Enhancing Multi-Tip Artifact Detection in STM Images Using Fourier Transform and Vision Transformers

T. Rodani [1,2], A. Ansuini [2], A. Cazzaniga [2]
1) Università degli Studi di Trieste, Trieste 2) AREA Sciente Park, Trieste

Background

Scanneling Tunneling Microscopy (STM) imaging can produce artifacts or distortions absent from the actual sample surface. These often stem from unpredictable tip-surface interactions that modify the tip's shape, causing highly nonlinear changes in the acquired data [1].

A common artifact is structure duplication, which occurs when the probe shape includes multiple atoms at the apex. This multi-tip or "ghosting" artifact creates duplicated signals, complicating data interpretation [2].

The Fourier transform, can unveil features like perturbations that are difficult to spot in the spatial domain but are prominent in the Fourier domain [3].

We address multi-tip artifacts using Fourier transform to decompose image content into frequencies. Combining FFT preprocessing with Vision Transformers (ViT) [4] significantly enhances model performance compared to grayscale-only input. Unlike traditional offline methods, our approach enables real-time, scalable classification.

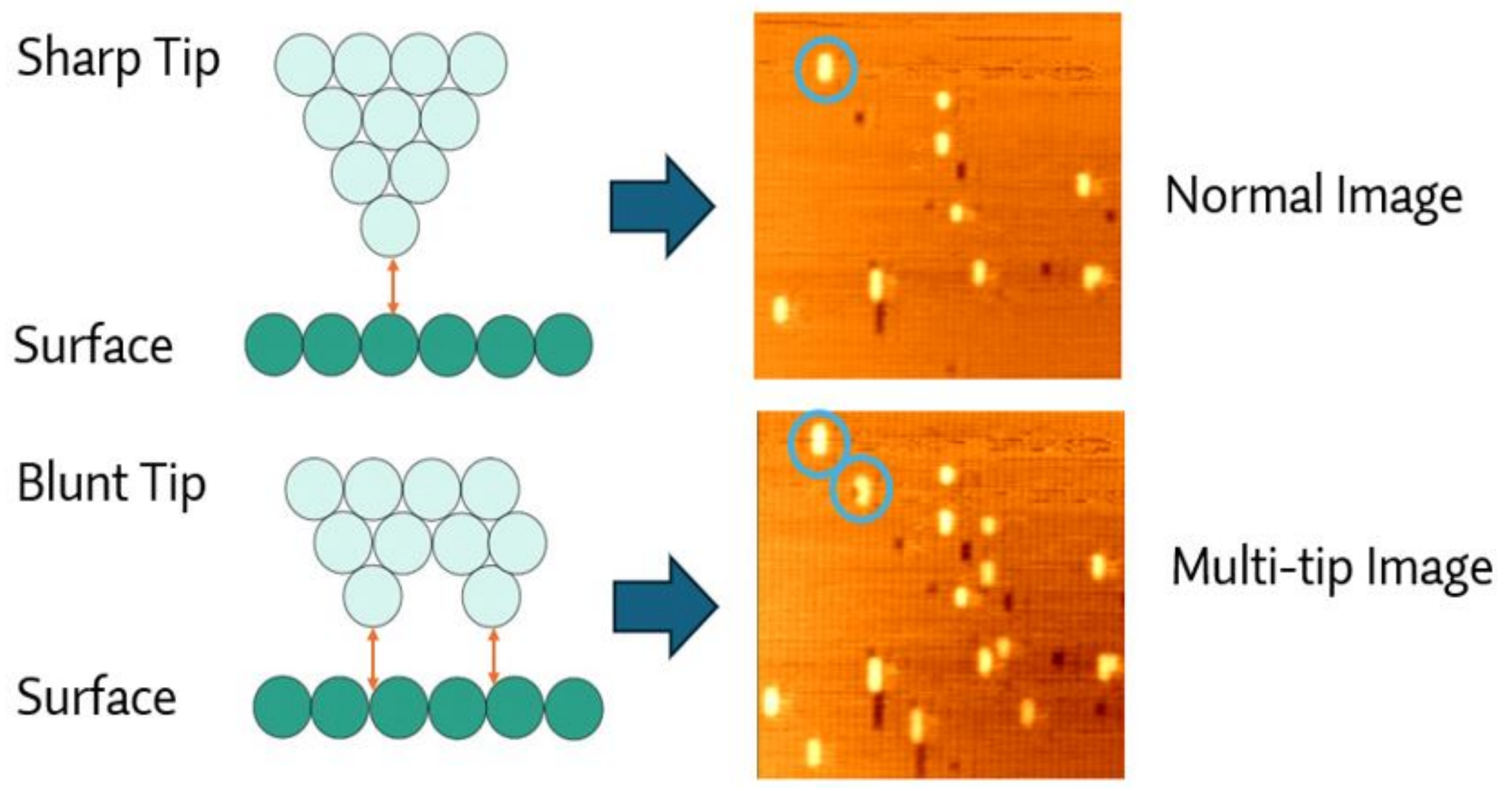

Figure 1. Tip quality impact on image: Top: Sharp tip - clear, normal image. Bottom: Blunt tip - multi-tip artifact, degraded quality.

Method

We created a synthetic dataset using STM images from an Omicron VT-STM microscope. The dataset contains 2080 grayscale images, with 82 labeled as containing multi-tip artifacts and 1998 as clean.

To balance the dataset, we created synthetic multi-tip images by summing clean images with N of their translations, N representing the number of tips.

We applied FFT to obtain amplitude and phase, saving each image as a three channel image (grayscale, amplitude, and phase). Figure 2 illustrates an example before and after applying the multi-tip artifact. The three columns represent grayscale, amplitude, and phase of the FFT transformation.The first row shows a clean image, while the second row shows the image with the synthetic multi-tip artifact.

