Supplementary Information

Discriminating Single-Molecule Binding Event from Diffraction-Limited Fluorescence

Yueming Yin[1], Nithin Pathoor[2], Kamal Kant Sharma[2], Shiwen Zhu[2], Iong Ying Loh[3], Yan Shan Ang[3], Shao Ren Sim[2], Lin Yue Lanry Yung[3], Thorsten Wohland[*1,2,4], and Lipo Wang[*1,5]

[1]Institute for Digital Molecular Analytics and Science (IDMxS), Nanyang Technological University, 59 Nanyang Drive, Singapore 636921
[2]Centre for Bioimaging Sciences, Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117557
[3]Department of Chemical and Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117585
[4]Department of Chemistry, National University of Singapore, 3 Science Drive 3, Singapore 117543
[5]School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798

Contents

*Corresponding authors: twohland@nus.edu.sg, elpwang@ntu.edu.sg

Supplementary Notes

1 Frame correlation definition

Given a video represented as a sequence of frames, we compute the Pearson correlation coefficient (PCC) between all pairs of non-zero frames. The correlation matrix is then accumulated across different video samples belonging to the same class. Let $\mathbf{x}_t$ and $\mathbf{x}_s$ be the vectorized pixel intensities of frames t and s. The Pearson correlation coefficient between two frames is defined as:

$$\rho_{t,s} = \frac{\sum_i (\mathbf{x}_{t,i} - \mu_t)(\mathbf{x}_{s,i} - \mu_s)}{\sqrt{\sum_i (x_{t,i} - \mu_t)^2} \sqrt{\sum_i (x_{s,i} - \mu_s)^2}} \tag{1}$$

where:

- $\mathbf{x}_{t,i}$ is the intensity of pixel i in frame t.
- μ_t and μ_s are the mean pixel intensities of frames t and s, respectively.
- The denominator normalizes the covariance, ensuring the correlation values are within the range $[-1, 1]$.

For each class label y, the accumulated frame correlation matrix is computed as:

$$C_{\mathrm{y}}(t, s) = \sum_{n \in \text{class y}} \rho^n(t, s) \tag{2}$$

where $\rho^n(t, s)$ is the correlation coefficient between frames t and s in the n-th video sample. Since different frame pairs appear with varying frequencies in the dataset, the final averaged correlation matrix is computed as:

$$\bar{C}_{\mathrm{y}}(t, s) = \frac{C_{\mathrm{y}}(t, s)}{N_{\mathrm{y}}(t, s) + \epsilon} \tag{3}$$

where:

- $N_{\mathrm{y}}(t, s)$ is the number of times the correlation between frames t and s was computed.
- ϵ is a small positive constant (e.g., 10^{-6}) to avoid division by zero.

This ensures that correlation values are normalized by their respective occurrence counts, providing an unbiased estimate of frame-wise dependencies.

2 Analysis of frame correlation patterns

A particularly notable deviation occurs around frame 320 in Domain 1, where frames exhibit high mutual correlation but lower correlation with subsequent frames. Upon examining the raw data (Fig. **3**), we found that among the few binding events lasting over 300 frames, the longest event (655 frames) displayed two closely positioned spots between frames 308 and 334. This interference caused a fluorescence pattern distinct from the typical single-fluorescence pattern observed in other frames, leading to a drop in correlation with later frames, especially when fewer binding events followed.

To illustrate these fluorescence features, we sampled representative inter-frame correlation images (Fig. **3**). Beyond the correlation variations, subtle differences in fluorescence images can also be observed. For example, Domain 1 (8nt-10nt) exhibits slightly greater fluctuations in peak positions and shapes compared to Domain 2 (10nt-10nt). Frames such as 198 and 620 in Domain 1 show a fluorescence shift toward the right, reducing their correlation with other

frames. The shorter binding duration in Domain 1 also minimizes the impact of photobleaching. In contrast, the first long binding event in Domain 2 undergoes noticeable photobleaching in its later stages, reducing the contrast between the fluorescence spot and the background. However, peak positions and shapes remain relatively stable compared to Domain 1, providing a basis for differentiation. The final two blue lines in the Domain 2 inter-frame correlations result from transient blinking (signal loss) near the end of the longest binding event.

3 Multi-class experiments

Settings. Single-molecule experiments often involve two or three target types. To evaluate the method's performance in multi-class settings, we introduced an additional 6nt-6nt R1 strand binding type. As shown in Main Text Fig. 5(a), we shorten the target strand from 8nt to 6nt. However, due to the short binding duration of a 6nt strand with the original 10nt P3 imager, we instead used a complementary 6nt R1 strand as both the docking and imager strand. Notably, the 6nt R1 imager carries the same ATTO532 dye to enable multiplexing.

Data analysis. Main Text Fig. 5(b) presents the distribution of experimental data, corrected for background and blinking, for three binding types across six experiments. We controlled the concentration so that each binding site exhibited a similar number of binding events within the same 20k frames (10 fps), with most sites showing 1–9 binding events. The detected binding sites for the three binding types were 19,457, 6,073, and 10,888, respectively. Their binding durations increased sequentially, primarily ranging from 0.1 to 10 seconds.

To further investigate the discriminability of binding events based on brightness, we plotted both the integrated intensity and the mean frame intensity distributions in Main Text Fig. S4b. The integrated intensity distribution primarily reflects the underlying distribution of binding durations, as expected given the relatively stable per-frame intensity within each binding event. Consequently, CNN models need to distinguish binding types by learning from the full fluorescence video dynamics rather than relying on static intensity metrics. This involves leveraging temporal features such as inter-frame correlations (Main Text Fig. 2b) and other temporally correlated statistics (Main Text Fig. 2c). For the 6nt R1 imager, the average frame intensity was lower (approximately 50 and 100 for training and testing sets, respectively), likely due to short binding durations that sometimes result in incomplete fluorescence accumulation within a single frame.

Workflow. Main Text Fig. 5(b) illustrates the workflow for image-based multi-class classification of binding types. When a binding event is detected (typically lasting 0.1–10 seconds), its fluorescence signals undergo background and blinking correction to eliminate background intensity variations and transient disappearances. The processed signals are then sliced, padded to a uniform length, and input into image-based binding classification models, including 3D ResNet-18 [41], Video Transformer [34], ED-TCN [21], SqueezeTime [46], and the proposed T2C CNN.

Results. After training, T2C CNN achieved the highest test accuracy (88.15±1.86%), followed by Video Transformer (68.88±2.54%). These models classify each detected signal into one of three binding types (6nt-6nt, 8nt-10nt, or 10nt-10nt). Using these predictions, we differentiated targets labeled with distinct docking strands in the final super-resolution (SR) image, generating a three-color SR image, as shown in Fig. **4**. T2C CNN consistently demonstrated superior classification accuracy across accumulated time points, with its advantage becoming more pronounced over longer durations. This makes it a powerful tool for accurate single-fluorophore, three-color SR imaging.

When multiple fluorophores are available, T2C CNN can first distinguish fluorophores frame by frame (will be discussed in Supplementary Note 8) and then leverage accurate single-fluorophore classification ability to multiply the number of identifiable targets. This significantly enhances both analysis capacity and speed.

4 Analysis on fluorescence spots and T2C CNN convolutional outputs from distinct binding events

Main Text Fig. 6d and 6e show that T2C CNN can accurately classify short binding events, including single-frame events. Does this imply that temporal information contributes little to the classification? To explore the answer, we tracked the response of T2C CNN's top kernels to fluorescence spots over time in Fig. **5** and **6**, corresponding to the DNA origami and cell experiments, respectively. Overall, we found that the accumulated convolutional output becomes more pronounced and distinguishable as more frames are included, indicating that temporal information does contribute to the prediction. However, accumulating outputs across frames within a binding event does not always enhance previously observed patterns. For example, in the cell experiment (lower right of Fig. **6**b), the Top-3 kernel's accumulated output at frame 3 for Domain 2 weakens the pattern accumulated in frames 1 and 2. This is consistent with the input fluorescence spot at frame 3, which appears slightly smaller and dimmer, with a noticeable change in shape and intensity distribution compared to the first two frames.

When comparing across domains, we found that the accumulated convolutional outputs generally share a similar overall style—due to similar input spot characteristics and the use of the same convolutional kernels—but differ in detail. For instance, in the Top-1 kernel's accumulated output in the DNA origami experiments (second row in Fig. **5**a and **5**b), both domains show low values at the center. However, Domain 1 exhibits a horizontally extended low-value region, while Domain 2 shows a lower-right clustered pattern, especially as more frames are accumulated. These distinct patterns are not readily apparent from the input fluorescence spots alone. This suggests that local mismatches between kernel patterns and specific regions of the fluorescence spots contribute to the observed convolutional differences.

When comparing across experiments, we observed that convolutional kernels in the DNA origami experiments (first three columns in Fig. **5**) contain more negative values than those in the cell experiments (first three columns in Fig. **6**). Kernels with clustered negative values can produce stronger responses to certain peripheral patterns in fluorescence spots. In the DNA origami experiments, these peripheral regions are typically cleaner and more circular. When these regions exhibit domain-specific differences, the T2C CNN is likely to focus on them rather than the more random center intensity. This observation may help explain the phenomenon in Main Text Fig. 6a, where the periphery appears to contribute more to T2C CNN's predictions. In contrast, the fluorescence spots in the cell experiments display more irregular or complex periphery and center intensity, leading to more variability. In this case, the Top-1 kernel tends to emphasize peripheral regions (more negative values), while the Top-2 and Top-3 kernels focus more on center intensities (more positive values).

When analyzing single-frame binding events (Fig. **7**), we found that their convolutional outputs are generally as weak as the first frame of longer events in Fig. **5** and **6**. Despite being weak, these outputs can still differ between domains, e.g., through differences in value distributions. The output patterns are driven by the fluorescence inputs and do not form consistent intra-domain features. However, the trained top kernels act as fixed templates that highlight differences in domain-specific spot characteristics. For example, the Top-1 kernel in the DNA origami experiments (second row in Fig. **7**a and **7**b) shows a peripheral gradient pattern, such as a medium–high–low intensity transition in the first row of the kernel. In the cell experiments, the Top-1 kernel (second row in Fig. **7**c and **7**d) exhibits a center-focused pattern, characterized by a high-intensity pixel located at the upper-right of two adjacent high-intensity pixels within the same row. These kernel patterns serve as templates to measure the degree of alignment in the input fluorescence spots. Higher convolutional layers and fully connected layers then use these measurements to abstract more high-level representations for final classification.

From the above analyses, we found that temporal information plays an important role in T2C CNN's decision-making by changing the final convolutional patterns. Misclassifications of rare long events may arise from diluted or inconsistent temporal patterns. Single-frame events,

while weaker in response, can still be distinguished by domain-specific convolutional patterns. The increased complexity and variability of fluorescence spots in cell experiments contribute to lower classification accuracy. In contrast, the cleaner and more regular peripheral regions of DNA origami spots encourage the model to focus on periphery-based patterns for more reliable classification.

5 Robustness test of image-based binding-type classification models

In single-molecule fluorescence binding experiments, nonspecific binding, background intensity noise, and camera defects are common issues, making it crucial to evaluate the robustness of various methods under noisy conditions. This section assesses the robustness of image-based binding-type classification models (3D ResNet, Video Transformer, ED-TCN, SqueezeTime, and the proposed T2C CNN) by simulating these three types of noise interference.

Robustness test settings. (a) Light interference with Poisson distributions: Nonspecific molecular binding events generate short, weak fluorescence signals, which can overlap with specific signals [45] and disrupt accurate classification. Considering the discrete nature of photon detection [30] and a radial intensity drop-off based on physical optics [44], we simulated interference signals using Poisson distributions with Gaussian decay. As shown in the first three images of Main Text Fig. 7(a), an artificial light spot is overlaid on the original signal, producing the third image with interference. Despite increasing the signal-to-noise ratio (SNR) (defined as the ratio of the image mean to its standard deviation [8]), the added spot is not fluorescence from molecular binding. To mimic experimental fluorescence images, we set the interference spot radius and Gaussian decay standard deviation to 5 and 1.8, respectively. The interference spot's peak intensity is varied from 0.1 to 1.2 times the original signal's peak intensity in increments of 0.1 or 0.2, with positions randomized near the original signal to test the robustness of the classification models.

