Perceptual Hashing Using Pretrained Vision Transformers

Jelle De Geest*, Patrick De Smet†, Lucio Bonetto†, Peter Lambert*, Glenn Van Wallendael*, Hannes Mareen*
*Ghent University – imec, IDLab, Department of Electronics and Information Systems,
Technologiepark-Zwijnaarde 122, 9052 Gent, Belgium
firstname.lastname@ugent.be, https://media.idlab.ugent.be
†National Institute of Criminalistics and Criminology (NICC), Vilvoordsesteenweg 100, 1120 Brussels, Belgium
firstname.lastname@just.fgov.be, https://nicc.fgov.be

***Abstract*—The rapid evolution of digital image circulation has necessitated robust techniques for image identification and comparison, particularly for sensitive applications such as detecting Child Sexual Abuse Material (CSAM) and preventing the spread of harmful content online. Traditional perceptual hashing methods, while useful, fall short when exposed to some common image transformations, or when images are doctored to avoid detection, rendering them ineffective for nuanced comparisons. Addressing this challenge, this paper introduces a novel pretrained vision transformer artificial intelligence (AI) model approach that enhances the robustness and accuracy of perceptual hashing. Leveraging a pretrained Vision Transformer (ViT-L/14), our approach integrates visual and textual data processing to generate feature arrays that represent perceptual image hashes. Through a comprehensive evaluation using a dataset of 50,000 images, we demonstrate that our method offers significant improvements in detecting similarities for certain complex image transformations, aligning more closely with human visual perception than conventional methods. While our method presents certain initial drawbacks such as larger hash sizes and high computational complexity, its ability to better handle perceptual nuances presents a forward step in the realm of image forensics. The potential applications of this research extend to law enforcement, digital media management, and the broader domain of content verification, setting the stage for more secure and efficient digital content analysis.**

***Index Terms*—Perceptual Hashing, Vision Transformer, Image Forensics.**

I. Introduction

In the current digital ecosystem, where images circulate with the click of a button, the need for methods to efficiently monitor, detect, and prevent the spread of harmful content has never been higher. Perceptual hashing plays an important role in this battle, offering a method to identify and compare images based on the similarity of their content. Swift detection of similar content can prevent the dissemination of illegal content, such as footage from terrorist attacks that is being uploaded by different sources. It also plays a role in the battle against child sexual abuse material (CSAM). Moreover, when hard drives are confiscated during criminal investigations, perceptual hashing enables the automated screening of big volumes of content without the need for manual review, which can both be time-consuming and traumatic for investigators.

Unlike traditional binary methods such as MD5 and SHA-based hashes [1], which can be entirely different for even slightly altered images, perceptual hashes align more closely with human visual perception; perceptual hashes computed for similar images should only be different by a few bits. Traditional pixel-based comparison methods fall short as they fail to recognize images that, while perceptually identical, have undergone changes like resizing or compression. Perceptual hashing aims to create hashes based on visual similarity rather than pixel-by-pixel accuracy and thus being more robust against changes.

There are several current prevalent hashing methods used for image identification [2] or other use cases [3]. pHash [4] utilizes a Discrete Cosine Transform (DCT) to capture image essence, offering sensitivity to content changes, but may falter with minor alterations [5]. Facebook's PDQ Hash [6], designed for rapid image matching, excels in performance yet struggles with varied image transformations [7]. Blockhash [8], an implementation of the Block Mean Based approach, uses overlapped blocking and rotation operations to enhance robustness against geometrical distortions. ColourHash matches images based on color distributions, featured in the Python ImageHash Library [9], thereby excelling in scenarios where color is key but potentially less effective in recognizing structural changes. Lastly, Wavehash, which applies the Discrete Wavelet Transform (DWT) instead of a DCT. Also featured in ImageHash [9].

Earlier research on using artificial intelligence (AI) for perceptual hashing studied the usage of deep neural networks to create hashes. These deep hashing [10] techniques are related to the work in this paper, yet are focused on finding optimal hashes for querying big databases, rather than on the human perception and forensics perspectives.

