Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

No-Reference Hallucination Assessment for AI-Reconstructed Fluorescence Microscopy Image

Bo Yan

byan@fudan.edu.cn

Fudan University https://orcid.org/0000-0001-5692-3486

Chenxi Ma

Fudan University

Weimin Tan

Fudan University https://orcid.org/0000-0001-7677-4772

Yuqi Sun

Fudan University

Hongju Fu

Fudan university

Article

Keywords:

Posted Date: August 6th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7026761/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

No-Reference Hallucination Assessment for AI-Reconstructed Fluorescence Microscopy Image

Chenxi Ma[1,2], Weimin Tan[1,2], Yuqi Sun[1], Hongju Fu[1], Bo Yan[1]*

[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China.
[2]These authors contributed equally.
*byan@fudan.edu.cn

2 July 2025

Fluorescence microscopy has become an essential tool in cell biology, enabling the visualization of subcellular structures and dynamic processes in living cells, tissues, and developing organisms. The integration of artificial intelligence (AI) based image restoration techniques has significantly enhanced imaging resolution and speed, advancing our understanding of cellular function and regulation in both physiological and pathological contexts. However, the inherent biases of AI models in training sets may lead to hallucinations, such as hallucinating artificial structures that do not exist in the original sample or removing real biological features, compromising the scientific authenticity and reliability of imaging-based discoveries. To address this critical challenge, we present the first systematic investigation and formal definition of AI-reconstructed hallucinations in fluorescence microscopy. We introduce HallAssess, a no-reference assessment method explicitly tailored for the reliable identification and quantification of AI-reconstructed hallucinations. By shifting the assessment from the high-quality (HQ) image domain to the low-quality (LQ) image domain, HallAssess effectively transforms a no-reference problem into a full-reference one. Our approach enables accurate hallucination quantification without requiring ground-truth HQ images by re-degrading AI-reconstructed images using an imaging model that simulates real-world image degradation processes, and then comparing them with the original LQ inputs. We validate HallAssess across multiple imaging modalities (SIM, confocal), diverse AI models, and common fluorescence microscopy image restoration tasks such as denoising and super-resolution. The results demonstrate its effectiveness in detecting AI-reconstructed hallucinations. Furthermore, we provide an open-access platform featuring an interactive web demo and a dynamic leaderboard, allowing researchers to evaluate hallucinations in fluorescence microscopy image restoration results and benchmark state-of-the-art methods under a standardized framework. This work provides a foundational tool for ensuring the reliability of AI-assisted imaging in life sciences, particularly in cell biology, where accurate interpretation of subcellular structures is essential for understanding cellular function and disease mechanisms.

Fluorescence microscopy has become an indispensable tool in cell biology research by providing detailed insights into dynamic biological structures (living cells, tissues, or developing organisms) and enabling the discovery of cellular functions and regulatory mechanisms [2, 3]. While techniques such as super-resolution microscopy (SRM) [4–12] surpass the diffraction limit, it often sacrifices temporal resolution and field-of-view [13]

Computational methods, particularly deep learning, have emerged as a transformative force in fluorescence microscopy. As highlighted in a viewpoint article [14], artificial intelligence (AI)-based analysis is currently revolutionizing the field of SRM, with additional denoising and deconvolution techniques offering broad benefits across any imaging technique.

In particular, AI-based fluorescence microscopy image restoration (FMIR) methods, ranging from classic deconvolution methods [15–18] to deep learning-based methods [19–36], have been widely developed to recover nanoscale details that are otherwise inaccessible through hardware improvements alone [37–39]. These computational techniques effectively break the traditional trade-offs between spatial resolution, imaging speed, light exposure, and imaging depth [40], thereby extending the physical limits of fluorescence microscopy through intelligent post-processing.

Although significant progress has been achieved, the authenticity and reliability of AI-reconstructed content [41] pose a considerable challenge in biomedical research. As stated in the viewpoint [14], AI-based post-processing methods may inherit the inherent biases of AI in training data. More specifically, most current FMIR models trained with Mean Squared Error (MSE) loss tend to learn the average features of the training dataset. Therefore, when presented with images that differ from the training data, these models may generate new, visually reasonable but factually incorrect texture details during inference. This discrepancy between visual plausibility and factual accuracy is what we refer to as AI-reconstructed hallucinations. For example, the hollow, ring-like structures of CCPs are destroyed by AI-based FMIR models. By examining a single FMIR result, we cannot determine whether it is truly accurate. Some recovered texture structures may appear reasonable, but an accurate evaluation cannot be provided without a reference (Fig. 1b). By understanding and addressing the causes of AI-reconstructed hallucinations, biologists and practitioners can develop more robust and reliable AI-based image restoration methods.

AI-reconstructed hallucinations become the main concern for scientists, as they can compromise the fidelity of images. Therefore, accurately identifying these hallucinations, like artifacts, distortions, or features that do not exist in the original input image or are not supported by the available data, is crucial for ensuring the robustness of subsequent research based on fluorescence microscopy. This claim is also approved by Reto Fiolka [14] that ‘it is important to understand the risks of potential artifacts’. However, it remains uncertain whether

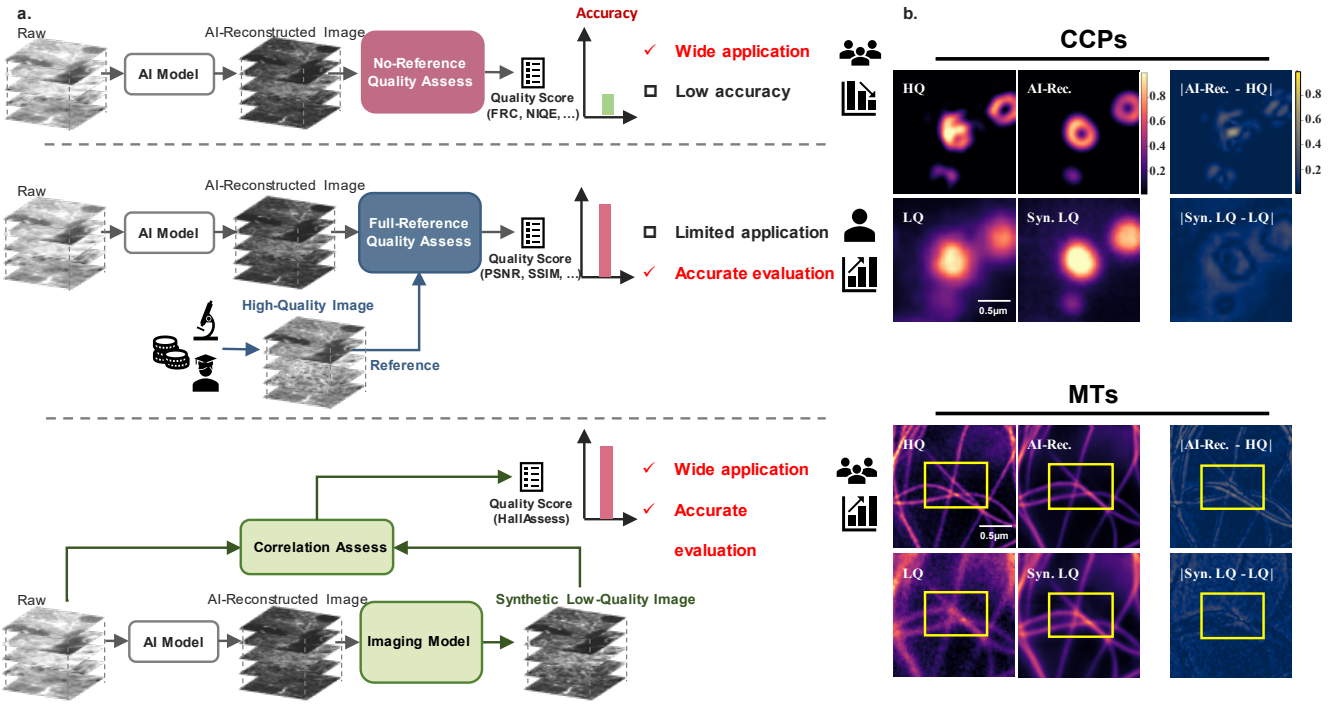


Figure 1 | The hallucination assessment of AI-reconstructed fluorescence microscopy images. **a**, Comparison between the no-reference assessment process, the full-reference assessment process, and the proposed approach, which turns a no-reference assessment into a full-reference assessment condition by taking the original low-quality image as reference and transferring the AI-reconstructed fluorescence microscopy image to the low-quality space with an imaging model. **b,** Demonstration of AI-reconstructed hallucinations in FMIR results for $\times 2$ upscaling on CCPs and MTs datasets [1]. The residual images (|AI-Rec. - HQ|) between AI-reconstructed images (AI-Rec.) and real high-quality images (HQ) can indicate the areas of real hallucination. The residual images (|Syn. LQ - LQ|) between synthetic low-quality images (Syn. LQ) and real low-quality images (LQ) are consistent with the |AI-Rec. - HQ|, demonstrating that the residual between synthetic low-quality images (Syn. LQ) and real low-quality images can also accurately indicate the hallucination areas. Scale bar: 0.5 μm.

the information conveyed by AI-reconstructed images is trustworthy. The measurement of AI-reconstructed hallucinations is a crucial problem that has never been investigated.

As illustrated in Fig. 1 a, AI models reconstruct a high-quality image given a low-quality imaging result. Evaluating the hallucination phenomena in this reconstructed image is inherently a no-reference problem. The 'no-reference' problem refers to scenarios without access to an original, high-quality reference image against which the reconstructed image can be compared. Instead, the evaluation must rely solely on the reconstructed image to determine the presence and extent of AI-reconstructed hallucinations.

To quantitatively assess the quality of restored fluorescence images, most current FMIR works adopt common full-reference or no-reference image quality assessment (IQA) methods from the natural image domain, which aim to establish a quantitative mapping between digital visual images and objective evaluation scores [42]. However, full-reference IQA methods, including peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) [43], and NanoJ-SQUIRREL [44], fundamentally depend on paired high-quality reference images - a requirement rarely fulfilled in practical microscopy workflows due to the inherent absence of ground-truth data in biological imaging. These full-reference IQA methods often involve the image registration step to align the restored image with the reference image, ensuring that corresponding features are properly matched. For example, NanoJ-SQUIRREL [44] generates pixel-wise error maps (RSE/RSP metrics) by comparing super-resolution images with their diffraction-limited counterparts using a resolution scaling function (RSF). While full-reference IQA methods offer high evaluation accuracy, they require a high-quality reference image and a complex image registration process, making them impractical for real-world imaging scenarios.