(b) Gaussian noises: Fluorescence images often suffer from Gaussian noise caused by detector electronic noise, such as thermal and readout noise [14]. We simulated Gaussian noise with a mean of 0 and variances ranging from 0.1 to 1.2 times the original signal's mean intensity. The fourth image in Main Text Fig. 7(a) displays an example with Gaussian noise variance set to 0.1 times the original signal's mean intensity, reducing the SNR from 3.94 to 3.79.

(c) Hot pixel noises: Prolonged camera usage can cause sensor defects, producing pixels with abnormally high intensity (hot pixels) [12]. To evaluate the robustness of classification models against hot pixel noise, we randomly introduced 1, 3, 5, 7, 9, 10, and 12 hot pixels with intensities equal to the original signal's peak intensity. The last image in Main Text Fig. 7(a) shows an example with 5 hot pixels, which mask parts of the original signal.

Robustness test results. Main Text Fig. 7(b) records the average classification accuracy of the tested models (3D ResNet-18 [41], Video Transformer [34], ED-TCN [21], SqueezeTime [46], and the proposed T2C CNN). For SqueezeTime and 3D ResNet-18, we used the default 48-frame input, which achieves comparable accuracy to the 512-frame input but with significantly higher efficiency.

(a) Light interference with Poisson distributions. Poisson noise has the most significant impact on classification performance, as it can result from specific or nonspecific binding events occurring in close proximity, causing signal overlap. When the interference spot's peak intensity exceeds 0.5 times that of the signal, all models show a marked performance decline. Increasing the interference intensity from 0.5 to 0.7 times results in accuracy drops of 2.78% to 12.54%. SqueezeTime (2.78%) and T2C CNN (2.88%) exhibit the greatest resilience. Both models employ temporal-to-channel transformations, highlighting the effectiveness of this operation in enhancing robustness against interference spots. One possible reason is that the temporal-to-channel transformation captures long-term fluorescence variations within channels, helping the model focus on the primary signal.

(b) Gaussian noises. The models display polarized performance under Gaussian noise. Video Transformer and ED-TCN, which extract image features before temporal operations, are the most vulnerable. In contrast, 3D ResNet-18 and T2C CNN, which feature residual connections and cross-block concatenation, show the strongest resistance. Residual connections and cross-block concatenation pass low-level features from earlier layers directly to deeper layers, helping retain structural and spatiotemporal information and making it easier to denoise the signal. This is because Gaussian noise typically affects fine-grained details rather than structural patterns. Conversely, models that first extract image features propagate Gaussian noise into later temporal operations, exacerbating its effect.

(c) Hot pixel noises. 3D ResNet-18 and SqueezeTime, both with large model parameters (>100 MB), show the highest resilience to hot pixel noise. Smaller models (T2C CNN, ED-TCN, and Video Transformer, with parameters from 1.65 to 7.80 MB) are more susceptible, with T2C CNN being the least affected among them. Large models can memorize fine-grained patterns, making them less sensitive to small-scale disruptions such as hot pixels. In contrast, smaller models, with limited parameter capacity, are more susceptible to pixel losses. Despite having the fewest parameters, T2C CNN stands out among small models due to its high parameter efficiency (long temporal kernel and equal-length stride), which enhances its resistance to hot pixel noise.

Based on the above robustness test results, we recommend maintaining the interference spot intensity below half of the signal spot intensity and using cameras with an effective pixel ratio above 93% (tolerating up to 7 hot pixels per 100 pixels) when applying the T2C CNN.

6 Hyperparameter analysis of the T2C CNN

To evaluate the impact of hyperparameter choices in T2C CNN, we analyzed its key hyperparameters, including network depth, width, and slice length, as shown in Main Text Fig. 8.

Experimental settings. (a) Depth: Depth is a critical attribute of deep neural networks, significantly affecting performance [37]. The proposed T2C CNN, as shown in Main Text Fig. 8, comprises eight hidden layers (convolutional or fully connected), evenly distributed across four blocks. To investigate the minimum required depth, each block was sequentially assigned 1 to 3 layers, increasing the total depth from 4 to 12 while maintaining the four-block structure. For the first block, which condenses spatial-temporal features, the convolutional kernel size was adjusted for configurations with one or three layers. Main Text Fig. 8b illustrates the corresponding architectures.

(b) Width: Like depth, network width (the number of convolutional kernels or hidden units) plays a critical role in CNN performance [23]. The default T2C CNN configuration uses 64 kernels or hidden units per layer. To explore the effect of width, we tested variations ranging from 16 to 512, including commonly used widths such as 16, 32, 48, 100, 128, 256, and 512.

(c) Slice length: The slice length, which determines the interval at which fluorescence frames are input into the T2C CNN, is critical for capturing long-term temporal patterns in fluorescence videos. As shown in Main Text Fig. 2e, most binding events occur within 50 seconds (500 frames), with the longest binding duration not exceeding 2000 frames before averaging. Therefore, the default slice length of 512 frames was varied by halving or doubling to explore alternatives of 64, 128, 256, 1024, and 2048 frames.

Experimental results (Main Text Fig. 8a). (a) Depth: Among the tested depths ranging from 4 to 12 layers, the top-performing configurations were depths of 8 (94.76%), 4 (94.00%), and 12 (93.27%), all multiples of 4. These results indicate that performance is optimal when each block in the T2C CNN contains an equal number of hidden layers, with two hidden layers per block yielding the best results, followed by one layer per block. Notably, at a depth of 4, the first block of T2C CNN required a larger convolution kernel (increased from "3 $\times$ 3" to "10 $\times$ 10") to condense all spatial-temporal features within a single hidden layer. This

adjustment increased the parameter count approximately 11-fold, significantly contributing to the model size.

(b) Width: Among the tested widths of 16, 32, 48, 100, 128, 256, and 512, a width of 128 slightly outperformed the default 64 (94.76% → 95.38%). However, this improvement came with a 2.62-fold increase in model size (1.65 MB → 4.32 MB) and a 2.04-fold increase in computational cost (0.48 GFLOPs → 0.98 GFLOPs). From an efficiency perspective, the default width of 64 strikes a good balance. If computational resources and latency are not constraints, a width of 128 is also recommended. Widths smaller than 64 or greater than 128 consistently underperformed compared to the default, with performance generally declining as the width became too wide.

Interestingly, even with an unusually small width of 16, T2C CNN maintained performance above 90%. This resilience can be attributed to the unique block design of the T2C CNN, as breaking the balance of hidden layers within blocks or removing blocks led to performance degradation, even with an optimal width of 64. Remarkably, this suggests that when a fluorescence video spanning 512 frames is condensed into 16 channel dimensions, most discriminative information is retained. This highlights the substantial redundancy in fluorescence videos, much of which stems from identical optical properties of the same fluorescence type. The key discriminative signals arise from subtle changes in the fluorescence microenvironment caused by binding to different targets. These subtle variations are condensed into a small set of discriminative features by the T2C CNN, which is fundamental to its functionality.

This also explains why performance significantly declined at widths of 256 and 512. The increased number of hidden nodes likely led to overfitting on redundant information, impairing generalization to test data from different experiments.

(c) Slice length: As the slice length input to T2C CNN increases up to 512 frames, classification accuracy improves significantly. This is because the longer slice length enables the T2C CNN to capture a more complete binding event, aiding in determining the binding type. However, beyond 512 frames, the performance plateaus, as only a small fraction of binding events extend beyond this duration. In practical applications, we recommend setting the slice length to encompass the majority of binding durations (e.g., the maximum binding time excluding outliers) to achieve optimal efficiency.

Comparison with baselines. This section compares the proposed T2C CNN with two closely related baseline methods: the classical temporal convolutional network ED-TCN [21] and SqueezeTime [46], which compresses the temporal dimension into the channel dimension. For a fair comparison, we used the same 512-frame input length for SqueezeTime as T2C CNN (its default input is a 16-frame or 48-frame video clip). This adjustment significantly increased SqueezeTime's model size (28.7 MB → 113.3 MB) and computational cost (5.5 GFLOPs → 19.56 GFLOPs). SqueezeTime achieved classification accuracies of 71.72% with the default 16-frame input and 72.98% with the 512-frame input on fluorescence videos, indicating that increasing the number of input frames yields only marginal improvement. This is likely because the repeated transformations between channel and temporal dimensions, along with weighted convolutions in SqueezeTime, risk losing discriminative spatiotemporal details, regardless of the input length.

For ED-TCN, we tested both its default 64-frame input and a 512-frame input matching T2C CNN. Its classification accuracies on fluorescence videos were 70.53% and 58.53%, respectively, demonstrating its unsuitability for long fluorescence video inputs. A key limitation is that ED-TCN's fragmented frame-wise spatial features hinder the capture of long-term spatiotemporal patterns in fluorescence videos. The top-performing ED-TCN and SqueezeTime models are placed in Main Text Fig. 8a.

In contrast, the proposed T2C CNN is tailored for fluorescence video analysis. Its shallow cross-layer connections retain multi-scale spatiotemporal features, while long-term spatial convolutions are employed to capture extended spatial frequency patterns. Additionally, the

integration of shallow cross-connected blocks with a pooling-free design facilitates efficient spatiotemporal information fusion. All spatiotemporal features are compressed into the channel dimension and jointly processed within the T2C layer. The individual and combined contributions of these three components are systematically evaluated in the Ablation Study section. For reference, 3D ResNet-18 and Video Transformer are included in Main Text Fig. 8a. These two models are widely adopted in computer vision for video classification and generally outperform lighter architectures such as SqueezeTime and ED-TCN due to their strong representational capacity. However, as discussed in the "Motivation" section of the Main Text Method section, they are still suboptimal for fluorescence video classification, which demands a domain-specific architectural design.

7 Ablation study of the T2C CNN

To investigate the impact of T2C CNN's architectural components, we conducted an ablation study focusing on its core building blocks and components. Three key components (**long-term spatial convolutions**, **skip concatenations**, and **pooling-free**) were selected for a detailed evaluation of their individual and combined contributions, as presented in Main Text Fig. 8 and Table **3**.

Experimental settings. (a) Blocks: T2C CNN consists of four blocks, with two being fundamental: the Temporal-to-Channel (T2C) block for condensing spatial-temporal features and the Feature Classification block. To assess the importance of the remaining blocks—the second (Hidden Transformation) and third (Multi-Scale Feature Fusion)—we individually removed each and evaluated the impact.

(b) Components: The proposed T2C CNN is specifically designed for fluorescence video analysis. It employs **long-term spatial convolutions** to capture extended spatial frequency patterns. The combination of **skip concatenations** and a **pooling-free** design enables effective spatiotemporal information fusion across the network. For evaluation, we compared the optimal configuration (slice length of 512, all skip concatenations enabled, no pooling) with a baseline configuration (slice length of 64, no skip concatenation, average pooling), corresponding respectively to the presence and absence of the three proposed components.

Experimental results (Main Text Fig. 8a). (a) Blocks: Removing the second and third blocks from the T2C CNN architecture reduced classification accuracy from 94.76% to 88.80% and 90.33%, respectively. This significant performance drop indicates that both blocks are integral components of T2C CNN. Among the two, the second block (Hidden Transformation) plays a more critical role. It consists of two 1×1 convolutional layers that transform the condensed global spatiotemporal features into more discriminative representations, facilitating the differentiation of binding types. Notably, when the third block (Multi-Scale Feature Fusion) operates without the transformed features from the second block, its fusion-based features result in lower accuracy (88.80%) compared to using only the transformed features from the second block (90.33%). This demonstrates the importance of the second block in preparing features for effective multi-scale fusion.

(b) Components: Table **3** presents a systematic evaluation of the key components of T2C CNN—long-term spatial convolutions, skip concatenations, and the pooling-free strategy—on both single-event and single-molecule classification accuracy, with all results averaged over five cross-validation trials.

Among individual components, long-term spatial convolutions contribute the most significant performance gain, boosting class-wise single-event and single-molecule accuracies to 84.14% and 88.86%, respectively, compared to the baseline's 72.01% and 79.87%. This highlights their pivotal role in capturing extended spatial-frequency patterns essential for distinguishing binding types in fluorescence videos. In contrast, skip concatenations (72.58% / 78.13%) and no-pooling (70.73% / 77.60%) result in slightly lower class-wise accuracies than the baseline, suggesting that their benefit emerges primarily when combined with other architectural elements to sup-

port multi-scale feature fusion and preserve spatiotemporal information. The overall single-event accuracy follows a similar trend, with long-term convolutions providing the most substantial standalone gain (89.07% vs. 72.95% baseline).

