Understanding and Predicting Temporal Visual Attention Influenced by Dynamic Highlights in Monitoring Task

Zekun Wu and Anna Maria Feit

***Abstract*—Monitoring interfaces are crucial for dynamic, high-stakes tasks where effective user attention is essential. Visual highlights can guide attention effectively but may also introduce unintended disruptions. To investigate this, we examined how visual highlights affect users' gaze behavior in a drone monitoring task, focusing on when, how long, and how much attention they draw. We found that highlighted areas exhibit distinct temporal characteristics compared to non-highlighted ones, quantified using normalized saliency (NS) metrics. We found that highlights elicited immediate responses, with NS peaking quickly, but this shift came at the cost of reduced search efforts elsewhere, potentially impacting situational awareness. To predict these dynamic changes and support interface design, we developed the Highlight-Informed Saliency Model (HISM), which provides granular predictions of NS over time. These predictions enable evaluations of highlight effectiveness and inform the optimal timing and deployment of highlights in future monitoring interface designs, particularly for time-sensitive tasks.**

***Index Terms*—Dynamic Highlight, Visual Saliency, Visual Attention, Gaze Behavior Analysis;**

I. Introduction

In high-stakes complex monitoring environments like air traffic control and network operations centers, visual highlights are widely used to direct attention, highlight critical information, and enhance task performance [1]–[6]. For instance, dynamic color changes in monitoring dashboards—such as those used in flight displays or drone operations—can effectively draw attention to critical trends, helping users promptly address safety-critical situations [4], [7]. While intuitively effective, the precise influence of such visual highlights on user attention—how quickly they are noticed, how long they hold attention, or how they impact awareness of other interface elements—remains poorly understood and difficult to predict during design.

Currently, deep neural networks are commonly used to predict user visual attention on GUIs [8]–[10]. These models typically use Convolutional Neural Networks (CNNs) to extract features from images and predict pixel-level saliency. However, this pixel-level prediction is insufficient in two respects, when it comes to capturing how dynamic visual highlights drive visual attention in GUIs: firstly, pixel-level saliency cannot directly indicate how much attention a highlighted UI element has drawn, which is especially important when comparing highlighted elements with non-highlighted elements on GUIs; secondly, current models are designed for either constantly changing contexts like video streams [11], [12] or static contexts like GUI screenshots [8], [9] or visualizations [13]. They struggle to capture the temporal characteristics of attention in mixed GUI environments where static graphical elements are combined with dynamic content such as visual highlights, and dynamic changes in the underlying task.

To address this gap, our study focuses on temporal element-level attention prediction: modeling how visual attention evolves in response to dynamic highlights over time. Achieving this requires not only predictive modeling but also a clear understanding of how dynamic visual highlights shape user behavior. We begin by analyzing the impact of visual highlights on user attention in a simulated drone monitoring task, where participants were asked to detect critical events with visual highlights indicating their occurrence. Gaze data collected during the task suggested that highlights lead to immediate and focused attention on the relevant elements. Based on this, we then aimed to predict user attention over time, both at the pixel and element levels. Since existing saliency models lack the capacity to process highlight dynamics explicitly, we developed the Highlight-Informed Saliency Model (HISM), which incorporates spatial and temporal features to predict normalized saliency (NS) for highlighted interface elements. Based on prior work and our observations, the following hypotheses guided our research:

Hypothesis 1 (H1). Visual highlights will lead to a rapid increase in user attention to the highlighted element, improving detection performance but potentially reducing situation awareness (SA) due to less exploration of the overall interface.

Hypothesis 2 (H2). Incorporating temporal features—such as the timing of the highlight—into the saliency model will enable more accurate predictions of when, how long, and how much attention is captured. Additionally, integrating task-specific state information will further enhance model performance, especially when highlights are not present.

In sum, this paper makes the following contributions:

- We examine how visual highlights affect user attention and SA in monitoring interfaces, using NS metrics to quantify both the rapid attraction of attention and its temporal evolution.

Zekun Wu and Anna Maria Feit are with Saarland University, Saarland Informatics Campus, Germany (e-mail: wuzekun@cs.uni-saarland.de; feit@cs.uni-saarland.de).

This project is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 389792660 – TRR 248 – CPEC, see https://perspicuous-computing.science

- We propose HISM, a novel saliency model that integrates temporal and spatial features to predict dynamic attention shifts, outperforming state-of-the-art pixel-level models.

Ultimately, by shedding light on the role of visual highlights and predictive modeling in GUI interaction, we hope to contribute towards the creation of more supportive, user-centric interfaces for monitoring tasks.

II. LITERATURE REVIEW

A. Visual Highlighting in Monitoring Interfaces

Monitoring tasks in dynamic interfaces such as multi-drone control panels require users to manage attention across multiple sources of information. As automation increases and system functionalities become more integrated, the interface itself can become visually dense and cognitively demanding, raising the risk of missed critical events [14], [15]. In such contexts, visual highlights—especially those using color for contrast—have proven effective in directing user attention to relevant elements [16], [17].

Nevertheless, without careful design of the timing, duration, and frequency of such cues, the implementation of visual highlights can mask users' perception of critical information rather than facilitating effective task completion [18]. When multiple visual highlights appear in rapid succession or remain on screen for too long, they can compete for attention and inadvertently obscure other essential information. This challenge is particularly pronounced in multitasking scenarios, where users must divide their attention across multiple interface elements while managing simultaneous tasks [19]. One well-documented instance of this problem is alarm floods, where multiple warning signals appear in close spatial and temporal proximity, overwhelming the user with dense and simultaneous notifications. As a result, users may struggle to prioritize information, leading to cognitive overload and loss of SA [20].

B. Navigating Monitoring Tasks: Situation Awareness and Gaze Behavior

Maintaining SA is fundamental in monitoring tasks: it requires perceiving and comprehending environmental elements, then projecting future states to guide decisions and actions [21]. In multi-drone monitoring, for example, SA includes understanding real-time sensor data as well as potential hazards [22], [23]. SAGAT [24] is often used to measure SA by freezing the task and querying the operator's knowledge, though it traditionally targets air traffic scenarios. Accordingly, we refined it for multi-drone settings where participants must also address critical alerts.

Evidence supports a strong correlation between SA scores and other physiological measures such as eye-tracking [25]. Researchers found that frequent fixations on important events predict higher SA for those events [26]. Furthermore, analysis of eye-tracking metrics reveals that higher fixation counts and longer dwell times on Areas of Interest (AOIs) containing information about environmental hazards correlate with higher SA. Zhang et al. [25] provide a comprehensive review of how different eye-tracking metrics relate to SA, concluding that compared to involuntary physiological responses to environmental stimuli (e.g., blink rate and pupil dilation), conscious eye movements like fixations and saccades exhibit more significant correlations with SA.

C. Predicting Visual Attention on GUIs: Temporal Saliency of Elements

Among the various approaches for predicting users' visual attention, one of the most prominent ones is saliency prediction [10], [27]. Such saliency prediction mostly produces pixel-level maps reflecting how likely each region of an image is to attract attention [8]. However, in GUI contexts, pixel-level saliency may not accurately represent user attention on structured elements such as buttons, text, or icons. User attention is shaped not only by visual features like size, shape, and color, but also by the semantic relevance of these elements within the task [13], [28]. When highlights are applied to specific elements, it is crucial to evaluate the entire targeted area and understand its relative saliency compared to the overall interface.

On the other hand, as human visual attention evolves over time [29], there is a growing interest in predicting the saliency at different viewing times [13], [30]. While these approaches have advanced our understanding of temporal visual attention, they are not effective in capturing the influence of dynamic highlights in static GUIs. This mixed visual stimulus demands a new saliency model that can fully utilize both the temporal information, such as when the highlight occurs, and the spatial information in the GUIs. In response to this gap, our work proposes a new saliency model with two branches to integrate the temporal information of the visual cue together with the spatial information from the GUIs.