This paper proposes to use a pretrained vision transformer (ViT) AI model to enhance the accuracy and robustness of perceptual image hashing methods thereby achieving a closer resemblance to human perception of similarity. As additional contributions. We demonstrate the areas where a pretrained ViT approach excels, showcasing its potential to significantly improve upon traditional hashing techniques. Furthermore, we

This work was funded by the Research Foundation – Flanders (FWO), IDLab (Ghent University – imec), Flanders Innovation & Entrepreneurship (VLAIO), and the Flemish Government. NICC participation was Funded by the European Union (Belgian Internal Security Fund project ISF-084-108).

discuss the limitations and drawbacks inherent to ViT models, providing a balanced view on their applicability and performance in the context of perceptual hashing. This analysis aims to offer comprehensive insights into how hashing methods utilizing vision transformers can advance the forensics field, while also acknowledging the challenges that accompany its implementation.

II. Limitations of prevalent hashing methods

While perceptual hashing has significantly advanced the field of digital forensics by providing tools for image comparison that are more aligned with human visual perception, these methods are not without their shortcomings.

The first notable flaw in existing perceptual hashing algorithms is their susceptibility to attacks that only slightly alter the pixel values. It has been proven [11] that by making some specific minimal changes, it is possible to retain a high degree of visual similarity while the hash values differ substantially, effectively evading detection.

The second flaw is that traditional perceptual hashing methods often fail to align with human perception when identifying similar images within large datasets. Despite the availability of visually more compatible matches, these algorithms can yield results that, to human observers, appear to be unrelated or mismatched. This discrepancy highlights a gap between algorithmic assessments of similarity and the intuitive judgments made by people, underscoring a critical area for improvement in perceptual hashing technologies.

A third flaw of traditional perceptual hashing methods is that they are very sensitive to certain common modifications. For example, by mirroring an image or rotating it by 90 degrees, traditional methods often fail to find a match. To address this issue, a real-life system would typically store (or compute) not only the hash of the original image but also the hashes of the image mirrored and rotated in the most common orientations. As a downside, adding hashes of transformed images significantly increases the database size.

The outlined flaws demonstrate that there is still room for improvement in the field of perceptual hashing.

III. Proposed Method

In light of the limitations inherent in existing perceptual hashing techniques, this paper introduces a novel approach leveraging the capabilities of pretrained vision transformers to enhance the robustness and accuracy of perceptual image hashing. Our method capitalizes on the strengths of vision transformers to process and interpret images in a manner that more closely resembles human visual perception, thus addressing the critical flaws identified in Section II. The cornerstone of our proposed method is a ViT model that integrates visual and textual data processing to generate hashes. This approach is predicated on the insight that the features of a multimodal model are representative of the content of an image as it needs to encapsulate a context-aware understanding of the image to be able to accept both text and images as input. This model's ability to comprehend both text and images presents a significant advantage. Rather than merely searching for content by comparing similarities with previously discovered items, it enables the approach of conducting searches using textual prompts to describe the sought-after content.

In the implementation of our ViT method, we employ the pretrained vision transformer ViT-L/14 [12], [13] developed by OpenAI, renowned for its exceptional performance in image understanding tasks [14], [15]. The Vision Transformer represents a significant shift away from traditional convolutional neural networks (CNNs) by applying the principles of transformers, originally designed for natural language processing, to visual data. This model segments an image into a series of patches and processes these patches sequentially, allowing it to capture both local and global image features effectively. A key component of the model is the attention mechanism which strategically emphasizes the patches of greatest relevance, thereby sharpening its capacity to detect and interpret complex patterns and connections in the visual data.

The model is capable of accepting both text and images as inputs. Images are first resized into a $226 \times 226 \times 3$ RGB pixel matrix. Upon processing, the model outputs a feature array of 768 floating point numbers, which serves as the perceptual hash for the given input. To assess the similarity between two images, we compute the cosine similarity between their respective feature arrays, defined by equation (1), in which $\mathbf{A}$ and $\mathbf{B}$ are feature arrays of the two images, $\cdot$ denotes the dot product, and $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are the Euclidean norms of the arrays.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}||||\mathbf{B}||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

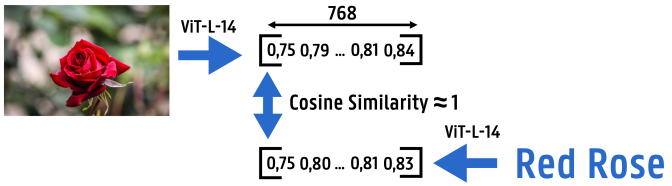
The whole process, along with the model's dual-input capability, is illustrated in Fig. 1. In Fig. 1a an image of a red rose is compared to the text-prompt "Red Rose" and is found to have a high similarity, as the cosine similarity between the two outputted vectors is close to 1. The same happens in Fig. 1b where the image of the red rose is compared to an image of a race car. This time the similarity is low, as the cosine similarity between the two outputted vectors is close to 0.