Regarding no-reference IQA methods, such as naturalness image quality evaluator (NIQE), Full Width at Half Maximum (FWHM), and Fourier ring correlation (FRC) based resolution measure [45–47] demonstrate complementary yet constrained applicability. As exemplified by a Fourier ring correlation (FRC) variant, rFRC [47], which identifies high uncertainty regions through rolling Fourier ring correlation analysis for 20 nm-scale resolution heterogeneity mapping combined with spatial-frequency verification through modified resolution-scaled error maps. While effective for resolution uncertainty localization, this method is primarily confined to the quality assessment for super-resolution microscopy. It does not address the spectrum of hallucination artifacts emerging in general fluorescence microscopy image restoration tasks. Furthermore, neither methodological paradigm adequately resolves the critical challenge of quantifying AI-reconstructed hallucinations in fluorescence microscopy image restoration.

Here, we proposed HallAssess, an imaging mechanism-informed hallucination assessment framework for AI-reconstructed fluorescence microscopy images. Unlike conventional approaches, our method leverages a low-quality imaging model to evaluate whether AI-reconstructed images

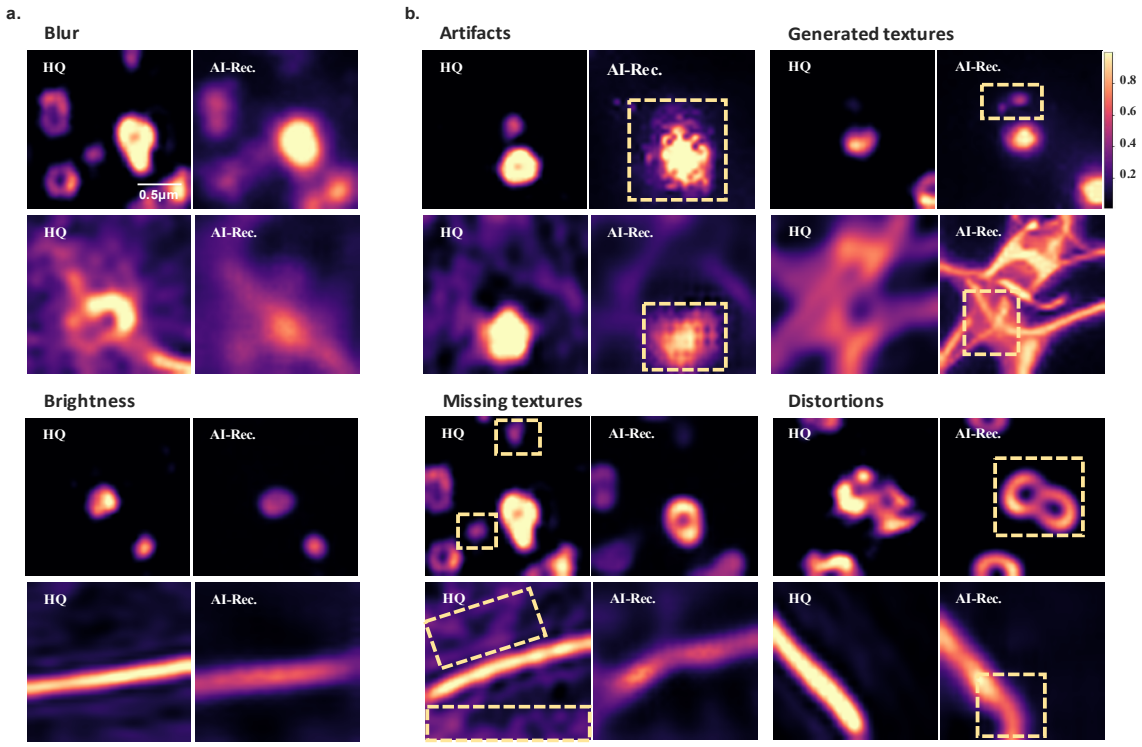


Figure 2 | Visualization of AI-reconstructed hallucinations. **a,** The blur and brightness changes in fluorescence microscopy image restoration results are not hallucinations, as the invalidation of the AI models mainly causes them. **b,** Four types of hallucinations. Artifacts refer to visual imperfections, such as noise, that do not exist in the real-world scenes. Generated textures refer to the nonexistent textures or details that the model hallucinates during the super-resolution process, which do not exist in the original high-resolution image. Missing textures refer to the loss of signal, where certain details or textures that are present in the original image are not accurately recovered or are entirely lost. Distortions refer to the alteration of shapes, where the visual appearance may seem enhanced but inaccurately represents the original forms and structures in the image. Scale bar: 0.5 μm.

remain faithful to realistic image formation processes. Specifically, we first introduced an imaging model that simulates a real-world low-quality imaging process, where degradation patterns are more straightforward to model, to produce degradations such as noise, blur, and artifacts typically present in low-quality fluorescence images. This imaging model was applied to the AI-reconstructed image to generate a synthetic low-quality version. By computing the correlation between this synthetic image and the original low-quality input, we quantified the degree of hallucination introduced by the AI reconstruction process (Figure 1a and Figure 3).

Our HallAssess not only identifies localized hallucinated regions in AI-reconstructed images but also provides a universal, interpretable quality score that reflects the overall fidelity of the reconstructed image. Importantly, our approach effectively transforms a no-reference assessment scenario into a pseudo full-reference one, combining the practicality of no-reference methods with the accuracy of full-reference metrics. As a result, HallAssess achieves performance close to that of full-reference evaluation, without requiring access to a high-quality reference image.

We validated the hallucination assessment capability of HallAssess across various microscopy modalities, image restoration tasks, and AI models. Our approach provides deeper insights into the reliability and limitations of fluorescence microscopy image reconstruction (FMIR) results, offering valuable guidance for future algorithm development and application design. Notably, HallAssess demonstrates significant practical value in scenarios where high-fidelity and accurate image reconstruction are essential, such as biomedical imaging and materials science, especially when ground-truth high-quality images are unavailable. This makes our method particularly suitable for real-world applications where reference-based evaluation is infeasible.

Results

Hallucination analysis

Before assessing and quantifying AI-reconstructed hallucinations in fluorescence microscopy image restoration, we first discussed what constitutes a hallucination in this context (Fig. 1b, Fig. 2). Due to the limited representational capacity of deep learning models, AI methods often struggle to accurately recover fine details from low-quality images (LQ). As a result, the reconstructed images ($AI\text{-}Rec.$) inevitably contain errors when compared to the corresponding high-quality ground truth (HQ), as measured by $|AI\text{-}Rec. - HQ|$ (Fig. 1b).

We did not consider blur or brightness changes as hallucinations; these artifacts are primarily due to the limited learning capability of AI models rather than the generation of entirely false information. Instead, we focused on hallucinated content—that is, artificial details or structures introduced by the model that do not exist in the original scene. This error can mislead scientific interpretation and, therefore, requires careful identification and evaluation. We categorized hallucinations into four types:

- **Artifacts**: Visual imperfections such as noise or streaks that are absent in real-world scenes.

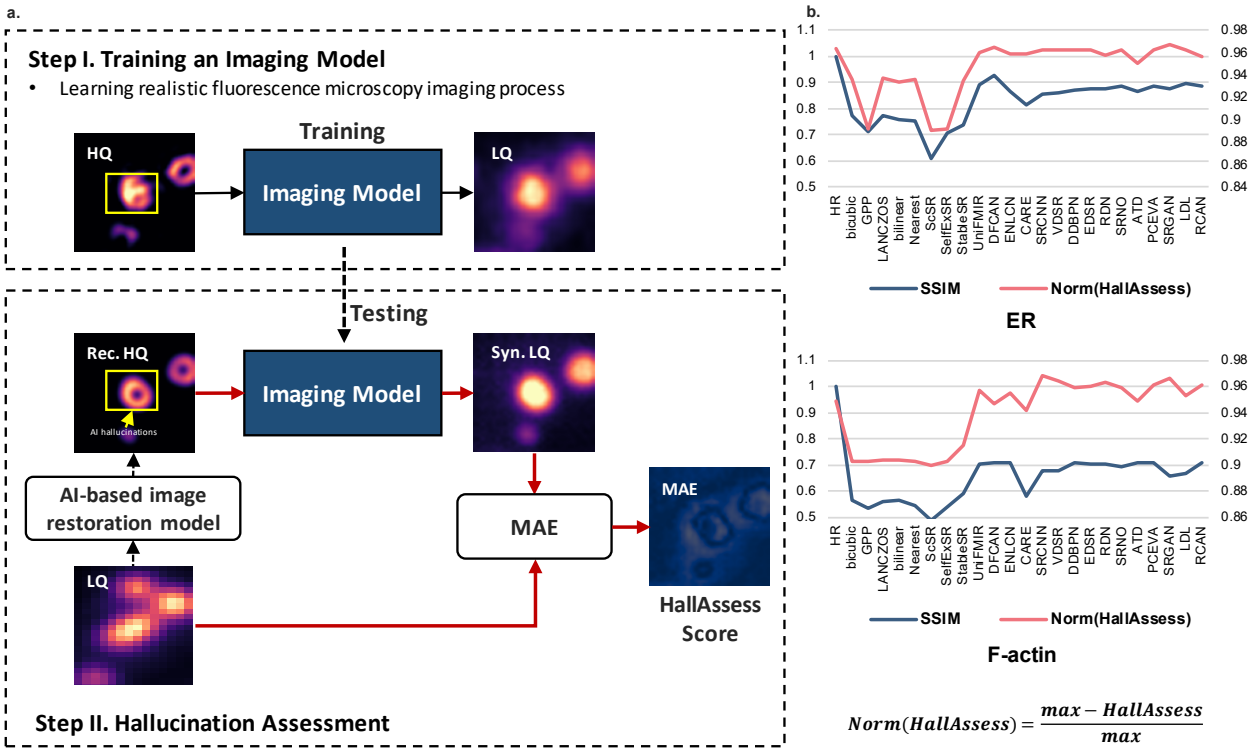


Figure 3 | Overview of the realistic imaging mechanism-informed hallucination assessment framework. a, An imaging model is constructed and trained on real high-quality (HQ) and low-quality (LQ) image pairs to learn a realistic low-quality imaging process, which is the reverse process of the image restoration process. Then, the AI-reconstructed fluorescence microscopy images (Rec. HQ) are fed into the imaging model to generate the reverse-inferred low-quality images (Syn. LQ). The hallucination artifacts are evaluated by measuring the correlation between the imaging model's outputs (Syn. LQ) with real low-quality data (LQ). **b,** The two line charts compare the full-reference SSIM and the proposed reference-free HallAssess metrics for measuring fluorescence microscopy image restoration (FMIR) results. The x-axis shows the results from 23 different super-resolution methods, while the y-axis represents the SSIM and the normalized HallAssess scores.