III. METHOD

A. Data Collection

To investigate how visual highlights impact user attention and SA in a dynamic monitoring task, we conducted a controlled user study using a simulated multi-drone monitoring interface (see Figure 1). Participants were tasked with continuously monitoring multiple drones, interpreting system alerts, and responding to critical situations in real-time. Throughout the experiment, eye-tracking data were recorded to analyze visual attention shifts and temporal gaze patterns in response to visual highlights.

1) Monitoring Interface: Informed by previous research [22], [31], [32] and resembling existing multi-drone monitoring interfaces [22], [33], [34], while being simple enough to be used in controlled lab studies and for collecting reliable gaze that, and easy to understand without prior experience in flying drones. As illustrated in Figure 1, the chosen interface combines icon-based elements, facilitating quick assimilation of drone parameters, with a map display to increase spatial awareness and task immersion. Each drone block features eight elements, showcasing core drone metrics along with a representative image (for simplicity, we refer to the combination of both the image and textual or numeric data as *icon* in the following). These were chosen to cover

different categories of data found to be relevant for drone monitoring [22]:

1) *Safety Level and Alert* Battery level, wind speed, rotor condition, and no-fly-zone warnings determine the drone's current safety status. These values indicate potential critical situations.
2) *Telemetry Data*: Horizontal speed, altitude, and distance to the destination are data collected in real-time by sensors of the drone, providing insights into the drone's current operational mode and task progression.
3) *Weather Data*: An icon about current weather conditions was incorporated to provide real-time environmental data, which might impact the drone's safety and task success.

The icons visually symbolize these core metrics, which makes it easy to interpret the corresponding sensor values displayed below. Throughout the study tasks described below, the icons retained a static visual representation. Only the underlying values are updated according to the simulated state of the drone. This ensured visual consistency across the study conditions, which is important to isolate the effect of visual highlighting on users' gaze behavior.

As shown in Figure 1, our design used a yellow background highlight to signal critical situations. After experimenting with various sizes and intensities, we determined that this struck the right balance between visibility and distraction prevention. We decided to focus on one established highlighting technique to thoroughly investigate the effect of visual cues on the users' task-driven attention during a monitoring task and to collect enough training data to develop deep learning-based visual saliency models that could predict this effect.

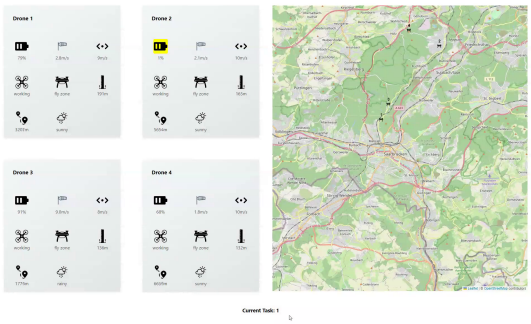


Fig. 1: Drone monitoring interface highlighting a low-battery critical situation. (A non-highlighted critical situation interface appears the same, but without the yellow mark.)

2) Task Design: Each participant was tasked with completing four drone monitoring tasks, as indicated by a 'current scene number' on the screen, each lasting 5 minutes. Participants were instructed to focus on two tasks: (1) identify critical situations and pressing the space bar to acknowledge their detection, and (2) at any time be aware of the state of all drones and correctly answer questions about icon values at random intervals. The design of each of these tasks is further described in the following.

Detecting Critical Situations: Participants were asked to identify all critical situations by observing any changes in the indicator values for each drone block. Upon detecting a critical situation, they should press the space bar to acknowledge detection. We separated each monitoring task into twenty 15-second intervals. These were initially randomized with respect to two aspects: (1) the occurrence of a critical situation (80% probability) and (2) highlighting of the corresponding critical icon (50% probability across critical situations). As a result, during the study, participants encountered 64 critical situations (each lasting for 15s), where one of the four safety level icons (battery level, wind speed, rotor condition, and fly-zone warning) transitioned into a critical range, as defined in Table I. In 33 of these cases, the corresponding icon was highlighted as described above.

TABLE I: Critical Situations and their Indicator Values

Critical Situation	Indicator Value	Icon
Low Battery	$<$10%	
Extreme Wind	$>$10 m/s	
Rotor Off	Off	
No-fly Zone Warning	No-fly	

Situation Awareness: In addition to monitoring the drones for critical situations, participants were instructed to stay aware of the state of all icons across all drones, as well as the ones not related to critical situations (horizontal speed, altitude, distance to target, and weather condition). To assess participants' SA for these icons, our task design incorporated a SAGAT-based questionnaire whose design followed the recommendations provided by Endsley [35]. It appeared at randomized intervals, ranging from 30 to 60 seconds (but kept the same across participants). As a result, there was a total of 20 questionnaires during the whole experiment, with 2 appearing during a non-critical situation, 13 during a non-highlighted critical situation, and 5 during a highlighted critical situation. Each questionnaire consisted of two questions, focusing on two icon values. The design shown in Figure 2 followed prior recommendations [35] to ensure that the SA assessment was short and as unintrusive to the monitoring task.

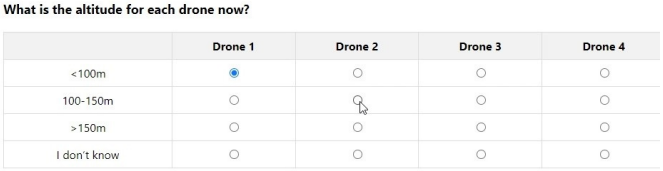


Fig. 2: Situation awareness question screenshot

3) Procedure: Each participant began the experiment with a detailed introductory video explaining the user interface and tasks. We used Tobii Pro Eye Tracker Manager's five-point calibration. Additionally, we introduced four check pages—one per task—to confirm tracking accuracy and filter out poor-quality data (details in supplemental material). After calibration, participants completed a brief 90-second practice

task (including all four critical situations and SA questions) to confirm their understanding of both the detection and question-answering procedures. They then undertook four tasks as detailed in the task design session. Breaks were allotted between each of them, during which the eye-tracker was recalibrated to maintain tracking quality. Upon completing the last task, participants were asked to fill out a post-study questionnaire to gather their demographic data, feedback, and perceptions of the study (see supplementary material). At the end, participants were paid 15€ as compensation for their time. The study procedure and task were approved by the university's ethics committee.

4) Participants: We recruited a total of 28 participants (5 female, 23 male) from Saarland University, aged between 20 and 37 years (median age = 26). Most participants had no prior experience flying drones, with only five indicating intermediate to advanced experience. The sample size was determined based on two considerations. First, an a priori power analysis using G*Power indicated that at least 27 participants were required to detect a medium effect size ($d = 0.5$) with 80% power at a 0.05 significance level using a two-tailed paired-sample t-test. Second, we performed a reliability check by repeatedly partitioning the participants into equal groups and computing the correlation between their fixation maps to ensure stability and representativeness of the gaze data. Further details of this validation procedure are provided in the supplementary material.

5) Hardware Setup: The experiment was conducted using a 24-inch desktop monitor (52 × 32.5 cm, 1920 × 1200 px resolution), with participants seated at a fixed viewing distance of 60 cm. This setup yielded visual angles of approximately 45.4° horizontally and 30.5° vertically. Each pixel corresponded to roughly 0.0271 cm, making the 25-pixel dispersion threshold used for fixation detection equivalent to a visual angle of approximately 0.65°. To capture high-fidelity gaze data, we used a Tobii Pro Fusion eye tracker operating at 250 Hz, positioned below the screen and angled upward to optimize accuracy. Calibration was performed using the Tobii Pro Eye Tracker Manager.

B. Data Analysis

1) Data filtering and preprocessing: Each gaze point was captured as an (x,y) screen coordinate, together with its corresponding timestamp. From the raw gaze points, fixations were identified based on specific criteria: low dispersion (25 px) and adequate duration (50 ms), using the fixation detection function from the PyGaze package [36]. We ensured data quality in two steps: first, we verified dataset comprehensiveness using correlation-coefficient checks; second, we excluded intervals with poor accuracy in gaze offsets. Full details appear in the supplementary materials.

