Advanced Intelligent Discovery

ADVANCED INTELLIGENT DISCOVERY Open Access

RESEARCH ARTICLE OPEN ACCESS

Real-Time Multicolor Fluorescence Microscopy via Cross-Channel Acquisition and Deep-Learning-Based Inference

Juan J. Huaroto | Yuxin Jin | Nicholas Roy | Sarthak Misra

Surgical Robotics Laboratory, Department of Biomechanical Engineering, University of Twente, Enschede, The Netherlands

Correspondence: Juan J. Huaroto (j.j.j.huarotosevilla@utwente.nl) | Sarthak Misra (s.misra@utwente.nl)

Received: 26 September 2025 | **Revised:** 7 February 2026 | **Accepted:** 25 February 2026

Keywords: machine learning | microrobotics | sequential imaging | spectral crosstalk

ABSTRACT

Monitoring real-time interactions in microrobotic and biological systems necessitates rapid multicolor fluorescence microscopy to resolve cellular, micro-agent, and environmental dynamics. Sequential acquisition is widely used to capture multiple fluorescence channels while minimizing spectral crosstalk and photobleaching. However, this approach limits imaging speed in proportion to the number of channels, reducing its suitability for capturing micro-agents and cellular interactions. This study introduces a real-time multicolor reconstruction framework that exploits cross-channel inputs (frames containing mixed spectral contributions) to enable simultaneous reconstruction of target fluorescence channels. The framework is evaluated on a custom-built three-channel fluorescence microscope, benchmarking two representative models: a standard supervised convolutional encoder–decoder with skip connections (U-Net) and an adversarially trained conditional model (pix2pix). Experimental validation is conducted in a microfluidic environment containing active functionalized structures (Coumarin 6-labeled electrospun magnetic fibers) and passive biological agents (CellTracker Red CMTPX-labeled HeLa cell spheroids), where external magnetic fields actuate the micro-agents to drive interactions with the spheroids. Prediction performance is evaluated in two- and three-channel settings, yielding high-fidelity reconstructions at 13.9 ms inference time. The proposed approach can increase the effective frame rate for three-color channels by up to 83%, enabling high-throughput multicolor imaging.

1 | Introduction

Multicolor fluorescence microscopy is a fundamental tool in modern bioimaging, enabling the simultaneous visualization of multiple biological structures through spectrally distinct fluorophores [1, 2]. In recent years, its capabilities have become increasingly valuable in medical microrobotics and related fields, where functional micro-agents are deployed to interact with living tissues [3], deliver therapeutics [4], detect toxins [5, 6], or perform precision biopsies within complex biological environments [7, 8]. Visualizing the spatiotemporal dynamics of such micro-agents, often involving multiple fluorophore-labeled agents, requires high spectral specificity and real-time imaging to capture rapid events such as binding, internalization, or mechanical manipulation [9]. Applications such as targeted cancer therapy, localized drug delivery, and minimally invasive interventions can benefit from multicolor imaging, enabling simultaneous monitoring of micro-agent motion, cellular responses, and environmental markers [10–12]. However, simultaneously achieving high temporal resolution and spectral discrimination for microrobotic systems interacting with biological samples remains challenging [13–15].

Multicolor fluorescence systems typically balance spectral specificity against acquisition time in four ways: (1) Sequential time-multiplexing alternates excitation and detection paths to acquire each channel separately [16]. This minimizes crosstalk but reduces the effective frame rate and risks temporal misregistration in rapidly moving scenes [17]. (2) Hyperspectral microscopy and Fourier

transform imaging spectroscopy (FTIS) resolve a per-pixel spectrum and separate fluorophores by spectral unmixing using measured libraries [18, 19]. In FTIS, the interferogram of each pixel is recorded by stepping the optical path difference, and its spectrum is recovered via inverse Fourier transform with phase/apodization corrections [20, 21]. These approaches typically require spectral calibration and are computationally demanding, which can constrain real-time use [22–24]. (3) Snapshot/compressive spectral imaging acquires one or a few coded measurements and reconstructs a spatial–spectral datacube (or band images) by solving an inverse problem—classically via iterative, sparsity-regularized solvers. This trades spatial/spectral resolution for speed and can be sensitive to motion and solver artifacts. (4) Optical multiplexing without scanning assigns spectral identity to engineered point spread function (PSF) shapes for computational classification [25]. Performance is sensitive to PSF morphology—which varies with depth, aberrations, and emitter density/crowding—making it less robust in dynamic, extended, or scattering scenes. Another optical-multiplexing route is simultaneous multichannel acquisition (e.g., beam splitters/multisensors), which demands precise interchannel registration and crosstalk calibration for software correction [26].

Beyond acquisition strategies, deep learning has broadened fluorescence workflows by extracting spatial and spectral structure from limited measurements [27]. AutoUnmix, a lightweight autoencoder, performs spectral unmixing of multispectral inputs with near-real-time GPU inference [28]. Generative adversarial network (GAN) models add perceptual/detail fidelity and distribution matching, enabling cross-modality super-resolution and 3D unpaired translation [29, 30]. Task-assisted GAN models couple resolution enhancement with downstream biological segmentation [31]. Learned PSF models support filter-free, color-to-PSF classification for multicolor localization microscopy [32]. Snapshot/confocal spectral systems have been accelerated with deep networks for fast band reconstruction [33], and hyperspectral spectrum imaging has likewise leveraged generative models (SpecGAN) for learned spectral recovery [34]. Filter-free fluorescence microscopes driven by deep learning models further reduce hardware complexity while computationally recovering color channels [35]. Despite the aforementioned advancements, most deep learning applications report spectral specificity on static samples, with less emphasis on real-time reconstruction during motion. Furthermore, while recent studies often target specialized optical platforms, routine multicolor microscopy mainly uses sequential acquisition in standard microscopes. Applying deep learning in such settings remains comparatively underexplored—particularly for microrobotic systems interacting with biological samples.

This study presents a real-time multicolor reconstruction framework that simultaneously predicts target fluorescence channels from cross-channel inputs (i.e., frames containing mixed spectral contributions). During framework deployment, it replaces sequential multichannel stacks by deriving target fluorescence channels from cross-channel frames. As a result, it shortens acquisition time and excitation dose while preserving channel-specific fidelity and temporal coherence. We benchmark the framework on a custom-built three-channel fluorescence microscope using two representative models under matched settings: a supervised U-Net and a pix2pix model with an identical U-Net generator and a PatchGAN discriminator. Validation is performed on a microfluidic platform under magnetic manipulation, comprising Coumarin 6-labeled electrospun magnetic fibers (active micro-agents) interacting with CellTracker Red CMTPX-labeled HeLa spheroids (passive agents). In this setting, the framework supports spatial analysis of spheroid position and micro-agent orientation while maintaining performance on unseen sequences and under reduced-excitation conditions. Overall, the proposed strategy provides a practical and scalable route to real-time multicolor fluorescence imaging in dynamic microsystems.

2 | Results and Discussions

The results are organized to introduce and evaluate the proposed framework for real-time multicolor reconstruction. We first outline the cross-channel acquisition and deep-learning-based inference approach, and motivate real-time constraints with a dynamic micro-agent–spheroid assay under magnetic manipulation. We then quantify the structural informativeness of the cross-channel inputs relative to each target channel using a gradient-based contrast similarity metric. Reconstruction fidelity and latency are assessed with quantitative metrics and visual comparisons. We further examine performance under reduced excitation, analyze fluorescence efflux over time, and probe spatial accuracy with a tracking-based analysis. We extend the workflow from two to three channels by incorporating an auxiliary background-fluorescence measurement. Finally, we test our approach on morphology-matched HeLa cell spheroids labeled with different fluorescent dyes.

2.1 | Cross-Channel Acquisition and Deep-Learning-Based Inference

In multicolor fluorescence microscopy, achieving spectral separation between fluorescence channels requires careful selection of labels that can be independently excited by tunable or discrete wavelengths. Each fluorescent label emits spectrally filtered fluorescence upon excitation and is directed to one or more detectors (e.g., cameras or photomultipliers) for image formation. To minimize spectral crosstalk and cumulative phototoxicity, channels are typically acquired sequentially using synchronized excitation sources and either a shared detector with tunable filters or multiple source–detector pairs. However, this sequential strategy can be limited in real-time multicolor fluorescence microscopy, as the effective frame rate decreases with the number of channels, and timing becomes more demanding as throughput increases. Figure 1A illustrates this conventional strategy, where staggered timing underscores the sequential nature of acquisition and each frame corresponds to a matched excitation–emission pair.

In this study, we introduce a cross-channel acquisition strategy coupled with deep-learning-based reconstruction. Instead of acquiring each fluorescence channel separately, cross-channel frames are captured using crossed excitation/emission paths (Figure 1B). This way, based on the fluorophore spectra and filter sets, the acquired images contain a mixed but informative contribution from the target channel. A deep-learning model then uses this mixed input to reconstruct the target channels simultaneously in real time. This reduces the number of exposures per acquisition cycle, increases the effective sampling rate, and maintains channel specificity through learned cross-channel mappings. Two representative deep-learning models are evaluated to validate the proposed framework. First, a supervised U-Net

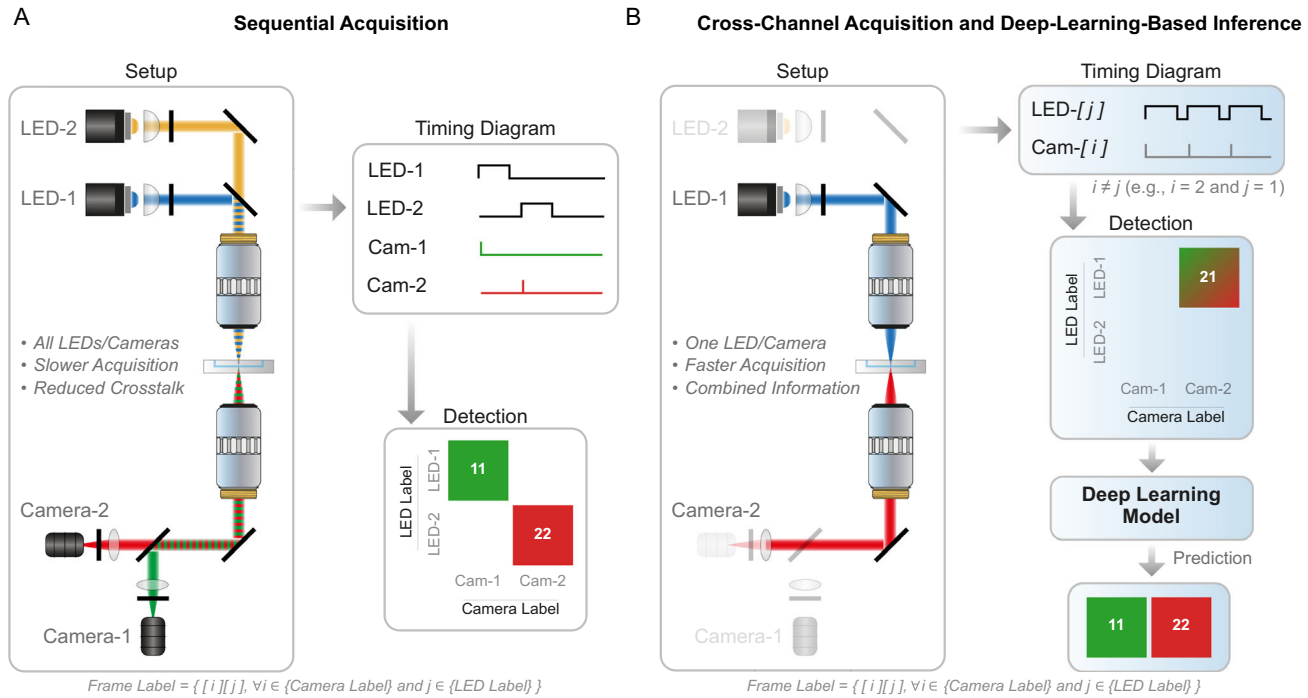

FIGURE 1 | Conceptual overview and comparison of sequential multicolor fluorescence acquisition and the proposed framework. (A) Sequential acquisition: each fluorescence channel is acquired independently using matching excitation–detection pairs, requiring multiple LED-camera cycles to obtain multicolor images with reduced crosstalk. (B) A predefined cross-channel excitation–detection combination captures mixed information used by a deep learning model (U-Net or pix2pix) to predict the fluorescence channels in real-time.

