten.ac.in[3],
22csaiml117@jssaten.ac.in[4], rohitpandey73551@gmail.com[5], 22csaiml033@jssaten.ac.in[6]}

JSS University, Noida, India[1,2,3,4,5,6]

Abstract. The increasing use of a generative AI like diffusion models has now set up this ongoing struggle that will not easily end between those generating fake images and those who try to identifying them. The very first techniques tried to find technical errors, like finding odd frequency patterns or leftover artifacts from the generation. These worked fine on older AI but once they are applied to newer generators, the methods seem not to work anymore. This review paper looks at the latest work with AI for detecting fake images and divides people's methods into two basic types. One side of researchers continue to focus on artifacts and hope to get better at catching tiny technical mistakes. The alternative group is going towards looking out for what are called semantic oddities. The middle approach is where something like usual Vision Transformer gets it done just by itself. This proposed prototype is named Global Relational Probing (GRP). The main idea is easy to get; an usual ViT gets tuned, and then the self-attention layers look for important high-level semantic mismatches. To check if this technique actually is effective, workflow was set up that is easy to follow by other people. All of it runs with open source programs and the dataset collections anyone can use, also relies on explainable AI so everybody can see what system decided and why. GRP is built to survive new kinds of threats, can work with many kinds of the image generators, and does not depend on strong computer power.

Keywords: AI-Generated Image Detection, Deepfake, Vision Transformer (ViT), Explainable AI (XAI), Semantic Inconsistency, Forensics, Generalization

1 Introduction

Generative AI has reached a stage where it gets pretty impossible to notice if a photograph is real or computer made, just by seeing it. Programs like DALL-E 3, Stable Diffusion and Midjourney are making pictures that come out really professional looking. Often they could be mistaken for

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2363243

someone who has a camera setup that costs a lot [1, 2] . However, there is negative aspect to it too. People are using these fake images in ways that disturb elections make up advertisements with celebrities that ruin their names, and overall contribute to mistrust among people [3, 4].

Experts working in digital forensics have spent extra time developing software that decides which images are real and which ones are artificially produced. At first it was like a game where different tools detected special mistakes noticeable in the older software such as funny pixel effects that come when GANs perform upsampling [5, 6]. This method worked for a period, but main difficulty is, those specific issues do not last long, every update corrects them and shifts problem somewhere else [7]. Hence, the detector might handle old versions but fails with new ones. It becomes close to impossible to find detector that can work for all future models.

Still, there is a particular problem that never seems to go away i.e. these models do not fully grasp the rules of physical reality. They only study millions of flat pictures, so they are able to do impressive things with details like skin or liquids. But if required to analyze more detailed objects in three dimensions or think about causes and effects in the images, they make mistakes and get things wrong [8]. Strange problems keep popping up, for example people have extra fingers, lighting directions look unrealistic, shadows break normal physics rules [9]. Sometimes people overlook such errors at first, but machine detectors can find them easily, these mistakes are reliable clues and they do not vanish when new model versions are released.

This is what our proposed research prototype describes where we talk about how the AI image detection field is going and come up with a pretty direct method to find these deeper errors missing from earlier studies. The focus here is on Global Relational Probing (GRP)—it basically means this paper believes that an ordinary Vision Transformer [10], without need for complex changes, could reveal large-scale relationship mistakes simply by employing its built-in attention ability.

This paper provides: 1) A detailed review of different detection techniques, starting from the artifact analysis and also approaches involving a meaning; 2) A clarified explanation of the GRP idea, and why filling this research gap is important; and 3) An actual test framework, which other scientists are able to repeat or alter to suit their own experiments.

2 Literature Survey: The State of AI Image Detection

Research on detecting AI generated images has evolved along three main directions, each based on different assumptions about what makes synthetic images detectable.

2.1 Category 1: The Artifact Hunters

The easiest way of solving the problem is to approach the problem as a simple binary classification task. This is working quite well with CNNs since they are simply good at local picking up patterns and texture stuff [11, 12]. we can determine whether something is synthetic or not using the high frequency noise patterns and subtle artifacts generators leave behind. This is alright when we are testing on the same generators which we trained on [**13**] but this is where the test fails- use a different generator and the performance is gone. A ProGAN-trained detector (such as the one in [13]) fundamentally cannot work at all once given images produced by Stable Diffusion (as in [1]

). Worse still, even basic operations such as JPEG compression or resizing an image may entirely conceal these artifacts of their existence even [5, 7]. And the result is the detectors that are easily broken.