For the subsequent analysis, gaze data were segmented into 15-second task intervals corresponding to the duration of potential critical situations, as described in subsection III-A. Additionally, the timestamps of participants' button presses, which indicated the detection of critical situations, were aligned with the gaze data timestamps. The dataset was further complemented with participants' responses to the SA questions.

2) Gaze and Performance Metrics: To assess the impact of visual highlighting on detection performance and SA, we analyze gaze behavior through two primary categories: engagement and exploration. Engagement metrics capture how users direct attention toward specific elements, while exploration metrics describe how users scan the interface and shift focus across different areas.

Engagement is measured by tracking *fixation count*, *fixation duration*, and *revisits*, which provide insights into how users allocate attention to critical areas such as icons indicating system alerts or SA-related queries. In contrast, exploration behavior is characterized by *mean saccade amplitude*, *scanpath length per second*, and *AOI transition rate*, which reflect broader search patterns and visual scanning strategies across the interface.

To further analyze visual attention distribution, we generated continuous saliency maps by synchronizing the collected gaze data with the UI's frame rate and extracting fixation details as specified in subsubsection III-B1. Fixation points from all participants were aggregated and processed using a Gaussian kernel, a commonly used technique in prior research, before normalizing the resultant saliency map within a [0,1] range.

3) Statistical Analysis: The analysis begins with a static approach, where gaze and performance metrics are aggregated over the entire 15-second task interval to assess the overall effects of highlighting. This is followed by a temporal approach, which examines how the influence of highlights evolves over time, providing deeper insights into the dynamic interaction between visual cues and user behavior.

For each analysis, statistical hypothesis testing was conducted after assessing normality using the Shapiro-Wilk test. For normally distributed continuous variables, we applied an independent two-sample t-test, reporting the test statistic t, degrees of freedom (df), and p-value (p). For non-normal or non-continuous variables, we used the Mann-Whitney U test (Wilcoxon rank-sum test), reporting the test statistic U, sample sizes n_1 and n_2, and the p-value p.We considered $p < 0.05$ as significant. All analyses were conducted at the participant level, with each data point representing the average of all trials per participant.

IV. Empirical Results

A. Influence of Highlighting on Detecting Critical Situations

TABLE II: Summary of Detection Performance (Mean and SD)

Detection Result	**Critical**	**Non-Critical**
Identified as critical (%)	85.68(1.77)	2.71(0.30)
Identified as non-critical (%)	14.32(1.77)	97.29(0.30)

Table II presents an overview of detection performance across participants. Specifically, a trial was labeled "non-critical" if the participant did not press the space bar before

the end of the 15-second interval.. Most critical situations were correctly identified (85.68%), while false alarms were minimal (2.71%). Given the negligible rates of false alarms, in the following, we focus only on the analysis of correctly detected critical situations.

TABLE III: Hit Rates and Response Time in Different Highlight Conditions(Mean and SD)

Metric	Highlight	No Highlight
Hit Rate (%)	96.18(0.40)	74.76(0.84)
Response Time (s)	1.33(0.31)	5.16(0.79)

As indicated in the first row of the Table III, participants detected significantly more critical situations when highlights were present (*U(28,28)* = 819.0, $p < .001$). Additionally, the second row of Table III shows that response times were significantly shorter with highlights, confirming that visual cues facilitate faster detection (*U(28,28)* = 0.0, $p < .001$). Together, these results confirm that visual highlighting enhances both accuracy and speed in detecting critical situations, aligning with results from previous findings [16], [37].

To better understand how highlights drive these improvements, we analyzed the gaze of the participants when engaging with the targeted icon in the case of a critical situation. Each AOI, as illustrated in Figure 3, encompassed the icon and its associated text, spanning $142\text{px} \times 128\text{px}$. We calculated the three engagement metrics – fixation count, fixation duration, and revisits – on the AOI related to a critical situation for those task intervals where participants successfully detected the critical situation.

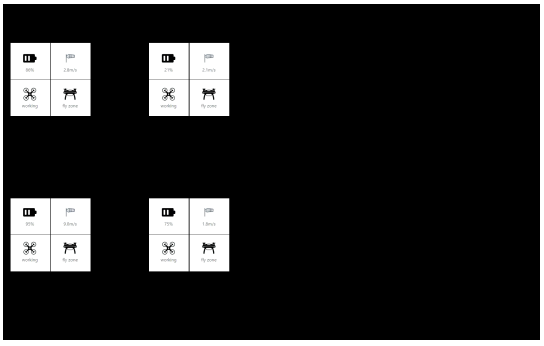


Fig. 3: AOIs related to the critical situation.

TABLE IV: Verification Gaze Metrics in Highlight and No-Highlight Conditions (Mean and SD)

Gaze Metric	Highlight	No Highlight
Fixation Count	6.37(3.76)	8.69(4.79)
Fixation Duration (s)	5.36(1.58)	8.02(2.58)
Revisits	2.18(0.52)	2.73(0.62)

Table IV shows that there were fewer fixations in highlighted trials compared to non-highlighted trials (*U(28,28)* = 519.0, $p < .01$). Fixation duration was also shorter with highlight ($t(27) = 4.58$, $p < .001$). Additionally, there were fewer revisits when highlights were present (*U(28,28)* = 580.5, $p < .001$).

These results suggest that visual highlights enable participants to notice critical situations more quickly, reducing the time required to process the information. However, they also lead to shorter fixation durations and fewer revisits, indicating that users may engage with the highlighted information less thoroughly.

B. Influence of Highlighting on Situation Awareness

TABLE V: Situation Awareness and Gaze Metrics under Different Highlight Conditions (Mean and SD)

Metric	Highlight	No Highlight
SA Score	0.49 (0.20)	0.51 (0.22)
Fixation Count	1.27 (1.32)	2.62 (1.96)
Fixation Duration (s)	0.10 (0.06)	0.19 (0.11)
Revisits	0.11 (0.27)	0.70 (0.36)
Mean Saccade Amplitude (px)	98.97 (37.33)	109.66 (41.36)
Scanpath Length per Second (px/s)	892.64 (336.73)	1028.71 (390.19)
AOI Transition Rate (/s)	2.54 (1.06)	3.08 (1.22)

Similarly, we analyze the SA and the relevant gaze behavior across the entire 15s task trial when critical situations happen. Our analysis revealed that SA scores were slightly lower in the highlight condition, compared to the no-highlight condition, as shown in the Table V, although this difference was not statistically significant. However, we found significantly different gaze behavior when calculating the gaze behavior metrics critical for maintaining SA. The participants' engagement with the SA-queried icons was significantly higher in the no-highlight condition. Specifically, fixation count, fixation duration, and revisits were significantly higher in the no-highlight condition (*U(28,28)* = 697.0, $p < .001$; $t(27) = 4.11$, $p < .001$; *U(28,28)* = 723.5, $p < .001$, respectively).

Additionally, exploration behaviors—characterized by broader search patterns—were reflected in significantly higher scanpath length per second (*U(28,28)* = 607.0, $p < .05$) and AOI transition rate (*U(28,28)* = 616.0, $p < .05$) in the no-highlight condition. These increases indicate more extensive visual scanning across the interface. However, mean saccade amplitude, another indicator of search breadth, did not show a significant difference between conditions (*U(28,28)* = 525.0, $p = 0.271$).

These findings indicate that while highlights accelerate the detection of critical icons, they do not necessarily increase engagement with other elements. Reduced fixations, durations, and revisits point to a potential narrowing of attention at the expense of broader SA. To explore this trade-off over time, we analyze saliency maps to see whether highlights broaden or tunnel user attention across the interface.