[36] is used for per-pixel regression: an encoder–decoder with skip connections that predicts each target fluorescence channel from the cross-channel input using an L_1 loss. Second, we assess pix2pix [37], configured with the same U-Net generator and a PatchGAN discriminator. U-Net and PatchGAN architectures are detailed in Figure S1 (Supporting Information). Training uses paired data consisting of the cross-channel frame (input) and ground-truth target channels obtained from sequential acquisitions. Pre- and postprocessing, as well as key hyperparameters, are kept identical across models to ensure a fair comparison (details in the Experimental Section). At deployment, U-Net and pix2pix generator models are exported and compiled with Torch-TensorRT into a unified pipeline, with identical input/output and batching paths.

2.2 | Real-Time Imaging of Micro-Agent–Spheroid Interactions

The dynamic interaction between magnetic micro-agents and biological targets presents a complex imaging scenario requiring high temporal and accurate spectral channel separation. Sequential fluorescence acquisition, while minimizing spectral crosstalk, is inherently limited by temporal resolution and risks spatial drift or misalignment across channels due to object motion during the acquisition process. Coumarin 6-labeled electrospun magnetic fibers serve as micro-agents, magnetically actuated to interact with CellTracker Red CMTPX-labeled HeLa cell spheroids (Figure 2A). Under external magnetic fields, the micro-agents exhibit controlled translational and angular motion, enabling physical contact, manipulation, and displacement of spheroids within the microfluidic environment. While these interactions can lead to partial spatial overlap at the interface between agents and spheroids, the system does not involve true fluorescence colocalization within the same diffraction-limited volume, as the fluorophores label distinct physical entities.

Fluorescence channels for micro-agents (M) and spheroids (C) are defined by specific LED-camera configurations designed to selectively excite and capture channel-specific information from each agent (Figure 2B). Overlap arises primarily from emission tail spillover between fluorophores due to partial spectral overlap in their excitation and emission spectra. Although excitation and emission filters (details can be found in the Experimental Section) are employed to suppress this crosstalk, residual spectral mixing and optical scattering persist, particularly under off-diagonal LED-camera combinations. These conditions result in cross-channel frames containing mixed yet structurally informative fluorescence signals.

In the context of the proposed framework, these cross-channel frames serve as inputs for reconstructing the target fluorescence channels. To acquire sequential and cross-channel frames, all cameras are hardware-triggered on every LED pulse (Figure 2B), producing four frames per LED sequence. Frames with matched excitation/detection (LED-M→Cam-M and LED-C→Cam-C) form the sequential frames, denoted MM and CC. The unmatched pairs (LED-M→Cam-C and LED-C→Cam-M) are the cross-channel frames, denoted CM and MC. Figure 2C shows representative sequential and cross-channel frames capturing micro-agent–spheroid interactions. The cross-channel frames embed mixed spectral and spatial features, providing the basis for computational inference of target fluorescence channels.

2.3 | Channel Reconstruction Assessment

To evaluate the feasibility of reconstructing the sequential fluorescence channels (i.e., ground truth) from cross-channel inputs, we

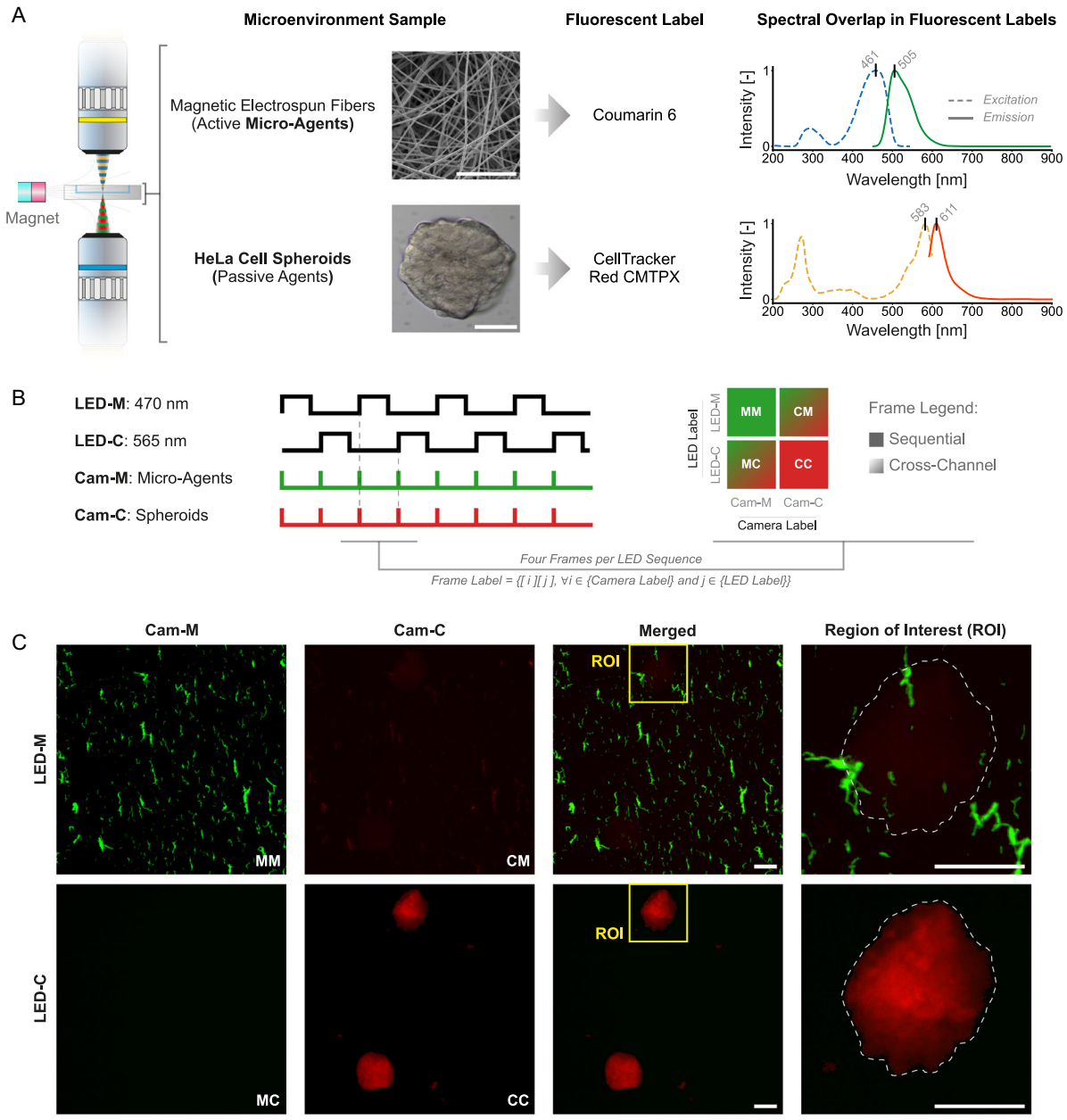

FIGURE 2 | Experimental system and acquisition design for multicolor fluorescence imaging of microrobotic and cellular microenvironments. (A) Schematic of the microfluidic sample environment containing magnetic micro-agents (made of electrospun fibers labeled with Coumarin 6) and HeLa cell spheroids (passive agents, labeled with CellTracker Red CMTPX), alongside fluorescence excitation and emission spectra of each label. (B) Timing diagram of LED excitation and camera detection for each fluorescence label, illustrating sequential and cross-channel acquisition frames. Frame labels denote excitation–detection pairings for micro-agents (M) and HeLa cell spheroids (C). (C) Example fluorescence images acquired under sequential and cross-channel excitation–detection combinations, shown by camera and LED pairing. Scale bars: 100 μm.

quantified their structural overlap using G-NCC. This metric captures the correlation between the spatial gradient fields, making it suitable for microscopy, where morphology, edges, and fine details are of paramount relevance. Unlike intensity-based metrics, G-NCC emphasizes structural alignment and is robust to global intensity variations caused by exposure time or fluorophore concentration. Figure 3A outlines the analysis workflow: the input cross-channel frames (CM and MC) are masked at varying percentiles, and their gradients are compared against the ground truth (MM and CC frames). The resulting G-NCC values are computed across percentile thresholds, and the maximum value is used to quantify the structural overlap between the input and each target channel. Figure 3B shows that CM has substantially higher structural overlap with MM than with CC, indicating strong morphological cues from the micro-agent channel. For both ground truths (MM and CC), CM also yields higher structural overlap compared to MC. This asymmetry highlights a directional bias in cross-channel information, driven by the specific excitation–emission characteristics of the fluorophores and the configuration of spectral filters. Although the acquisition geometry appears symmetric, the resulting structural content is not, due to factors such as unequal emission tails, filter leakage, and differential brightness. This

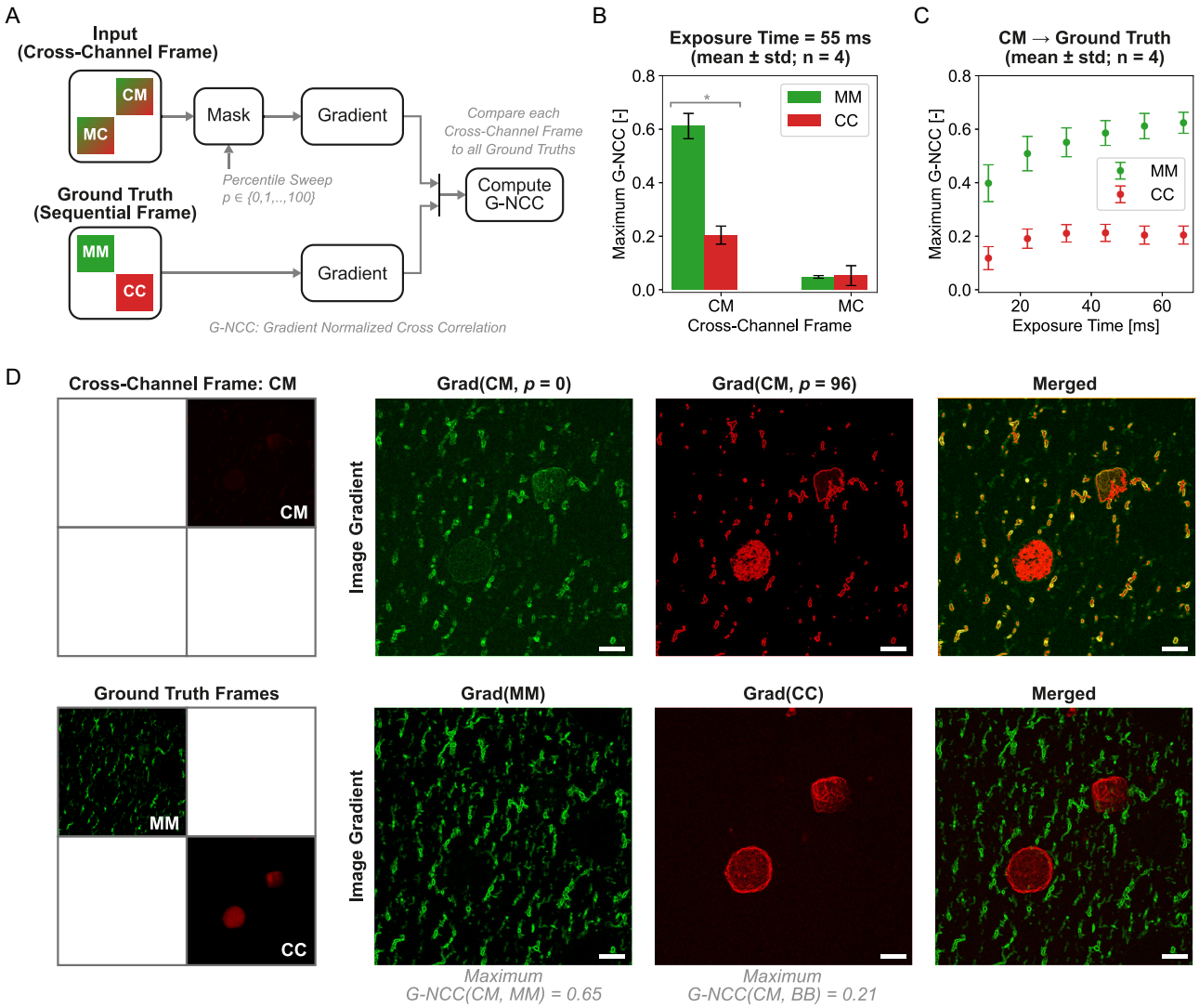

FIGURE 3 | Gradient-based cross-channel analysis for evaluating channel reconstruction. (A) Workflow diagram: gradients are computed from cross-channel input frames and sequential ground-truth frames, followed by gradient-normalized cross-correlation (G-NCC) between input and ground-truth gradients. (B) Bar plot of maximum G-NCC values across all cross-channel frames (mean ± SD, $n = 4$) at 55 ms exposure. (C) Dependence of maximum G-NCC values on exposure time (mean ± SD, $n = 4$) for CM input frame against each ground-truth channel. (D) Example analysis for cross-channel frame CM: gradient images of CM input (top row) and ground-truth channels MM and CC (bottom row), with corresponding maximum G-NCC values. Gradient visualizations are contrast-enhanced by applying a luminance bias toward the upper dynamic range. Scale bars: 100 μm.

asymmetry suggests that cross-channel selection should be guided by acquisition balance and the actual information content captured under each configuration.