IV. Evaluation

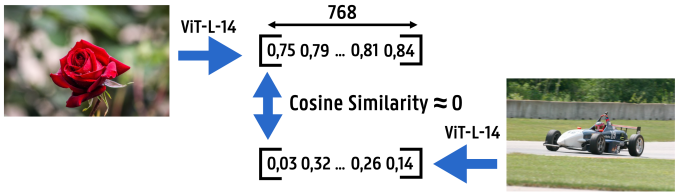
This section is organised as follows. First, Section IV-A describes the experimental setup, followed by the results in Section IV-B. Finally, open issues and future research directions are discussed in Section IV-C.

A. Experimental Setup

To asses the performance of our proposed ViT perceptual hashing method, we designed a comprehensive experimental setup that aims to replicate real-world scenarios where modifications on images are common. The computer used in this experiment is a single desktop computer with an Intel Core i9-9900K CPU and a NVIDIA RTX 2080 SUPER GPU. We used the DISC21 dataset [16] created by Meta for the



(a) A comparison between an image input and a textual prompt, demonstrating a high similarity score.



(b) An image-to-image comparison, highlighting a case of low similarity.

Fig. 1: Illustration of the ViT-L/14 model's processing and hash generation capabilities for both text and image inputs.

(a) Original Image

(b) Blur (c) Compression (d) Color_distortion

(e) Scale (f) Rotate (g) Crop_rotation

Fig. 2: Illustration of the different transformations applied to an example image.

Image Similarity Challenge at NeurIPS'21. Utilizing a 50,000 image partition of this dataset, we generated a comprehensive database of hashes using pHash [4] and our proposed method. pHash is used as a reference here, as it is a popular perceptual hashing method that is widely used.

To simulate real-world attacks and assess the resilience of our proposed method, we selected a subset of 1,000 images from within the 50,000-image database and an additional 1,000 images from another part of the dataset, ensuring no overlap with the database images. These 2,000 images were then subjected to a series of transformations (rotation, scaling, blurring, mirroring, color distortion, compression, and crop rotation) to create altered versions that represent potential attacks. Fig. 2 illustrates these transformations, and a more detailed description of the transformations can be found in Table I. After applying these transformations, we generated hashes for the altered images and compared them with the hashes from the original dataset. This process allowed us to conduct a targeted evaluation of our framework's ability to identify similarities for each specific type of digital manipulation across both sets of images.

The performance of our method in detecting and matching these modified images against the original database was quantitatively measured by generating Receiver Operating Characteristic (ROC) curves and calculating the Area Under the Curve (AUC values) for each type of transformation. To get an indication of how well a method resembles human vision, we force the method to find a match a database where there is no match exists. This process gives an indication of how a method perceives similarity.

B. Experimental Results

The evaluation of our proposed ViT perceptual hashing method yielded insightful findings across various performance metrics, including processing time, disk space utilization, and accuracy in matching images.

TABLE I: Summary of the image transformations used in the experiments.

Distortion	Description
blur-2	Gaussian blur with radius 2
compression-30	Saved as a JPEG with compression factor 30
color_distortion-0_5	Reduces the color saturation by 50%
scale-0_5	Rescales the image to half the original size
rotate-90	Rotates the image by 90 degrees
crop_rotation-15	Rotates the image 15 degrees and crops the image to remove the generated pixels that are necessary when doing a rotation that is not a multiple of 90 degrees.

1) Computational Performance Metrics: Table II presents a computational performance comparison using a database of 50,000 images. The database generation with ViT-L-14 is observed to be only twice as slow as pHash, which is reasonable given the complexity of the model. However, the larger file size and longer time required for calculating similarities with ViT-L-14 can be attributed to the use of floating point numbers in the hashes. These floating point numbers, while providing a richer and more nuanced representation, are inherently larger and more computationally intensive to process than the simpler bit operations utilized by pHash. Future work could investigate the impact of reducing the hash length, such as by quantizing the floating point numbers.

2) Accuracy Analysis: Fig. 3 gives an indication of how the ViT-L/14 approach and pHash perceive similarity, as there were no matches present but the methods were forced to find

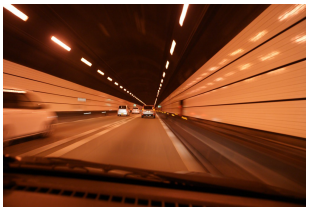
TABLE II: Computational performance comparison between ViT-L-14 and pHash for a database of 50,000 images.