C. Temporal Visual Attention Analysis

To analyze the participants' temporal visual attention changes, we generated saliency maps for the interface frames and visualized them at specific timestamps. As shown at the

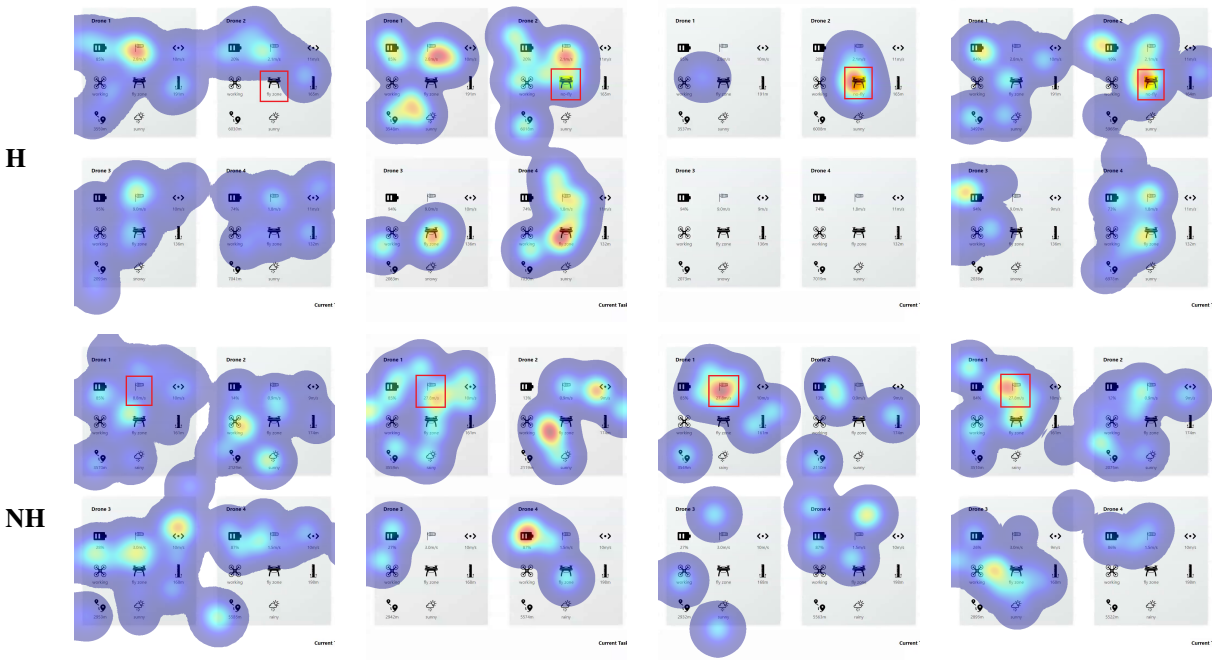


Fig. 4: Shifts in attention during a critical situation (CS) with (Top) and without (Bottom) highlights. Icons indicating a critical situation are marked in red (not visible during the study).

top of Figure 4, the saliency maps reveal a clear tunneling effect caused by the visual highlight, which directs nearly all user attention to the highlighted icon shortly after its appearance. This effect is particularly evident in the initial moments (T1 and T2), where the highlight effectively draws attention to the critical situation icon. However, this focused attention comes at the cost of reduced engagement with other parts of the interface. As the participant identifies the critical situation, their attention begins to shift back toward the broader interface, gradually recovering the perception of the global picture. In contrast, in the no-highlight condition (bottom of Figure 4), this effect is less pronounced. Without the visual cue, attention is initially more distributed, and participants take longer to focus on the critical situation. As a result, their gaze behavior remains more varied over time, suggesting a broader search strategy that may help maintain awareness of other elements on the interface. This contrast further underscores the trade-off of visual highlights: while they facilitate rapid detection, they may temporarily suppress exploration and reduce engagement with non-highlighted areas.

Building on these observations, we analyzed the normalized saliency (NS) of a critical icon over time [13], [28]. The NS of a visual element is defined as the saliency of that element relative to the saliency of all elements within the GUI. The element saliency can be computed by integrating the pixel-level saliency density over the area covered by an element. Formally, we define the normalized saliency (NS) of an element e_k at time slice t_i as below:

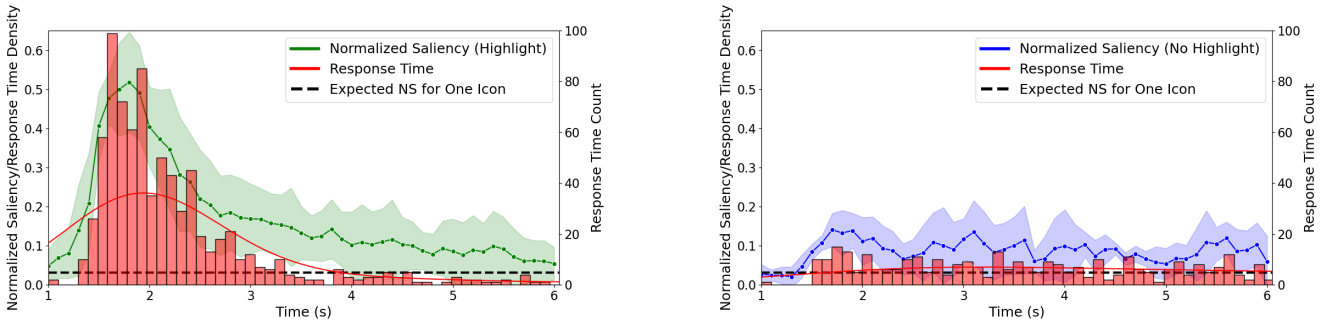
$$NS(e_k, t_i) = \frac{S(e_k, t_i)}{\sum_{j=1}^{n} S(e_j, t_i)} \quad (1)$$

where $S(e_k, t_i)$ is the saliency of element e_k at timestamp t_i. This allows us to track the saliency of the potentially highlighted element (e_h) over the entire temporal trajectory $T = \{t_1, ..., t_m\}$.

Having defined NS as a measure of relative attention distribution, we now use it to quantify the saliency dynamics of the critical situation icon against all other icons over time. Specifically, we compute NS at 0.1-second intervals, enabling precise, time-resolved analysis of how attention dynamically shifts in response to the critical situation.

Figure 5a illustrates the NS trends for the icon with highlight during the first five seconds following the onset of a critical situation. We observe a rapid surge in NS within the first second, peaking at values around 0.5. This sharp increase corresponds with the intense visual attention drawn by the highlight, as depicted in the heatmap at T1 in Figure Figure 4. However, this peak in visual attention is only temporary; the NS begins to decline shortly after reaching the peak, specifically after 0.6 to 1 second. This pattern of NS change aligns closely with the distribution of response times across these intervals, which also shows a rapid increase followed by a decrease. The average response time (M = 1.33 seconds) closely follows the NS peak, indicating that NS is an effective indicator of the timing at which users respond to highlighted information.

By contrast, in the no-highlight condition, NS remains low and stable throughout the first five seconds, as shown in blue in Figure 5b. Unlike the sharp peak observed with highlights, NS only slightly increases to 0.1 following the onset of the critical situation, indicating a more gradual detection process. Across the five-second window, participants had a consistent



(a) Highlight (b) No Highlight

Fig. 5: Comparison of NS Changes Over Time (a) with Highlight. (b) without Highlight.

probability of detecting the situation at any given timestamp, without a dominant moment of focus. Notably, by the final 0.5 seconds, NS values in the highlight condition (Figure 5a) rapidly diminish, approaching those in the no-highlight condition (Figure 5b). This suggests that after the initial highlight-driven attention spike, users gradually redistribute their gaze, bringing NS closer to the expected random distribution level of 1/32.

We calculated the NS for the SA queried icons over a six-second range, including one second before the critical situation began. As shown in Figure 6, the NS of the critical situation icon (green) is already slightly higher than that of the SA queried icons (gray) in the 0 to 1 second range, indicating that participants allocated more attention to the SA-related icons when no highlight was present. Following the onset of the critical situation (1 to 6 seconds), there is a clear shift in attention from the SA icons to the highlighted icon. This is evident as the NS of the SA icons decreases while the NS of the critical situation icon surges, before gradually returning to its previous level as the NS of the critical situation icon declines. Statistical analysis confirmed a significant negative correlation between the NS values of the critical situation and SA-queried icons ($r = -0.24$, $p < .001$), reinforcing the observed shift in visual attention.