2.2 Category 2: The Hybrid Architects

It was found that CNNs alone were not working well, and so people began combining them with transformers to extract the local detail as well as the global context knowledge. This can be done by models such as GenConViT [14] and CViT models which have convolutional layers, followed up by passing to ViT parts, which are able to look at the bigger picture, when necessary, because they can do just that by doing so [15]. DaViT further extends it, it even has these dual attention mechanisms operating at the same time, across spatial location as well as channels in the feature map, [16]. These hybrid configurations are better than pure CNNs, but they also complicate the architecture. And on a day-to-day basis, they end up doing the same artifact detection.

2.3 Category 3: The Semantic Reasoners

The other more recent approach has changed gears to what AI generators actually do not do well i.e. maintaining things consistent with the way the real world operates. Semantic tests Benchmarks such as AnomReason [2] are used to test whether detectors can identify semantic violations or objects which are physically impossible. In order to address these more difficult problems, researchers began to employ Vision Language Models and Large Language Models capable of reasoning about situations where things do not make sense in their contexts [12, 17]. This strategy is conceptually sound and it usually works fairly well. However, there is a drawback because these models are computationally costly. It renders them unfeasible as far as real time detection is concerned or dealing with bounded resources. It is also simply redundant due to ultimately just doing binary classification.

2.4 The Unexplored Gap: Native ViT Capabilities

The gap between all three of these approaches is something that has not been given a lot of attention (Figure 1). Most studies remain at artifact identification, or pass directly to complicated VLM reasoning, yet there exist an intermediate. The questions such as do standard Vision Transformers without modifications or hybrid forms already have sufficient relational reasoning embedded in them to detect semantic inconsistencies? arise. That is what we really do not know. Other studies such as AnomReason have used ViTs as baseline classifiers but they have not specifically studied whether such models could already contain the semantic understanding that we require to have good detection.

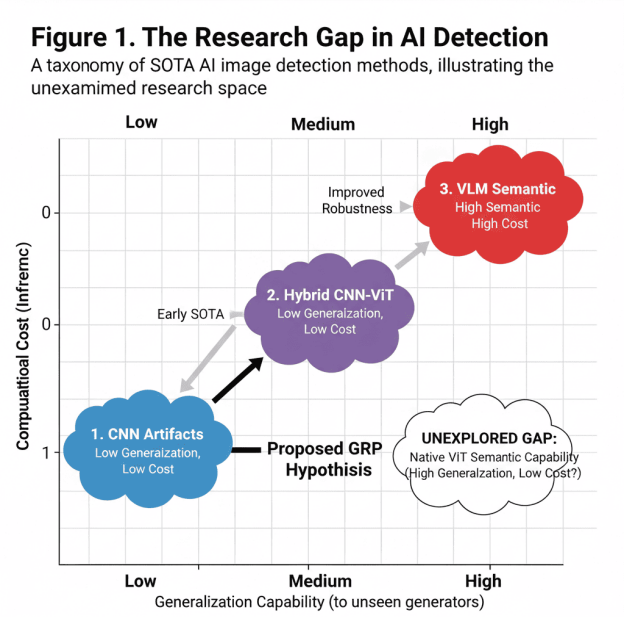


Fig. 1. The middle ground. The field between artifact detection and VLM reasoning, bypassing the question of standard ViTs being able to perform it in the first place, was the field.

3 The GRP Hypothesis: Global Relational Probing

This section presents a focused research premise we call *Global Relational Probing (GRP)*. The core idea is : Vision Transformers might already have the ability to detect semantic inconsistencies just by virtue of how they're designed, without extra reasoning modules or modifications.

3.1 Architectural Foundation

Primarily, a ViT is a global relational learner. The distinction that is most important compared to CNNs is evident in how images are processed. The CNNs develop the knowledge bit by bit as views a bit more of the image than the previous layer, and slowly increase the receptive field. ViTs however are inherently learning long range dependencies on the entire image initially, i.e. at the first level, each patch of the image (so-called token), can directly attend to all other patches of the image.