Figure 3C presents a systematic evaluation of G-NCC as a function of exposure time, using CM as the input for reconstructing the target fluorescence channels. For both CM–MM and CM–CC pairs, structural overlap drops noticeably below 22 ms, indicating that noise becomes prominent and degrades the ability to recover meaningful spatial features at low exposure. This supports a 22 ms practical lower bound for reliable reconstruction. Beyond this point, CM–CC reaches a plateau around 33 ms, while CM–MM continues to increase more gradually, with no apparent saturation within the tested range. Notably, G-NCC remains relatively stable between 22 and 66 ms across both channel pairs, supporting the decision to train the models using only 22 ms exposure data. This selection balances temporal resolution and structural fidelity, enabling real-time imaging performance without compromising reconstruction accuracy. The G-NCC framework enables a pretraining diagnostic to evaluate channel informativeness and guide model design. In different imaging systems employing automated filter wheels or tunable optics, similar cross-channel relationships can be characterized in advance. It is worth noting that the model learns the mapping defined by the supervision it receives. Consequently, sufficient separation in the target channels is essential at the training stage, as strongly mixed targets do not provide the supervisory structure required to learn or enforce channel-separated outputs.

2.4 | Evaluation

Building on the channel reconstruction assessment, performance is evaluated on an independent test set (750 frames), disjoint

from training and early-stop validation datasets (details can be found in the Experimental Section). We compare U-Net against pix2pix (U-Net generator and PatchGAN discriminator) predictions, both trained using the same L_1 loss weight. Thereby, any performance gap isolates the effect of the adversarial term used in pix2pix. In both cases, CM serves as the input to predict MM (micro-agents) and CC (spheroids) channels. Figure 1A shows one representative frame, including the input, ground truth, and model predictions for MM and CC channels. Quantitative analysis is performed using SSIM, PSNR, and GCS between

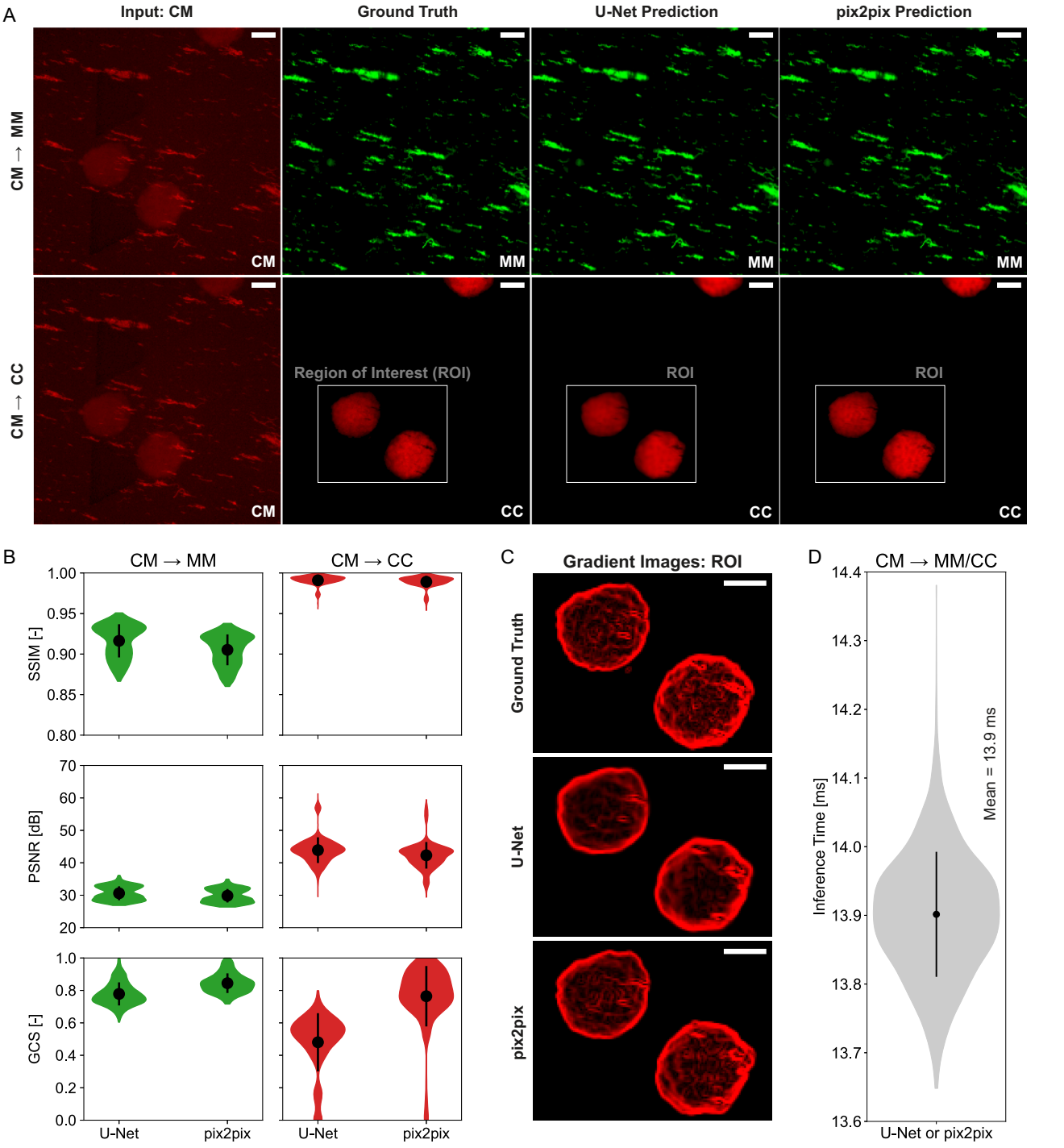

FIGURE 4 | Framework performance using U-Net and pix2pix architectures. (A) Representative frame reconstructions for each target channel (MM: micro-agents; CC: spheroids) from a single CM input (cross-channel frame). Columns show input, ground truth, and predictions. Gamma correction ($\gamma = 0.5$) is applied to the input CM to improve the visualization. (B) Quantitative evaluation of structural similarity index metric (SSIM), peak signal-to-noise ratio (PSNR), and GLCM-based contrast similarity (GCS) across 750 frames. Violin plots show distribution, mean (·), and standard deviation (vertical black line). (C) Gradient images of regions of interest (ROIs) presented in panel (A), comparing ground truth and model predictions. (D) Inference time distribution of U-Net and pix2pix over 750 frames, representing simultaneous reconstruction of both MM and CC channels. Because pix2pix uses a U-Net generator, the inference time is effectively identical for both models (mean ≈13.9 ms). Scale bars: 100 μm.

ground truth and predictions (Figure 4B). Our results show consistent reconstruction capabilities as reflected by SSIM (CM→CC >0.95; CM→MM >0.87) and PSNR (>25 dB for both models) values across models. In comparison to U-Net, pix2pix yields higher GCS values on CC predictions without a significant reduction of SSIM and PSNR. This suggests that the adversarial term primarily improves perceptual/texture contrast while preserving intensity-based fidelity. For MM, differences are minimal as micro-agent morphology presents lower intrinsic texture (i.e., more edge-dominated) relative to CC. Figure 4C illustrates the texture differences using gradient images of a region of interest in channel CC. Consistent with these observations, Figure S3 shows that even the introduction of a small adversarial weight (λ_{adv}) systematically increases GCS while leaving SSIM and PSNR virtually unchanged. The inference time distribution is identical for the two models (i.e., mean ≈13.9 ms) because pix2pix uses a U-Net generator (Figure 4D). This low-latency performance supports real-time deployment at standard video frame rates, enabling live reconstruction of magnetically actuated micro-agents interacting with biological targets in microfluidic environments.

The framework's robustness under limited excitation is evaluated across four irradiance levels. We tested U-Net and pix2pix by progressively dimming LED-M and LED-C in tandem (Figures S5A and B, Supporting Information), from the training settings of 62.5 and 13.5 mW cm^{-2} down to 18.9 and 3.8 mW cm^{-2}, respectively. With a fixed L_1 weight ($\lambda_1 = 100$) and $\lambda_{adv} = 0.05$ for pix2pix, both models produced stable MM and CC reconstructions across the full range, including irradiances outside training (Figure S5C, Supporting Information). We further assessed real-time performance on an independent dataset acquired at 30.7 and 6.5 mW cm^{-2} for LED-M and LED-C (Video S1), where the framework maintained high reconstruction accuracy without retraining, indicating generalization to dimmer conditions.

Following the variable-irradiance test, we evaluate the framework's capability to track a natural fluorescence decay due to dye efflux. This setting isolates temporal intensity changes from illumination effects while tracking the fluorescence change of a HeLa cell spheroid labeled with CellTracker Red CMTPX. Figure 5A shows a time-lapse sequence of ground truth and U-Net/pix2pix merged predictions. The predictions for micro-agents (green) and spheroid (red) channels qualitatively preserve structural integrity. Moreover, Figure 5B shows that normalized fluorescence change ($\Delta F/F_0$) in the spheroid channel closely follows the monotonic ground-truth trend across the tested models. Root mean square errors are reported for U-Net: 0.58% and pix2pix: 1.55%. Notably, the models track the trend and slope of the fluorescence decay $F(t)$ over time while minimizing the excitation dose on the spheroid. In practice, the spheroid dye is not excited directly, which minimizes photodamage and bleaching and isolates the efflux measurement.

2.5 | Fluorescence-Guided Real-Time Tracking

To assess the framework's suitability for real-time applications, we evaluate its spatial accuracy on dynamic fluorescence sequences using U-Net and pix2pix (Video S2, Supporting Information). Figure 6A shows frame-by-frame comparisons between ground truth and pix2pix predictions for the MM and CC channels across representative time stamps. Colored markers indicate spheroid centroids for direct prediction–ground-truth comparison. Positional errors along frame axes (u and v) are shown in Figure 6B, demonstrating subcellular-scale localization accuracy. Figure 6C shows angular error in micro-agent orientation, confirming directional consistency across the sequence. Root-mean-square errors for spheroid position and micro-agent orientation using U-Net and pix2pix are presented in Table 1. Independent of the models used, the input channel (CM) is acquired in 22 ms, and the inference of MM and CC takes ≈13.9 ms. This way, the effective dual color frame rate increased by 22% with respect to fully sequential dual-color acquisition at 44 ms. It is worth noting that for tracking applications, precise position and orientation matter more than texture realism. U-Net is less computationally demanding and less prone to nonconvergence than pix2pix while still providing accurate centroid and pose estimates. When visual fidelity and realism are critical, pix2pix is preferable.