We conclude that the NS metric is a robust and precise indicator of user attention on specific targets relative to other AOIs. Compared to gaze metrics such as AOI statistics or frame-level saliency maps, the NS metric captures the relative visual saliency of the highlighted element in comparison to all other elements on the interface. It provides a single, interpretable value that reflects the proportion of user attention allocated to the highlight relative to the rest of the interface. It measures how quickly, how much, and for how long visual highlights attract attention, and reveals how these shifts in attention influence detection responses and overall SA, making it a more effective tool for temporal visual attention analysis.

V. Predicting Temporal Attention with Dynamic Highlighting

In this section, we explore whether we are able to predict these empirical observations regarding the impact of visual highlights on the temporal visual attention of users. We began

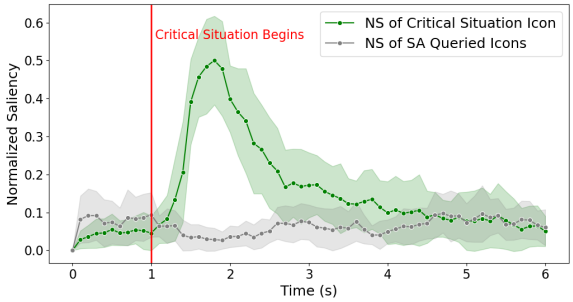


Fig. 6: Comparative NS changes for the critical situation icon and the SA queried icons, highlighting the shift in user attention.

by applying existing saliency models, including the bottom-up ITTI saliency map model [38], and data-driven approaches [11], [39], to estimate pixel-level attention. We then shifted to a temporal, element-level prediction task, aiming to directly model the normalized saliency (NS) of the highlighted icon over time (as described in IV-C). To support this, we introduced two key input representations to capture dynamic attention changes arising from both bottom-up and task-driven processes: a highlight vector, indicating when the highlight is active or absent across time steps, and a task vector, encoding the drone state values of the targeted icon. By combining these inputs with temporally-aware model architectures such as LSTMs and transformer encoders, our model significantly outperforms state-of-the-art baselines, enabling accurate predictions of attention dynamics in critical monitoring tasks.

A. Temporal Pixel-Level Saliency Prediction

1) Saliency Dataset Generation: For the first part of the temporal visual attention prediction, we focused on how users' gaze behavior changes during the transition from a non-critical to a critical situation. We analyzed gaze data from one second before to five seconds after the onset of each critical situation to capture visual attention dynamics and the influence of visual highlights. Specifically, we synchronized the gaze data with the UI's frame rate (1/24 per second) and extracted fixation details as described in Section III-B1. Following

TABLE VI: Saliency Map Prediction Results under Different Highlight Conditions (Mean)

Model	Highlight	AUC ↑	SIM ↑	NSS ↑	CC ↑
ITTI	H	0.484	0.140	0.740	0.139
	NH	0.461	0.142	0.467	0.107
SimpleNet	H	**0.914**	**0.570**	**2.874**	**0.717**
	NH	**0.896**	**0.569**	**1.913**	**0.651**
TASED-Net	H	0.876	0.451	2.166	0.535
	NH	0.860	0.475	1.645	0.511

established methods [8], [11], continuous saliency maps were created by merging fixation points from all participants during 42ms and applying a Gaussian kernel with a window size of 35px to smooth them [40]. We thus obtained 4,608 smoothed saliency maps (of which 60% were used for training, 10% for validation, and the remaining 30% for testing), with each map corresponding to a specific frame of the monitoring interface. The paired saliency maps and corresponding UI frames formed the input and output data used for training and evaluating our models.

2) Evaluation Metrics: We assessed the performance of pixel-level saliency models with four commonly used metrics [40]: *AUC (Area Under the ROC Curve)* measures the model's ability to classify fixation points across varying thresholds, calculating true and false positive rates across thresholds; *NSS (Normalized Scanpath Saliency)* calculates the mean normalized saliency at fixation points, penalizing false positives; *SIM (Similarity)* measures the intersection of two normalized saliency distributions as histograms, with a score of one indicating perfect overlap and zero indicating no overlap, and is sensitive to missing values and Gaussian blur, favoring partial matches; and *CC (Pearson's Correlation Coefficient)* assesses the linear correlation between predicted and ground-truth saliency maps, symmetrically penalizing false positives and negatives and remaining resilient to linear transformations. In the development of our model, we also incorporated the Kullback-Leibler divergence (KL) into the loss function to evaluate the discrepancy between the predicted saliency and ground truth distributions.

3) Model Choice and Implementation: We evaluated three models: ITTI [38], a bottom-up model using low-level visual features; SimpleNet [39], an encoder–decoder with ResNet backbone; and TASED-Net [27], which integrates spatiotemporal features using 3D convolutions.

For the training process of the two data-driven saliency models, while SimpleNet was originally trained using KL [41] and CC [39], we extended the loss function to include SIM, resulting in the combined loss: 10KL – 3CC – 2SIM. This formulation improved alignment with ground truth and reduced false positives. Since TASED-Net requires sequences of frames (we used 32 per batch), its application was limited by the dataset size. To address this, we froze the encoder and fine-tuned only the decoder. We also modified its original loss to KL – 0.5CC – 0.1SIM, following prior work [42], which improved predictive accuracy and robustness in our task.

4) Results: Table VI presents the performance metrics of the three saliency models under two conditions: when a critical situation was highlighted and when it was not. Across all models, performance declined in the no-highlight condition, though the extent of the drop varied. Notably, there is a clear performance gap between the traditional ITTI model and the data-driven approaches—SimpleNet and TASED-Net—with both learning-based models achieving substantially better results. Among them, SimpleNet slightly outperformed TASED-Net across all quantitative metrics.

However, a different picture emerges in the qualitative analysis of how well each model captures attention shifts triggered by highlights. As shown in Figure 7, ITTI produces dispersed saliency maps across all timestamps, failing to emphasize any specific interface elements. SimpleNet, while achieving higher metric scores, generates nearly identical predictions across highlighted frames (T1, T2, T3), as it processes each frame independently and lacks temporal awareness. In contrast, TASED-Net demonstrates a promising capacity to incorporate temporal context. Its predictions clearly shift from broadly distributed attention at the onset of the critical situation (T1) to a concentrated focus on the highlighted icon one second later (T2), closely mirroring the actual attention dynamics observed in Figure 4.

Although TASED-Net is able to reflect the initial delay in detecting highlighted GUIs elements, there are several key limitations in using pixel-level saliency models to analyze how visual highlights modulate user attention over time. First, pixel-level saliency metrics provide a static, frame-by-frame evaluation of prediction accuracy but fail to capture how attention dynamically shifts in response to visual highlights. While SimpleNet achieves higher metric scores in isolated evaluations, its predictions remain largely invariant across timestamps, failing to reflect the gradual transition of visual focus. Second, these models do not allow for a direct comparison of the relative saliency of the highlight against other critical visual elements, such as drone icons, making it difficult to assess the user's SA within the broader interface context. Finally, crucial temporal aspects—such as how quickly a highlight captures attention and how long this attention is sustained—cannot be directly inferred from pixel-level saliency maps, further limiting their ability to characterize the dynamic effects of visual highlights. To address all these limitations, we transition to temporal normalized saliency prediction in the next section.

B. Temporal Normalized Saliency Prediction

1) Task Description: We formally define the *temporal normalized saliency prediction task* as follows. Given a *spatial input* $S' \in \mathbb{R}^{h \times w \times c}$, where S' is a stacked image combining the global interface view S^g and the local highlight region S^e, and *temporal inputs* $v_t \in \{-1, 0, 1\}^T$ and $c_t \in \mathbb{R}^T$, where v_t is a binary highlight vector indicating the presence (1), absence (–1), or padding (0) of highlights over the past T time steps, and c_t encodes the corresponding drone state values for the targeted icon— the task is to predict the *output* $\hat{\text{NS}}(e, t) \in [0, 1]$, representing the normalized saliency of the highlighted element e at timestamp t.