	Perceptual Hashing Method			
Measure	Proposed (ViT-L-14)		pHash	
Time to generate database	31min 39s	(37.98 ms/image)	13min 10s	(15.80 ms/image)
File size	86.63 MB	(1.77 kB/image)	0.78 MB	(0.16 kB/image)
Time to calculate similarities	0.2 seconds	(4 µs/image)	0.01 seconds	(0.2 µs/image)

(a) Original Image

(b) pHash: 1st Match

(c) pHash: 2nd Match



(d) ViT: 1st Match

(e) ViT: 2nd Match

Fig. 3: False positives encountered when using the ViT-L/14 approach and pHash to find a match in database where there is no match present.

the most similar match.

Fig. 6 presents the ROC curves for the transformations blur, compression, color distortion, and scaling, where pHash outperforms the ViT-L/14 approach. These transformations predominantly maintain the general orientation of the image. The ViT-L/14 method's relative underperformance in these cases is attributed to its tendency to generate more false positives than pHash. This is illustrated by Fig.4, which showcases a selection of mismatches by ViT-L/14. Despite these images not being exact matches, they often bear a close resemblance to the originals, suggesting that the matched images possess perceptual similarities closely aligned with human visual processing. This contrasts with the type of false positives generated by pHash, as depicted in Fig. 3.

TABLE III: AUC comparison, of the different transformations, between ViT-L-14 and pHash.

	Perceptual Hashing Method	
Transformation	Proposed (ViT-L-14)	pHash
blur-2	0.946	0.999
compression-30	0.966	0.999
color_distortion-0_5	0.982	0.999
scale-0_5	0.964	0.999
rotate-90	0.6780	0.000
crop_rotation-15	0.811	0.049

Conversely, Fig. 7 illustrates the ROC curves for transformations that alter the image's general orientation (rotation and crop rotation) where the ViT-L/14 method outshines pHash. Here, the general orientation of the image is changed and the ViT-L/14 approach can handle this well, unlike pHash which struggles with these transformations.

In addition to the ROC curve analysis, Table III provides a quantitative comparison through AUC values, offering a deeper insight into the performance of both pHash and ViT-L/14 across the discussed transformations. The AUC values further substantiate the qualitative observations from the ROC curves, providing a comprehensive evaluation of each method's effectiveness in various scenarios.

While this paper does not delve into text-based image search capabilities, it remains an intriguing direction for future research. Fig. 5 highlights the potential for expanding the scope of our analysis to include text as an additional modality for image retrieval and comparison.

C. Open issues and future research

Despite the promising results, our method requires further research. Notably, the feature vectors, or hashes, generated by the ViT-L-14 model are substantially larger than those of traditional algorithms such as pHash, given that they consist of an array of floating-point numbers. This increase in size has implications for storage and for the speed of hash retrieval and comparison. Furthermore, the computational speed of our method has not yet been optimized. One possible avenue for optimization involves determining the optimal batch size, since computation is carried out in batches on the GPU. Moreover, the lengths of the hashes, currently utilizing full floating-point numbers, play a crucial role in computational efficiency. If

(a) Image 1: Original (b) Image 1: 1st FP (c) Image 1: 2nd FP

(d) Image 2: Original (e) Image 2: 1st FP (f) Image 2: 2nd FP

(g) Image 3: Original (h) Image 3: 1st FP (i) Image 3: 2nd FP

Fig. 4: False positives (FP) when using the ViT-L/14 approach (Originals were found as well).

(a) Best match (b) Second best match

Fig. 5: Best and second best match when looking for images using the text-prompt "Boat".

reducing the precision to only a few bits does not significantly impact accuracy, this adjustment could enable the use of more efficient calculations, thereby potentially enhancing processing speed without sacrificing performance. Another unexplored aspect crucial for certain privacy-sensitive applications may be whether these hashes could be reversed back into their original images, thus leaking sensitive information. These research opportunities offer interesting avenues to pursue further improvement of the ViT approach for perceptual hashing.

V. Conclusion

In this paper, we presented a novel pretrained vision transformer model approach to enhance the robustness and accuracy of perceptual hashing. By integrating a vision transformer, our method has shown potential in more closely aligning with human perception of image similarity, especially in the face transformations where traditional methods struggle. Through extensive experimentation with a dataset of 50,000 images, we demonstrated that while the vision transformer approach does not always outperform traditional methods in conventional metrics, it excels in handling several 'harder' transformations. Additionally, mismatched images have similar content than the original image, in contrast to mismatches in traditional hashing methods that show significantly different content. Moving forward, refining computational efficiency, minimizing storage, mitigating hash reversibility, and leveraging text-based searches are key areas for development, aiming to enhance the practicality and applicability of our approach.