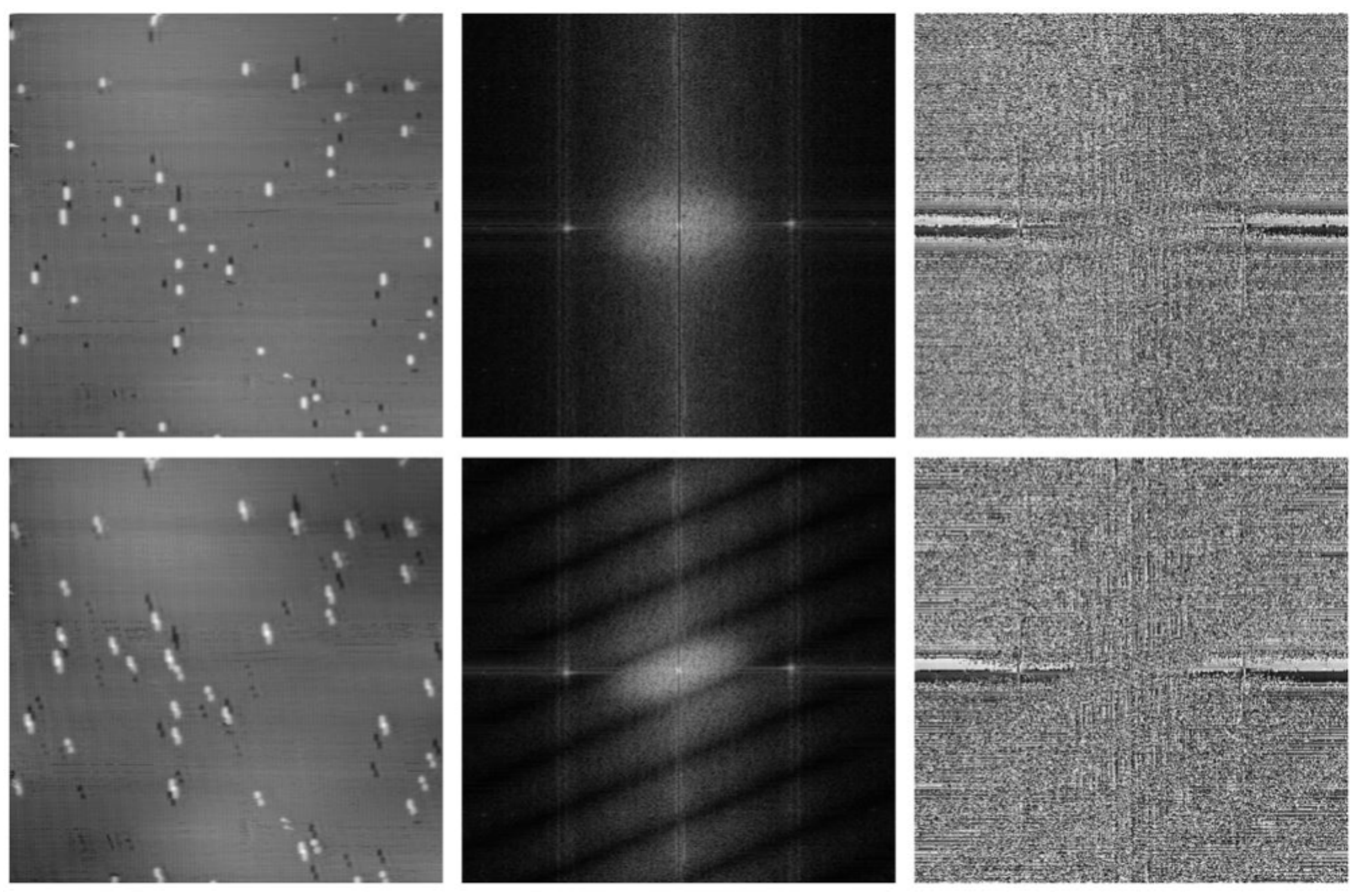

Figure 2: STM image before and after applying the multi-tip artifact.

We fine-tuned various pre-trained neural network architectures on the synthetic dataset, specifically ResNet [5] with 18 and 50 layers, and Vision Transformer Base (ViT-B) with 16 and 32 layers.

Results

Table 1 summarizes our experiment results, with best outcomes in bold. ViT-B16 and ViT-B32 models show significant performance gains using our FFT-based preprocessing compared to grayscale-only input. This improvement can be attributed to ViT robustness against high-frequency noise, unlike ResNet, which is more vulnerable.
The Multi-Head Self-Attention (MSA) in ViT performs low-pass filtering, which enhances the FFT-based method's advantage in preserving high-frequency components [6].

We tested each network using individual components of our method: grayscale, amplitude, and phase. Table 2 results show that the amplitude component contribute most to performance gains.
This finding aligns with amplitude's known role in preserving image feature geometry.

We conducted ablation experiments to explore the validity of our approach. The number of tips in the synthetic dataset generation does not affect performance using our FFT method.
Additionally, we studied the effect of increased ranges of translation vectors in the synthetic dataset generation. Using wider ranges has a detrimental effect on the model performance.

MODEL	ACCURACY	FFT-BASED