- **Generated textures**: Nonexistent textures or details created by the model during image restoration calculations that are not present in the original high-resolution image.

- **Missing textures**: Loss of genuine structural information, where certain features present in the original image are either poorly recovered or completely omitted.

- **Distortions**: Structural deformations where shapes or morphologies appear altered or enhanced but inaccurately represent the true forms in the original image (Fig. 2).

No-reference hallucination assessment method

Inspired by the hallucination detection methods [48–51] in large language models (LLMs), which capture information overlap (consistency) between a reverse-inferenced question of the LLM's output and its source question, the proposed HallAssess was designed to assess hallucinations by calculating the correlation between the source low-quality image (LQ) and the low-quality version ($Syn.LQ$) of FMIR results ($Rec.HQ$), reconstructed by applying a realistic low-quality imaging model to the AI-reconstructed image. Specifically, we first constructed a realistic low-quality imaging model to transform high-quality images into realistic low-quality images. This imaging model learned the reverse inference of the image restoration process, which aims to transform low-quality images into realistic, high-quality images by training on realistic datasets, composed of high-quality and low-quality fluorescence microscopy image pairs, to simulate real-world noise, blur, and artifacts in conditions under which low-quality images are produced. Then, given an AI-reconstructed fluorescence microscopy image ($Rec.HQ$), we fed it into the imaging model to generate a "re-degraded" version, that are a synthetic low-quality image ($Syn.LQ$), of the reconstructed image. By calculating the correlation between the synthetic low-quality image ($Syn.LQ$) and the original low-quality image (LQ), we can assess whether the AI-reconstructed images maintain fidelity with the actual low-quality imaging processes (Fig. 2a).

This approach directly links the validity of image reconstruction to experimental reproducibility. Specifically, hallucinated structures—those not present in the original scene—are unlikely to reproduce authentic low-quality (LQ) image patterns when passed through a realistic imaging model. The closer the synthetic LQ image matches the original LQ image, the fewer hallucinations are present in the AI-reconstructed high-quality (HQ) image. This indicates that the model has successfully minimized the introduction of artifacts or unrealistic details. The core idea is grounded in the principle that the real HQ image and its corresponding LQ version should convey consistent information beyond differences in clarity and visual appearance. In other words, the LQ image should not contain features that were not originally present. Therefore, if an AI-reconstructed image contains hallucinated or artificial content, the resulting synthetic LQ image will deviate from the original LQ input, revealing inconsistencies in both visual and statistical properties.

This method provides a principled way to assess the fidelity

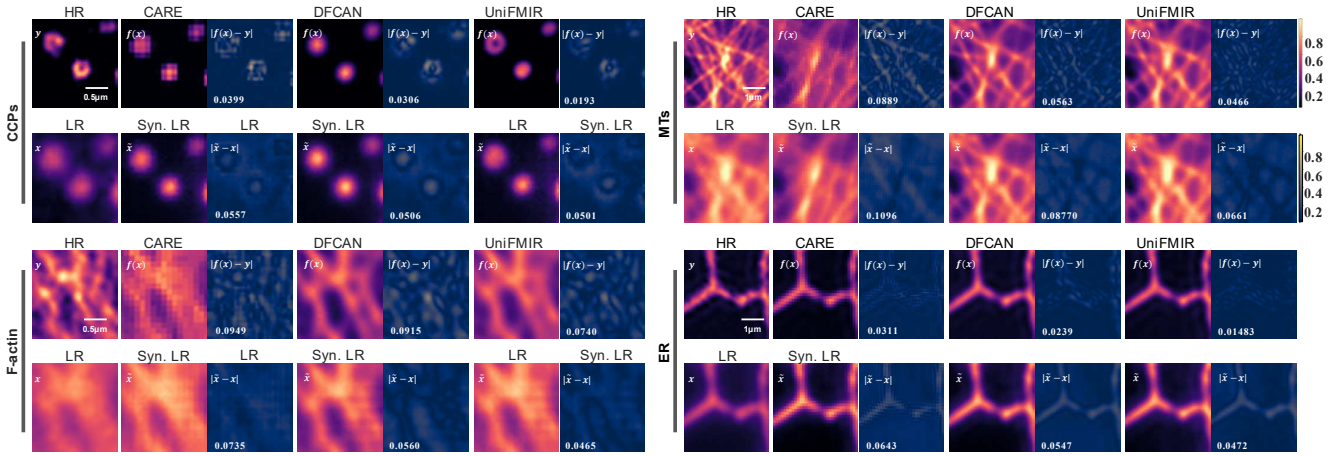


Figure 4 | Applying the proposed no-reference hallucination assessment approach on ×2 super-resolution (SR) results of fluorescence microscopy images from the BioSR datasets [1] (CCPs, F-actin, MTs, ER). The first line shows the ground-truth (GT) high-resolution (HR) images, outputs of FMIR models (CARE [40], DFCAN [1], UniFMIR [36]), and the residual images ($r_1 = |f(x) - y|$) between HR and the FMIR results. The second line shows the original low-resolution (LR) images, synthetic LR images from the corresponding FMIR results, and the residuals ($r_2 = |\tilde{x} - x|$) between them. The MSE and HallAssess scores are shown on the two residual images, r_1 and r_2, respectively. Scale bar: 0.5 μm for CCPs and F-actin datasets, 1 μm for MTs and ER datasets.

of AI-based reconstructions by evaluating how well they adhere to the characteristics of the original data. Notably, AI models may generate diverse outputs from the same input. However, reconstructions containing hallucinations are less likely to yield synthetic LQ images that resemble the original when processed through the imaging model. In contrast, when the synthetic LQ image closely aligns with the original, it confirms that the AI model has preserved essential structural and statistical features, thereby validating the reliability and quality of the reconstruction.

Furthermore, the proposed method enables quantitative evaluation of hallucinations in fluorescence microscopy image restoration results. It provides an objective scoring mechanism to assess the performance of FMIR (Fluorescence Microscopy Image Restoration) without requiring high-quality images as references. In this work, we employ the computation function of Mean Absolute Error (MAE) to derive the HallAssess score—a metric specifically designed to measure the extent of hallucination in AI-reconstructed images.

Our HallAssess effectively combines the strengths of both no-reference and full-reference image quality assessment paradigms. By shifting the evaluation from the high-quality domain to the low-quality domain (Fig. 1a), it achieves high accuracy without requiring a reference high-quality image y. Given an AI-reconstructed fluorescence microscopy image $f(x)$, our method generates a residual map that highlights potential artifacts or hallucinated regions, enabling precise and interpretable quality assessment of FMIR results.

AI-reconstructed hallucination assessment dataset

There is no dataset to assess hallucination in AI-reconstructed fluorescence microscopy images. Therefore, we adopted several representative FMIR models to generate AI-reconstructed fluorescence microscopy images $f(x)$ to evaluate hallucinations in images. The hallucination assessment datasets are composed of the original low-quality (LQ) images, the AI-reconstructed images of various FMIR models, and the original high-quality (HQ) images from the existing FMIR datasets (BioSR [1], Denoising_Tribolium [40], and Denoising_Planaria [40]). Note that the original high-quality images were only used to calculate the ground-truth hallucination assessment results.

First, we constructed the hallucination assessment datasets on the SR results. Specifically, we adopted 6 traditional super-resolution methods (bicubic, GPP, LANCZOS, bilinear, Nearest, ScSR [52], GPP [53]) and 17 deep-learning models, including convolutional neural network (CNN)-based models (SRCNN [54], VDSR [55], SelfExSR [56], EDSR [57], RCAN [58], RDN [59], ENLCN [60], PCEVA [61], DBPN [62], SRNO [63], ATD [64], CARE [40], DFCAN [1]), generative adversarial network (GAN)-based (SRGAN [65], LDL [66]), transformer-based model (UniFMIR [36]), and diffusion-based model (StableSR [67]). In addition to two fluorescence microscopy image super-resolution (SR) models (DFCAN [1], UniFMIR [36]), we re-trained SR models on the BioSR dataset [1]. Then, we super-resolved low-resolution (LR) images x from the test set of the BioSR dataset (MTs, F-actin, CCPs, ER) [1]. As for the hallucination assessment dataset on the denoising results, we also adopted representative fluorescence microscopy image denoising models (CARE [40], GVTNet [24], UniFMIR [36]) to generate the denoised results on the Denoising_Tribolium and Denoising_Planaria datasets [40].