When components are combined, their effects are complementary. Integrating long-term convolutions with skip concatenations increases class-wise single-event accuracy to 86.33%, while pairing long-term convolutions with no-pooling pushes it to 87.34%. In contrast, combining skip concatenations with no-pooling without long-term convolutions yields only a modest 72.51%, reinforcing the central importance of long-term spatial convolutions.

The full T2C CNN, incorporating all three components, achieves the best performance: 94.76% class-wise single-event accuracy, 96.99% single-molecule accuracy, and 96.78% overall accuracy. Notably, it also exhibits the lowest standard deviations across all metrics ($\pm$0.47%, $\pm$0.46%, and $\pm$0.50%, respectively), indicating stable and consistent predictions across cross-validation folds. This demonstrates that the proposed architectural combination of existing elements enables not only optimal spatiotemporal information fusion and feature preservation, but also robust generalization, tailored specifically for high-precision fluorescence video analysis.

8 Single-frame discrimination of different dye-labeled binding events

As a fluorescence classification model, T2C CNN can be used for multiplexing using the wavelength dependence of the emission PSF. This is an alternative to multi-fluorophore experiments using spectral separation [3, 13]. Spectral isolation typically reduces the signal by avoiding spectrally overlapped regions, and increases noise due to cross-channel bleed-through [26]. Some of these issues can be solved by sequential imaging, but only at the expense of longer measurement times [29]. And spectral unmixing strategies require a sufficient signal to properly separate different fluorophores [49]. We demonstrated the feasibility of this approach by applying T2C CNN to two fluorophore data sets (green and red) without spectrally isolating their emission, but using purely spatial features of the emission profiles as differentiating factors. The success of this strategy allows us to perform the classification of two or more spectrally different signals in any SMLM method, such as PALM and STORM in addition to DNA-PAINT.

To discriminate different dye-labeled molecules from diffraction-limited images, many efforts have been made, including spectral imaging and unmixing [15, 2], fluorescence lifetime imaging microscopy (FLIM) [20, 11], super-resolution fluorescence microscopy (SMFM) [16], and machine learning methods. Spectral imaging and unmixing separate fluorophores based on their emission spectra, with the reliance on spectrally resolved microscopy [15, 2]. Fluorescence lifetime differences between dyes have been used to discriminate them in FLIM, but require prolonged observations [20, 11]. Super-resolution techniques like PALM [5] and STORM [31] are frequently used to distinguish fluorophores based on their blinking and emission properties. However, themselves generally cannot discriminate different dyes in a single frame because these techniques rely on temporal separation, where individual fluorophores are stochastically activated and imaged one at a time across many frames [4, 35, 43]. Machine learning, or deep learning, methods have been employed to act as automatic SMFM trace selector [22], distinguish the true signal of fluorescently labeled molecules from background fluorescence and noise [42, 10], classify binding events based on blinking kinetic [22, 40], spectrum [47], or statistical features [7], reconstruct a super-resolution image [9], de-overlap fibrils through fluorescence lifetime imaging [27], among others [1, 24]. However, there has been limited work focused on distinguishing different dyes from single-frame diffraction-limited fluorescent spots. Therefore, we apply T2C CNN to classify the dyes corresponding to these fluorescent spots.

Here, we aim to leverage the point spread function (PSF) characteristics of emission wavelengths to distinguish multicolor data at the single fluorophore level. In this experiment, we used the same 8-nucleotide docking and 10-nucleotide imager sequences, where the imager was alternately tagged with green (Atto532) and red (Atto647) dyes (as shown in Fig. **8**a). We collected 20,000 frames for each dye, using the binding events from the first 10,000 frames for training

and the last 10,000 frames for testing. To avoid duplicated frames from the same binding site, we only selected the first frame of each site for classification by the model. For comparison, we evaluated several commonly used CNNs on the dataset, including (1) lightweight models: MobileNetV2 [32], SqueezeNet [19], ShuffleNetV2 [25], EfficientNetV2 [39], and (2) standard-weight models: ResNet-18 [48], ResNet-50 [18], VGG16 [36], and DenseNet-121 [17]. We measured the classification accuracy, computational cost, and model sizes for T2C CNN and these commonly used CNNs across five cross-validation experiments. The results are presented in Fig. **8**b.

Our T2C CNN demonstrated superior classification accuracy (92.88%), comparable to larger models like VGG16. This outcome confirms the ability to differentiate multicolor data at the single fluorophore level by analyzing the PSF patterns of emission wavelengths. This finding lays the groundwork for future multicolor microscopy techniques that do not require wavelength-specific analysis. Notably, single-frame fluorophore classification is a relatively easier task compared to classifying binding types from single-fluorophore videos. This difference in difficulty stems from the fact that PSF characteristics differ more prominently across fluorophores with distinct emission wavelengths (Fig. **8**a), whereas in the binding-type classification task, the same fluorophore is used and the emission patterns are only subtly affected by different DNA strand bindings (Main Text Fig. 1). As a result, models like ResNet-18 achieve higher performance in multicolor classification (Fig. **8**) than in the more challenging single-fluorophore task (Main text Fig. 3). Importantly, the T2C CNN achieved this accuracy with a very small computational cost—only 0.02 Giga Floating Point Operations per Second (GFLOPs)—which is even lower than ShuffleNetV2 (0.07 GFLOPs). This efficiency stems from T2C CNN's compact architecture: it has only six convolutional layers, one-third the number in ResNet-18, with the last four being 1×1 convolutions—requiring one-ninth the operations of ResNet-18's 3×3 convolutions. Additionally, its hidden transformations operate on a 64×64 width—1/64 of ResNet-18's 512×512 transformations—and its fully connected layers are only 192×64, approximately 1/40 of ResNet-18's 512×1000 configuration. Consequently, the total computational cost of T2C CNN is less than one-thousandth of ResNet-18, making it deployable on lightweight computing devices.

9 T2C CNN blocks

1. **Temporal2Channel block:** This block integrates the spatiotemporal information of a video by transforming a temporal sequence of grayscale fluorescence images into a channel-rich image representation and applying convolutions with large strides. This design effectively captures long-term spatiotemporal variations in fluorescence videos. The block is defined as follows:

$$l_1 = \mathrm{T2C}(f, \omega_1, \gamma_1, \beta_1), \tag{4}$$

$$l_2 = \mathrm{ReLU} \circ \mathrm{BN} \circ \mathrm{Conv}^{(2\mathrm{D})}(l_1, \omega_2, \gamma_2, \beta_2), \tag{5}$$

where

$$\mathrm{Conv}^{(2\mathrm{D})}(l, \omega)(i, j, d) = \sum_{u=0}^{\mathrm{W_{kernel}}-1} \sum_{v=0}^{\mathrm{H_{kernel}}-1} \sum_{c=0}^{\mathrm{C_{in}}-1} \omega_d(u, v, c) \cdot f(s_w \cdot i + u - p_w, s_h \cdot j + v - p_h, c). \tag{6}$$

Here, l_i, ω_i, γ_i, and β represent the output, weight, and two learnable parameters of batch normalization (BN) for the i-th layer, respectively. The convolutional kernels in these two layers have strides of 2 and 5, respectively, to capture the spatial features of diffraction-limited fluorescent spots over a long time span. They abstract the 512-frame (51.2s) temporal features of the same binding event into 64-dimensional channel features.

2. **Hidden transformation block:** This block transforms the hidden layer features element-wise by pointwise convolution (1×1 kernel) with an equal-dimensional output. These transformations improve feature representation for binding type classification. They are defined as:

$$l_3 = \mathrm{ReLU} \circ \mathrm{BN} \circ \mathrm{Conv}^{(2\mathrm{D})}(l_2, \omega_3, \gamma_3, \beta_3), \tag{7}$$

$$l_4 = \mathrm{Cat}\left[\mathrm{ReLU} \circ \mathrm{BN} \circ \mathrm{Conv}^{(2\mathrm{D})}(l_3, \omega_4, \gamma_4, \beta_4),\ l_2\right]. \tag{8}$$

The output of l_4 is the concatenation (denoted by Cat[·]) of current convolution result with the output from l_2, which is the previous Temporal2Channel block's output. This operation provides multi-scale convolutional features for subsequent feature fusion.

3. **Multi-scale feature fusion block:** This block integrates features before and after transformations and prepares them for the final step of binding type classification. The definitions are as follows:

$$l_5 = \mathrm{ReLU} \circ \mathrm{BN} \circ \mathrm{Conv}^{(2\mathrm{D})}(l_4, \omega_5, \gamma_5, \beta_5), \tag{9}$$

$$l_6 = \mathrm{Flatten}\left(\mathrm{Cat}\left[\mathrm{ReLU} \circ \mathrm{BN} \circ \mathrm{Conv}^{(2\mathrm{D})}(l_6, \omega_5, \gamma_5, \beta_5),\ l_4\right]\right). \tag{10}$$

The function "Flatten(·)" reshapes the input tensor into a one-dimensional vector for each sample, making it compatible with the subsequent fully connected (FC) layer.

4. **Feature classification block:** This block consists of two fully connected (FC) layers: the first refines the features processed by the preceding blocks through fusion, transformation, and abstraction, while the second performs the final classification. They are defined as follows:

$$l_7 = \mathrm{ReLU} \circ \mathrm{BN} \circ \mathrm{FC}(l_6, \omega_7, b_7, \gamma_7, \beta_7). \tag{11}$$

Then, the activated hidden features are passed to the final fully connected layer for classification:

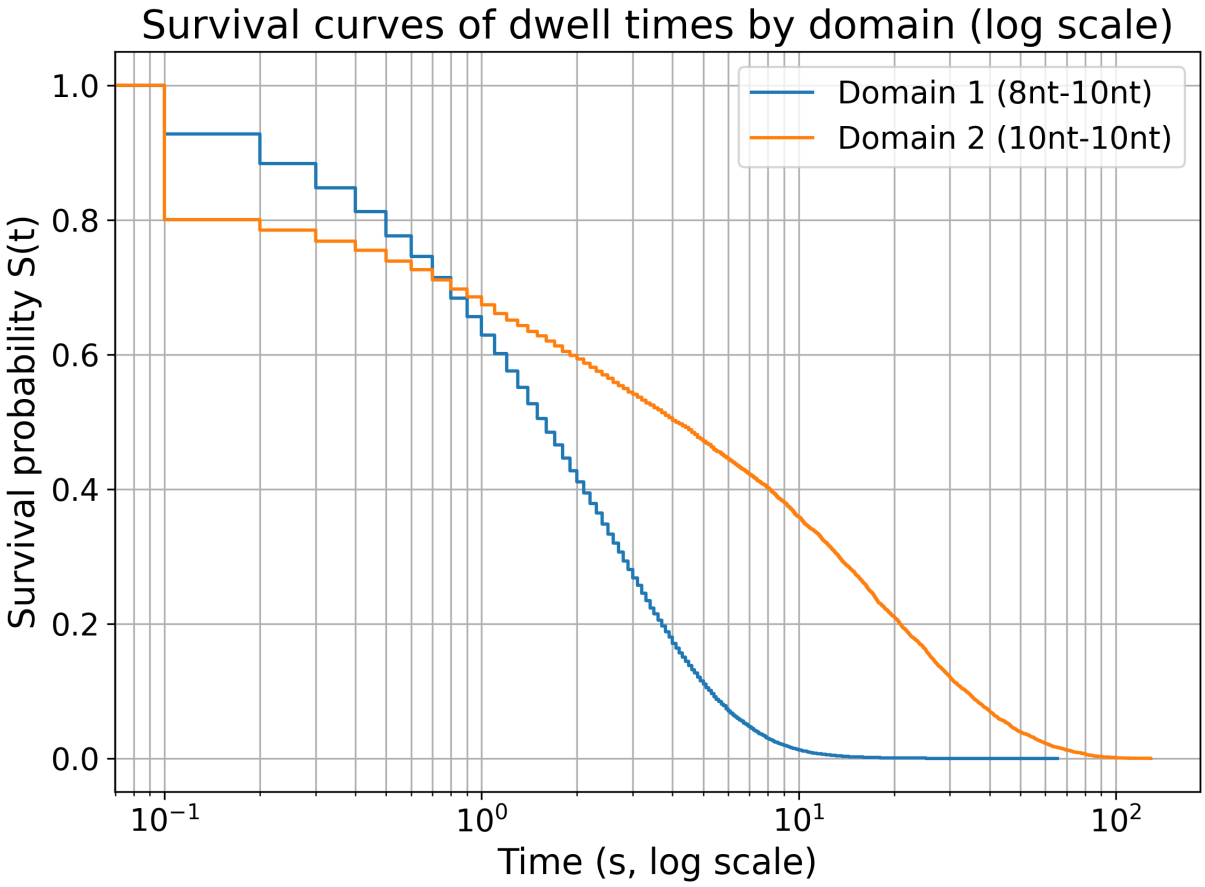
$$\hat{y}_{\mathrm{slice}} = \mathrm{Softmax} \circ \mathrm{FC}(l_7, \omega_8, b_8), \tag{12}$$

$$\hat{y} = \frac{1}{N_{\mathrm{slice}}} \sum_{\mathrm{slice}} \hat{y}_{\mathrm{slice}}. \tag{13}$$

The output $\hat{y}_{\mathrm{slice}}$ from the final layer represents the predicted probabilities for each temporal slice, normalized using the Softmax function [6]. For multiple temporal slices from the same binding event, we take the average ($\hat{y}$) of their predictions.