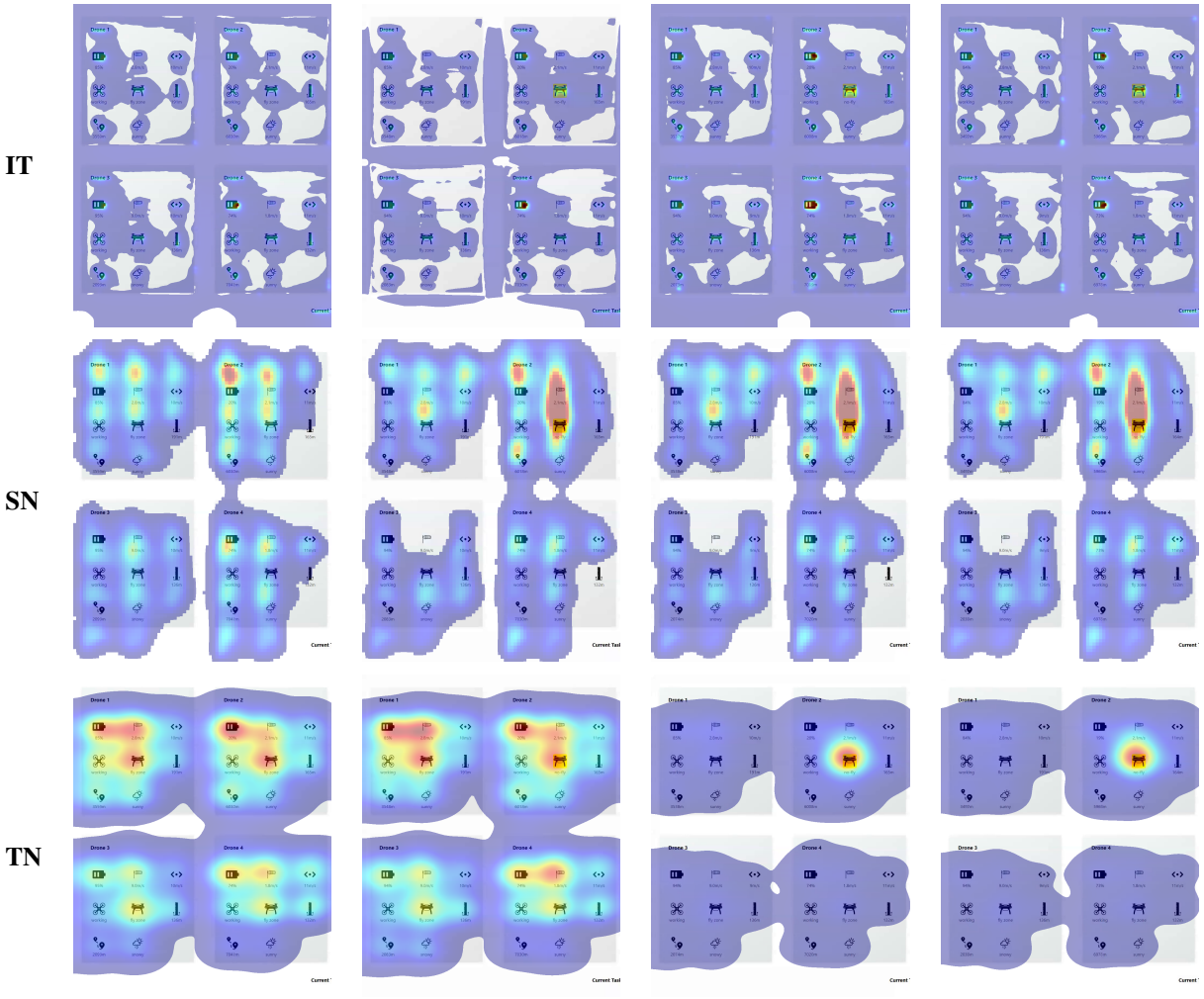


Fig. 7: Saliency model predictions for a highlighted critical situation (CS). The top row shows ITTI (IT), the middle row shows SimpleNet (SN), and the bottom row shows TASED-Net (TN).

Formally, the model learns the following mapping:

$$f : (S', v_t, c_t) \mapsto \hat{\text{NS}}(e, t)$$

The model is trained to minimize the discrepancy between the predicted value $\hat{\text{NS}}(e, t)$ and the ground truth $\text{NS}(e, t)$, enabling accurate temporal modeling of attention dynamics driven by visual highlights.

2) HISM Model Structure: HISM consists of two main components: a *spatial branch* and a *temporal branch*, as shown in Figure 8. The spatial branch processes a stacked image that combines the global interface view and the local highlight region, allowing the model to capture spatial characteristics such as local contrast and layout context.

To capture temporal attention dynamics, we implemented **three variants** of the temporal branch:

- **HISM (LSTM)**: uses an LSTM [43] to process the highlight vector, which encodes the presence, absence, or padding of highlights over the past time steps.
- **HISM (TranEnc)**: replaces the LSTM with a Transformer Encoder [44], using the same highlight vector as input.
- **HISM (TranEnc + TaskVec)**: extends the Transformer Encoder variant by concatenating the highlight vector with a task vector that encodes the drone state values of the targeted icon at each timestamp.

The task vector introduces top-down information that complements the highlight signal and helps the model make accurate predictions, especially in the absence of visual cues.

The architecture of HISM (illustrated in Figure 8) includes the following key steps:

1) **Spatial Processing**: The stacked image S', containing both the global GUI and the binary mask of the highlighted region, is processed through a ResNet50 backbone pre-trained on saliency prediction tasks [45]. The spatial features are extracted via a global average pooling layer.
2) **Temporal Processing**: We implemented three temporal variants: one using an LSTM and two using Transformer Encoders. While all variants take the highlight vector, v^t, as input, the final variant also incorporates a task vector, c_t, representing the drone state at each time step. These

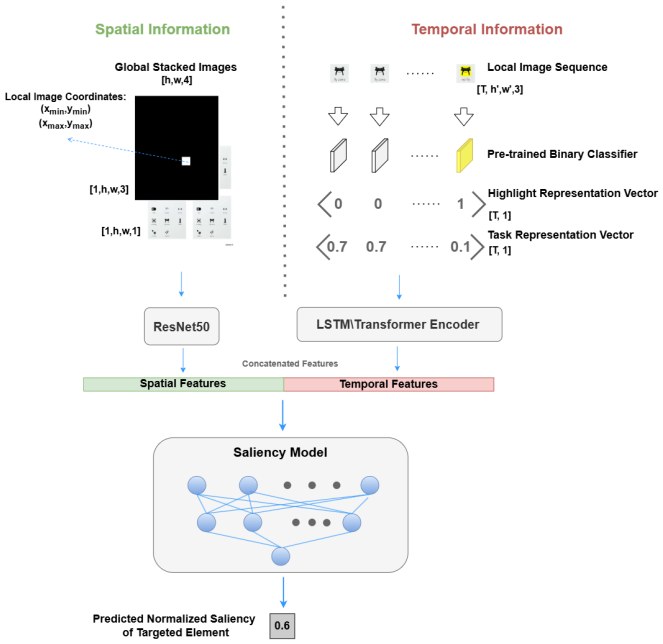


Fig. 8: The overall architecture of the HISM.

alternatives allow us to compare the impact of model architecture and input richness on attention prediction performance.

3) **Feature Fusion and Prediction**: The spatial and temporal features are concatenated and passed through a fully connected neural network consisting of three layers. These layers use ReLU activations and dropout regularization to predict the NS ($\hat{\text{NS}}(e,t)$) of the highlighted UI element for the given time slice. Since NS is a continuous scalar value, the model is optimized using the Mean Squared Error (MSE) loss function, which minimizes the difference between predicted and ground truth NS values. (Detailed training procedures are provided in the supplemental material.)

This dual-branch architecture allows HISM to effectively leverage both the static spatial properties and the dynamic temporal characteristics of highlights. By integrating these components, HISM captures the evolving saliency of a UI element, enabling accurate prediction of how and when a highlight attracts attention.