Hallucination assessment in super-resolution results

We first demonstrate the HallAssess on the hallucination assessment of super-resolution results. Given an AI-reconstructed fluorescence microscopy image $f(x)$, we calculated the residual $r_1 = |f(x) - y|$ between $f(x)$ and the high-resolution (HR) image y to identify areas with visual artifacts and hallucination that need further improvement. Compared with the residual ($r_2 = |\tilde{x} - x|$) between synthetic low-resolution (LR) $\tilde{x}$ and the real LR image x, we found that the two residuals (r_1, r_2) are consistent for one image $f(x)$. It highlights that, even without a high-resolution (HR) reference image, our method can produce residual maps that closely match the residual maps obtained by using real high-quality reference images, demonstrating comparable perfor-

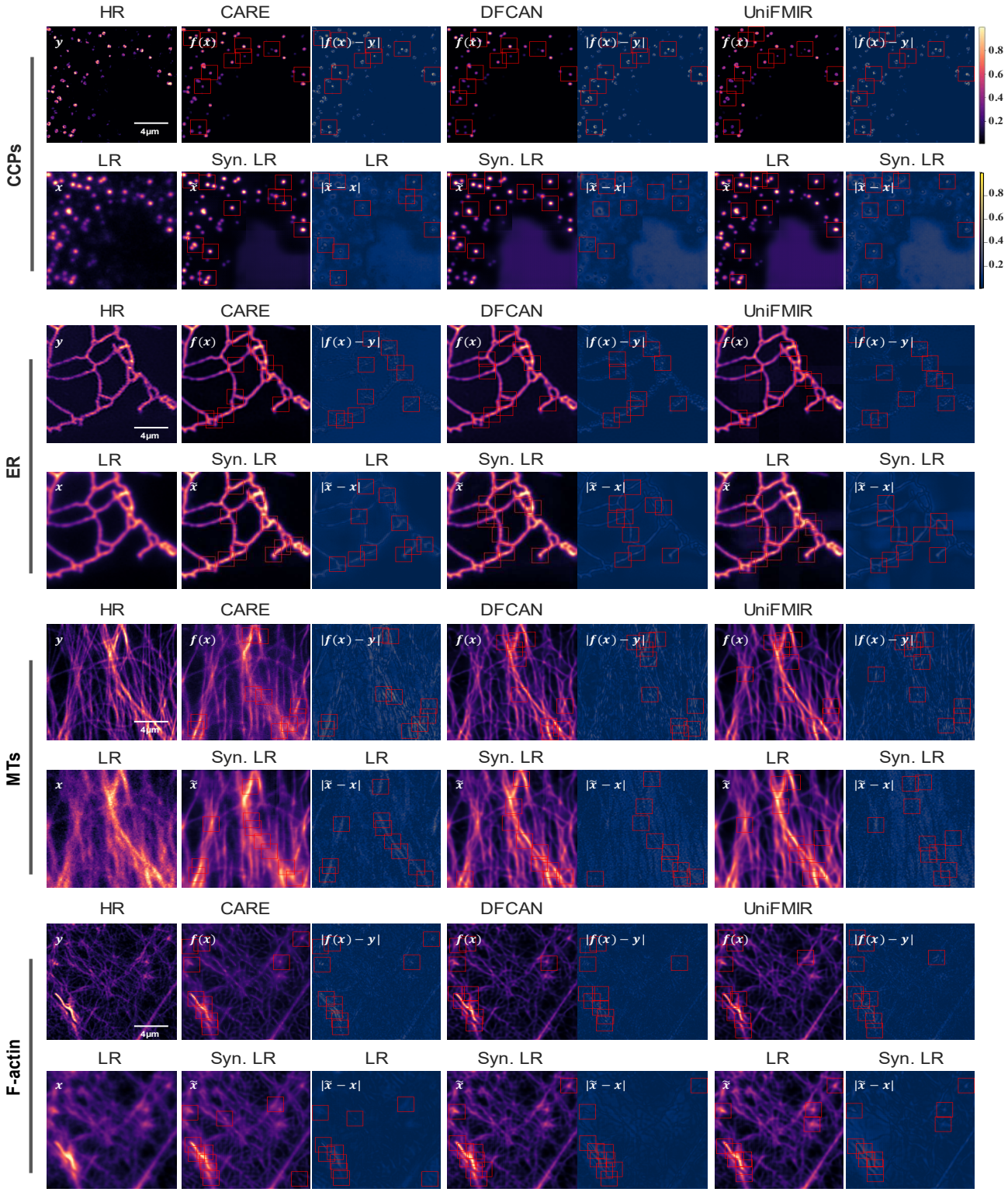


Figure 5 | Comparison between residual images calculated using the HR images and the residual images calculated using the original LR images. The synthetic LR images are obtained by applying the well-trained imaging model on $\times 2$ SR results of CARE [40], DFCAN [1], UniFMIR [36] from the BioSR datasets [1] (CCPs, F-actin, MTs, ER). The red rectangular boxes in the SR results, synthetic LR images, and residual images highlight the regions with the maximum residual values. The areas with large residuals between the SR and HR images largely overlap with those with large residuals between the synthetic LR images and the original LR images. Scale bar: 0.5 μm for CCPs and F-actin datasets, 1 μm for MTs and ER datasets.

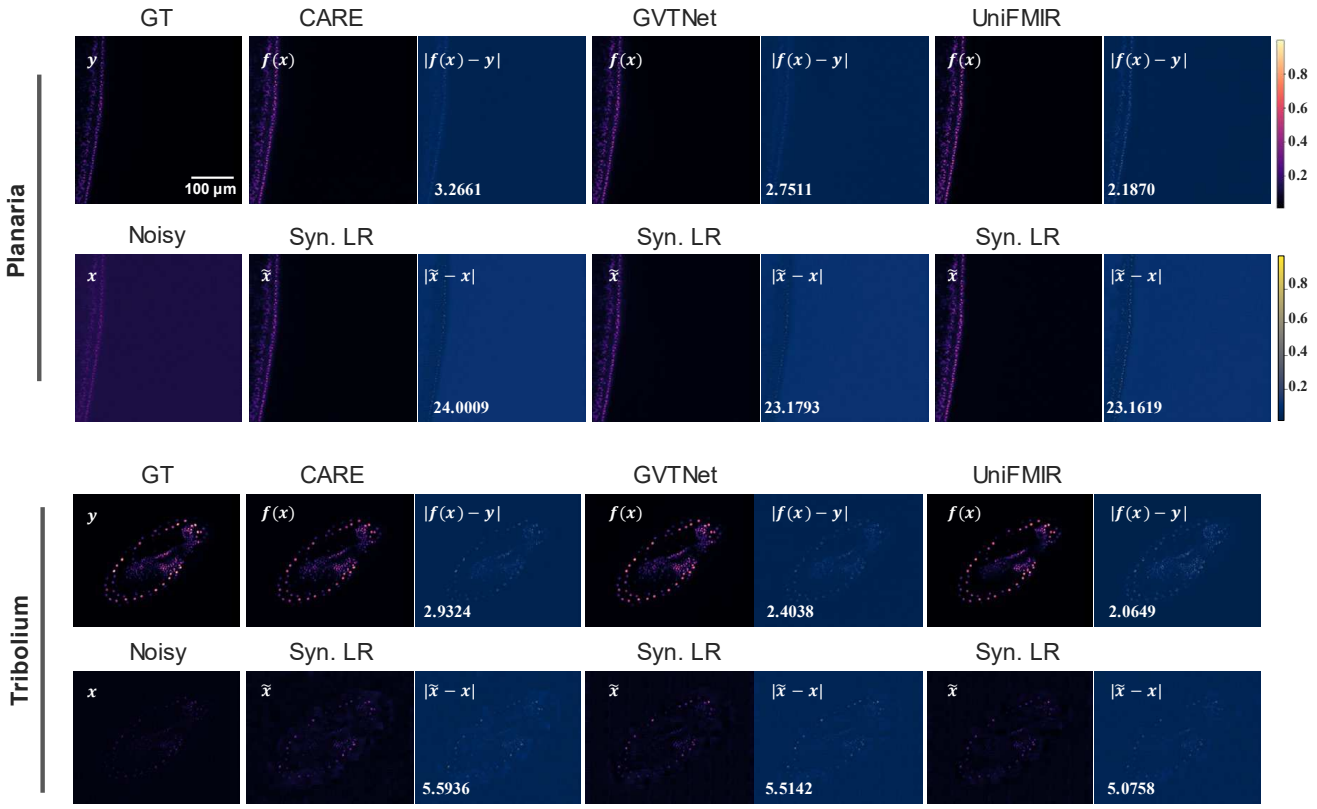


Figure 6 | Applying the proposed no-reference HallAssess on content-aware 3D image denoising results of developing castaneum embryos and flatworm (S. Mediterranea) from the Denoising_Planaria dataset (C2) and Denoising_Tribolium dataset (C1) [40]. The first line shows the ground-truth (GT) images, outputs of three denoising models (CARE [40], GVTNet [24], UniFMIR [36]), and the residual images between GT and the denoising results. The second line shows the original noisy images, synthetic low-quality (LQ) images from the corresponding denoising results, and the residuals between them. Scale bar: 100 µm.

mance to reference-based IQA methods (Fig. 1a). Our approach demonstrated its efficacy in highlighting the presence of visual hallucinations in AI-reconstructed images, helping researchers improve the FMIR method to mitigate the errors in the reconstructed results (Fig. 4).

The HallAssess scores were derived by quantifying the hallucinations with the MAE$(\tilde{x}, x)$ between the synthetic LR image $\tilde{x}$ and the real LR image x. Then, to provide an objective validation for the HallAssess, we regarded MAE$(f(x), y)$ between the AI-reconstructed image $f(x)$ and the HR image y as the ground truth (GT) scores. We observed that the distribution of the MAE$(\tilde{x}, x)$ scores is consistent with that of MAE$(f(x), y)$, demonstrating that the evaluation of hallucination in the LR image domain can also reflect hallucination in the HR image domain. The MAE$(\tilde{x}, x)$ scores of different SR models (CARE [40], DFCAN [1], UniFMIR [36]) have the same rank as the MAE$(f(x), y)$ scores. The UniFMIR has the lowest MAE$(f(x), y)$ scores, demonstrating the best performance on SR, and also has the lowest MAE$(\tilde{x}, x)$ scores (Fig. 4). To further demonstrate the performance of our HallAssess, we compared the HallAssess with SSIM [43], which is commonly used to evaluate the structural similarity between $f(x)$ image restoration accuracy by taking a real HR image y as a reference. The distributions of scores between SSIM and HallAssess provided insights into the performance of our HallAssess. Notably, our method, which does not require GT references, performs comparably to the full-reference SSIM (Fig. 3b).

Local hallucination assessment

The MAE$(\tilde{x}, x)$ score reflects the global hallucination performance. To further assess local hallucinations, we highlighted regions with maximum residual values in the residual maps r_1 and r_2, as indicated by red rectangular boxes. Specifically, these regions were 32×32 patches centered on the pixels with the highest residual values in the residual images. The residuals r_1 and r_2 were computed using high-resolution (HR) images y and original low-resolution (LR) images x as ground truth, respectively. By comparing these highlighted regions, we observed a significant overlap between areas of large residuals derived from the SR–HR comparison $r_1(f(x), y)$ and those from the synthetic LR–ground truth LR comparison $r_2(\tilde{x}, x)$. This indicates that our method can accurately reflect the quality of the super-resolution results by analyzing the correlation between synthetic and ground truth LR images (Fig. 5).