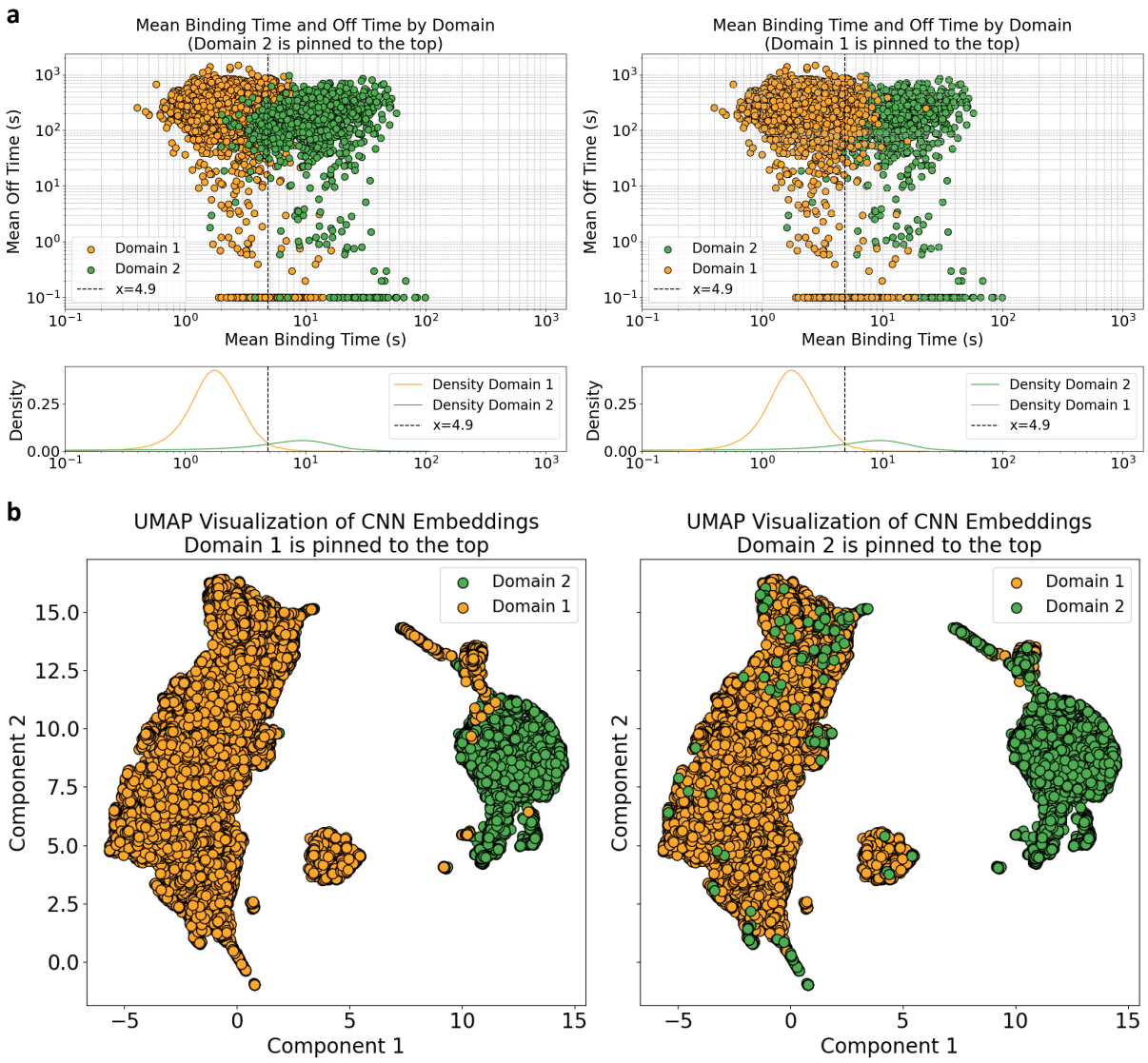
Supplementary Figures

Supplementary Figure 1: Survival curves of dwell times by domain



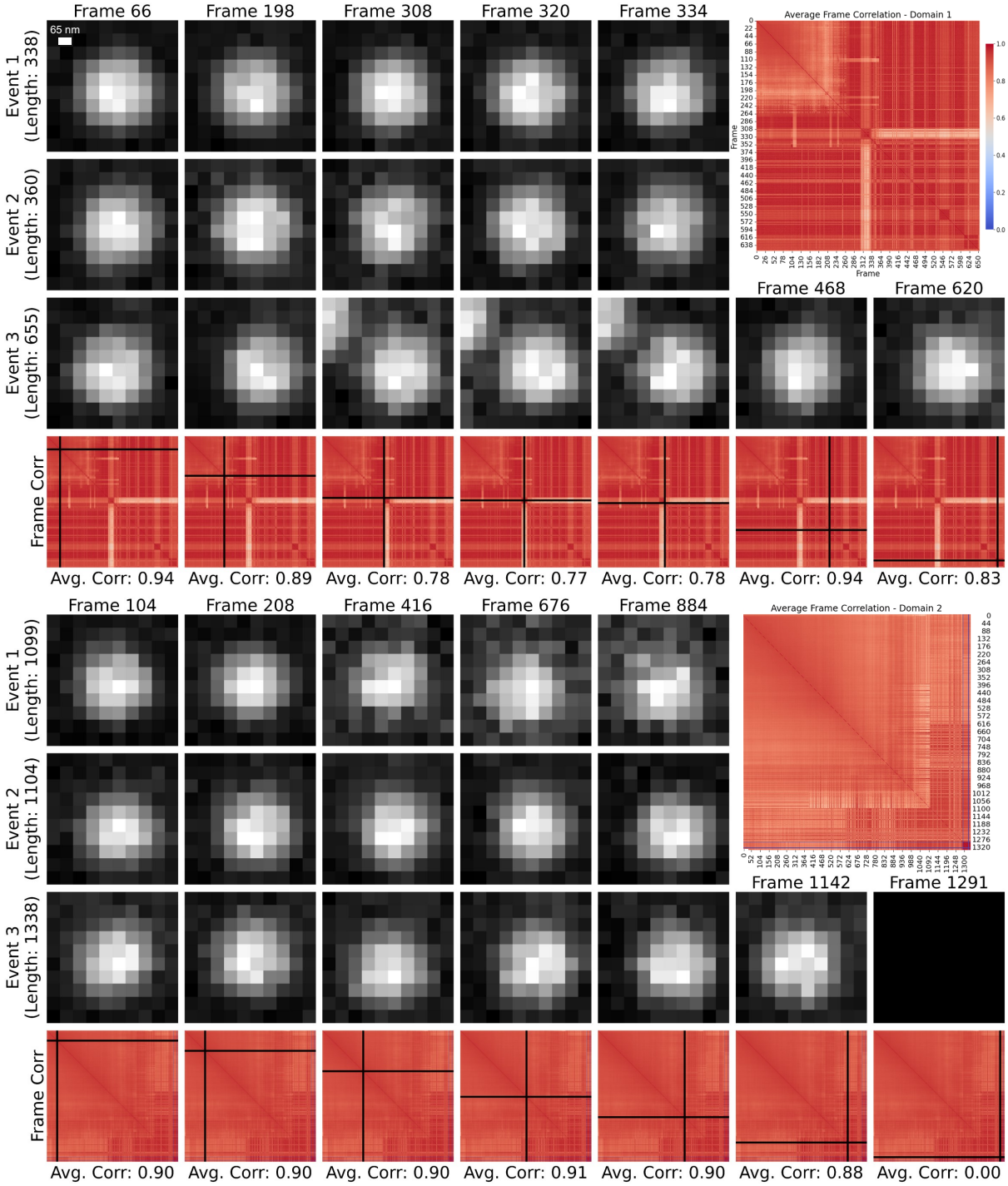
Supplementary Figure 1: **Survival curves of dwell times by domain.** Survival curves were derived from the raw experimental data corresponding to Main Text Figure 1. Source data are provided as a Source Data file.

Supplementary Figure 2: Binding kinetics and CNN embedding visualization



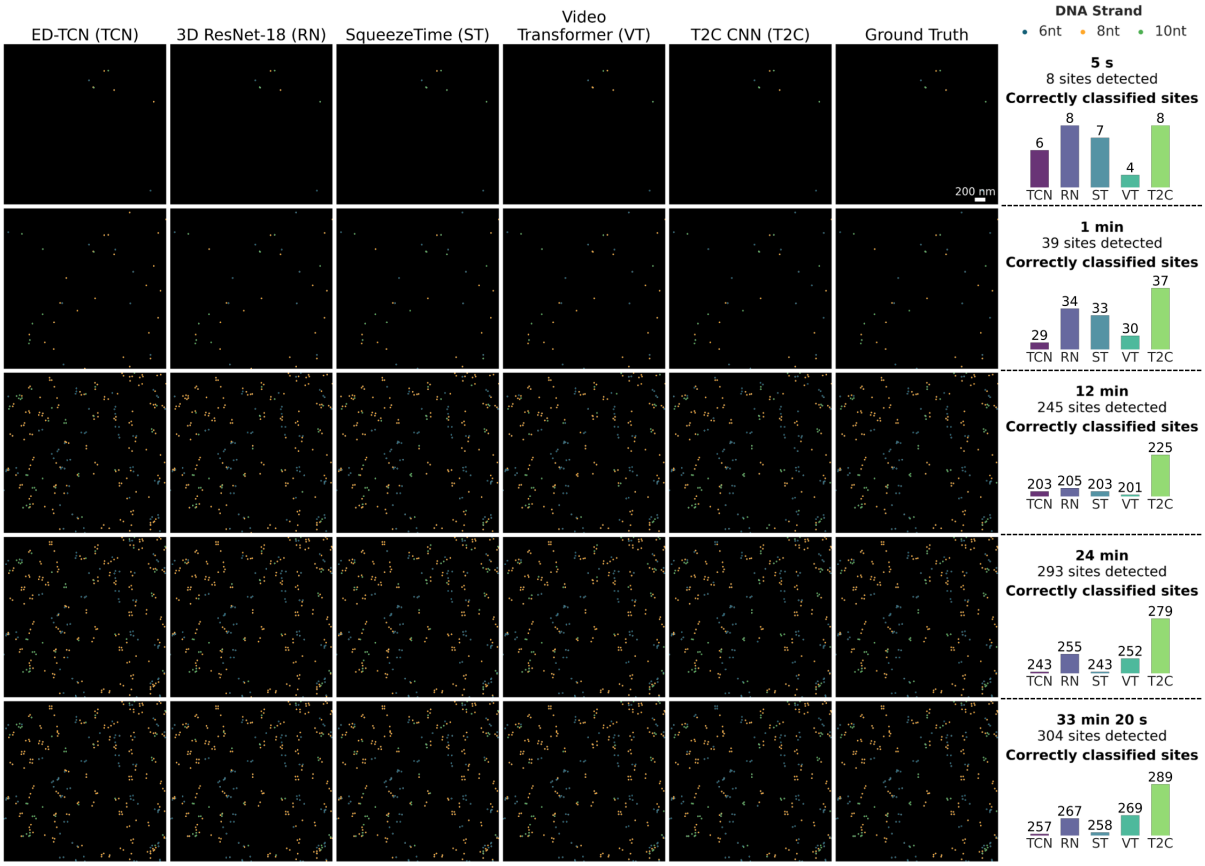
Supplementary Figure 2: **Binding kinetics and CNN embedding visualization.** Scatter plots show the **a) mean binding time and off time** (n=6,160 independent binding sites) and **b) A 2D uniform manifold approximation and projection (UMAP) plot of T2C CNN embeddings** (n=25,530 independent binding events) with different domain types plotted last to ensure visibility. Each subplot highlights one domain by rendering it above the other, enabling clearer inspection of the distribution overlap. Source data are provided as a Source Data file.