3) Baselines & Evaluation Metrics: To assess the performance of the HISM model on the temporal NS prediction task, we compared its prediction results with the ground-truth data and the results generated from the three previous pixel-level saliency models: ITTI, SimpleNet, and TASED-Net. As none of these models is designed to directly predict the NS of the highlighted element, we first predict the saliency map for the corresponding timestamps and then calculate the NS for the targeted element using Equation 1.

We evaluated the models using two metrics: *MSE (Mean Squared Error)*, which measures the average squared difference between predicted and ground-truth values, with lower values indicating better predictions; and *MAE (Mean Absolute Error)*, which calculates the average absolute difference between predicted and ground-truth values, providing a direct measure of prediction accuracy.

TABLE VII: Model Performance Evaluation under Different Highlight Conditions (Mean)

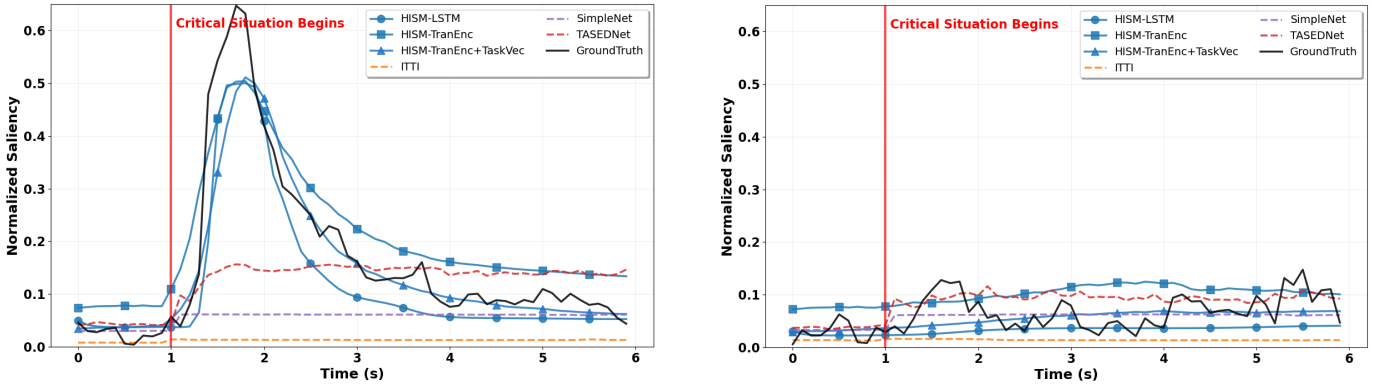
Model	Highlight	MSE ↓	MAE ↓
ITTI	H	0.0534	0.1502
	NH	0.0047	0.0515
SimpleNet	H	0.0411	0.1255
	NH	0.0028	0.0407
TASED-Net	H	0.0302	0.1145
	NH	0.0054	0.0589
HISM(LSTM)	H	0.0062	0.0558
	NH	0.0030	0.0400
HISM(TranEnc)	H	0.0052	0.0527
	NH	0.0037	0.0521
HISM(TranEnc+TaskVec)	H	**0.0048**	**0.0472**
	NH	**0.0022**	**0.0345**

4) Results: As indicated in Table VII, we observe several notable patterns in model performance on the temporal normalized saliency (NS) prediction task. First, the highlight condition poses greater prediction challenges compared to the no-highlight condition. This is due to the dynamic nature of attention under highlighting—NS rises sharply to a peak and then quickly declines—introducing more temporal variance. In contrast, NS remains relatively low and stable in the absence of highlights, making it easier to predict. As a result, all models show higher prediction errors under the highlight condition, unlike the pixel-level saliency task, where highlighting improved performance.

Second, under the highlight condition, all three HISM variants—**HISM (LSTM)**, **HISM (TranEnc)**, and **HISM (TranEnc + TaskVec)**—consistently outperform the baseline models (ITTI, SimpleNet, and TASED-Net), demonstrating the advantage of directly modeling temporal saliency with appropriate architectural support. However, in the no-highlight condition, this advantage diminishes: SimpleNet surpasses both HISM (LSTM) and HISM (TranEnc), likely due to the lower temporal complexity in these cases.

Third, incorporating the task vector provides a substantial performance boost. **HISM (TranEnc + TaskVec)** achieves the lowest error in both conditions, highlighting the value of combining highlight signals with task-relevant information. This suggests that modeling both bottom-up (highlight-driven) and top-down (task-driven) cognitive processes leads to more accurate attention prediction. Paired t-tests confirm that HISM (TranEnc + TaskVec) performs significantly better than all other models in the highlight condition for both MSE and MAE ($p < .05$).

Finally, the prediction curves (see Figure 9) further validate these findings. Under the highlight condition (Figure 9a), all HISM variants accurately capture the sharp rise in NS, peaking around 0.5 within the first second. In contrast, SimpleNet and TASED-Net show only minor changes and fail to reflect the peak or subsequent decline. Under the no-highlight condition (Figure 9b), NS changes are more random, yet all HISM variants still track a slight increase after the critical situation.



(a) Highlight (b) No Highlight

Fig. 9: The average prediction results of NS on the targeted element from our HISM-LSTM, HISM-TranEnc, HISM-TranEnc+TaskVec, the ITTI, the SimpleNet, and the TASED-Net on the test dataset (a) Highlight. (b) No Highlight.

Among them, HISM (TranEnc + TaskVec) yields the most accurate predictions.

VI. Discussion & Conclusion

In this paper, we investigated how visual highlights influence user attention in monitoring tasks, helping users focus on critical information more efficiently and accurately. To quantify the temporal effects of visual highlights, we introduced NS as a metric that captures when, how much, and for how long a highlighted element draws attention. Our findings revealed that NS not only reflects user engagement with the highlighted information at different timestamps but also provides insights into overall attention, which is closely related to SA. We proposed a new saliency model, HISM, to directly predict the NS of highlighted elements by integrating temporal and spatial information. HISM outperforms traditional pixel-level models by accurately inferring the NS of highlighted elements, enabling precise modeling of user attention dynamics.

A. Supporting Human Oversight with Adaptive Interfaces

Visual highlighting is widely used to direct attention efficiently, but it comes with potential trade-offs, especially in multitasking environments. While our findings reinforce that highlights can accelerate critical information detection, they also reveal unintended side effects, such as reduced engagement with other elements on the interface, which can impact SA. This aligns with concerns seen in alarm floods, where excessive cues lead to cognitive overload and reduced user oversight. Our study highlights the need to balance attention guidance and interface adaptivity to avoid narrow visual tunneling while ensuring efficient task performance. Future adaptive interface designs should incorporate models like HISM to dynamically adjust visual cues based on user engagement patterns, ensuring that highlights support rather than hinder broader SA.

B. Limitations and Future Work

While our study focused on yellow highlights, the HISM framework is expected to generalize to a broader range of visual cues, including color changes, blinking effects, and shape modifications. However, we recognize that such abrupt onset cues may appear rigid or intrusive in certain safety-critical contexts. Future work should therefore not only investigate how various highlighting techniques influence attention distribution and task performance, but also explore more ecologically grounded alternatives. Inspired by Hansen's work [46] on configural displays and optical invariants, promising directions include the use of *figural goodness*, *dynamic figural deformation*, and *container-based representations*—approaches that subtly guide attention by leveraging perceptual principles rather than sudden visual changes. These alternatives could offer more cognitively compatible ways to direct user attention while maintaining broader situational awareness. Incorporating such mechanisms into predictive frameworks like HISM may enhance both usability and robustness across real-world monitoring interfaces. At the same time, we acknowledge that HISM currently predicts the normalized saliency (NS) of a single element within a short detection window, while extending predictions to the full distribution of attention across the interface and to longer temporal horizons remains an open challenge. Moreover, our evaluation was conducted with 28 university participants; future work could examine expert populations and incorporate richer gaze signals, such as saccade sequences, to better capture domain-specific attentional strategies.

References

[1] D. Kern, P. Marshall, and A. Schmidt, "Gazemarks: Gaze-based visual placeholders to ease attention switching," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 2093–2102. [Online]. Available: https://doi.org/10.1145/1753326.1753646

[2] Y. Yang, M. Götze, A. Laqua, G. C. Dominioni, K. Kawabe, and K. Bengler, "A method to improve driver's situation awareness in automated driving," *Proceedings of the human factors and ergonomics society Europe*, pp. 29–47, 2017.