RESNET18	52.43	×
RESNET18	57.07	√
RESNET50	58.84	×
RESNET50	65.24	√
VIT-B/32	78.96	×
VIT-B/16	88.11	×
VIT-B/16	97.25	√
VIT-B/32	**97.86**	√

Table 1. Classification accuracies of different models with and without our FFT-method on the test set.

MODEL	IMAGE	AMPLITUDE	PHASE
RESNET18	52.43	**60.06**	53.04
RESNET50	58.84	**62.50**	51.21
VIT-B/16	88.11	**97.56**	66.15
VIT-B/32	78.96	**97.86**	68.90

Table 2. Classification accuracies of different models for each component of our FFT-method on the test set.

In conclusion, our FFT-based method, combined with Vision Transformers, significantly enhances multi-tip artifact detection in STM images, offering robust, real-time classification for large datasets.

References

[1] WK Lo and JCH Spence. "Investigation of STM image artifacts by in-situ reflection electron microscopy". In: Ultramicroscopy 48.4 (1993), pp. 433– 444.
[2] EJ Van Loenen et al. "Evidence for tip imaging in scanning tunneling microscopy". In: Applied physics letters 56.18 (1990), pp. 1755–1757.
[3] Ioannis Pitas. Digital image processing algorithms and applications. John Wiley & Sons, 2000
[4] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: arXiv preprint arXiv:2010.11929 (2020).
[5] Kaiming He et al. "Deep residual learning for image recognition". In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 770–778.
[6] Namuk Park and Songkuk Kim. "How do vision transformers work?" In: arXiv preprint arXiv:2202 06709 (2022).

SCAN ME