Supplementary Figure 3: Abnormal correlation patterns



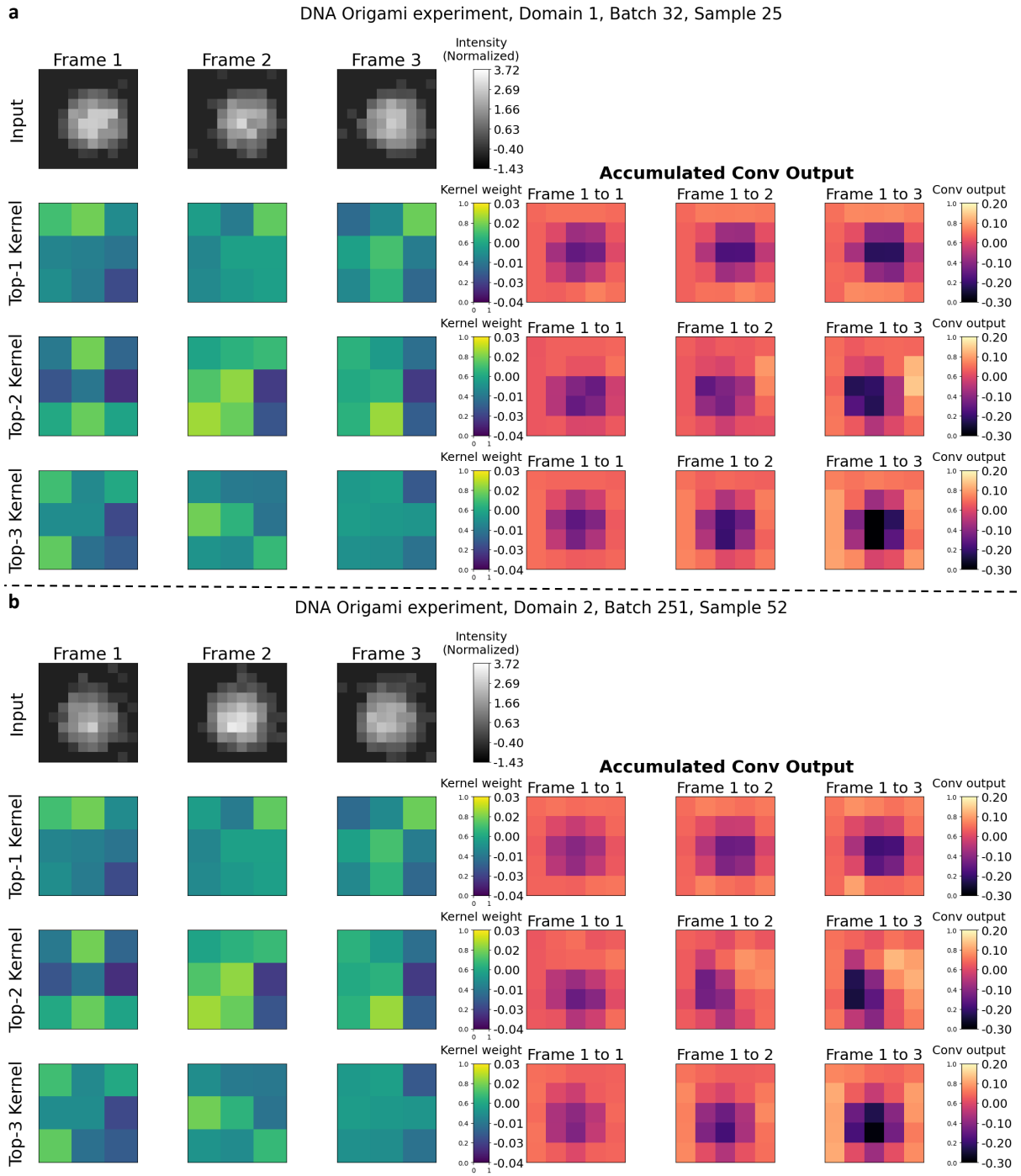
Supplementary Figure 3: **Sampled fluorescence spots from the Top-3 longest binding events in each domain, selected based on typical inter-frame correlation patterns.** Domain 1 is shown above and Domain 2 below, with inter-frame correlation maps displayed in the upper right corner of each domain, highlighting the temporal fluorescence variations associated with binding stability. The same phenomenon was observed in five randomly selected sets of frames. Source data are provided as a Source Data file.

Supplementary Figure 4: Multi-class predictions



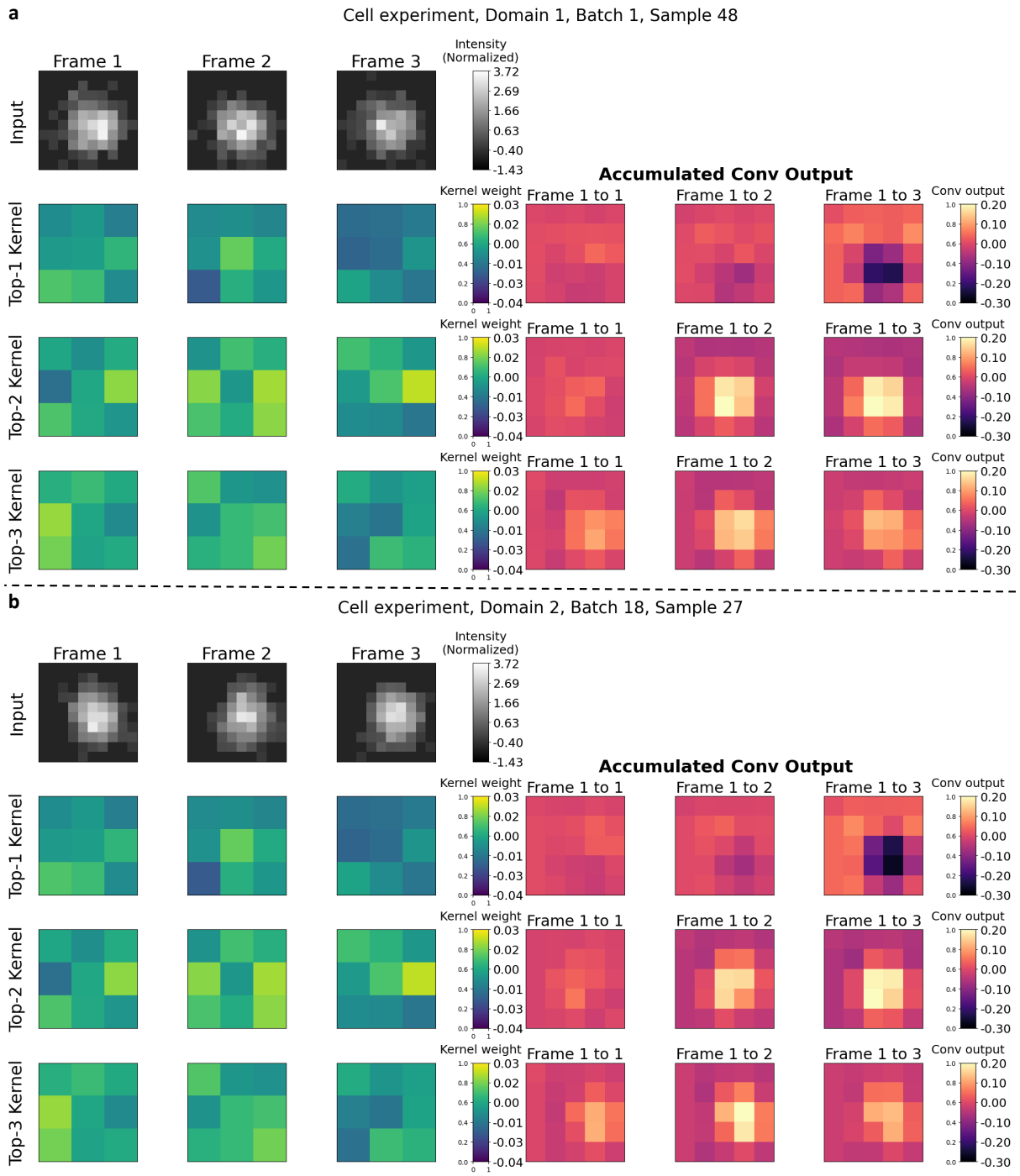
Supplementary Figure 4: **Multi-Class Predictions.** Pseudo-colored super-resolution images reconstructed at 5 s, 1 min, 12 min, 24 min, and 33 min 20 s using image-based binding type classification models. The numbers of detected binding sites and correctly classified sites are shown on the right. Note that due to the higher concentration of 6nt and 8nt, they exhibited an increasing number of observed binding sites over time. Source data are provided as a Source Data file.

Supplementary Figure 5: Example fluorescence spots and temporally accumulated T2C-CNN convolutional outputs corresponding to distinct binding events in DNA origami experiments



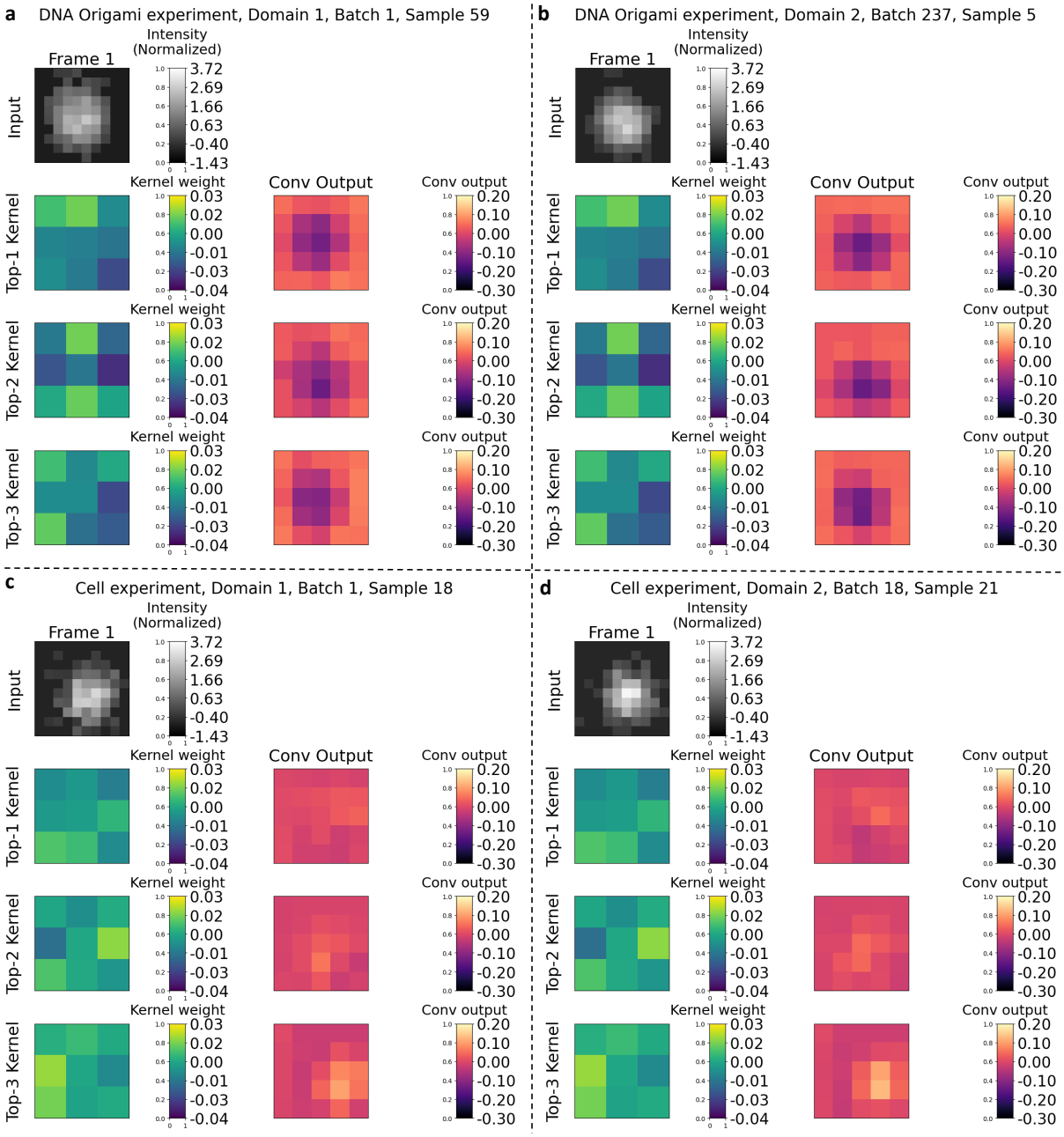
Supplementary Figure 5: Example fluorescence spots and temporally accumulated T2C-CNN convolutional outputs corresponding to distinct binding events in DNA origami experiments. The T2C CNN applies distinct convolutional kernel groups synchronously to each frame within a predefined slice length. Convolutional (Conv) outputs from kernels within the same group are summed across accumulated frames of an event. The top-3 kernel groups were selected globally for visualization, each showing distinct responses to fluorescence spots associated with different domain events. Subfigures **a** and **b** present representative examples from Domain 1 and Domain 2, respectively. The right and left legend labels indicate the original and normalized value ranges, respectively, which are consistent across both domains for comparability. Fluorescence spot intensities were mean–standard deviation normalized within each event. For clearer visualization, a global range truncation was applied to both fluorescence spot intensities and convolutional output values. Similar results were observed in five randomly selected sets of fluorescence spots.