The potential implications of our work for the field of image forensics are substantial. As the method is developed further, it could provide a powerful tool for managing and navigating the vast quantities of visual data that define the digital age. It has the potential to set a new standard for content verification, digital safety, and the ethical use of multimedia content, contributing to a more secure and trustworthy digital environment.

References

[1] Wikipedia webpage on secure hashing algorithms. [Online]. Available: https://en.wikipedia.org/wiki/Secure_Hash_Algorithms

[2] S. McKeown and W. J. Buchanan, "Hamming distributions of popular perceptual hashing techniques," *Forensic Science International: Digital Investigation*, vol. 44, p. 301509, 2023, selected papers of the Tenth Annual DFRWS EU Conference. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281723000100

[3] H. Mareen, N. Van Kets, P. Lambert, and G. Van Wallendael, "Fast fallback watermark detection using perceptual hashes," *Electronics*, vol. 10, no. 10, p. 1155, 2021.

[4] Phash website and source code. [Online]. Available: https://phash.org/

[5] C. Zauner, "Implementation and benchmarking of perceptual image hash functions," Master's thesis, University of Applied Sciences Hagenberg, 2010. [Online]. Available: https://www.phash.org/docs/pubs/thesis_zauner.pdf

[6] Facebook github site for pdq. [Online]. Available: https://github.com/facebook/ThreatExchange

[7] J. Dalins, C. Wilson, and D. Boudry, "PDQ & TMK + PDQF – a test drive of facebook's perceptual hashing algorithms," 2019. [Online]. Available: https://arxiv.org/abs/1912.07745

[8] B. Yang, F. Gu, and X. Niu, "Block mean value based image perceptual hashing," in *2006 International Conference on Intelligent Information Hiding and Multimedia*, 2006, pp. 167–172.

[9] J. Buchner, "Image hash," https://github.com/JohannesBuchner/imagehash, Year Published or Last Updated, accessed: Date Accessed.

[10] X. Luo, H. Wang, D. Wu, C. Chen, M. Deng, J. Huang, and X.-S. Hua, "A survey on deep hashing methods," *ACM Trans. Knowl. Discov. Data*, vol. 17, no. 1, feb 2023. [Online]. Available: https://doi.org/10.1145/3532624

[11] S. Jain, A.-M. Cretu, and Y.-A. de Montjoye, "Adversarial detection avoidance attacks: Evaluating the robustness of perceptual hashing-based client-side scanning," in *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 2317–2334. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/jain

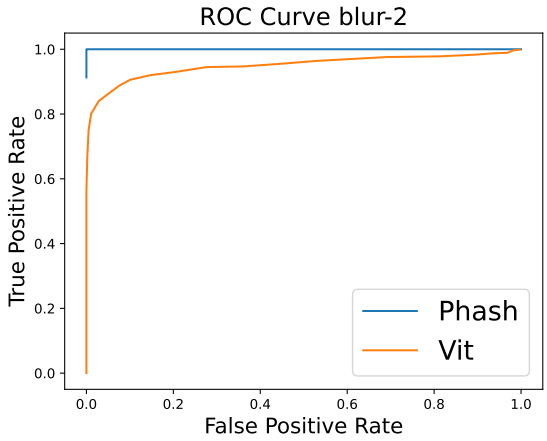
[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[13] OpenAI, "CLIP-ViT-L/14 model," https://huggingface.co/openai/clip-vit-large-patch14, accessed: March 24, 2024.

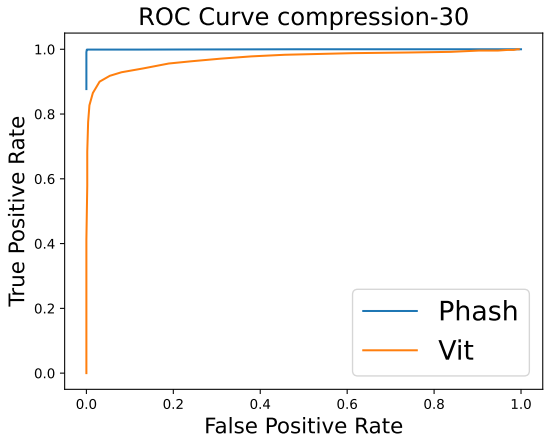
[14] X. Dong, J. Bao, T. Zhang, D. Chen, S. Gu, W. Zhang, L. Yuan, D. Chen, F. Wen, and N. Yu, "Clip itself is a strong fine-tuner: Achieving 85.7

[15] Y. Zhang, Y. Shao, X. Zhang, W. Wan, J. Li, and J. Sun, "Clip pre-trained models for cross-modal retrieval in newsimages 2022," in *Proceedings of the MediaEval workshop, CEUR-WS. org*, 2022.