Hallucination assessment in 3D denoising results

We further evaluated the performance of our HallAssess in assessing the results of 3D fluorescence microscopy image denoising. Specifically, we applied it to denoised outputs generated by state-of-the-art methods, including CARE [40], GVTNet [24], and UniFMIR [36]. The input noisy images are taken from two widely used biological datasets: the Denoising_Tribolium dataset, consisting of developing Tribolium castaneum embryos [40], and the Denois-

ing_Planaria dataset, containing images of the flatworm Schmidtea mediterranea [40]. The corresponding residual images (Fig. 6) visually demonstrate that our HallAssess is capable of reflecting the restoration quality across all regions of interest, including both structurally complex and relatively homogeneous areas. Moreover, a consistent pattern emerges between the residuals $r_1(f(x), y)$, which compare the denoised outputs with high-resolution ground truth images, and $r_2(\tilde{x}, x)$, which assesses the fidelity of synthetic low-resolution images relative to original noisy inputs. This strong consistency suggests that HallAssess effectively captures the structural integrity and noise suppression performance of the denoising algorithms without requiring access to ground truth data, a key advantage in real-world applications where such references may not be available.

Due to the diversity of the imaging modality and degradation types in fluorescence microscopy images, such as noise contamination and low resolution, different reconstruction processes are needed, leading to different visual characteristics. Therefore, we separately validated our HallAssess on assessing the FMIR results of super-resolution and denoising tasks, the two most representative problems in the fluorescence microscopy image restoration field.

Discussion

Analysis of AI-reconstructed hallucination

This work addresses the critical challenge of assessing AI-reconstructed hallucinations in fluorescence microscopy images. As AI-based image restoration techniques have gained prominence, concerns about the reliability and fidelity of AI-reconstructed images have also emerged. AI-based fluorescence microscopy image restoration (FMIR) methods may introduce artifacts, such as 'hallucinating' false structures or artificially removing image structures, depending on the bias present in training sets [14]. We defined the phenomenon where the restored image contains visually plausible but inaccurate texture details as AI-reconstructed hallucinations in FMIR results. Scientists are increasingly cautious about the hallucinations introduced by these AI methods, which can compromise the authenticity and reliability of the reconstructed images.

The causes of AI-reconstructed hallucinations can be attributed to several key factors:

Overfitting: K.P. and C.S. [14] demonstrated that using AI in image restoration can lead to artifacts, often misconstrued as novel findings, owing to encountering unseen image data during training. Therefore, the model may overfit the training data, learning to generate details specific to the training set but not generalizing well to new, unseen images;

Data Limitations: Insufficient or biased training data can lead the model to learn incorrect or incomplete representations of the image features;

Loss Function Design: The choice of the loss function can influence the model's behavior. Loss functions, that focus solely on pixel-wise differences, may not capture higher-level structural or perceptual consistency, leading to hallucinations.

Contributions

This work delves into the phenomenon of AI-reconstructed hallucinations in fluorescence microscopy images. It analyzes the significance of accurate hallucination assessment for ensuring the robustness of subsequent research based on fluorescence microscopy. Here, we introduced HallAssess to accurately identify AI-reconstructed hallucinations in fluorescence microscopy image restoration results by distinguishing fake and genuine information. HallAssess can quantitatively measure visual artifacts produced during the restoration, enabling a more objective image quality assessment. The quality score of an FMIR result can be directly calculated without requiring a high-quality image aligned with the image as the reference to be evaluated.

HallAssess leverages the strengths of both no-reference and full-reference image quality assessment methods from the natural image domain, achieving near-full-reference accuracy under no-reference conditions. HallAssess effectively transfers the need for full-reference assessment to a no-reference condition. This approach identifies hallucination areas and calculates an overall quality score for the entire image.

Our comprehensive evaluation allows for a detailed comparison and analysis of the proposed approach's effectiveness and robustness. This provides a substantial practical advance in the field, advancing the assessment of AI-reconstructed hallucinations tailored to real-world conditions. HallAssess alleviates the challenge of lacking reference images, offering a more holistic understanding of the types and sources of visual artifacts. Consequently, our approach contributes significantly to the development of more reliable image restoration solutions and aids scientists in comprehending current FMIR performance, thereby facilitating further improvements in these techniques.

This methodology provides a robust framework for assessing the accuracy of AI-based reconstructions by evaluating their adherence to the original data's properties. Notably, AI models can produce varied outputs from identical inputs. However, reconstructions containing hallucinations are less likely to yield synthetic LQ images that match the original LQ images when degraded through a realistic imaging model. On the other hand, when the synthetic LQ image closely aligns with the original LQ image, it confirms that the AI model has retained essential structural and statistical features, thereby validating the reconstruction's reliability and quality.

By grounding our evaluation in a realistic LQ imaging mechanism, we aim to provide a more practical and pertinent assessment tool. This approach allows us to evaluate the fidelity of AI-reconstructed fluorescence microscopy images by comparing the 're-degraded' version (the synthetic LQ image) with the original LQ image. Thus, we ensure that the AI model's outputs are not only visually appealing but also scientifically accurate and reproducible.

Reasons for assessing in LQ domain

The central idea of this study is to explore whether the hallucinations within the low-quality (LQ) domain can reflect similar issues in the high-quality (HQ) domain. The underlying principle of our approach is that both the HQ and its corresponding LQ versions should convey consistent information beyond mere differences in resolution and visual clarity. In other words, we posit that images reconstructed by AI models, if they contain hallucinated or non-existent features, will not degrade into forms consistent with the original LQ im-

ages when subjected to realistic imaging degradation models. This inconsistency reveals potential flaws in the reconstruction process. Conversely, when the synthetic LQ images closely resemble their original counterparts, it indicates that the AI model has preserved the intrinsic properties of the source data, thereby affirming the fewer hallucinations present in the HQ reconstruction. This suggests that the AI model has minimized the introduction of artifacts or unrealistic details, thus enhancing the reliability of the reconstruction.

Limitations and future work

Our approach relies on the accuracy of the learned imaging model, which simulates the real degradation process from high-quality (HQ) to low-quality (LQ) images. While this mapping is generally simpler than the image restoration task, it still depends on the learning capability of the AI model. As such, any inaccuracies or biases in the training data or model architecture may limit the fidelity of the simulated degradation, thereby affecting the reliability of the hallucination assessment. To address these limitations, we will focus on incorporating the mechanism of microscopic imaging into the design and training of the degradation model in future work. By integrating domain-specific knowledge, such as point spread function (PSF) modeling, noise characteristics, and optical aberration correction, we aim to enhance the realism and generalization of the synthetic degradation process. This physics-informed approach will not only improve the interpretability of the model but also strengthen its ability to reflect true imaging conditions, leading to a more robust and trustworthy evaluation framework for AI-based image reconstruction methods.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (NSFC, 62472102 to Y.B., 62372117 to W.M.T., and 62402120 to C.X.M.) and Natural Science Foundation of Shanghai (Grant No.: 24ZR1490400). The computations in this research were performed using the CFFF platform of Fudan University.

Author contributions

W.T. and B.Y. conceived a framework and technique for researching hallucination assessment in AI-reconstructed fluorescence microscopic images. C.M., W.T., and B.Y. designed and implemented the algorithm. C.M. designed and performed the validation experiments. Y. S. organized the codes and models. C.M., W.T., and B.Y. analyzed the validation results. B.Y. supervised the research. All authors had access to the study data and contributed to writing the paper.

Competing interests

The authors declare no competing interests.

References

[1] C. Qiao et al. Evaluation and development of deep neural networks for image super-resolution in optical microscopy. *Nature Methods* pp. 1–9.

[2] D. David. Deep-learning for microscopy. *Nature Cell Biology* **24**, p. 1321 (2022).

[3] Milestones in light microscopy. *Nature Cell Biology* **11**, p. 1165 (2009).

[4] S. S. H. Z. e. a. Sau, A. Overlapping nuclear import and export paths unveiled by two-colour minflux. *Nature* **640**, p. 821–827 (2025).

[5] E. Callaway. 'democracy in microscopy': cheap light microscope delivers super-resolution images. *Nature* **616**, pp. 417–418 (2023).

[6] S. J. Holden, S. Uphoff and A. N. Kapanidis. Daostorm: an algorithm for high- density super-resolution microscopy. *Nature Methods* **8** no. 4, p. 279 (2011).

[7] S. Tonzani. Tirf: imaging at the cellular edge: Tirf microscopy. *Nature Cell Biology* **11** no. S1, pp. S16–S16 (2009).

[8] K. I. Willig et al. Sted microscopy with continuous wave beams. *Nature Methods* **4** no. 11, pp. 915–918 (2007).

[9] G. G. V. D. W. J. e. a. Shi, X. Super-resolution microscopy reveals that disruption of ciliary transition-zone architecture causes joubert syndrome. *Nature Cell Biology* **19**, p. 1178–1188 (2017).

[10] K. B. M. K. e. a. Mennella, V. Subdiffraction-resolution fluorescence microscopy reveals a domain of the centrosome critical for pericentriolar material organization. *Nature Cell Biology* **14**, p. 1159–1168 (2012).

[11] F. A. H. T. e. a. Schermelleh, L. Super-resolution microscopy demystified. *Nature Cell Biology* **21**, p. 72–84 (2019).

[12] G. G. V. D. W. J. e. a. Shi, X. Super-resolution microscopy reveals that disruption of ciliary transition-zone architecture causes joubert syndrome. *Nature Cell Biology* **19**, p. 1178–1188 (2017).

[13] C. J. C. L. e. a. Andrews, B. Imaging cell biology. *Nature Cell Biology* **24**, p. 1180–1185 (2022).

[14] K. Prakash, D. Baddeley and C. e. a. Eggeling. Resolution in super-resolution microscopy — definition, trade-offs and perspectives. *Nature Reviews Molecular Cell Biology* **25**, p. 677–682 (2024).

[15] Z. S. H. Z. e. a. Zhao, W. Enhanced detection of fluorescence fluctuations for high-throughput super-resolution imaging. *Nature Photonics* **17**, p. 806–813 (2023).

[16] W. H. Richardson. Bayesian-based iterative method of image restoration. *Journal Of The Optical Society Of America A-optics Image Science And Vision* **62**, p. 55–69 (1972).

[17] M. e. a. Arigovindan. High-resolution restoration of 3d structures from widefield images with extreme low signal-to-noise-ratio. *Proceedings of the National Academy of Sciences of the United States of America* **110**, p. 17 344–17 349 (2013).

[18] S. e. a. Preibisch. Efficient bayesian-based multiview deconvolution. *Nature Methods* **11**, p. 645–648 (2014).