Supplementary Figure 6: Example fluorescence spots and temporally accumulated T2C-CNN convolutional outputs corresponding to distinct binding events in cell experiments



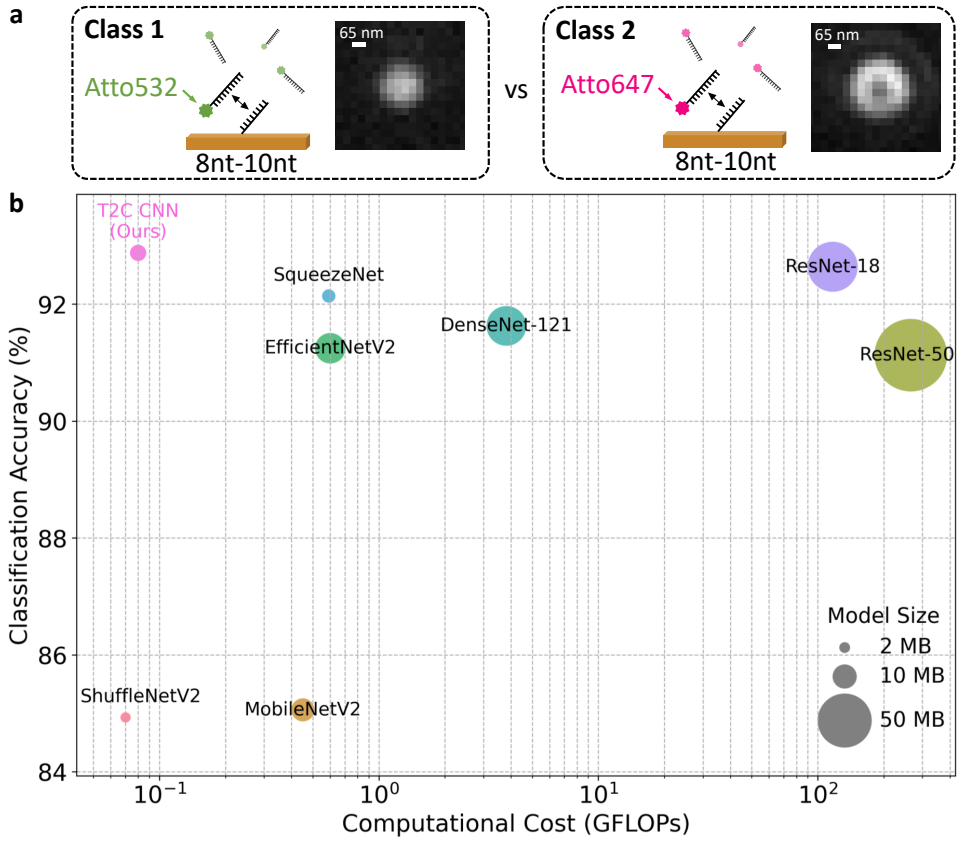
Supplementary Figure 6: **Example fluorescence spots and temporally accumulated T2C-CNN convolutional outputs corresponding to distinct binding events in cell experiments.** The T2C CNN applies distinct convolutional kernel groups synchronously to each frame within a predefined slice length. Convolutional (Conv) outputs from kernels within the same group are summed across accumulated frames of an event. The top-3 kernel groups were selected globally for visualization, each showing distinct responses to fluorescence spots associated with different domain events. Subfigures **a** and **b** present representative examples from Domain 1 and Domain 2, respectively. The right and left legend labels indicate the original and normalized value ranges, respectively, which are consistent across both domains for comparability. Fluorescence spot intensities were mean–standard deviation normalized within each event. For clearer visualization, a global range truncation was applied to both fluorescence spot intensities and convolutional output values. Similar results were observed in five randomly selected sets of fluorescence spots.

Supplementary Figure 7: Example fluorescence spots and T2C CNN convolutional outputs from distinct single-frame binding events



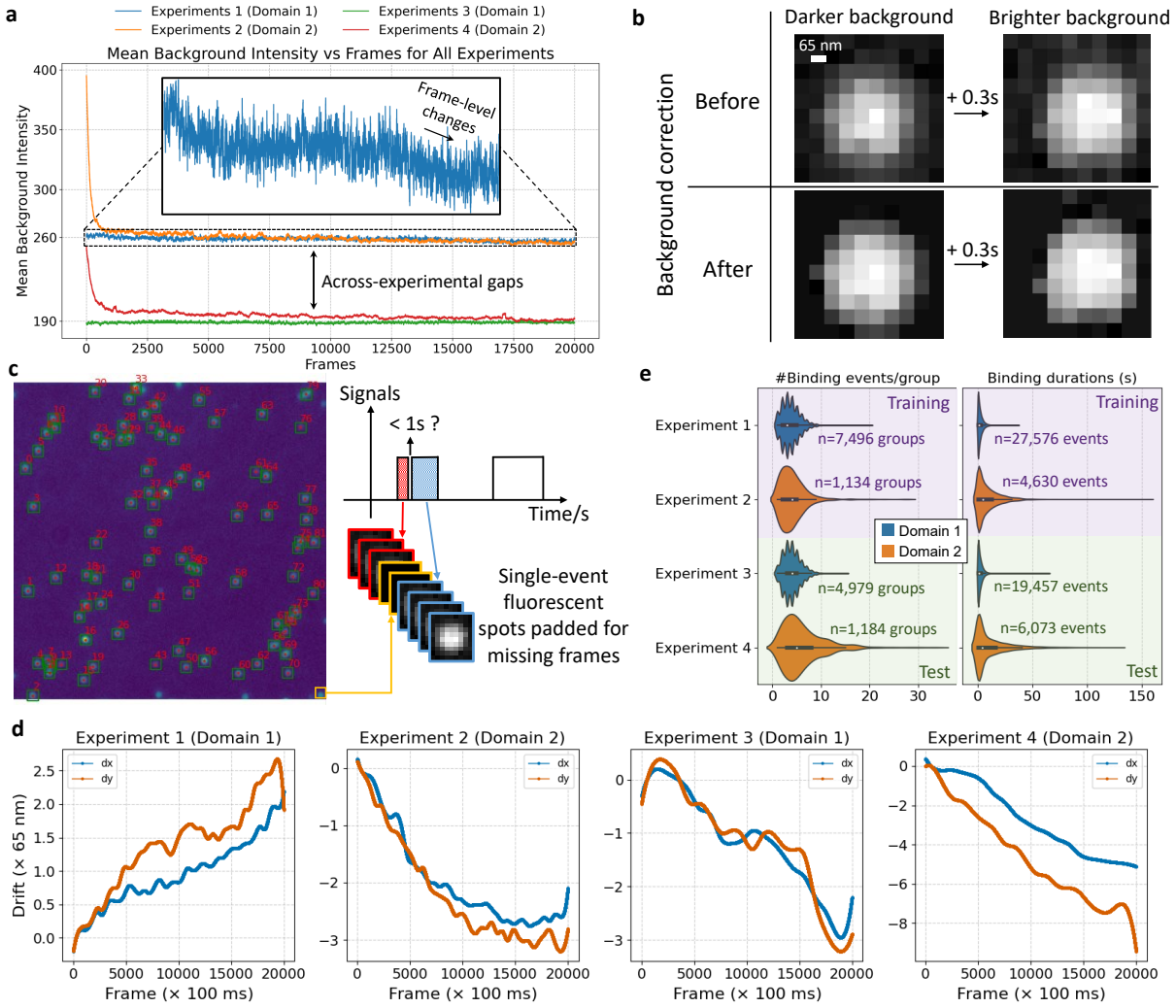
Supplementary Figure 7: **Example fluorescence spots and T2C CNN convolutional outputs from distinct single-frame binding events.** The T2C CNN applies distinct convolutional kernel groups synchronously to each frame within a predefined slice length. Convolutional (Conv) outputs from kernels within the same group are summed across all frames of an event. The top-3 kernel groups were selected globally for visualization, each showing distinct responses to fluorescence spots associated with different domain events. Subfigures **a** and **b** present representative examples from Domain 1 and Domain 2 in DNA origami experiments, respectively, while subfigures **c** and **d** show examples from cell experiments. The right and left legend labels indicate the original and normalized value ranges, respectively, which are consistent across both domains for comparability. Fluorescence spot intensities were mean–standard deviation normalized within each event. For clearer visualization, a global range truncation was applied to both fluorescence spot intensities and convolutional output values. Similar results were observed in five randomly selected sets of fluorescence spots.