[3] B. B. De Koning, H. K. Tabbers, R. M. Rikers, and F. Paas, "Towards a framework for attention cueing in instructional animations: Guidelines for research and design," *Educational Psychology Review*, vol. 21, pp. 113–140, 2009.

[4] W.-C. Li, A. Horn, Z. Sun, J. Zhang, and G. Braithwaite, "Augmented visualization cues on primary flight display facilitating pilot's monitoring

performance," *International Journal of Human-Computer Studies*, vol. 135, p. 102377, 2020.

[5] S. Lallé, D. Toker, and C. Conati, "Gaze-driven adaptive interventions for magazine-style narrative visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 6, pp. 2941–2952, 2019.

[6] M. Gingerich and C. Conati, "Constructing models of user and task characteristics from eye gaze data for user-adaptive information highlighting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 29, no. 1, 2015.

[7] Z. Wu and A. M. Feit, "Enhancing user gaze prediction in monitoring tasks: The role of visual highlights," in *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*, 2024, pp. 1–3.

[8] Y. Jiang, L. A. Leiva, H. Rezazadegan Tavakoli, P. RB Houssel, J. Kylmälä, and A. Oulasvirta, "Ueyes: Understanding visual saliency across user interface types," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–21.

[9] L. A. Leiva, Y. Xue, A. Bansal, H. R. Tavakoli, T. Korodhlu, J. Du, N. R. Dayama, and A. Oulasvirta, "Understanding visual saliency in mobile user interfaces," in *22nd International conference on human-computer interaction with mobile devices and services*, 2020, pp. 1–12.

[10] C. Fosco, V. Casser, A. K. Bedi, P. O'Donovan, A. Hertzmann, and Z. Bylinskii, "Predicting visual importance across graphic design types," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 2020, pp. 249–260.

[11] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2018, pp. 4894–4903.

[12] Z. Li, Y. F. Cheng, Y. Yan, and D. Lindlbauer, "Predicting the noticeability of dynamic virtual elements in virtual reality," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613904.3642399

[13] Y. Wang, A. Bulling *et al.*, "Scanpath prediction on information visualisations," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[14] T. Cheng, I. B. Utne, B. Wu, and Q. Wu, "A novel system-theoretic approach for human-system collaboration safety: Case studies on two degrees of autonomy for autonomous ships," *Reliability Engineering & System Safety*, vol. 237, p. 109388, 2023.

[15] C. D. Wickens, *Engineering Psychology and Human Performance. Routledge*, 2021.

[16] A. Das, Z. Wu, I. Skrjanec, and A. M. Feit, "Shifting focus with hceye: Exploring the dynamics of visual highlighting and cognitive load on user attention and saliency prediction," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. ETRA, pp. 1–18, 2024.

[17] A. M. Feit, L. Vordemann, S. Park, C. Berube, and O. Hilliges, "Detecting relevance during decision-making from eye movements for ui adaptation," in *ACM Symposium on Eye Tracking Research and Applications*, ser. ETRA '20 Full Papers. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3379155.3391321

[18] Y. Wan and N. Sarter, "Attention limitations in the detection and identification of alarms in close temporal proximity," *Human factors*, vol. 66, no. 1, pp. 234–257, 2024.

[19] C. Wickens and A. Colcombe, "Dual-task performance consequences of imperfect alerting associated with a cockpit display of traffic information," *Human factors*, vol. 49, no. 5, pp. 839–850, 2007.

[20] Y. Wan, "Identifying and overcoming attention limitations in the detection and identification of alarms in close temporal proximity," Ph.D. dissertation, 2019.

[21] F. Zhou, X. J. Yang, and J. C. De Winter, "Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving," *IEEE transactions on intelligent transportation systems*, vol. 23, no. 3, pp. 2284–2295, 2021.

[22] S. Sun, "Exploring the interface to aid the operator's situation awareness in supervisory control of multiple drones," Master's thesis, 2022. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-323073

[23] M. Choi, J. Houle, and T. L. Wickramarathne, "On the development of quantitative operator situational awareness assessment methods for small-scale unmanned aircraft systems," in *2022 25th International Conference on Information Fusion (FUSION)*. IEEE, 2022, pp. 1–8.

[24] M. R. Endsley, *Direct measurement of situation awareness: Validity and use of SAGAT*. Lawrence Erlbaum Associates Publishers, 2000, pp. 147–173. [Online]. Available: https://psycnet.apa.org/record/2000-02240-005

[25] T. Zhang, J. Yang, N. Liang, B. J. Pitts, K. O. Prakah-Asante, R. Curry, B. S. Duerstock, J. P. Wachs, and D. Yu, "Physiological measurements of situation awareness: a systematic review," *Human factors*, p. 0018720820969071, 2020.

[26] K. Moore and L. Gugerty, "Development of a novel measure of situation awareness: The case for eye movement analysis," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 54, no. 19. SAGE Publications Sage CA: Los Angeles, CA, 2010, pp. 1650–1654.

[27] K. Min and J. J. Corso, "Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2394–2403.

[28] P. Gupta, S. Gupta, A. Jayagopal, S. Pal, and R. Sinha, "Saliency prediction for mobile user interfaces," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1529–1538.

[29] V. J. Traver, J. Zorío, and L. A. Leiva, "Glimpse: A gaze-based measure of temporal salience," *Sensors*, vol. 21, no. 9, p. 3099, 2021.

[30] B. Aydemir, L. Hoffstetter, T. Zhang, M. Salzmann, and S. Süsstrunk, "Tempsal-uncovering temporal information for deep saliency prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6461–6470.

[31] S. Schirmer, C. Torens, J. C. Dauer, J. Baumeister, B. Finkbeiner, and K. Y. Rozier, *A Hierarchy of Monitoring Properties for Autonomous Systems*. [Online]. Available: https://arc.aiaa.org/doi/abs/10.2514/6.2023-2588

[32] M. D. Gregorio, M. Romano, M. Sebillo, G. Vitiello, and A. Vozella, "Improving human ground control performance in unmanned aerial systems," *Future Internet*, vol. 13, no. 8, p. 188, 2021.

[33] DJI, "Mavic 3 - downloads," 2023, accessed: August 26, 2023. [Online]. Available: https://www.dji.com/de/mavic-3/downloads

[34] FlytBase, "Drone delivery system: Everything you need to know," 2023, accessed: August 26, 2023. [Online]. Available: https://www.flytbase.com/blog/drone-delivery-system

[35] M. R. Endsley, "Direct measurement of situation awareness: Validity and use of sagat," *Situation awareness analysis and measurement*, vol. 10, pp. 147–173, 2000.

[36] E. S. Dalmaijer, S. Mathôt, and S. Van der Stigchel, "Pygaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments," *Behavior research methods*, vol. 46, pp. 913–921, 2014.

[37] A. M. Feit, L. Vordemann, S. Park, C. Berube, and O. Hilliges, "Detecting relevance during decision-making from eye movements for ui adaptation," in *ACM Symposium on Eye Tracking Research and Applications*, 2020, pp. 1–11.

[38] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[39] N. Reddy, S. Jain, P. Yarlagadda, and V. Gandhi, "Tidying deep saliency prediction architectures," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 241–10 247.

[40] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 3, pp. 740–757, 2018.

[41] S. Shin, S. Chung, S. Hong, and N. Elmqvist, "A scanner deeply: Predicting gaze heatmaps on visualizations using crowdsourced eye movement data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 396–406, 2022.

[42] Q. Chang and S. Zhu, "Temporal-spatial feature pyramid for video saliency detection," *arXiv preprint arXiv:2105.04213*, 2021.

[43] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[45] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 262–270.

[46] J. P. Hansen, "Representation of system invariants by optical invariants in configural displays for process control," in *Local applications of the ecological approach to human-machine systems*. CRC Press, 2018, pp. 208–233.