CNNs due to their construction are more inclined towards local texture patterns and local features. ViTs on the one hand cannot help but be forced by their structure to have more globalized perspective on the image. It is this intrinsically acquired awareness of the world that we believe makes them appropriate to the detection of semantic incompatibilities.

3.2 Formal Hypothesis

This paper state the GRP hypothesis as:

The GRP Hypothesis: On a typical pre trained Vision Transformer fine tuned on binary real vs. synthetic classification tasks, the paper finds that it can naturally learn to detect global semantic or physical inconsistencies. The driver of this is its self attention as a *global relational probe* in that it learns to emphasize relationships between semantically related but logically counterintuitive patches. Eg. it'll notice the discrepancy between a light source and an unrealistic shadow it throws.

3.3 Theoretical Basis: Receptive Field vs. Global Context

The actual way CNNs perceive images, initially they begin with simple cases, the first levels are simply picking out the edges and corners then as we delve deeper they are combined into a texture then a pattern then finally into an entire object.

The problem with this approach is that we are required to reason about things that are separated far away. Eg. if there is a lamp in the upper left corner and a shadow in the lower right. Before a CNN can even begin to compare them, both of these regions lamp and shadow must come to the same receptive field [18]. Until that occurs they could be as well in other pictures. The network is looking at them individually hence there are no means of determining whether they are consistent or not.

ViTs do not need to go through this bottleneck. By attending to oneself they are able to bridge those two remote regions at the very beginning [19]. It is possible to observe both the lamp and the shadow in layer one. Immediately the network can begin to reason.

We believe this is enormous in the semantic inconsistencies. CNNs are left with this hierarchical build up strategy they must observe the parts prior to being able to reason concerning the whole. ViTs are not limited to the extent that they can examine relationships in the world in the first place. Upon learning this it appeared to be worth exploring how this architectural benefit can in fact be converted into an improved ability to detect those physical impossibilities that AI generators generate, which triggered the entire GRP notion.

4 Proposed Methodology

For our experiments we wanted to keep three things in mind i.e. make it simple to understand, ensure other researchers can actually replicate what we did, and only use publicly available resources so there's no barriers to entry.

4.1 Models and Baselines

In this study we're basically comparing two approaches to see if our hypothesis holds up:

- **Primary Model (ViT):** We're using a common pre-trained Vision Transformer model namely `google/vit-base-patch16-224` [20]. This is what we are testing the GRP hypothesis

on. What's relevant here is we're using it natural vanilla, from the box. No architectural changes or no hybrids to see what a regular ViT can do on its own.

- **CNN Baseline:** We require an existing strong CNN that captures the classical methods for artifact detection for illustration. We choose for this to go either with ResNet50[**23**] or EfficientNet-B4[21], both are well known and act as a good benchmark to compare ourselves against. These essentially correspond to what occurs when detection is based on local texture patterns and low level features.

4.2 Datasets and Task

- **Task:** Binary classification, real vs synthetic. We are only fine tuning the final layers of each model nothing advanced with no pixel annotations or fancy loss functions. The argument goes that models can work out if they can do so using that which they have an innate knowledge of.
- **Training Data:** We're using large datasets like **GenImage** which has around a million images from like seven different generators, or **DRAGON** that focuses on diffusion models. The real images come from COCO .
- **Test Set 1 – Generalization:** We test on generators the models never saw during training. Like train on SD v1.5, test on v2.1. This shows us whether the detector learned something general or it just memorized specific generator quirks [22].
- **Test Set 2 – Robustness:** We take Test Set 1 and throw common distortions at it JPEG compression (quality 50), Gaussian blur ($\sigma = 3$), and resizing. If the model only learned artifacts this should really hurt performance.
- **Test Set 3 – Probing:** Most important one we created a *Flaw Set* with about 200 images that got semantic problems like impossible shadows, extra limbs, wrong reflections [20]. If GRP actually works the ViT should perform well here.

We fine tune for around 20 epochs using AdamW (learning rate 1e-5) and binary cross entropy loss and we measure performance using **AUC-ROC** and **F1-Score**.