Supplementary Figure 8: Single-frame classification of different dye-labeled binding events



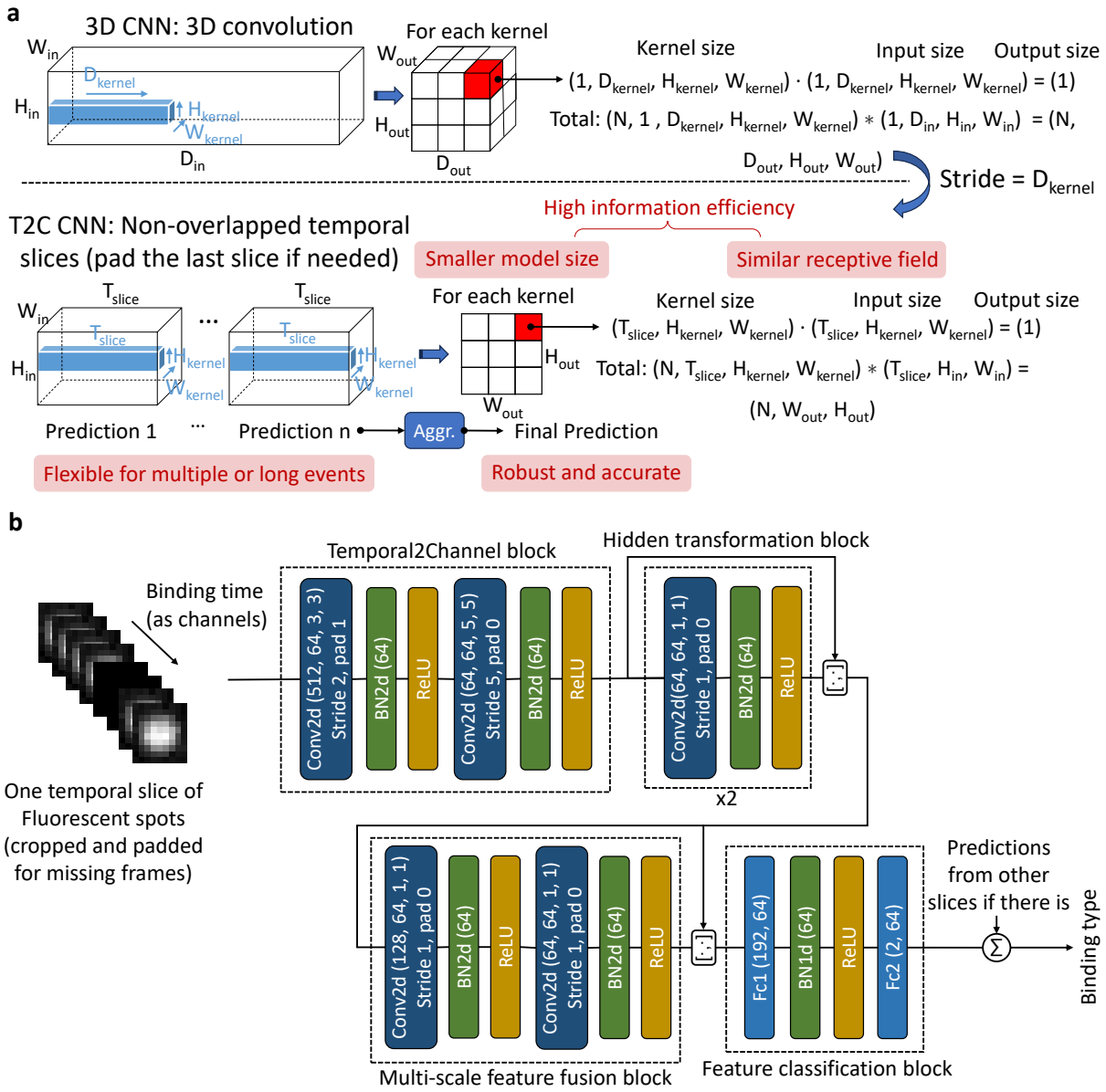
Supplementary Figure 8: **Single-frame classification of different dye-labeled binding events.** **a)** Experimental setup and example images of two dye types: green (Atto532) and red (Atto647), both using the same 8nt-10nt binding. **b)** Comparison of classification accuracy, computational cost, and model sizes for single-frame discrimination of dye-labeled binding events. GFLOPs (Giga Floating Point Operations per Second) quantify the computational power required, indicating the hardware demands for processing. Source data are provided as a Source Data file.

Supplementary Figure 9: Data preprocessing and distribution



Supplementary Figure 9: **Data preprocessing and distribution.** **a)** Background intensity of the microscope videos in each experiment. **b)** An example of background intensity correction. **c)** Illustration of fluorescence blinking correction. **d)** Drift trajectories of the slides read by Picasso [33] in each experiment. **e)** The quantity and duration distribution of binding events in each experiment. Violin plots show the full distribution of the data. The embedded boxplots indicate the 25th–75th percentiles (bounds of the box), the median (center line), and the minima and maxima (whiskers). Source data are provided as a Source Data file.

Supplementary Figure 10: The temporal-to-context convolutional neural network (T2C CNN)



Supplementary Figure 10: The temporal-to-context convolutional neural network (T2C CNN). a) Comparison of T2C CNN with 3D CNN. **b)** T2C CNN architecture. "Aggr.", "$[\cdot, \cdot]$", and "$\sum$" denote aggregation, concatenation, and summation, respectively.

Supplementary Tables

Supplementary Table 1: Classification accuracy on scrambled fluorescence image sequences under varying information constraints

Supplementary Table 1: Classification accuracy on scrambled fluorescence image sequences under varying information constraints. Source data are provided as a Source Data file.

Experiment	Method type[1]	Scramble type	Contained information[2]				Accuracy (%)[3]
			Lengths	Images	Images in event	Image order	
DNA Origami	Length-based	–	✓	–	–	–	83.88±0.21
	Image-based	Random-length	–	✓	–	–	78.98±2.04
		Cross-event	✓	✓	–	–	87.66±4.73
		In-event	✓	✓	✓	–	89.47±1.41
		–	✓	✓	✓	✓	94.76±0.47
HER2 Cell	Length-based	–	✓	–	–	–	66.34±0.37
	Image-based	Random-length	–	✓	–	–	62.64±1.18
		Cross-event	✓	✓	–	–	70.25±3.31
		In-event	✓	✓	✓	–	71.59±1.96
		–	✓	✓	✓	✓	74.09±4.54

[1] We take the proposed T2C CNN and probability density function (PDF) as representatives of image-based and length-based methods, respectively. [2] This indicates the type of information retained or lost in different scrambling strategies. "In-event": Scrambling image sequences within an event destroys temporal order but retains event integrity. "Cross-event": Cross-event scrambling mixes frames from different events, disrupting both temporal order and event-level coherence, while preserving individual event lengths. "Random-length": random-length scrambling groups frames arbitrarily, eliminating both temporal and event-level information. The length-based method uses only the lengths of multiple events as input. [3] For image-based methods, we report 5-fold class-wise accuracy for single-event classification (requires ∼5 s of measurement). For length-based methods, we report 5-fold class-wise accuracy for single-molecule classification (requires ∼10 min of measurement), using all observed events per molecule. Accuracy is reported as mean ± standard deviation across 5 folds.

Supplementary Table 2: Hyperparameter evaluation summary for baseline models

Supplementary Table 2: Hyperparameter evaluation summary for baseline models.

Video Transformer	Depth (width=128)[1]	1	2	**3**[*]	4	5	6	7
	Accuracy[2]	66.74	70.16	**71.91**	69.53	71.61	71.07	69.81
	Width (depth=3)	16	32	48	64	100	**128**	256
	Accuracy	71.01	67.23	70.21	71.18	71.78	**71.91**	70.48
SqueezeTime	Depth (width=512 or 2048)[3]	10	18	34	**50**	101	152	200
	Accuracy	64.20	67.07	67.61	**71.27**	67.87	69.01	66.28
	Width (depth=50)	64	128	204	612	**2048**	3072	4096
	Accuracy	56.01	47.52	55.65	70.38	**71.27**	67.15	63.48
ED-TCN	Depth (width=64)	2	**4**	6	8	10	12	14
	Accuracy	70.78	**73.56**	73.34	57.37	72.58	67.25	71.28
	Width (depth=4)	16	32	48	**64**	100	128	256
	Accuracy	68.51	70.12	68.24	**73.56**	69.24	70.38	69.45
3D ResNet	Depth (width=512 or 2048)	10	**18**	34	50	101	152	200
	Accuracy	74.80	**79.73**	78.35	76.47	73.40	79.15	73.25
	Width (depth=18)	32	64	128	256	**512**	1024	2048
	Accuracy	75.02	78.20	75.02	73.68	**79.73**	76.58	77.61

[1]Model depths and widths were evaluated near their default configurations. [2]Class-wise accuracies for single-event classification were recorded. [3]For 3D ResNet and the ResNet-based method (SqueezeTime), a width of 512 was used when the depth was below 50; otherwise, a width of 2048 was used. [*]Bolded values indicate the optimal hyperparameters identified in the trials.

Supplementary Table 3: Ablation study of T2C CNN components

Supplementary Table 3: Ablation study of T2C CNN components[1]. Source data are provided as a Source Data file.

Settings	Long-term spatial convolutions	Skip concatenations	No pooling	Class-wise accuracy (%)[2]		Overall single-event accuracy (%)[3]
				Single-event	Single-molecule	
Baseline	–	–	–	72.01±2.76	79.87±2.70	72.95±2.33
Individual component	–	–	✓	70.73±1.69	77.60±2.66	74.66±3.59
	–	✓	–	72.58±3.36	78.13±3.96	76.22±3.09
	✓	–	–	84.14±3.80	88.86±4.07	89.07±2.44
Combined components	–	✓	✓	72.51±3.82	79.86±3.19	74.96±4.56
	✓	✓	–	86.33±5.16	91.11±4.21	90.70±3.65
	✓	–	✓	87.34±1.89	92.02±2.38	91.59±1.75
All components	✓	✓	✓	94.76±0.47	96.99±0.46	96.78±0.50

[1]Each component is removed or included to assess its individual and joint contribution to classification accuracy. [2]Class-wise accuracy at the single-event and single-molecule levels. Single-molecule predictions are derived by confidence-based voting over associated single-event predictions. [3]Overall single-event accuracy is calculated across all events without class-wise averaging. Reported values are mean ± standard deviation over five cross-validation trials.

Supplementary Table 4: 5-fold single-event class-wise accuracies of all tested methods and their average ranks

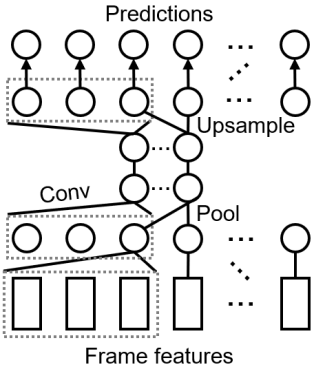
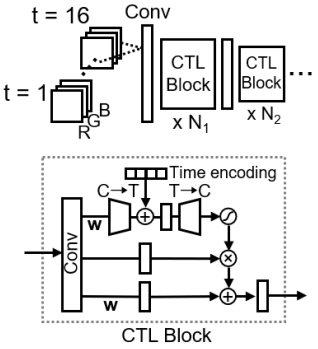
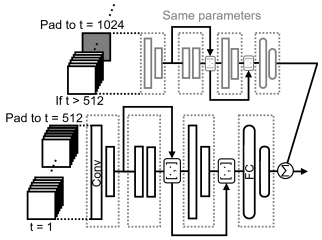
Supplementary Table 4: 5-fold single-event class-wise accuracies of all tested methods and their average ranks.

Methods	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average rank (↑)[1]
T2C CNN (ABC[2])	94.76	94.28	94.36	95.60	94.78	1
T2C CNN (AC)	90.91	86.68	86.85	87.01	85.26	2.8
PDF (All event)	83.88	84.42	84.26	84.42	84.42	4
T2C CNN (AB)	79.35	89.59	86.33	93.97	82.40	4.4
PDF (2 event)	83.80	83.90	83.92	83.90	83.90	5
T2C CNN (A)	81.84	80.70	88.11	89.38	80.69	5.2
PDF (3 event)	82.09	82.18	82.00	82.18	82.18	6.6
3D ResNet-18	80.17	72.58	82.79	80.45	83.72	7.4
PDF (1 event)	75.27	75.27	75.16	75.27	75.27	9.6
T2C CNN (B)	70.22	72.64	69.69	78.98	71.38	12
Video Transformer	74.67	70.85	69.76	72.44	70.85	12.4
T2C CNN (BC)	68.66	68.01	73.26	78.40	74.22	12.4
T2C CNN None	71.61	70.00	69.55	71.61	77.28	12.6
T2C CNN (C)	70.49	69.75	70.35	73.97	69.09	13.2
SqueezeTime	71.72	69.18	71.27	69.55	66.34	13.6
ED-TCN	62.19	63.63	76.01	67.43	71.84	13.8

[1]Ranks were computed within each fold using Friedman ranking and then averaged across folds. These ranks were used in the Friedman test and post-hoc analyses reported in the main text. All 16 methods and their variants were sorted in ascending order based on their average ranks. [2]The letters "A", "B", and "C" following "T2C CNN" indicate the inclusion of long-term spatial convolutions (A), skip concatenations (B), and no-pooling (C) components, respectively. Combinations of these letters represent variants that include the corresponding components. The suffix "None" denotes the variant in which none of these three components are included.

Supplementary Table 5: Comparison with methods most related to the proposed T2C CNN

Supplementary Table 5: Comparison of methods that are most related to the proposed T2C CNN.