[19] Y. Wu, X. Han and Y. e. a. Su. Multiview confocal super-resolution microscopy. *Nature* **600**, pp. 279–284 (2021).

[20] T. L. Waller, L. Machine learning for 3d microscopy. *Nature* **523**, p. 416–417 (2015).

[21] A. Dance. Sharper signals: how machine learning is cleaning up microscopy images. *Nature* **589**, pp. 318–319 (2021).

[22] X. Li, Y. Wu and Y. e. a. Su. Three-dimensional structured illumination microscopy with enhanced axial resolution. *Nature Biotechnology* **41**, p. 1307–1319 (2023).

[23] C. J et al. Three-dimensional residual channel attention networks denoise and sharpen fluorescence microscopy image volumes. *Nature Methods* **18** no. 11, p. 678–687 (2021).

[24] Z. Wang, Y. Xie and S. Ji. Global voxel transformer networks for augmented microscopy. *Nature Machine Intelligence* **3** no. 2, pp. 161–171 (2021).

[25] L. Jin et al. Deep learning enables structured illumination microscopy with low light levels and enhanced speed. *Nature communications* **11** no. 1, pp. 1–7 (2020).

[26] L. von Chamier et al. Democratising deep learning for microscopy with zerocostdl4mic. *Nature communications* **12** no. 1, pp. 1–18 (2021).

[27] A. R. Durmaz et al. A deep learning approach for complex microstructure inference. *Nature communications* **12** no. 1, pp. 1–15 (2021).

[28] X. Li et al. Real-time denoising enables high-sensitivity fluorescence time-lapse imaging beyond the shot-noise limit. *Nature Biotechnology* **41** no. 2, pp. 282—292 (2022).

[29] C. Qiao et al. Rationalized deep neural network for sustained super-resolution live imaging of rapid subcellular processes. *Nature Biotechnology* **41**, pp. 367—-377 (2023).

[30] Z. Wang, L. Zhu and H. e. a. Zhang. Real-time volumetric reconstruction of biological dynamics with light-field microscopy and deep learning. *Nature Methods* **18**, p. 551–556 (2021).

[31] R. L. Belthangady, C. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nature Methods* **16**, p. 1215–1225 (2019).

[32] Y. Wu and H. Shroff. Faster, sharper, and deeper: structured illumination microscopy for biological imaging. *Nature Methods* **15**, pp. 1011–1019 (2018).

[33] H. Wang et al. Deep learning enables cross-modality super-resolution in fluorescence microscopy. *Nature Methods* **16** no. 1, pp. 103–110 (2018).

[34] R. Chen, X. Tang and Y. e. a. Zhao. Single-frame deep-learning super-resolution microscopy for intracellular dynamics imaging. *Nature Communications* **14** no. 2854.

[35] Y. Xu, A. Graves and G. e. a. Coste. Cross-modality supervised image restoration enables nanoscale tracking of synaptic plasticity in living mice. *Nature Methods* **20**, p. 935–944 (2023).

[36] C. Ma et al. Pretraining a foundation model for generalizable fluorescence microscopy-based image restoration. *Nature Methods* **21**, p. 1558–1567 (2024).

[37] S. e. a. Preibisch. Efficient bayesian-based multiview deconvolution. *Nature Methods* **11**, p. 645–648 (2014).

[38] N. e. a. Gustafsson. Fast live-cell conventional fluorophore nanoscopy with imagej through super-resolution radial fluctuations. *Nature Communications* **7** no. 12471.

[39] M. Arigovindan et al. High-resolution restoration of 3d structures from widefield images with extreme low signal-to-noise-ratio. *Proceedings of the National Academy of Sciences of the United States of America* **110** no. 43.

[40] M. Weigert, U. Schmidt and T. e. a. Boothe. Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nature Methods* **15** no. 12, pp. 1090–1097 (2018).

[41] X. Yu et al. Fake artificial intelligence generated contents (faigc): A survey of theories, detection methods, and opportunities. *ArXiv* **abs/2405.00711**.

[42] H. Zhu et al. Adaptive image quality assessment via teaching large multimodal model to compare. In *Conference on Neural Information Processing Systems (NeurIPS)* (2024).

[43] Z. Wang et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13** no. 4, pp. 600–612 (2004).

[44] S. Culley, D. Albrecht and C. e. a. Jacobs. Quantitative mapping and minimization of super-resolution optical imaging artifacts. *Nature Methods* **15**, p. 263–266 (2018).

[45] M. C. T. D. A. D.-a. S. V. Koho, G. Tortarolo and G. Vicidomini. Fourier ring correlation simplifies image restoration in fluorescence microscopy. *bioRxiv* **151** no. 3, pp. 535–583 (2019).

[46] R. Nieuwenhuizen, K. Lidke and M. e. a. Bates. Measuring image resolution in optical nanoscopy. *Nature Methods* **10**, p. 557–562 (2013).

[47] H. X. Y. J. e. a. Zhao, W. Quantitatively mapping local quality of super-resolution microscopy by rolling fourier ring correlation (2023).

[48] K. C. Alex Wang and M. Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020).

[49] H. H. Esin Durmus and M. Diab. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020).

[50] R. A. E. N. I. S. Or Honovich, Leshem Choshen and O. Abend. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (2021).

[51] S. L. B. P. J. S.-A. W. Thomas Scialom, Paul-Alexis Dray and P. Gallinari. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (2021).

[52] J. Yang et al. Image super-resolution via sparse representation. In *IEEE Transactions on Image Processing (TIP)*, vol. 19 (2010).

[53] J. Sun, Z. Xu and H.-Y. Shum. Image super-resolution using gradient profile prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2008), pp. 1–8.

[54] C. Dong et al. Image super-resolution using deep convolutional networks. *IEEE Trans Pattern Anal Mach Intell* **38** no. 2, pp. 295–307 (2014).

[55] J. Kim, J. K. Lee and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016), pp. 1646–1654.

[56] J.-B. Huang, A. Singh and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 5197–5206.

[57] B. Lim et al. Enhanced deep residual networks for single image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPRW)*, vol. 1 (2017), p. 4.

[58] Y. Zhang et al. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision (ECCV)* (2018).

[59] Y. Zhang et al. Residual dense network for image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[60] B. Xia et al. Efficient non-local contrastive attention for image super-resolution. In *Association for the Advancement of Artificial Intelligence (AAAI)* (2022).

[61] J. Chao et al. Equivalent transformation and dual stream network construction for mobile image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 14 102–14 111. doi:10.1109/CVPR52729.2023.01355.

[62] M. Haris, G. Shakhnarovich and N. Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[63] M. Wei and X. Zhang. Super-resolution neural operator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 18 247–18 256.

[64] L. Zhang et al. Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 2856–2865.

[65] C. Ledig et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 105–114.

[66] J. Liang, H. Zeng and L. Zhang. Details or Artifacts: A Locally Discriminative Learning Approach to Realistic Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 5647–5656.

[67] J. Wang et al. Exploiting diffusion prior for real-world image super-resolution. In *International Journal of Computer Vision* (2024).

[68] J. Liang et al. Swinir: Image restoration using swin transformer. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (2021), pp. 1833–1844. doi:10.1109/ICCVW54120.2021.00210.

[69] M. G. L. Gustafsson. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. *Journal of Microscopy* **198** no. 2, pp. 82–87 (2000).

[70] R. Heintzmann and M. Gustafsson. Subdiffraction resolution in continuous samples. *Nature Photonics* **3**, p. 362–364 (2009).

[71] K. I. Mortensen et al. Optimized localization analysis for single-molecule tracking and super-resolution microscopy. *Nature Methods* **7**, pp. 377–381 (2010).

[72] B. E, P. G. H and e. a. Sougrat R. Imaging intracellular fluorescent proteins at nanometer resolution. *Betzig E , Patterson G H , Sougrat R ,et al.* **5793** no. 313, p. 362–364 (2006).

Methods

Data preparation process

To generate the experimental data, that is, the fluorescence microscopy images to be evaluated, we first collected public datasets, including BioSR [1], Denoising_Planaria, and Denoising_Tribolium datasets [40], and preprocessed images in the datasets with the data processing method in UniFMIR [36]. Then, we acquired AI-reconstructed fluorescence microscopy images using existing image restoration methods.

To generate super-resolved images, we super-resolved the low-resolution (LR) fluorescence microscopy images from the BioSR [1] dataset using current deep-learning-based and non-deep-learning-based image restoration methods. In specific, thee deep-learning-based methods include convolutional neural network (CNN)-based models (SRCNN [54], VDSR [55], SelfExSR [56], EDSR [57], RCAN [58], RDN [59], ENLCN [60], PCEVA [61], DBPN [62], SRNO [63], ATD [64]), generative adversarial network (GAN)-based (SRGAN [65], LDL [66]), transformer-based model (SwinIR [68]), and diffusion-based model (StableSR [67]). The non-deep-learning-based natural image restoration methods include interpolation-based (bilinear, bicubic, LANCZOS, Nearest), sparse-coding-based (ScSR [52]), and edge-based (GPP [53]). We further experimented with the evaluation of 3D denoising results. To generate denoised images, we adopted the denoising models (CARE [40], GVTNet [24], SwinIR [68]), trained on the Denoising_Tribolium and Denoising_Planaria datasets [40].

Hallucination assessment

The proposed HallAssess assesses the hallucinations by calculating the correlation between the source low-quality image and the low-quality image obtained by applying a realistic imaging model to the AI-reconstructed image. Here is the detailed workflow of our approach:

1. Development of an imaging model. We first constructed a realistic low-quality imaging model to transform high-quality images into realistic low-quality images. Thus, the low-quality imaging model can be regarded as a reverse inference of the image restoration process, which aims to transform low-quality images into realistic, high-quality images. This imaging model was trained on realistic datasets, composed of high-quality and low-quality fluorescence microscopy image pairs, to simulate real-world noise, blur, and artifacts in conditions under which low-quality images are produced.

2. Low-quality image generation for full-reference condition. Once the imaging model is established, we fed AI-reconstructed fluorescence microscopy images $f(x)$ into this model to generate "re-degraded" versions, that are the synthetic low-quality images $Imaging(f(x))$, of these reconstructed images. These "re-degraded" images simulate the realistic degradation that would occur under actual imaging conditions. By pairing the original low-quality image x with the generated low-quality versions of the outputs of different fluorescence microscopy image restoration models, we can effectively transform the no-reference assessment scenario into a full-reference one, where the original low-quality image serves as the reference. This strategy enables more reliable and quantifiable evaluation of image restoration quality.