Hypothesis 1 (Performance): Assuming GRP is correct then ViT would be able to perform better on both generalization and robustness tests than CNN. This improved performance implies that semantic reasoning was employed by the ViT and not artifact detection.

5 Validation via Explainable AI

The trick is that to demonstrate GRP is real, it is not sufficient to get good accuracy numbers, we need to look under the hood and understand what the ViT is doing when performing predictions.Eg. a model might achieve high accuracy due to a variety of reasons, we need to actually see what ViT is doing when making predictions.

To make sure that it is relational reasoning that is making the predictions of the ViT, we must analyze the behavior of its attention mechanism. To this we apply special explainable AI procedures

of the transformer [23]. These XAI methods allow us to visualise what the model is paying attention to in the image and how various areas relate with each other in the attention patterns of the model. Assuming our hypothesis is right we would observe the attention between semantically related and inconsistent parts.

5.1 Probing Methodology

To see what's actually happening inside these models we need some visualization techniques:

- **ViT Visualization:** For the ViT we're using stuff like *Attention Rollout* and *Transformer LRP*. What these do is they take attention maps from all the layers and combine them to show which parts of the image is actually influencing the final decision like what the `[CLS]` token is looking at when it decides.
- **CNN Baseline Visualization:** For the CNN we just use Grad CAM [**25**] which is standard for visualizing CNNs. It is the closest thing we have for comparing with convolutional networks.

5.2 Hypothesis

Apart from getting good accuracy numbers the real confirmation needs to come from what we see in the attention maps. If the ViT's attention is highlighting regions that are connected relationally and they correspond to the semantic inconsistencies then that is our proof.

What we're expecting to see:

- **ViT Outcome:** For images from the Flaw Set that gets classified as synthetic the attention maps should show something specific. We should be seeing attention spanning across multiple regions that are semantically linked but does not make sense all together. Eg. if we have got a light source and the object its lighting and then some shadow that doesn't make physical sense the attention should be linking all three of these [**12**]. That's what relational reasoning looks like.
- **CNN Outcome:** On the other hand when we look at the Grad CAM heatmaps for CNN we're expecting something completely different. They are probably highlighting surface textures, edge patterns, maybe artifacts. But they are not going capture the semantic relationships between regions that are far apart. This would basically shows CNNs are still doing artifact detection and not really reasoning about semantics.

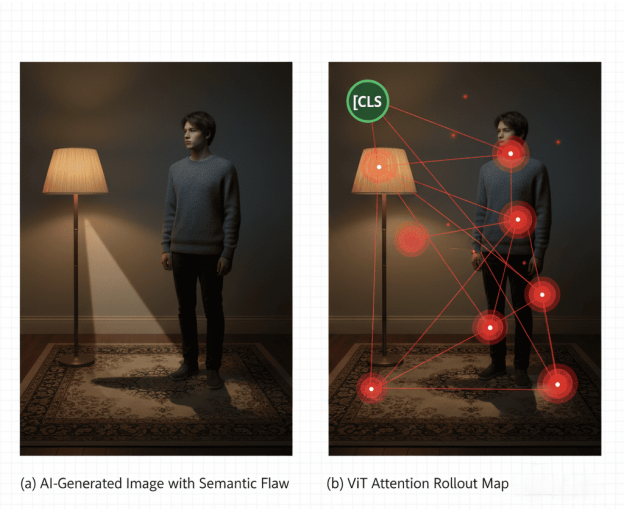


Fig. 2. GRP validation conceptual illustration by using XAI. Relational awareness The attention map of ViT is supposed to bridge semantically interrelated and inconsistent patches (e.g., lamp person shadow).