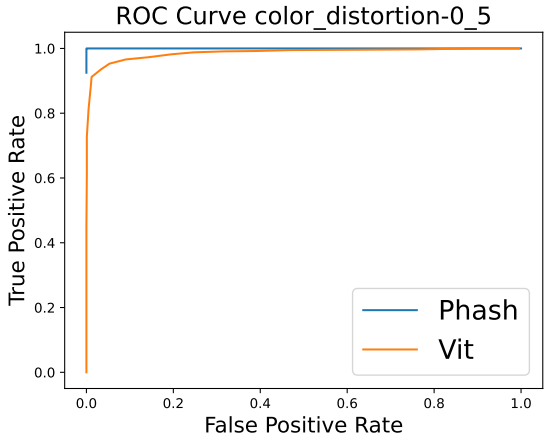
[16] M. Douze, G. Tolias, E. Pizzi, Z. Papakipos, L. Chanussot, F. Radenovic, T. Jenicek, M. Maximov, L. Leal-Taixé, I. Elezi, O. Chum, and C. C. Ferrer, "The 2021 image similarity dataset and challenge," 2022. [Online]. Available: https://arxiv.org/abs/2106.09672



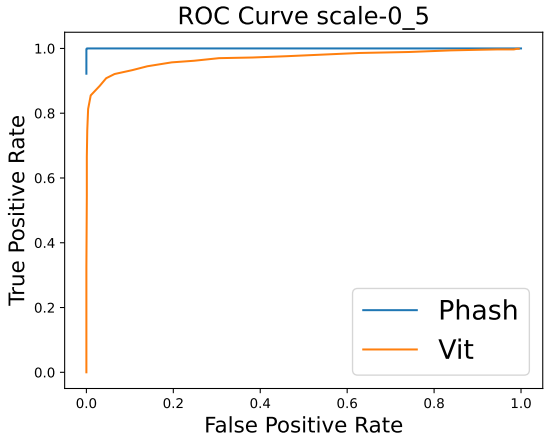
(a)



(b)

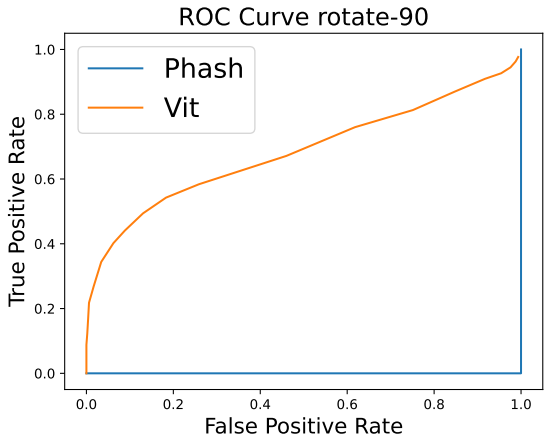


(c)

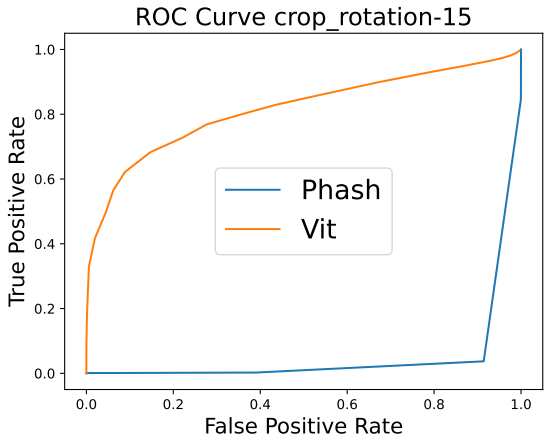


(d)

Fig. 6: ROC curves of transformations for which pHash outperforms the ViT approach.



(a)



(b)

Fig. 7: ROC curves of transformations for which the ViT approach outperforms pHash.