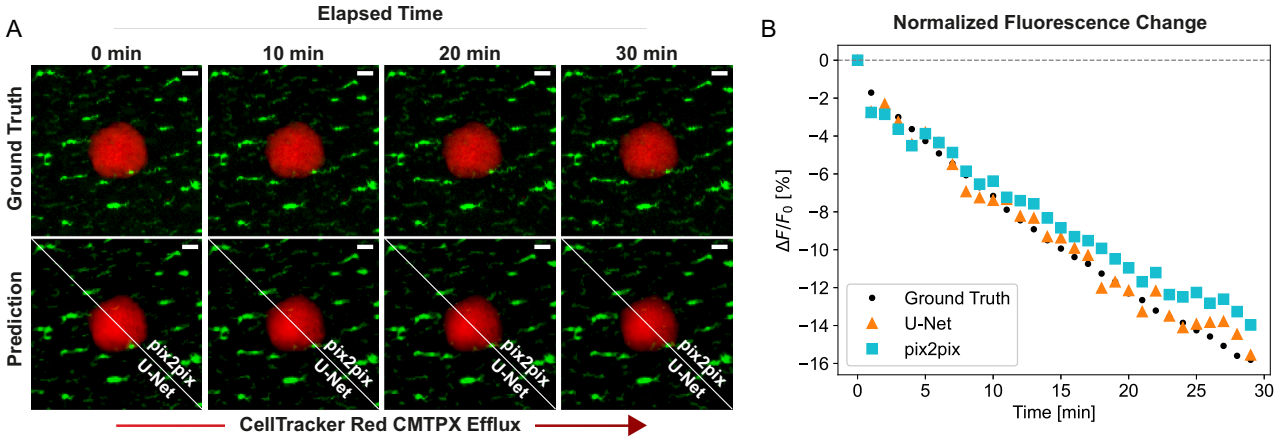

FIGURE 5 | Temporal fluorescence decay and model tracking of dye efflux. (A) Time-lapse fields of view at 0, 10, 20, and 30 min showing a HeLa cell spheroid labeled with CellTracker Red CMTPX within an aqueous solution containing micro-agents (green). Top: ground truth; bottom: merged predictions using U-Net and pix2pix. The gradual loss of red intensity reflects CMTPX efflux. (B) Normalized fluorescence change ($\Delta F/F_0$) over time, computed on a centered 128 × 128 pixels region of interest with F_0 taken at $t = 0$ min. Each 1-min time point represents the average of three consecutive acquisitions. Predictions closely follow the ground truth (black circles), capturing the monotonic decay dynamics across the full 30-min sequence. Scale bars: 50 μm.

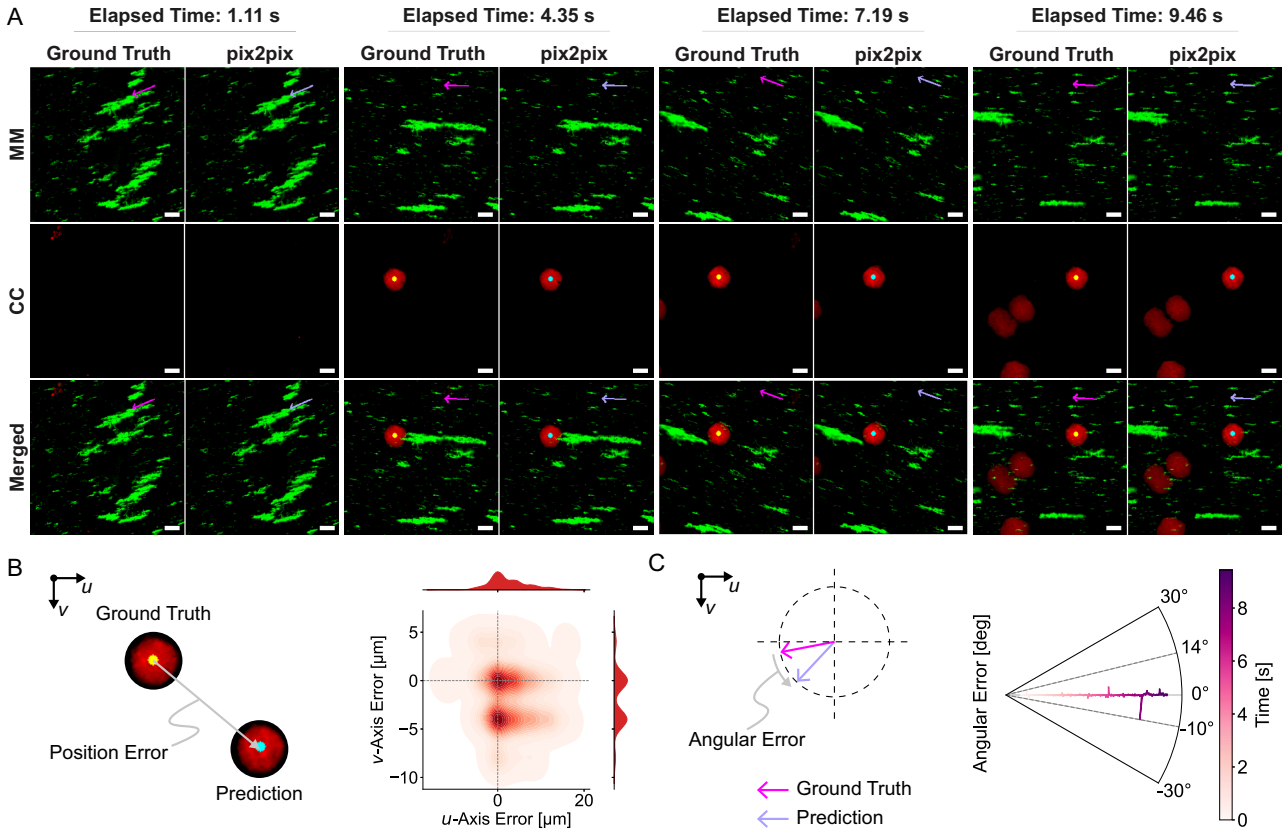

FIGURE 6 | Assessment of spatial accuracy on unseen frames. (A) Frame-by-frame comparison of ground truth and pix2pix predicted fluorescence channels for micro-agents (MM, green) and HeLa cells (CC, red) at selected time points. The merged view illustrates channel overlays, showing preserved structure and spatial consistency in predicted frames. Colored markers indicate the detected positions of spheroids in the ground truth (yellow) and predictions (cyan), while arrows show the orientation of magnetic micro-agents in the ground truth (magenta) and predictions (pink), enabling error quantification in a tracking-like test scenario. (B) Left: schematic illustration of spheroid position error as the vector displacement between predicted and ground-truth centers. Right: joint histogram of 2D localization errors along the frame axes (u and v) across test frames ($n = 350$), with marginal distributions highlighting error spread. (C) Left: schematic of angular error between predicted and ground-truth micro-agents orientation. Right: polar plot showing angular error (in degrees) as a function of elapsed time. Dashed lines indicate the maximum and minimum deviations observed during the sequence. Scale bars: 100 μm. (Video S2, Supporting Information).

TABLE 1 | Aggregate root-mean-square (RMS) errors for spheroid radial localization and micro-agents orientation across models.

Error metric	U-Net	pix2pix
Spheroid position—radial RMS, μm	5.79	6.15
Micro-agents orientation—RMS, deg	1.53	1.05

2.6 | Three-Color Fluorescence Microscopy

Building on the real-time two-channel results, we extend the framework to three-color fluorescence microscopy. We add a third channel (LED-B/Camera-B) and introduce indocyanine green (ICG), which labels the microfluidic channel and the surrounding medium, providing context beyond the primary agents (Figure 7A). The BB channel frames also carry indirect structural cues about micro-agents and spheroids, likely due to scattering by embedded iron-oxide nanoparticles and the gradual uptake/swelling of ICG in spheroids over time.

We evaluate two input configurations for three-channel operation: (1) CM→(MM, CC), using an auxiliary BB frame acquired sequentially for contextual visualization only; and (2) BB→(MM, CC) to test how informative BB is as an alternative input (Figure 7B,C). For CM-based reconstruction, CM and BB are acquired sequentially while MM/CC inference overlaps BB exposure, yielding 44 ms per three-color frame (50% higher frame rate relative to fully sequential acquisition at 66 ms). In contrast, BB-based reconstruction requires only 36 ms per three-color frame (83% higher frame rate relative to fully sequential). Video S4 shows CM- and BB-based reconstructions using U-Net and pix2pix. Predictions at two timestamps with pix2pix are shown in Figure 7D, demonstrating three-channel fluorescence imaging. Overall, pix2pix reconstructions are consistent, as reflected by the SSIM values (Figure 7E). Mean SSIM and PSNR values for U-Net and pix2pix are reported in Table 2. CM-based reconstruction achieves higher similarity in MM, likely because BB frames do not carry the same fine-grained information for MM as they do for CC. From an image-formation standpoint, BB→MM underperforms because it relies on indirect scattering contrast (e.g., from iron-oxide–doped fibers and gradual ICG uptake) rather than direct spectral crosstalk with MM. These cues are weaker, geometry-dependent, and biased toward low spatial frequencies, rendering the recovery of MM's high-frequency features ill-posed.

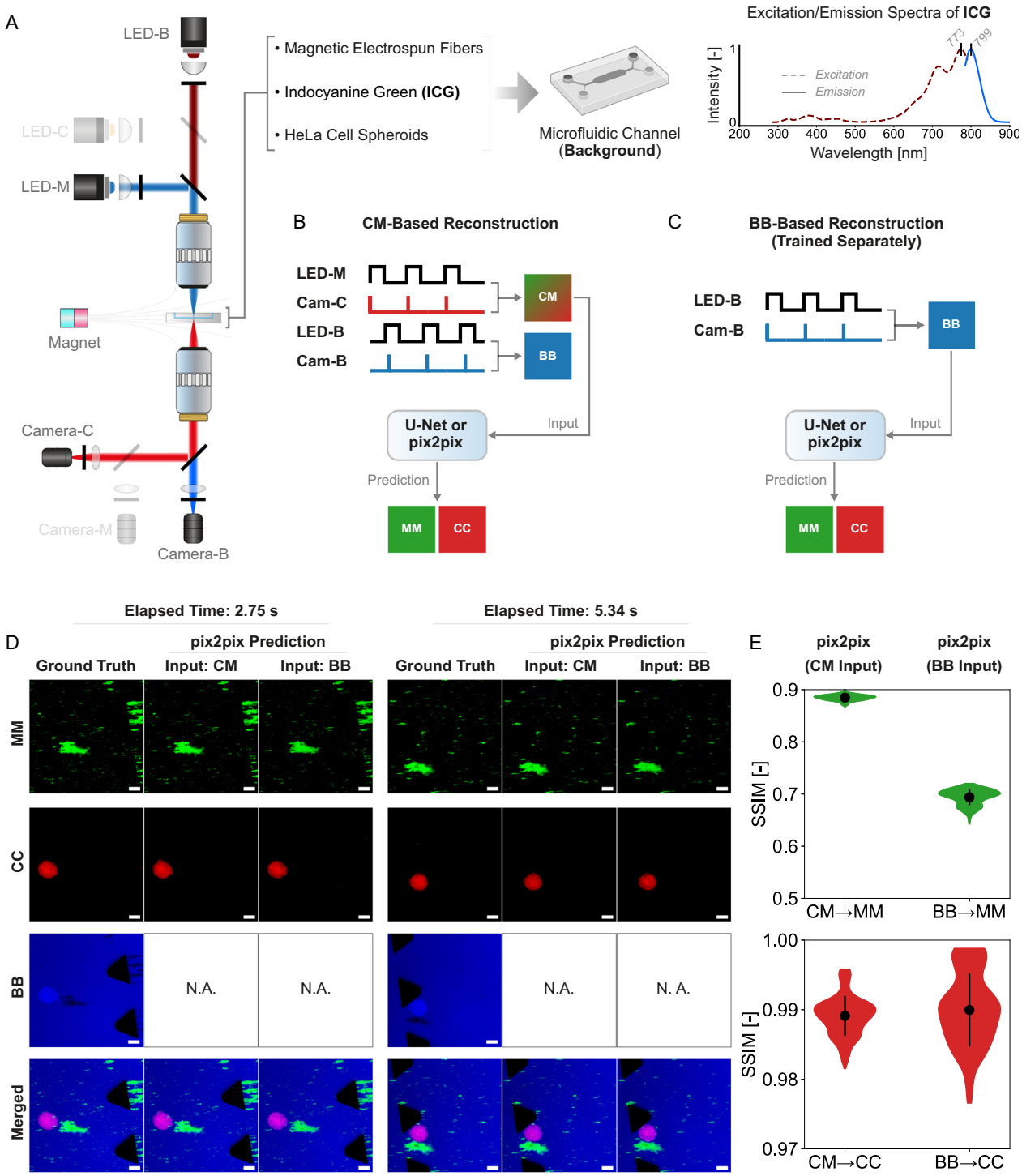

FIGURE 7 | Real-time inference within a three-channel fluorescence microscopy workflow. (A) Fluorescence imaging setup used to acquire the framework input frame (CM: LED-M/Cam-C). An additional acquisition (BB: LED-B/Cam-B) captures indocyanine green (ICG) fluorescence from the surrounding medium, providing contextual information on the microfluidic environment. Right: excitation and emission spectra of ICG. (B) CM-based reconstruction of MM and CC channel frames. The BB frames are acquired sequentially for visualization but are not used during inference. Inference is performed in parallel with BB acquisition. (C) BB-based reconstruction of MM and CC. This configuration evaluates the informativeness of BB as an alternative reconstruction source. (D) Representative predictions from both models compared to ground-truth MM and CC frames. (E) SSIM comparison for MM and CC reconstructions using CM or BB as model input ($n = 350$ test patches). Scale bars: 100 μm. (Video S3, Supporting Information).