Feature	ED-TCN [21]	SqueezeTime [46]	T2C CNN (Ours)
Input	Pre-prepared image features	Short (16 frames), fixed-length video clips	Variable-length video
Channel composition	Stack image features in chronological order	Stack original image channels in chronological order	Stack single image channel in chronological order
Channel aggregation	Aggregated by a hierarchy of temporal convolutions	Aggregated by applying temporal-adaptive weights	Sliced and aggregated by equal-length temporal kernels
Architecture uniqueness	 Uses hierarchical temporal sequence convolutions ("Conv"), pooling, and upsampling on frame features for fine-grained action segmentation or detection.	 Short video clips are processed through hierarchical and iterative channel-to-time ("C → T") and time-to-channel ("T → C") spatial convolutions, along with channel weighting ($\mathbf{w}$) and temporal encoding, to capture sequence temporal dynamics.	 Processes variable-length videos via spatial convolutions on large channelized temporal slices; utilizes four evenly distributed blocks with cross-connected convolutional and fully connected layers to capture subtle spatial frequency variations over a broad temporal domain.
Fitness to fluorescence spot videos	Fragmented associations between frame-level features hinder the capture of long-term variations in fluorescence videos: 1. Feature-based frames: Challenging to perceive the spatiotemporal characteristics of fluorescence videos. 2. Hierarchical temporal convolutions: Low-level detail features are less likely to be retained in high-level representations. 3. Pooling and upsampling: Impede the preservation of subtle fluorescence features and the restoration of detailed fluorescence characteristics.	Repeated conversions between channel-to-time and time-to-channel with weighted convolutions risk losing critical spatiotemporal details: 1. Weighted convolutions: May disrupt the original temporal variation patterns in fluorescence videos. 2. Iterative channel-to-time conversions: In hidden layers, reducing channel dimensions to match time length may not adequately represent temporal sequence information. 3. Multi-level convolutions without cross-links: Pose risks of over-parameterization [28] and gradient vanishing [38] during training.	A small number of cross-connected, long-term spatial convolutions enhance the ability to capture long-term detail variations in fluorescence spot videos: 1. Shallow cross-connections: Help preserve multi-scale spatiotemporal features in fluorescence videos. 2. Long-term spatial convolutions: Facilitate capturing long-term frequency patterns in fluorescence videos. 3. Pooling-free design: Enables all spatiotemporal features to be condensed into channels through large-stride convolutions, maximizing spatiotemporal information integration.

Supplementary References

References

[1] Tim Albrecht, Gregory Slabaugh, Eduardo Alonso, and SM Masudur R Al-Arif. Deep learning for single-molecule science. *Nanotechnology*, 28(42):423001, 2017.

[2] Kazi Saima Banu, Maricarmen Lerma, Sharif Uddin Ahmed, and Jorge L Gardea-Torresdey. Hyperspectral microscopy-applications of hyperspectral imaging techniques in different fields of science: a review of recent advances. *Applied Spectroscopy Reviews*, 59 (7):935–958, 2024.

[3] Mark Bates, Bo Huang, Graham T Dempsey, and Xiaowei Zhuang. Multicolor super-resolution imaging with photo-switchable fluorescent probes. *Science*, 317(5845):1749–1753, 2007.

[4] Mark Bates, Sara A Jones, and Xiaowei Zhuang. Stochastic optical reconstruction microscopy (storm): a method for superresolution fluorescence imaging. *Cold Spring Harbor Protocols*, 2013(6):pdb–top075143, 2013.

[5] Eric Betzig, George H Patterson, Rachid Sougrat, O Wolf Lindwasser, Scott Olenych, Juan S Bonifacino, Michael W Davidson, Jennifer Lippincott-Schwartz, and Harald F Hess. Imaging intracellular fluorescent proteins at nanometer resolution. *science*, 313(5793):1642–1645, 2006.

[6] Christopher M Bishop. Pattern recognition and machine learning. *Springer google schola*, 2:1122–1128, 2006.

[7] Daniela M Borgmann, Sandra Mayr, Helene Polin, Susanne Schaller, Viktoria Dorfer, Lisa Obritzberger, Tanja Endmayr, Christian Gabriel, Stephan M Winkler, and Jaroslaw Jacak. Single molecule fluorescence microscopy and machine learning for rhesus d antigen classification. *Scientific Reports*, 6(1):32317, 2016.

[8] Jerrold T Bushberg and John M Boone. *The essential physics of medical imaging*. Lippincott Williams & Wilkins, Philadelphia, PA, 2011.

[9] Rong Chen, Xiao Tang, Yuxuan Zhao, Zeyu Shen, Meng Zhang, Yusheng Shen, Tiantian Li, Casper Ho Yin Chung, Lijuan Zhang, Ji Wang, et al. Single-frame deep-learning super-resolution microscopy for intracellular dynamics imaging. *Nature Communications*, 14(1): 2854, 2023.

[10] Silvia Colabrese, Marco Castello, Giuseppe Vicidomini, and Alessio Del Bue. Machine learning approach for single molecule localisation microscopy. *Biomedical optics express*, 9 (4):1680–1691, 2018.

[11] Rupsa Datta, Tiffany M Heaster, Joe T Sharick, Amani A Gillette, and Melissa C Skala. Fluorescence lifetime imaging microscopy: fundamentals and advances in instrumentation, analysis, and applications. *Journal of biomedical optics*, 25(7):071203–071203, 2020.

[12] Jozsef Dudas, Linda M Wu, Cory Jung, Glenn H Chapman, Zahava Koren, and Israel Koren. Identification of in-field defect development in digital image sensors. In *Digital Photography III*, volume 6502, pages 319–330. SPIE, 2007.

[13] Pablo A Gómez-García, Erik T Garbacik, Jason J Otterstrom, Maria F Garcia-Parajo, and Melike Lakadamyali. Excitation-multiplexed multicolor superresolution imaging with fm-storm and fm-dna-paint. *Proceedings of the National Academy of Sciences*, 115(51): 12991–12996, 2018.

[14] SA Haider, A Cameron, P Siva, D Lui, MJ Shafiee, A Boroomand, N Haider, and A Wong. Fluorescence microscopy image noise reduction using a stochastically-connected random field model. *Scientific reports*, 6(1):20640, 2016.

[15] Tokuko Haraguchi, Takeshi Shimi, Takako Koujin, Noriyo Hashiguchi, and Yasushi Hiraoka. Spectral imaging fluorescence microscopy. *Genes to Cells*, 7(9):881–887, 2002.

[16] Bo Huang, Mark Bates, and Xiaowei Zhuang. Super-resolution fluorescence microscopy. *Annual review of biochemistry*, 78(1):993–1016, 2009.

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[18] Brett Koonce and Brett Koonce. Resnet 50. *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pages 63–72, 2021.

[19] Brett Koonce and Brett Koonce. Squeezenet. *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pages 73–85, 2021.

[20] Joseph R Lakowicz, Henryk Szmacinski, Kazimierz Nowaczyk, Klaus W Berndt, and Michael Johnson. Fluorescence lifetime imaging. *Analytical biochemistry*, 202(2):316–330, 1992.

[21] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[22] Jieming Li, Leyou Zhang, Alexander Johnson-Buck, and Nils G Walter. Automatic classification and segmentation of single-molecule fluorescence time traces with deep learning. *Nature Communications*, 11(1):5833, 2020.

[23] X Li. Exploring the effect of depth and width of cnn models on binary classification of dogs and cats. *Applied and Computational Engineering*, 47(1):147–158, 2024.

[24] Xiaolong Liu, Yifei Jiang, Yutong Cui, Jinghe Yuan, and Xiaohong Fang. Deep learning in single-molecule imaging and analysis: recent advances and prospects. *Chemical Science*, 13(41):11964–11980, 2022.

[25] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[26] Dragan Maric, Jahandar Jahanipour, Xiaoyang Rebecca Li, Aditi Singh, Aryan Mobiny, Hien Van Nguyen, Andrea Sedlock, Kedar Grama, and Badrinath Roysam. Whole-brain tissue mapping toolkit using large-scale highly multiplexed immunofluorescence imaging and deep neural networks. *Nature communications*, 12(1):1550, 2021.

[27] Fanjie Meng, Janghyun Yoo, and Hoi Sung Chung. Single-molecule fluorescence imaging and deep learning reveal highly heterogeneous aggregation of amyloid-β 42. *Proceedings of the National Academy of Sciences*, 119(12):e2116736119, 2022.

[28] Eshaan Nichani, Adityanarayanan Radhakrishnan, and Caroline Uhler. Do deeper convolutional networks perform better? In *International Conference on Machine Learning*, 2021.

[29] Susanne CM Reinhardt, Luciano A Masullo, Isabelle Baudrexel, Philipp R Steen, Rafal Kowalewski, Alexandra S Eklund, Sebastian Strauss, Eduard M Unterauer, Thomas Schlichthaerle, Maximilian T Strauss, et al. Ångström-resolution fluorescence microscopy. *Nature*, 617(7962):711–716, 2023.

[30] Chandrasekhar Roychoudhuri, Katherine Creath, and A Kracklauer. The nature of light: what is a photon? SPIE, 2005.

[31] Michael J Rust, Mark Bates, and Xiaowei Zhuang. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm). *Nature methods*, 3(10):793–796, 2006.

[32] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[33] Joerg Schnitzbauer, Maximilian T Strauss, Thomas Schlichthaerle, Florian Schueder, and Ralf Jungmann. Super-resolution microscopy with dna-paint. *Nature protocols*, 12(6): 1198–1228, 2017.

[34] Javier Selva, Anders S Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B Moeslund, and Albert Clapés. Video transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12922–12943, 2023.

[35] Paul D Simonson, Eli Rothenberg, and Paul R Selvin. Single-molecule-based super-resolution images in the presence of multiple fluorophores. *Nano letters*, 11(11):5090–5096, 2011.

[36] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.

[37] Shizhao Sun, Wei Chen, Liwei Wang, Xiaoguang Liu, and Tie-Yan Liu. On the depth of deep neural networks: A theoretical view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[38] A Supani, Y Andriani, and H Indarto. Enhancing deeper layers with residual network on cnn architecture: A review. In *Proceedings of the 6th FIRST 2022 International Conference (FIRST-ESCSI 2022)*, volume 14, page 449. Springer Nature, 2023.

[39] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[40] Johannes Thomsen, Magnus Berg Sletfjerding, Simon Bo Jensen, Stefano Stella, Bijoya Paul, Mette Galsgaard Malle, Guillermo Montoya, Troels Christian Petersen, and Nikos S Hatzakis. Deepfret, a software for rapid and automated single-molecule fret data classification using deep learning. *Elife*, 9:e60404, 2020.

[41] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[42] Allison Chia-Yi Wu and Scott A Rifkin. Aro: a machine learning approach to identifying single molecules and estimating classification error in fluorescence microscopy images. *BMC bioinformatics*, 16:1–8, 2015.

[43] Ke Xu, Sang-Hee Shim, and Xiaowei Zhuang. Super-resolution imaging through stochastic switching and localization of single molecules: an overview. *Far-Field Optical Nanoscopy*, pages 27–64, 2015.

[44] Amnon Yariv and Pochi Yeh. Optical waves in crystal propagation and control of laser radiation. 1983.

[45] Laura C Zanetti-Domingues, Christopher J Tynan, Daniel J Rolfe, David T Clarke, and Marisa Martin-Fernandez. Hydrophobic fluorescent probes introduce artifacts into single molecule tracking experiments due to non-specific binding. *PloS one*, 8(9):e74200, 2013.

[46] Yingjie Zhai, Wenshuo Li, Yehui Tang, Xinghao Chen, and Yunhe Wang. No time to waste: Squeeze time into channel for mobile video understanding. *arXiv preprint arXiv:2405.08344*, 2024.

[47] Zheyuan Zhang, Yang Zhang, Leslie Ying, Cheng Sun, and Hao F Zhang. Machine-learning based spectral classification for spectroscopic single-molecule localization microscopy. *Optics letters*, 44(23):5864–5867, 2019.

[48] Yi Zhao, Xinchang Zhang, Weiming Feng, and Jianhui Xu. Deep learning classification by resnet-18 based on the real spectral dataset from multispectral remote sensing images. *Remote Sensing*, 14(19):4883, 2022.

[49] Timo Zimmermann. Spectral imaging and linear unmixing in light microscopy. *Microscopy Techniques: -/-*, pages 245–265, 2005.