3. Hallucination assessment. With the paired low-quality images, we can visually identify regions affected by AI hallucinations by computing the residual between the original and synthetic low-quality images. Additionally, we calculated a quantitative score to evaluate the correlation between these two image sets using various similarity metrics. In this study, we employ the Mean Absolute Error (MAE) as the HallAssess score to assess whether the AI-reconstructed images maintain fidelity with the actual low-quality imaging processes. This approach enables both qualitative visualization and quantitative evaluation of the degree of hallucination in AI-reconstructed fluorescence microscopy images and the fidelity of restored images.

$$MAE = \frac{1}{N}\sum_{i=1,j=1}^{H,W} |x(i,j) - Imaging(f(x))(i,j)|, \quad (1)$$

where H, W denote the height and width of the image x, $x(i,j)$ denotes the $(i,j)-th$ pixel in x.

Current image quality assessment (IQA) methods

There is no method for assessing the AI-reconstructed hallucinations in fluorescence microscopy images. The straightforward solutions for measuring AI-reconstructed fluorescence microscopy images are current resolution measurement or image quality assessment (IQA) methods, including full-reference IQA metrics (PSNR, SSIM [43]) and no-reference IQA metrics (NIQE, FWHM, FRC [45, 46]), to measure the performance of different FMIR methods. However, the practicability of full-reference methods is limited, as they require high-quality images as a reference to calculate the quality score of an image, and there is currently no universal FMIR evaluation standard. In addition, these traditional image quality assessment (IQA) methods do not focus on measuring AI-reconstructed hallucinations in FMIR results. The specific calculations of these methods are introduced below.

The peak signal-to-noise ratio (PSNR) measures the pixel-level similarities between a restored image $f(x)$ and a GT image y. A higher PSNR denotes a higher restoration accuracy of $f(x)$, and PSNR can be calculated by:

$$\begin{aligned} MSE &= \frac{1}{N}\sum_{i=1,j=1}^{H,W}(y(i,j) - f(x)(i,j))^2, \\ PSNR &= 10log_{10}(\frac{L^2}{MSE}), \end{aligned} \quad (2)$$

where H, W denote the height and width of the image y, $y(i,j)$ denotes the $(i,j)-th$ pixel in y and L denotes the maximum pixel value of the image.

The structural similarity index measure (SSIM) [43] measures the structure-level similarities between $f(x)$ and y. A higher SSIM denotes a higher restoration accuracy of $f(x)$. The SSIM is given by:

$$SSIM(f(x), y) = \frac{(2\mu_{f(x)}\mu_y + c_1)(2\sigma_{f(x)y} + c_2)}{(\mu_{f(x)}^2 + \mu_y^2 + c_1)(\sigma_{f(x)}^2 + \sigma_y^2 + c_2)}, \quad (3)$$

where $\mu_{f(x)}, \mu_y$ is the mean intensity of image $f(x), y$, $\sigma_{f(x)}^2, \sigma_y^2$ is the variance of image $f(x), y$, $\sigma_{f(x)y}$ is the covariance of images $f(x)$ and y, $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are constants for numerical stability with L being the dynamic range of the pixel values.

NanoJ-SQUIRREL [44] enables physics-anchored absolute error assessment, requiring simultaneous acquisition of wide-field/SR images. Calculate the resolution-scaled error (RSE) and resolution-scaled Pearson coefficient (RSP), representing the root mean square error and the Pearson correlation coefficient between the reference and resolution-scaled images.

The Naturalness Image Quality Evaluator (NIQE) is based on the premise that natural scene statistics (NSS) models can characterize the statistical properties of natural images to evaluate the quality of natural images without a reference image. The NIQE score can be obtained by following the steps: Extracting a set of statistical features from the input image $f(x)$; Using a pre-trained multivariate Gaussian model to fit the feature distribution of natural images; Computing the distance between the feature vector of the $f(x)$ and the model of natural image features to obtain a quality score.

The optical resolution, the inverse of the imaging system's spatial bandwidth, is a basic metric for representing the microscopy instrument's capability to distinguish between two close objects in an image and is critical for determining the level of detail observed in a sample. The diffraction limit theory proposed by German scientist Ernst Abbe revealed the resolution limit of traditional optical microscopy caused by light's diffraction effect and finite aperture resolution. The resolution is fundamentally limited by the wavelength of light (λ) and the imaging lens's numerical aperture (NA), and it can be calculated by Eq.(4). Advances in microscopy techniques, such as optimization of the microscopy hardware (SIM [69, 70], super-resolution microscopy [71, 72]) and computational-based FMIR algorithms [36, 40], have pushed past this classical limit, allowing for the visualization of features at the nanoscale that were previously unobservable.

$$d = \frac{0.61 \cdot \lambda}{NA}. \tag{4}$$

Generally, the optical resolution of a microscope is predicted by resolution prediction approaches, such as the Fourier ring correlation (FRC)-inspired resolution measure [45, 46] and the Full Width at Half Maximum (FWHM). The Fourier Ring Correlation (FRC) is a straightforward and objective method to measure effective image resolution [45, 46]. We adopted the public codes for the analysis of image resolution of FRC-based fluorescence microscopy. The Fourier ring correlation (FRC) variant, rFRC [47], remains confined to super-resolution contexts, employing rolling FRC analysis for 20 nm-scale resolution heterogeneity mapping combined with PANEL modules' spatial-frequency verification. While effective for resolution uncertainty localization, this spectral SNR estimation paradigm fails to address the spectrum of hallucination artifacts emerging in general fluorescence microscopy restoration tasks. FRC [45] measures the resolution of two-dimensional (2d) and 3d localization microscopy images.

The Full Width at Half Maximum (FWHM) quantifies the sharpness or resolution of an image, particularly in fields such as astronomy, microscopy, and imaging science. It measures the width of a peak in a point spread function (PSF) at the point where the intensity is half of the peak maximum. For a given intensity profile $y(i)$ of a point source in an image y, the FWHM can be determined by finding the points on either side of the peak where the intensity drops to half of its maximum value. Mathematically, this can be expressed as $y(i_1) = y(i_2) = y_{max}/2$, where $y(i_1)$ and $y(i_2)$ are the positions corresponding to the half-maximum intensities. The FWHM is then calculated as the absolute difference between these two positions, $FWHM = |y(i_2) - y(i_1)|$.

Data availability

The example images involved in the experiments can be found on GitHub (https://github.com/cxm12/HallAssess). The imaging models are trained on publicly available datasets by taking the HR/clean images as input and the LR/noisy images as input. The 3D denoising datasets can be downloaded from https://publications.mpi-cbg.de/publications-sites/7207/. The SR dataset can be downloaded from https://doi.org/10.6084/m9.figshare.13264793. The AI-reconstructed images, which serve as the HallAssess testing image, can be obtained by applying an AI-based fluorescence microscopy image restoration method on the original low-quality (LQ) image.

Code availability

The codes and models associated with some example images are freely available at https://github.com/cxm12/HallAssess (Supplementary Figure 3). We also provide a live demo for HallAssess at http://fdudml.cn:7871/evaluation. Users can follow the steps in the 'Instructions' section of the network to run the demo locally. This newly built interactive software platform allows users to use our method freely and easily.

We also established an online leaderboard for fluorescence microscopy image restoration (http://fdudml.cn:7871/leaderboard, Supplementary Figure 4) to provide a convenient and intuitive tool, allowing researchers to easily track and clearly understand the current advancements in existing fluorescence microscopy image restoration methods. This platform offers low-quality fluorescence microscopy images for testing in real-world applications. Users can submit their restoration results. Then, we will score and rank their results using our HallAssess and update the leaderboard statistics.

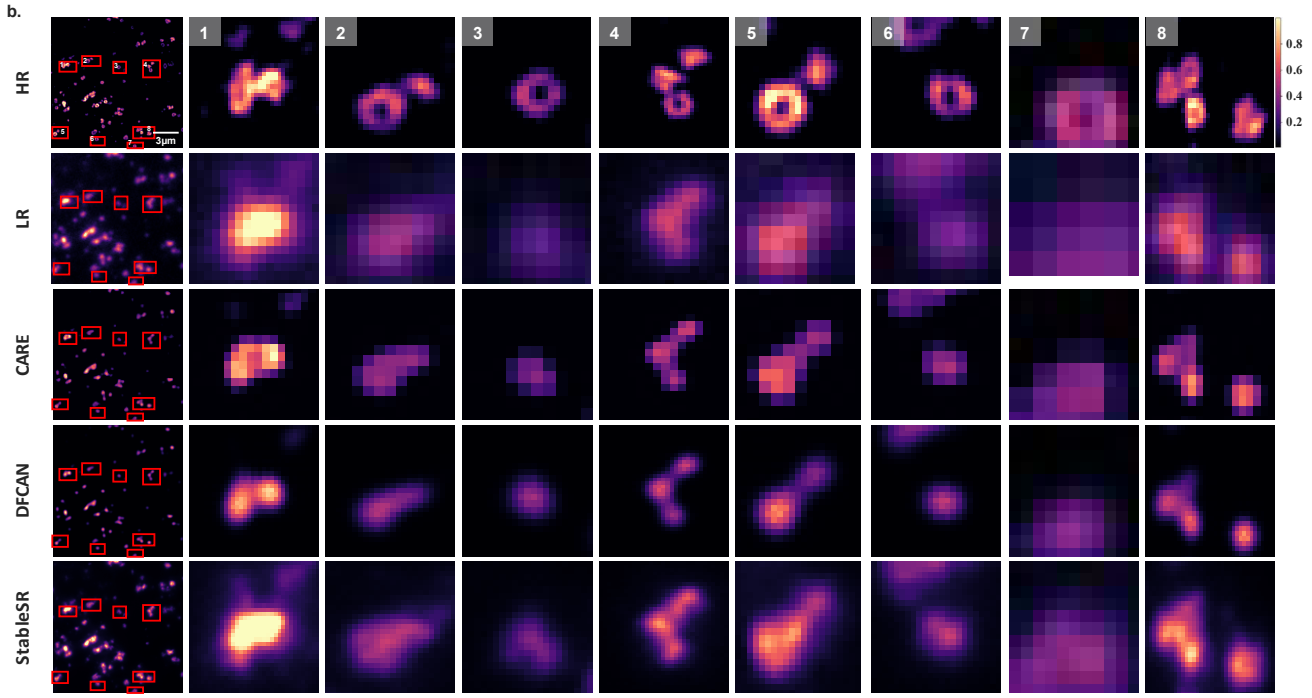
Supplementary information

Supplementary Figure 1 presents examples of AI-reconstructed hallucinations in the same region, as observed in FMIR results for ×2 upscaling on the CCPs dataset.