In both proposed input configurations, the framework runs in real-time with comparable inference latency (13.9 ms for two output channels) as the deployment pipeline is identical across models and inputs. Although BB is a feasible alternative input, ICG's lower photostability and higher susceptibility to photobleaching than the other labels can reduce quality during extended imaging [9, 38]. Thus, under these settings, BB is most useful as a contextual channel and for exploratory reconstructions, while

TABLE 2 | Target channels (MM and CC) reconstruction performance across inputs (CM and BB) and models in a three-channel microscopy workflow.

	Input: CM				Input: BB			
	MM		CC		MM		CC	
Metric (mean)	**U-Net**	**pix2pix**	**U-Net**	**pix2pix**	**U-Net**	**pix2pix**	**U-Net**	**pix2pix**
SSIM (–)	0.89	0.88	0.98	0.98	0.72	0.69	0.98	0.98
PSNR, dB	29.69	29.26	37.04	37.83	23.32	22.83	37.49	39.21

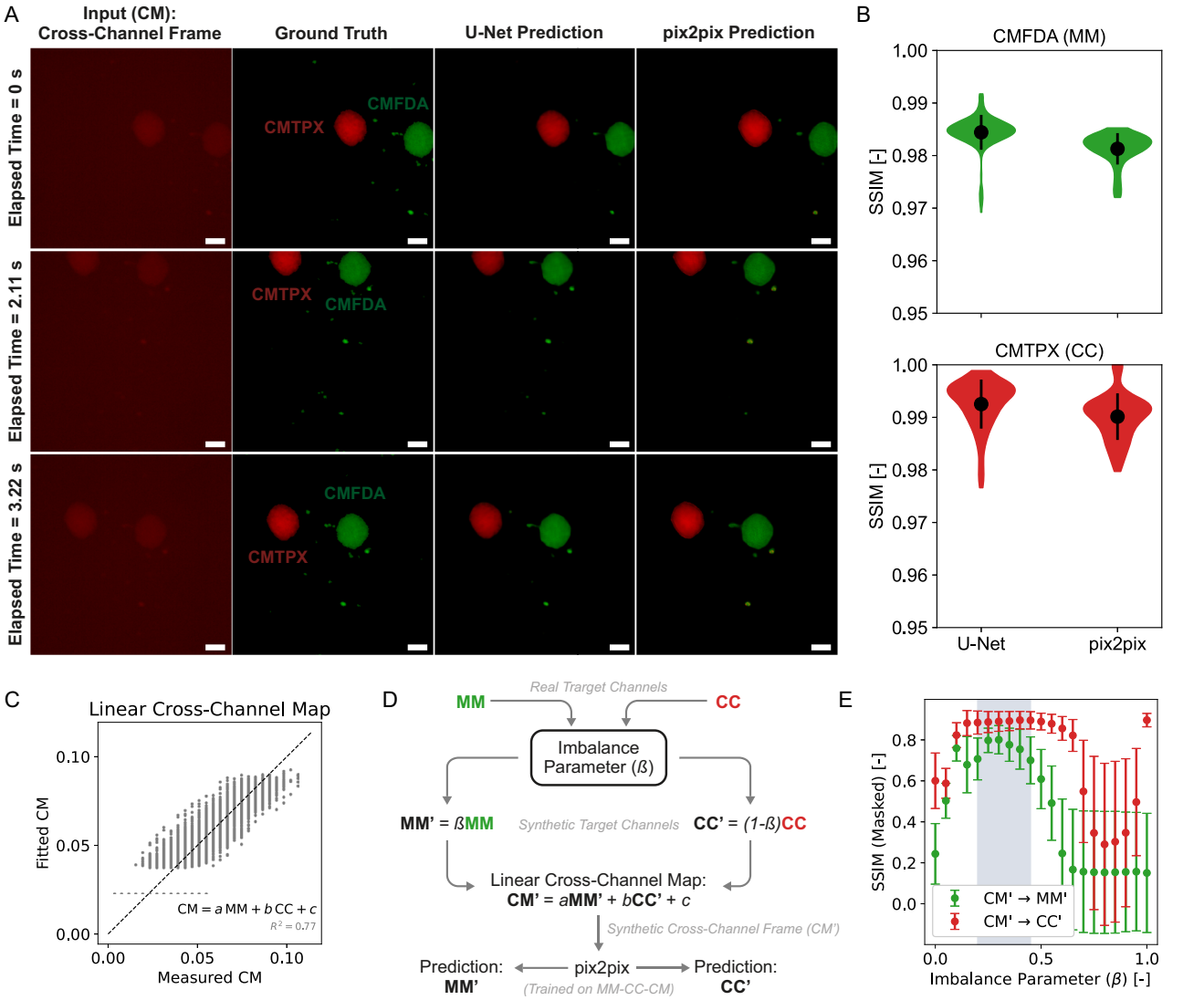

FIGURE 8 | Inference under similar morphology with intensity-based discrimination. (A) Spheroid samples are stained with two different dyes (green CMFDA and red CMTPX). Left to right: cross-channel input frame, ground truth, and model predictions at three representative frames. Despite morphology-matched spheroids, both models successfully recover distinct channels based on the intensity differences shown in input frames. (B) SSIM distributions for CMFDA (MM) and CMTPX (CC) reconstructions using U-Net and pix2pix ($n = 116$). Dots denote means and bars standard deviations. Scale bars: 100 µm. (Video S4, Supporting Information). (C) Linear cross-channel map fitted from real tri-channel measurements (MM, CC, CM), relating the measured cross-channel intensity (CM) to a linear combination of the corresponding target channels. (D) Schematic of the synthetic imbalance framework. Real target channels (MM, CC) are scaled using an imbalance parameter (β) to generate synthetic target channels (MM′ and CC′), which are combined via the empirical cross-channel map to form a synthetic cross-channel frame (CM′). A pix2pix model trained on real MM, CC, and CM data is then used to predict MM′ and CC′ from CM′. (E) Masked SSIM between pix2pix predictions and synthetic targets as a function of the imbalance parameter (β). Dots and error bars indicate mean values and standard deviations across samples, respectively. The gray shaded region highlights the empirically identified well-conditioned operating regime.

CM-based input is preferred when maximizing MM fidelity and robustness is the priority.

Following the three-channel workflow, we next consider the portability of the framework to new dye sets and microscope architectures. Because the proposed approach relies on informative cross-channel signals, successful deployment requires coarse awareness of cross-channel coupling and accurate channel registration to ensure that the acquired inputs encode recoverable information. As the number of channels increases, acquisition of paired training data becomes the primary bottleneck, particularly for frame-synchronous schemes using a single detector or filter turret, where switching speed limits throughput.

Within these constraints, several practical strategies can be employed, alone or in combination. (1) Clean, per-channel targets can be acquired on stationary samples to establish ground truth under representative optical conditions. (2) During continuous cross-channel acquisition, occasional fully separated target frames can be interleaved and registered to nearby frames to provide sparse supervision. (3) Clean targets can be forward-mixed using empirically measured cross-channel responses to synthesize cross-channel inputs matched to deployment conditions.

2.7 | Performance with Morphology-Matched Spheroids

When morphology differs across fluorescence channels, the proposed framework can utilize shape and texture cues alongside cross-channel intensity information to separate target channels. In such settings, reconstruction is typically robust to moderate noise and clutter. Here, we instead consider a morphology-matched regime in which shape-based discrimination is minimized, and reconstruction relies primarily on the intensity information contained in the cross-channel input.

To this end, we evaluate the framework on HeLa cell spheroids stained with two spectrally distinct dyes (green CMFDA and red CMTPX). In this configuration, the two channels share nearly identical morphology, and the discriminative information arises mainly from relative intensity differences rather than structural features. U-Net and pix2pix reconstruct the target channels with minimal false merging or cross-contamination (Figure 8A and Video S4). Quantitatively, full-frame SSIM values are ≥ 0.97 for both channels (MM and CC) (Figure 8B), indicating reliable intensity-based separation under morphology-matched conditions.

To probe the framework's behavior under controlled intensity imbalance, we construct a synthetic evaluation scenario based on an empirically fitted cross-channel map (Figure 8C). Intensity imbalance is introduced by scaling the target channels using a parameter (β). The resulting scaled targets (MM' and CC') are then combined through the cross-channel map to generate a synthetic cross-channel input frame ($\mathrm{CM}' = \mathrm{aMM}' + \mathrm{bCC}' + \mathrm{c}$). This synthetic input is processed by a pix2pix model trained on MM, CC, and CM data (Figure 8D), allowing assessment of reconstruction behavior outside the training intensity distribution.

Reconstruction accuracy quantified by masked SSIM (details in the Experimental Section) shows a nonmonotonic dependence on β, with MM' and CC' peaking within a well-conditioned regime ($\beta \approx 0.2$–0.45) (Figure 8E). Across all β, CC' is reconstructed more robustly than MM', reflecting the CC-dominant mixture statistics present in the measured training data. Representative reconstructions and ROI-averaged scatter plots illustrating this transition are shown in Figures S6A and B. As β approaches ~ 0.52, the relative contributions of MM' and CC' to CM' become comparable ($a\beta \approx b(1-\beta) \Leftrightarrow \beta = \frac{a+b}{a} \approx 0.52$), leading to a gradual loss of reconstruction fidelity for both channels. This transition is accompanied by an increasing spatial redistribution of the predicted signal, as quantified in Figure S6C.

Notably, this transition is continuous, reflecting a gradual loss of identifiability as relative signal contributions depart from the training distribution. This behavior is consistent with an intensity-based inversion constrained by empirically learned information relationships rather than categorical channel assignment. In practice, fluorescence channel reconstruction is further limited when the target channel contributions to the cross-channel frames approach the background offset captured by the constant term (c) in the empirical map. As either $a\beta$MM or $b(1-\beta)$CC becomes comparable to this offset, the corresponding signal is no longer reliably encoded in CM', defining a practical operating regime governed by relative signal balance and absolute information level.

From a hardware perspective, incomplete spectral separation is a recognized and often unavoidable limitation in several established fluorescence imaging modalities. In live-cell multicolor imaging, spectral overlap between fluorescent proteins and organic dyes is common, particularly under high-speed acquisition, where sequential excitation or filter switching is impractical [39]. Similar constraints arise in Förster resonance energy transfer experiments, where donor bleed-through and acceptor cross-excitation must be explicitly modeled due to unavoidable crosstalk [40]. Reduced-excitation and low-phototoxicity imaging accentuate crosstalk effects by operating close to the noise floor [2]. In such regimes, crosstalk is typically tolerated as a trade-off for temporal resolution or signal efficiency, motivating computational strategies that exploit informative cross-channel coupling.