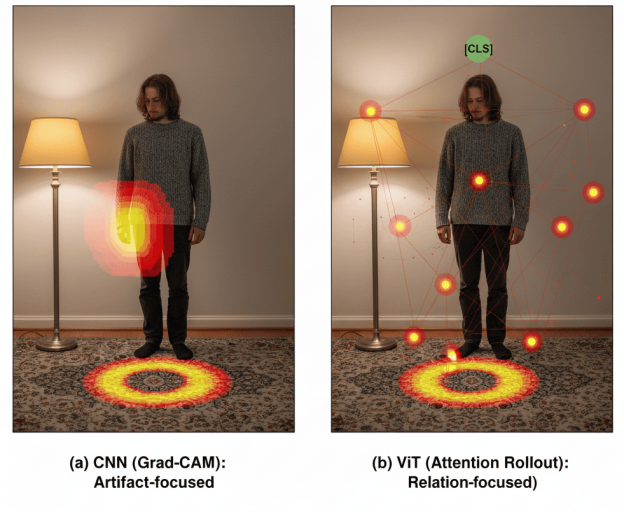


Fig. 3. Predicted difference between (a) CNN Grad-CAM maps on artifact (b) ViT attention rollout maps on the same flawed image on semantic relation.

6 Discussion, Limitations, and Future Work

6.1 GRP Impact

If the whole architecture works and the GRP hypothesis gets validated it would become important for digital forensics and detecting AI generated content. Here's why we think it matters:

1. **Cat Mouse Cycle:** The field is presently in a rut of this never ending loop. The detectors train on the artifacts of the new generator and the next version of the generator breaks and falls out. However when we change our attention not on those transient pixel level artifacts but on a higher level semantic reasoning perhaps we could stabilize the situation to some degree. The point with GRP is to aim at what all generators have a hard time with, i.e. actual grounded knowledge of how the physical world functions.

2. **Efficient and Accessible:** The other thing we can demonstrate given the same is that a standard ViT can easily generalize with only slight fine tuning which grows massive to use in practice. No costly GPUs would be needed, or huge computational budgets. WE would be able to run these detectors using a smaller hardware and thus it is accessible to more people.

3. **Established a New Baseline:** GRP might become a new standard benchmark of the field should it be validated in the model, Semiotic reasoning methods would be given a substance to base future research on. And also can be interpreted not being efficient.

6.2 Limitations and Disentanglement

The question that we must deal with here is how do we know that the ViT is actually performing a semantic reasoning task as opposed to merely picking up on the artifacts of the semantic level?

The possible thing that may occur is that the ViT may become aware of the fact that the backgrounds of fake images are often smoother or some form of color gradient. These are memorized associations which do not necessarily reflect actual logical thinking on an actual basis [24]. The model of memorizing patterns in a higher level and not thinking about physics or semantics. That would be putting the entire thing of what we are attempting to demonstrate to waste.

To cope with this we must be very cautious with regard to the way in which we curate the Flaw Set and be careful not to come up with pictures with the regularities that may tip things towards giving away. The data must compel the algorithm to make use of purely the real violations of physics or logic, and not the patterns of this resembles a typical AI image.

On top of that we might also do further validation by employing the use of causal probing techniques [25]. These approaches would assist us in determining whether the attention patterns that we are observing are reflecting of semantically significant relations or whether they are merely coinciding on superficial patterns that happen to be associated with.

6.3 Future Research Directions

Assuming GRP does the job there we can do this in some interesting directions:

- **Automated Attention Pattern Analysis:** At this moment, relational patterns of attention maps are being checked manually. Suppose we automate this and educate a meta-classifier which identifies these automatically. This would have made it more scalable and may improve performance.
- **Model Distillation:** Large ViTs are useful but impractical to implement. This relational reasoning could be condensed to smaller models such as MobileViT that would run on a phone or a small machine.
- **Temporal and Multimodal Extensions:** GRP can be applied to video, as well. ViViT style architectures with ViViT style are used to detect temporal inconsistencies that represent objects of unrealistic motion changing. This effectively makes it applicable in practice of deepfake detection. Detect audio visual mismatches lip sync being off 65 [12]. Videos with AI do not fare well in this case because cross-modal consistency is difficult, GRP may apply well in this case.

7 Conclusion

The artifact based techniques are losing its value with improved generators. However, the other semantic solutions require colossal architecture that is not as achievable by most individuals. This paper attempted to reach a pragmatic middle ground, which employs something ViTs already possesses implemented within their capability to attempt to reason about between image parts relationships. This paper has presented the Global Relational Probing (GRP) hypothesis. The paper believes a well trained ViT is capable of detecting semantic and physical inconsistencies using the attention mechanism, with no special changes, and have provided a testing framework of this one which can be reproduced and uses an explainable AI. In case GRP is right it would be a true move forward. We would have detectors that are both generalizable by generators, and efficient enough to support plausible hardware as well as understandable enough that you know why they are making decisions the way they did. The major strength lies in the fact that these models help eliminate deep flaws in the operation of the generative systems i.e. their inability to develop an understanding of the real world but instead pursue the transient artifacts that evolve with each iteration. The success of this is yet to be witnessed. The region requires the detection means to be in pace with the technology of the generation and one element of the solution is that of GRP.