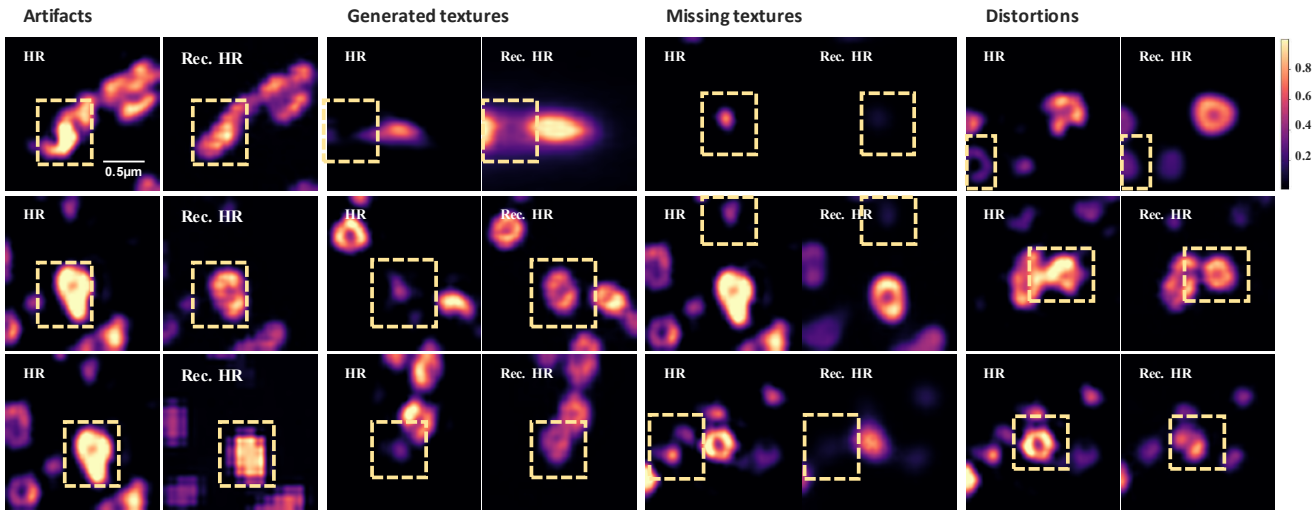
Supplementary Figure 2 shows a visualization of four types of hallucinations, including artifacts, distortion, generated texture, and missing texture, in FMIR results.

Supplementary Figure 3 shows the web-based assessment platform, which provides researchers with an intuitive interface for hallucination assessment in AI-reconstructed fluorescence microscopy images. This no-reference assessment system eliminates the dependency on ground-truth images while delivering immediate quality scores through automated HallAssess computations.

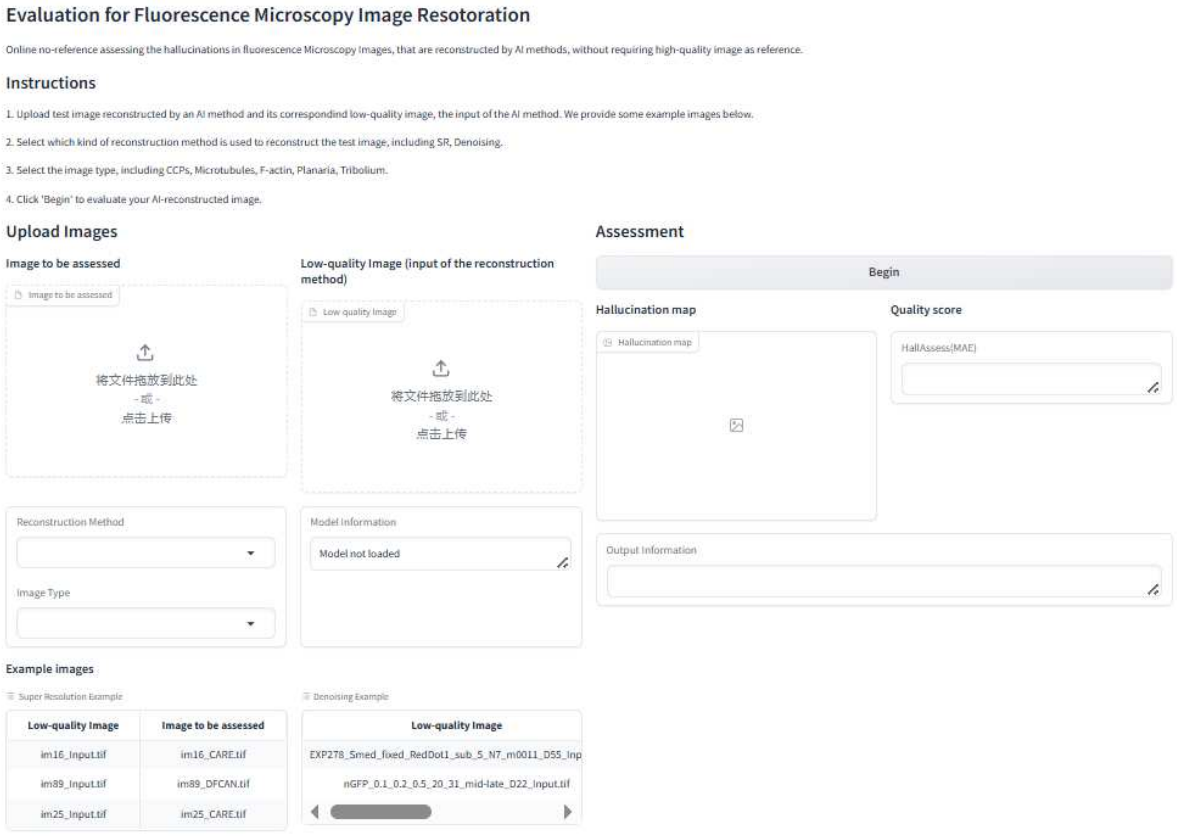
Supplementary Figure 4 illustrates the interactive leaderboard, which provides test images and can automatically update method rankings by assessing submissions with the HallAssess score. This dynamic benchmarking tool supports comparative analysis between emerging deep learning approaches.



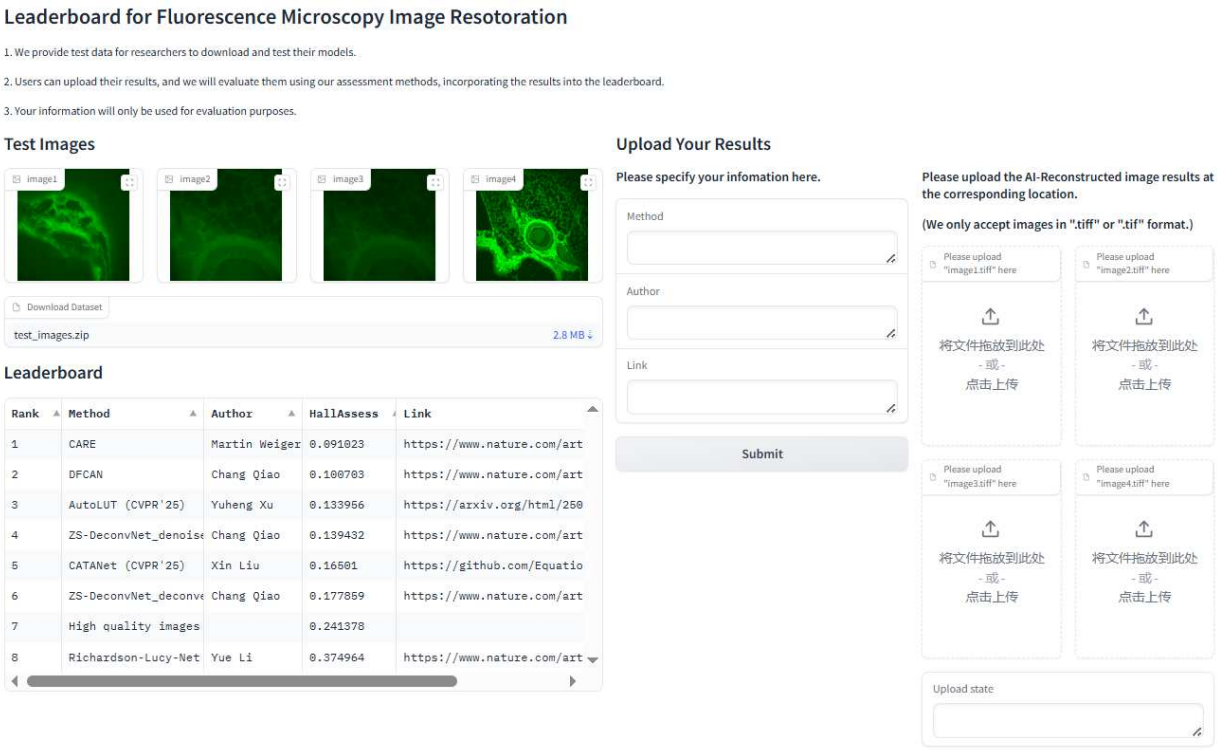
Supplementary Figure 1 | Demonstration of AI-reconstructed hallucinations of the same region across different FMIR results for ×2 upscaling on CCPs dataset [1]. Compared with the high-resolution image, the low-resolution image lost most details, making it difficult for AI-based FMIR models to restore real textures.



Supplementary Figure 2 | Visualization of Four Types of AI-reconstructed Hallucinations: Artifacts, Distortion, Generated Texture, and Missing Texture in FMIR Reconstruction.



Supplementary Figure 3 | Interface demonstration of the fluorescence microscopy image restoration assessment platform. Users can upload their fluorescence microscopy image restoration results. We can assess the AI-reconstructed hallucinations in these images without requiring high-quality images as references.



Supplementary Figure 4 | Interactive leaderboard interface for fluorescence microscopy image restoration. We provide low-quality testing data and a real-time submission portal for uploading restoration results. The auto-refreshing ranking panel can display a comparative analysis of state-of-the-art approaches, evaluated by our HallAssess.