3 | Conclusions

This study introduces a real-time, deep learning-based framework for multicolor fluorescence microscopy that uses cross-channel acquisition to reduce reliance on fully sequential channel capture. The framework is evaluated with two real-time inference architectures (U-Net and pix2pix), enabling high-fidelity imaging of dynamic microsystems while preserving spectral specificity and structural detail. Our approach increases the multicolor frame rate by 22% in two-color mode and up to 83% in three-color mode, relative to fully sequential acquisition. The framework covers two- and three-color micro-agent–spheroid imaging, reduced-excitation imaging, normalized fluorescence change due to dye efflux, real-time tracking, and inference under morphology-matched scenarios. These results highlight the ability of cross-channel learning to recover missing target fluorescence channels when informative correlations are present. Future work will extend the framework to higher channel counts and fully parallel multichannel inference, while accounting for conditioning limits imposed by cross-channel coupling and signal imbalance. Incorporating label-free contrast mechanisms or multimodal imaging modalities could further broaden applicability, from high-content screening to in vivo imaging, where rapid, multiplexed acquisition is critical.

4 | Experimental Section

4.1 | Cell Culture and Spheroid Preparation

HeLa cells are cultured in Dulbecco's modified Eagle's medium (11-965-092, Fisher Scientific Ltd. Canada), supplemented with 10% fetal bovine serum (F7524-500ML, Sigma–Aldrich, USA), and 1% penicillin–streptomycin (15-140-122, Thermo Fisher Scientific Inc., USA). During the culture period, the cells are maintained at 37°C in a humidified atmosphere containing 5% carbon dioxide. In the 2D cell culture, the medium is changed every 48 h. When cells reach 80% confluency, the cells are trypsinized, counted, and resuspended in cell culture medium at a concentration of 2×10^6 cells/mL. For spheroid preparation, 0.2 mL of cell suspension is placed on top of an agarose mold containing 1500 microwells with a depth and a diameter of 200 μm. This results in spheroids which, on average, include approximately 270 cells. Solutions of 10 μM CellTracker Red CMTPX and Green CMFDA are prepared in a serum-free medium and subsequently incubated for 30 min with the spheroids, which have been compacted for 3 days under cultivation conditions. Afterward, the working solutions are removed, and the culture medium is added again. HeLa cell spheroids stained with CMTPX and CMFDA are fixed for 15 min with 4% formaldehyde (F8775-25 ML, Sigma–Aldrich, USA) at room temperature for long-term storage. This is followed by washing twice with Dulbecco's phosphate-buffered saline.

4.2 | Preparation of Magnetic Micro-Agents

Magnetic Electrospun fibers are fabricated using the electrospinning technique. The two polymer solutions for electrospinning consist of polystyrene pellets (430102-1KG, Sigma–Aldrich, USA) as a carrier polymer, Coumarin 6 (442631-1G, Sigma–Aldrich, USA) and Nile Red (72485-1G, Sigma–Aldrich, USA) as hydrophobic fluorophores, iron oxide (Fe_3O_4) nanoparticles (637106-25G, Sigma–Aldrich, USA) as a magnetic material, and anhydrous *N*,*N*-Dimethylformamide (DMF) as a solvent (227056-1L, Sigma–Aldrich, USA). The polymer solutions are prepared using a 45.0% (29.8% polystyrene, 15% Fe_3O_4, and 0.2% Coumarin 6) weight-to-volume ratio in DMF. Each polymer solution is homogenized by stirring for 24 h using a roller mixer (LLG-uniRoller 6, LLG-Labware, Germany). For electrospinning, the polymer solution is placed in a plastic syringe and connected through Teflon tubing to an 18-gauge blunt-tip needle (TS18SS-15, Adhesive Dispensing Ltd., UK). Each solution (0.04 mL) is electrospun at once using an accelerating voltage of 14 kV, a feed rate of 2.5 mL/h, and a needle tip-to-collector distance of 16 cm. Randomly oriented fiber meshes are collected on grounded aluminum foil at 22°C and 30% humidity and dried at room temperature for 12 h to remove the solvent residue. Deposited fiber meshes are cut into 2–3 mm pieces using a scalpel and then immersed in a glass vial containing 1% (volume/volume) Tween 80 (P4780-100ML, Sigma–Aldrich, USA) in Milli-Q water. Fluorescent and magnetic micro-agents are obtained by grinding the fibers using a sonicator (model 2510, Branson, USA) for 2 h. Finally, the resulting micro-agents are magnetically collected at the bottom of the vial using a permanent magnet, allowing for the removal of the fluorescent supernatant, which is then replaced with 4 mL of fresh Milli-Q water containing 1% Tween 80.

4.3 | Microenvironment Sample

For two-channel experiments, 7 μL each of micro-agent solution and HeLa spheroid suspension are used. An additional 7 μL of indocyanine green (ICG) solution (250 mg/mL in phosphate-buffered saline) is included for three-channel experiments, maintaining a 1:1:1 volumetric ratio. The microfluidic channels are fabricated using a standard soft lithography process [41]. Negative molds of microfluidic channels (height ≈187 μm) are prepared on a silicon wafer in the cleanroom using SU-8 photoresist. A mixture of 10:1 Sylgard 184 polydimethylsiloxane (PDMS) and the curing agent is poured onto the wafer and cured in the oven at 70°C overnight. The cured PDMS layer is gently removed from the wafer, and inlet and outlet ports are punched. Finally, the microfluidic channels are obtained by bonding the PDMS layer to a microscope slide using plasma oxidation.

4.4 | Excitation and Emission Spectra

Excitation and emission spectra of fluorescent labels are measured using a spectrofluorometer (FP-8300, Jasco, Japan) in the range of 200 nm and 900 nm with a 1 nm data interval. For spectrum analysis, 2.5 μM indocyanine green (I2633-25MG, Sigma–Aldrich, USA) solution is prepared by dissolving 1.5 μg indocyanine green in 700 μL fetal bovine serum (16000044, Thermo Fisher Scientific, USA). The solution is placed in a water bath at 37°C for 2 h to bind indocyanine green to proteins in the culture medium [42]. A 10 μM working CellTracker Red CMTPX (Thermo Fisher Scientific, USA) solution is prepared in a serum-free medium by mixing 50 μg CMTPX dye in the vial with 7.29 μL dimethyl sulfoxide (D2650-100ML, Sigma–Aldrich, USA). 2 μL of CMTPX working solution is diluted in 700 μL of Milli-Q water. A solution of 4 μg of Coumarin 6 (442631-1G, Sigma–Aldrich, USA) is prepared in 700 μL dimethylformamide (227056-1L, Sigma–Aldrich, USA). All fluorescent solutions are placed in different quartz cuvettes (CV10Q700F, Thorlabs, USA) for analysis.

4.5 | Multicolor Fluorescence Microscope

All experiments presented in this study are carried out using a custom-built multicolor fluorescence microscope [38]. The microscope consists of excitation and emission units. The excitation unit generates discrete light beams with the center wavelengths of 470, 565, and 780 nm for sharp excitation of the fluorophores using three individual narrow-spectrum light-emitting diodes (LEDs) (M470L3, M565L3, M780LP1, Thorlabs, USA). LEDs are collimated using 20 mm focal length aspheric condenser lenses (ACL2520U-A, ACL2520U-B, Thorlabs, USA) and coupled with filters (ET470/40x, ET572/35x, ET775/50x, Chroma, USA) to select the excitation wavelengths. Three discrete and collimated light beams are combined using two dichroic mirrors (T660lpxrxt, T510lpxrxt, Chroma, USA) for the excitation of the fluorophores using a single optical path. The incoming light is focused and then collimated using bi-convex and plano-convex lenses. Images of the LED chips are obtained on the front focal plane of the plano-convex lens. A protective silver mirror (PF10-03-P01, Thorlabs, USA) directs the collimated light. A 10× long working distance objective lens (Plan Apo, Mitutoyo, Japan) is employed to focus the excitation light on the sample. The objective is placed so that its back focal

plane intersects with the image plane of the LED chips. Emitted light from fluorophores is collected using a 5× long working distance objective (Plan Apo, Mitutoyo, Japan) and transmitted toward the emission unit by a protective silver mirror (PF20-03-P01, Thorlabs, USA). The emission unit separates the emitted fluorescence light into spectral components for individual image formation of microfluidic channels, HeLa cells, and micro-agents using two dichroic mirrors (DMLP567L and DMLP805L, Thorlabs, USA). Individual fluorescence light beams are directed onto tube lenses (ITL200, Thorlabs, USA) to form images onto complementary metal–oxide–semiconductor (CMOS) cameras (CS135MUN, DCC3240N, Thorlabs, USA). Emission filters (ET845/55m, ET623/60m, ET520/40m, Chroma, USA) are placed before the cameras to block excitation light and select fluorescence wavelengths.

4.6 | Dataset Acquisition and Preprocessing

We employed hardware triggering to enable the simultaneous acquisition of sequential and cross-channel fluorescence frames. Two programmable signal generators (33510B, Keysight Inc., USA) are used in lock mode to ensure synchronization among signal outputs. The first generator is connected to a demultiplexer integrated circuit (74HC4052, Texas Instruments, USA) to drive the sequential LED excitation [38]. The second generator is used to send simultaneous trigger pulses to all cameras. All cameras are set to 0 gain to ensure consistent image acquisition conditions. The generated signals are validated using a high-resolution four-channel oscilloscope (DSOX3024A, Keysight Inc., USA). We mitigate camera-floor bias in the ground-truth channels via intensity thresholding applied before augmentation and training; cross-channel inputs are kept raw.

4.7 | Model Training and Deployment

For training, fluorescence images acquired at 22 ms exposure and native resolution 1024×1024 are resized to 256×256. Each resized frame is split into four nonoverlapping 128×128 tiles, which are spatially shuffled to increase variability. Models are trained on both the full resized frames and the shuffled-tile variants; random gamma jitter is applied only to the tile variants, while raw resized frames are left unchanged (Figure S2, Supporting Information). The dataset is split 4:1 into training and validation. In total, 3750 paired images (input–output) are extracted from 15 continuous video sequences spanning five distinct samples that include different spheroid sizes/morphologies and diverse micro-agent actuation/aggregation patterns (Table S1, Supporting Information). Three sequences (750 pairs) are reserved for validation; the remaining 3000 pairs are doubled to 6000 via the tile-shuffle augmentation for training.

We evaluate the proposed framework with two architectures: a plain U-Net and a pix2pix model (U-Net generator with a PatchGAN discriminator). Training is performed on an NVIDIA RTX A6000 GPU (48 GB VRAM), using 45 GB of system RAM and an 8-core CPU, with batch size 16 and learning rate 1×10^{-4}. Early stopping monitors the mean SSIM on the validation set (750 images) with a patience of six epochs and a minimum improvement threshold of 5×10^{-4}. A detailed list of training parameters is provided in Table S2.

The U-Net is optimized purely with an L_1 reconstruction loss:

$$\mathcal{L}_{\mathrm{U-Net}} = \lambda_1 \left\| \mathbf{y} - G(\mathbf{x}) \right\|_1 \tag{1}$$

with $\lambda_1 = 100$, where G maps the cross-channel input $\mathbf{x}$ to a predicted output $\hat{\mathbf{y}} = G(\mathbf{x})$, which is compared against the ground-truth target y.