References

[1] Robin Rombach et al. “High-Resolution Image Synthesis with Latent Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 10684–10695. DOI: 10.1109/CVPR52688.2022.01042.

[2] Chuan Tan et al. “Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 12105–12114.

[3] Sheng-Yu Wang et al. “CNN-Generated Images Are Surprisingly Easy to Spot... for Now”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 8695–8704.

[4] Luisa Verdoliva. “Media Forensics and DeepFakes: An Overview”. In: *IEEE Journal of Selected Topics in Signal Processing* 14.5 (2020), pp. 910–932.

[5] Hany Farid. “Seeing Is Not Believing”. In: *IEEE Spectrum* 46.8 (2009), pp. 44–48.

[6] Alexey Dosovitskiy et al. “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations (ICLR)*. 2021.

[7] Joel Frank et al. “Leveraging Frequency Analysis for Deep Fake Image Recognition”. In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2020.

[8] Chuan Tan et al. “Semantic Visual Anomaly Detection and Reasoning in AI-Generated Images”. In: *arXiv* (2025). URL: https://arxiv.org/abs/2510.10231.

[9] Chuan Tan et al. “Rethinking the Up-Sampling Operations in CNN-Based Generative Network for Generalizable Deepfake Detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 28130–28139.

[10] Tero Karras et al. “Progressive Growing of GANs for Improved Quality, Stability, and Variation”. In: *International Conference on Learning Representations (ICLR)*. 2018.

[11] B. Deressa et al. “GenConViT: Deepfake Video Detection Using Generative Convolutional Vision Transformer”. In: *Applied Sciences* 15.12 (2025), p. 6622. DOI: 10.3390/app15126622.

[12] Mingxing Tan and Quoc V. Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2019.

[13] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[14] Ming Zhu et al. “GenImage: A Million-Scale Benchmark for Detecting AI-Generated Images”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023.

[15] Ilya Loshchilov and Frank Hutter. “Decoupled Weight Decay Regularization”. In: *International Conference on Learning Representations (ICLR)*. 2019.

[16] Micah K. Johnson and Hany Farid. “Exposing Digital Forgeries by Detecting Inconsistencies in Lighting”. In: *Proceedings of the ACM Multimedia and Security Workshop*. 2005.

[17] Q. Liu et al. “Identifying Image Composites through Shadow Matte Consistency”. In: *IEEE Transactions on Information Forensics and Security* 6.3 (2011), pp. 1111–1122.

[18] Davide Cozzolino et al. “Raising the Bar of AI-Generated Image Detection with CLIP”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2024.

[19] Hila Chefer, Shir Gur, and Lior Wolf. “Transformer Interpretability Beyond Attention Visualization”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 782–791.

[20] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.

[21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the Knowledge in a Neural Network”. In: *arXiv* (2015). URL: https://arxiv.org/abs/1503.02531.

[22] H. Lin et al. “DeepFake Detection with Multi-Scale Convolution and Vision Transformer”. In: *Digital Signal Processing* 134 (2023), p. 103895. DOI: 10.1016/j.dsp.2022.103895.

[23] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *Computer Vision – ECCV 2014*. Springer, 2014, pp. 740–755. DOI: 10.1007/978-3-319-10602-1_48.

[24] Shruti Agarwal et al. “Protecting World Leaders Against Deep Fakes”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2019, pp. 38–45.

[25] Anurag Arnab et al. “ViViT: A Video Vision Transformer”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.

[26] Samira Abnar and Willem H. Zuidema. “Quantifying Attention Flow in Transformers”. In: *arXiv* (2020). URL: https://arxiv.org/abs/2005.00928.

[27] Sachin Mehta and Mohammad Rastegari. “MobileViT: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer”. In: *International Conference on Learning Representations (ICLR)*. 2022.