For pix2pix training, the discriminator D distinguishes real pairs $(\mathbf{x}, \mathbf{y})$ from fake pairs $(\mathbf{x}, G(\mathbf{x}))$ and minimizes

$$\mathcal{L}_D = -\mathbb{E}_{\mathbf{x},\mathbf{y}}[\log D(\mathbf{x}, \mathbf{y})] - \mathbb{E}_{\mathbf{x}}\left[\log\left(1 - D(\mathbf{x}, G(\mathbf{x}))\right)\right] \tag{2}$$

The generator uses a nonsaturating adversarial objective and $\mathcal{L}_{\mathrm{U-Net}}$:

$$\mathcal{L}_G = \lambda_{\mathrm{adv}} \mathbb{E}_{\mathbf{x}}[-\log D(\mathbf{x}, G(\mathbf{x}))] + \lambda_1 \left\| \mathbf{y} - G(\mathbf{x}) \right\|_1 \tag{3}$$

with $\lambda_1 = 100$ and $\lambda_{\mathrm{adv}} \in \{0.05, 0.1, 0.2, 0.4, 1.0\}$. For comparison, the U-Net corresponds to the $\lambda_{\mathrm{adv}} = 0$ case but is trained separately with $\mathcal{L}_{\mathrm{U-Net}}$. All trainings are stable, with smoothly decreasing generator losses and steadily improving validation SSIM until early stopping (Figure S4).

For the morphology-matched spheroids labeled with red CMTPX and green CMFDA, we curated new datasets containing 1268 paired images for training and 317 for validation and early stopping. Augmentations are applied as described in Figure S2, resulting in a total training dataset of 2536 paired frames.

For deployment, the independently trained generators (Input Channel → MM/CC) are wrapped into a single PyTorch module that fans out a shared input and returns both outputs concurrently. The composite model is compiled with Torch-TensorRT at FP16 for an input of $1 \times 1 \times 256 \times 256$. Inference runs on the RTX A6000, and latency is measured over the validation set with CUDA synchronization. Please refer to the GitHub repository (https://github.com/JuanJJHS/Cross-Channel-Acquisition-and-Deep-Learning-Based-Inference) for the scripts used for preprocessing, augmentation, and training.

4.8 | Brightfield and Scanning Electron Microscopy

Brightfield images are acquired using an inverted microscope (Axiovert A1, Carl Zeiss AG, Germany) for morphological inspection of HeLa cell spheroids. Scanning electron microscopy (SEM) is performed using a JSM-7200F (JEOL, Japan) to visualize the morphology of magnetic electrospun fibers at 450× magnification, with a working distance of 7.3 mm and an accelerating voltage of 5 kV.

4.9 | Irradiance Measurements

LED irradiance per channel is measured using a photodiode sensor (S130VC, Thorlabs, USA) connected to a digital power meter console (PM100D, Thorlabs, USA). Reported values are averaged over 10 consecutive measurements to ensure accuracy and repeatability.

4.10 | Quantitative Metrics

All metrics are computed on native grayscale images (no histogram matching). When ROIs are used, the side length and

location are reported in the figure captions. Let y denote the ground-truth image and $\hat{y}$ the predicted image, both quantized to 256 gray levels. Unless noted, all metrics are computed over all N pixels in the image, where i indexes pixel positions.

- G-NCC: By operating on gradients, G-NCC emphasizes structural similarity while being insensitive to uniform brightness shifts. The percentile sweep focuses the comparison on progressively stronger structural edges, with the reported score corresponding to the most correlated percentile. G-NCC is computed as follows:

$$\mathrm{G-NCC}_p(g,\hat{g}) = \frac{\sum_{i\in M_p}(g_i-\overline{g}_p)\left(\hat{g}_i-\overline{\hat{g}}_p\right)}{\sqrt{\sum_{i\in M_p}(g_i-\overline{g}_p)^2}\sqrt{\sum_{i\in M_p}\left(\hat{g}_i-\overline{\hat{g}}_p\right)^2}}$$

  where g and $\hat{g}$ are the gradient magnitude maps of y and $\hat{y}$, respectively. M_p is the set of pixels whose gradient magnitude in g or $\hat{g}$ exceeds the p-th percentile. The quantities $\overline{g}_p$ and $\overline{\hat{g}}_p$ denote the mean values of g and $\hat{g}$ over M_p, respectively. The final score is taken as

$$G-\mathrm{NCC} = \max_p\ (\mathrm{G-NCC}_p)$$

- SSIM: This metric correlates with perceived quality by jointly assessing luminance, contrast, and structure. We compute SSIM using the following equation:

$$\mathrm{SSIM}(y,\hat{y}) = \frac{(2\mu_y\mu_{\hat{y}}+C_1)(2\sigma_{y\hat{y}}+C_2)}{(\mu_y^2+\mu_{\hat{y}}^2+C_1)(\sigma_y^2+\sigma_{\hat{y}}^2+C_2)}$$

  with local means $\mu_y, \mu_{\hat{y}}$, variances $\sigma_y^2, \sigma_{\hat{y}}^2$, and covariance $\sigma_{y\hat{y}}$ computed using a Gaussian 11×11 window. The constants $C_1=(K_1L)^2$, $C_2=(K_2L)^2$ with $K_1=0.01$, $K_2=0.03$, $L=255$. These values follow the standard SSIM formulation and help stabilize the metric when local means or variances are close to zero.
- PSNR: This standard interpretable error metric in dB complements SSIM by emphasizing absolute, pixel-wise fidelity. PSNR is defined as follows:

$$\mathrm{PSNR}(y,\hat{y}) = 10\log_{10}\left(\frac{255^2}{\mathrm{MSE}(y,\hat{y})}\right)\ [\mathrm{dB}]$$

$$\mathrm{MSE}(y,\hat{y}) = \frac{1}{N}\sum_i (y_i-\hat{y}_i)^2$$

- GCS: This metric complements SSIM and PSNR, providing microtexture similarity and perceptual realism. Let $P_{ij}(z)$ be the normalized gray-level co-occurrence matrix (GLCM) of image z computed with offset distance $d=1$ pixel and angle 0°, 256 gray levels, symmetric counting, and normalization $\sum_{i,j}P_{ij}(z)=1$. GLCM contrast is computed as follows:

$$C(z) = \sum_{i,j}(i-j)^2\,P_{ij}(z).$$

  We compare contrast via:

$$\mathrm{GCS}(y,\hat{y}) = \max\left(0,\ 1-\frac{|\,C(\hat{y})-C(y)\,|}{C(y)+\varepsilon}\right),$$

  where $\varepsilon = 10^{-6}$. GCS $\in[0,1]$; 1 indicates perfect contrast match.

- Normalized fluorescence change ($\Delta F/F_0$): This is a dimensionless measure used in fluorescence imaging to track relative intensity dynamics independent of absolute gain. For a region containing N pixels (full frame or an ROI), let:

$$F(t) = \frac{1}{N}\sum_{i\in R} z_i(t),$$

  where $z_i(t)$ is the grayscale intensity at time t. We compute the normalized fluorescence change as follows:

$$\frac{\Delta F}{F_0}(t) = \frac{F(t)-F_0}{F_0}\times 100\,\%$$

  where the baseline F_0 is the fluorescence intensity reference.

4.11 | Cross-Channel Map and Intensity Imbalance Analysis

We estimated a linear cross-channel map using measured target channels (MM and CC) and their corresponding cross-channel frames (CM). The model is fitted using 116 paired frames acquired under identical imaging conditions. A linear least-squares regression is performed to estimate the average coupling between channels, yielding the following relationship: $\mathrm{CM} = a\,\mathrm{MM} + b\,\mathrm{CC} + c$, with fitted coefficients $a=0.1241$, $b=0.1334$, and a constant offset $c=0.02327$, which accounts for background signal and residual noise. This empirical model is subsequently used to generate synthetic cross-channel inputs for controlled evaluation. Synthetic target channels are obtained by globally scaling the measured images using an imbalance parameter β, such that $\mathrm{MM}' = \beta\,\mathrm{MM}$ and $\mathrm{CC}' = (1-\beta)\,\mathrm{CC}$. The corresponding synthetic cross-channel frame is then computed as $\mathrm{CM}' = a\,\mathrm{MM}' + b\,\mathrm{CC}' + c$. These synthetic inputs are used exclusively for evaluation and are not included during model training.

Reconstruction accuracy under intensity imbalance is quantified using a masked SSIM. ROIs are defined from the measured ground-truth target channels by percentile thresholding to exclude background-dominated regions. Specifically, for measured target channels, binary ROI masks are defined as

$$\mathcal{R}_{\mathrm{MM}}(i) = \begin{cases} 1, & \mathrm{MM}(i) \ge P_{98}(\mathrm{MM}) \\ 0, & \text{otherwise} \end{cases}$$
$$\mathcal{R}_{\mathrm{CC}}(i) = \begin{cases} 1, & \mathrm{CC}(i) \ge P_{98}(\mathrm{CC}) \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where i indexes pixels and $P_{98}(\cdot)$ denotes the 98th percentile of pixel intensities. The resulting ROI masks are computed once from the measured channels and applied identically to the corresponding synthetic targets $(\mathrm{MM}', \mathrm{CC}')$ and their predictions prior to metric computation within the ROI.

To quantify the redistribution of predicted signal between channel-specific regions, a signal leakage ratio (ρ) is computed. For the predicted MM' ($\widehat{\mathrm{MM}'}$), this ratio is defined as

$$\rho_{\mathrm{MM}} = \frac{\left\langle \widehat{\mathrm{MM}'} \cdot \mathcal{R}_{\mathrm{CC}} \right\rangle}{\left\langle \widehat{\mathrm{MM}'} \cdot \mathcal{R}_{\mathrm{MM}} \right\rangle} \quad (5)$$

whereas for the predicted CC′ ($\widehat{CC'}$) it is defined as

$$\rho_{CC} = \frac{\left\langle \widehat{CC'} \cdot \mathcal{R}_{MM} \right\rangle}{\left\langle \widehat{CC'} \cdot \mathcal{R}_{CC} \right\rangle} \tag{6}$$

Here, $\langle \cdot \rangle$ denotes the mean over pixels within the ROI. This ratio provides a direct and directional measure of cross-channel signal leakage in the predicted outputs as a function of the imposed intensity imbalance.

Acknowledgments

We want to thank Mr. Nick Helthuis for his assistance with the scanning electron microscope. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation program under Grant #866494 (project—MAESTRO).

Funding

This work was supported by the HORIZON EUROPE European Research Council (866494).

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. S. Andersson-Engels, J. Johansson, and S. Svanberg, "Bioimaging and Two-Dimensional Spectroscopy," *SPIE* 1205 (1990): 179–189.

2. P. P. Laissue, R. A. Alghamdi, P. Tomancak, E. G. Reynaud, and H. Shroff, "Assessing Phototoxicity in Live Fluorescence Imaging," *Nature Methods* 14 (2017): 7–661.

3. N. Senthilnathan, C. M. Oral, A. Novobilsky, and M. Pumera, "Intelligent Magnetic Microrobots with Fluorescent Internal Memory for Monitoring Intragastric Acidity," *Advanced Functional Materials* 34 (2024): 29–2401463.

4. Y. Wang, J. Shen, S. Handschuh-Wang, M. Qiu, S. Du, and B. Wang, "Microrobots for Targeted Delivery and Therapy in Digestive System," *ACS Nano* 17 (2022): 1–27.

5. Y. Zhang, L. Zhang, L. Yang, et al., "Real-Time Tracking of Fluorescent Magnetic Spore–based Microrobots for Remote Detection of *C. Diff* Toxins," *Science Advances* 5, no. 1 (2019): eaau9650.

6. A. Neettiyath and M. Pumera, "Micro/Nanorobots for Advanced Light-Based Biosensing and Imaging," *Advanced Functional Materials* 35, no. 8 (2025): 2415875.

7. A. Shakoor, T. Luo, S. Chen, M. Xie, J. K. Mills, and D. Sun, *2017 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2017), 5397–5402.

8. M. Yang, X. Guo, F. Mou, and J. Guan, "Lighting Up Micro-/nanorobots With Fluorescence," *Chemical Reviews* 123, no. 7 (2022): 3944.

9. M. Kaya, F. Stein, P. Padmanaban, et al.," Visualization of Micro-agents and Surroundings by Real-time Multicolor Fluorescence Microscopy," *Scientific Reports* 12 (2022): 1–13375.

10. C. K. Schmidt, M. Medina-Sánchez, R. J. Edmondson, and O. G. Schmidt, "Engineering Microrobots for Targeted Cancer Therapies From A Medical Perspective," *Nature Communications* 11, no. 1 (2020): 5618.

11. C. Xin, D. Jin, Y. Hu, et al.,"AgNWs/$Ti_3C_2T_x$ MXene-based Multi-responsive Actuators For Programmable Smart Devices," *ACS Nano* 15 (2021): 11–18048.

12. H. Chen, Y. Li, Y. Wang, et al.," Flexible Fibrous Electrodes For Implantable Biosensing," *ACS Nano* 16, no. 4 (2022): 6118.

13. M. M. Frigault, J. Lacoste, J. L. Swift, and C. M. Brown, "Live-Cell Microscopy – Tips and Tools," *Journal of Cell Science* 122, no. 6 (2009): 753.

14. A. Stylianou, V. Gkretsi, V. Huntošová, and A. Theodosiou, in *Bioimaging in Tissue Engineering and Regeneration: Advanced Microscopy and Preclinical Imaging* (Springer, 2024), 1–28.

15. S. Mukhtar, A. Arbabi, and J. Viegas, (IEEE Access, 2025).

16. C. M. Browning, S. Mayes, S. A. Mayes, T. C. Rich, and S. J. Leavesley, "Microscopy Is Better in Color: Development of a Streamlined Spectral Light Path for Real-Time Multiplex Fluorescence Microscopy," *Biomedical Optics Express* 13 (2022): 3751.

17. J. Ohn, J. Yang, S. E. Fraser, R. Lansford, and M. Liebling, "High-speed Multicolor Microscopy of Repeating Dynamic Processes," *Genesis* 49, no. 7 (2011): 514.

18. T. Zimmermann, *Spectral Imaging and Linear Unmixing in Light Microscopy* (Springer Berlin Heidelberg, 2005), 245–265.

19. Y. Fu, Y. Zheng, H. Huang, I. Sato, and Y. Sato, "Hyperspectral Image Super-Resolution With a Mosaic RGB Image," *IEEE Transactions on Image Processing* 27, no. 11 (2018): 5539.

20. L. Uguen, R. Piedevache, G. Russias, S. Helmer, D. Tregoat, and S. Perrin, "Single-Pixel-Based Hyperspectral Microscopy," *Applied Physics Letters* 125 (2024): 7.

21. Z. Wei, K. Wu, Y. Qin, H. Cheng, and C. Gu, *Fifteenth International Conference on Information Optics and Photonics (CIOP 2024)* (SPIE, 2024), 617–621.

22. M. M. Antony, C. S. Sandeep, and M. V. Matham," Hyperspectral Vision Beyond 3D: A Review," *Optics and Lasers in Engineering* 178 (2024): 108238.

23. Q. Li, Y. Yang, M. Tan, et al., "Rapid Detection of Single Bacteria Using Filter-Array-Based Hyperspectral Imaging Technology," *Analytical Chemistry* 96, no. 43 (2024): 17244.

24. Q. Lv, K. Liang, C. Tian, et al., "Unveiling Thymoma Typing Through Hyperspectral Imaging and Deep Learning," *Journal of Biophotonics* 17 (2024): 11–e202400325.

25. Y. Shechtman, L. E. Weiss, A. S. Backer, M. Y. Lee, and W. Moerner, "Multicolour Localization Microscopy by Point-Spread-Function Engineering," *Nature Photonics* 10, no. 9 (2016): 590.

26. C. Yang, V. Hou, L. Y. Nelson, and E. J. Seibel, "Mitigating Fluorescence Spectral Overlap in Wide-Field Endoscopic Imaging," *Journal of Biomedical Optics* 18, no. 8 (2013): 086012.

27. J. Mao and H. He, "Deep Learning in Fluorescence Imaging and Analysis," *Journal of Intelligent Medicine* 1, no. 1 (2024): 42.

28. Y. Jiang, H. Sha, S. Liu, P. Qin, and Y. Zhang, "AutoUnmix: An Autoencoder-Based Spectral Unmixing Method for Multi-Color Fluorescence Microscopy Imaging," *Biomedical Optics Express* 14, no. 9 (2023): 4814.

29. H. Wang, Y. Rivenson, Y. Jin, et al., "Deep Learning Enables Cross-Modality Super-Resolution in Fluorescence Microscopy," *Nature Methods* 16, no. 1 (2019): 103.

30. H. Zhuge, B. Summa, J. Hamm, and J. Q. Brown, "Deep Learning 2D and 3D Optical Sectioning Microscopy Using Cross-Modality Pix2Pix cGAN Image Translation," *Biomedical Optics Express* 12, no. 12 (2021): 7526.

31. C. Bouchard, T. Wiesner, A. Deschênes, et al., "Resolution Enhancement with a Task-Assisted GAN to Guide Optical Nanoscopy

Image Analysis and Acquisition," *Nature Machine Intelligence* 5, no. 8 (2023): 830.

32. E. Hershko, L. E. Weiss, T. Michaeli, and Y. Shechtman, "Multicolor Localization Microscopy and Point-Spread-Function Engineering by Deep Learning," *Optics Express* 27, no. 5 (2019): 6158.

33. S. Liu, W. Zou, H. Sha, et al., "Deep Learning-Enhanced Snapshot Hyperspectral Confocal Microscopy Imaging System," *Optics Express* 32, no. 8 (2024): 13918.

34. H. Sha, H. Li, Y. Zhang, and S. Hou, "Deep Learning-Enhanced Single-Molecule Spectrum Imaging," *Apl Photonics* 8 (2023): 9.

35. B. Dai, S. You, K. Wang, et al., "Deep Learning–enabled Filter-Free Fluorescence Microscope," *Science Advances* 11, no. 1 (2025): eadq2494.

36. O. Ronneberger, P. Fischer, and T. Brox, "Medical image computing and computer-assisted intervention–MICCAI," 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, 234–241 (Springer, 2015).

37. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image Translation with Conditional Adversarial Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2017), 1125–1134.

38. M. Kaya, F. Stein, J. Rouwkema, I. S. Khalil, and S. Misra, "Serial Imaging of Micro-Agents and Cancer Cell Spheroids in a Microfluidic Channel Using Multicolor Fluorescence Microscopy," *PLoS One* 16, no. 6 (2021): e0253222.

39. N. C. Shaner, P. A. Steinbach, and R. Y. Tsien, "A Guide to Choosing Fluorescent Proteins," *Nature Methods* 2 (2005): 12–909.

40. A. Hoppe, K. Christensen, and J. A. Swanson, "Fluorescence Resonance Energy Transfer-Based Stoichiometry in Living Cells," *Biophysical Journal* 83, no. 6 (2002): 3652.

41. D. Qin, Y. Xia, and G. M. Whitesides, "Soft Lithography for Micro- and Nanoscale Patterning," *Nature Protocols* 5, no. 3 (2010): 491.

42. B. Jung, V. I. Vullev, and B. Anvari, "Revisiting Indocyanine Green: Effects of Serum and Physiological Temperature on Absorption and Fluorescence Characteristics," *IEEE Journal of Selected Topics in Quantum Electronics*, 20, 2013, 149–157.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Supporting Figure S1:** Architectures used in the real-time reconstruction framework. (A) U-Net with a stem (first encoder block), six further encoder blocks, a bottleneck, seven decoder blocks, and a final transposed convolution that produces a $1 \times 256 \times 256$ reconstruction ($G(x)$) from input cross-channel frame (x). To maintain pixel-level registration between the input and the reconstruction, all two-dimensional convolutions (Conv2d) and transposed convolutions (ConvTranspose2d) use a kernel ($k = 4$), stride ($s = 1$), and padding ($p = 1$) (reflect padding for Conv2d). The U-Net architecture features skip connections, which are channel-wise concatenations: $U2 \leftarrow [U1 \| D7]$, $U3 \leftarrow [U2 \| D6]$, ..., $Final \leftarrow [U7 \| D1]$. (B) PatchGAN inputs are concatenated pairs—real ($[x \| y]$) and fake ($[x \| G(x)]$). With strides ($s = 1, 2, 2, 1, 1$), it outputs a patch logit map $61 \times 61 \times 1$. Binary cross-entropy with logits is computed element-wise and averaged over patches within each mini-batch. **Supporting Figure S2:** Data augmentation used to train U-Net and pix2pix. From each paired frame set—co-registered images of the same field of view consisting of the cross-channel input (CM) and its targets (MM, CC)—we generate two training sam ples. (1) Resized-only: resize to 256×256 pixels. (2) Tiled–shuffled: take the resized frame, split into four 128×128 tiles, randomly permute their order (re-sampled for each source frame), and scale by $\alpha \sim U(0.1, 1.0)$. The same α is applied to the input and its paired target to preserve correspondence. Both outputs (resized-only and tiled–shuffled) are stored and used for training. For intensity scaling notation, "$\sim$" means "is drawn from", and $U(a, b)$ denotes the continuous uniform distribution on $[a, b]$. **Supporting Figure S3:** Effect of adversarial weight on reconstruction quality. Violin plots of structural similarity index metric (SSIM), peak signal-to-noise ratio (PSNR), and contrast similarity (GCS) for (A) CM→MM and (B) CM→CC. All models use an L_1 loss with a reference weight of 100. pix2pix model adds an adversarial loss weight ($\lambda_{adv} \in \{0.05, 0.1, 0.2, 0.4, 1.0\}$). The U-Net baseline (leftmost) corresponds to $\lambda_{adv} = 0$. Increasing λ_{adv} improves the texture reconstruction fidelity and slightly reduces the SSIM and PSNR. In channel CC, this improvement is more notorious even at small adversarial weights as the spheroids contain more texture features compared to the micro-agent channel (MM). All pix2pix reconstructions presented in this study used $\lambda_{adv} = 0.05$, selected as a trade-off between GCS, PSNR, and SSIM. Violin envelopes show the score distribution for the 3 sequences used for validation (750 paired frames). Black dots and bars are the mean ± standard deviation. **Supporting Figure S4:** Training and validation performance for micro-agent (CM→CC) and spheroid models (CM→CC) using U-Net ($\lambda_{adv} = 0$) and pix2pix ($\lambda_{adv} = 0.05$) configurations. The first and third rows show generator loss curves—weighted sum of L_1 loss (λ_1=100) and adversarial loss (λ_{adv})—plotted across training epochs. The second and fourth rows show the corresponding validation SSIM curves used for early stopping. Vertical dashed gray lines mark the selected early-stopping checkpoints. Loss spikes reflect mini-batch variability and the inherent adversarial dynamics under mixed precision. **Supporting Figure S5:** Fluorescence channel reconstructions under progressive LED dimming. (A) Representative CM input frames, ground truth, and composed reconstructions for MM (micro-agents) and CC (spheroids) at four irradiance sets (I–IV). Gamma correction ($\gamma = 0.5$) is applied to input frames to improve visualization. Scale bars: 100 μm. (B) LED irradiance for sets I–IV, showing tandem dimming. (C) Relative fluorescence change ($\Delta F/F_{ref}$) for CM, MM, and CC frames. Using U-Net ($\lambda_1 = 100$, $\lambda_{adv} = 0$) and pix2pix ($\lambda_1 = 100$, $\lambda_{adv} = 0.05$), we accurately reconstruct MM and CC and track per channel $\Delta F/F_0$ across irradiance levels. **Supporting Figure S6:** Intensity imbalance analysis using synthetic cross-channel inputs. (A) Representative synthetic cross-channel inputs (CM′), corresponding synthetic target channels (MM′ and CC′), and pix2pix predictions for increasing intensity imbalance (β). Scale bars: 100 μm. (B) Scatter plots of region-of-interest (ROI)–averaged intensities comparing pix2pix predictions with corresponding synthetic targets for representative β values. (C) Signal leakage ratio as a function of β, quantifying redistribution of predicted signal into the opposite channel's region. Ratios remain near zero in the well conditioned regime ($\beta = 0.2$–0.45) and increase sharply across the transition regime ($\beta = 0.52$), indicating a progressive loss of identifiability as the intensity mixture becomes ill-conditioned. **Supporting Table S1:** Acquisition settings per data set used for training and validation. All sequences were recorded at 1024×1024 with 0 dB camera gain and 22 ms exposure time. "FOV Motion" indicates stage-driven field of view motion during acquisition. **Supporting Table S2:** General training hyper parameters used for training U-Net and pix2pix models.