Synergistic Bleeding Region and Point Detection in Laparoscopic Surgical Videos

Jialun Pei[1] Zhangjun Zhou[2] Diandian Guo[1] Zhixi Li[2,3]
Jing Qin[2] Bo Du[4*] Pheng-Ann Heng[1]

[1] The Chinese University of Hong Kong [2] The Hong Kong Polytechnic University
[3] Southern Medical University [4] Wuhan University

Abstract

*Intraoperative bleeding in laparoscopic surgery causes rapid obscuration of the operative field to hinder the surgical process and increases the risk of postoperative complications. Intelligent detection of bleeding areas can quantify the blood loss to assist decision-making, while locating bleeding points helps surgeons quickly identify the source of bleeding and achieve hemostasis in time to improve surgical success rates. To fill the benchmark gap, we first construct a real-world laparoscopic surgical bleeding detection dataset, named **SurgBlood**, comprising 5,330 frames from 95 surgical video clips with bleeding region and point annotations. Accordingly, we develop a dual-task synergistic online detector called **BlooDet**, enabling simultaneous detection of bleeding regions and points in laparoscopic surgery. The baseline embraces a dual-branch bidirectional guidance design based on Segment Anything Model 2. The mask branch detects bleeding regions through adaptive edge and point prompt embeddings, while the point branch leverages mask memory to induce bleeding point memory modeling and captures point motion direction via inter-frame optical flow. By coupled bidirectional guidance, our framework explores spatial-temporal correlations while exploiting memory modeling to infer the current bleeding status. Extensive experiments indicate that our method outperforms 13 counterparts in bleeding detection. Code and data are available at https://github.com/PJLallen/SurgBlood.*

1. Introduction

Minimally invasive surgery has revolutionized clinical healthcare by reducing patient trauma and accelerating postoperative recovery [31, 39]. However, intraoperative bleeding is an emergency that significantly impacts surgical safety and efficiency in endoscopic surgery [12, 13]. Rapid changes in the amount and speed of bleeding can severely obscure the surgical field, delaying the surgeon's response and reducing the success rate of surgery. More prolonged

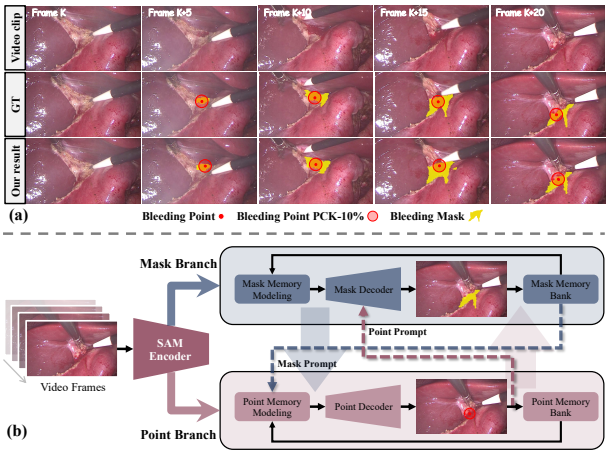

Figure 1. (a): Illustration of bleeding detection task with samples in SurgBlood and predictions of our solution. (b): The proposed BlooDet performs dual-branch bidirectional guidance for synergistic bleeding region and point detection.

bleeding increases the risk of organ damage and postoperative complications [10]. Thus, utilizing computer-assisted techniques to detect bleeding regions and localize bleeding points holds significant clinical value. In one respect, detecting bleeding areas can quantify blood loss, providing timely support for intraoperative decision-making. In another respect, precise bleeding point localization enables surgeons to control hemorrhage promptly to ensure safety.

Despite the popularity of laparoscopic surgery, automated detection of bleeding regions and bleeding points still faces numerous challenges [38]. Due to the narrow field of view under laparoscopy and unstable lighting conditions, the anatomical structures are incompletely exposed, which increases the difficulty of extracting discriminative representations. Additionally, the rapid accumulation and flow of blood can change tissue appearance and infiltrate surrounding tissues, reducing the availability of low-level visual clues and complicating the detection of bleeding regions. The bleeding points may also be buried by blood or obscured by tissues, making it difficult to quickly locate and continuously track [27]. Beyond these challenges, intelli-

*Corresponding author. (dubo@whu.edu.cn)

gent bleeding warning involves detecting bleeding regions and locating bleeding points during dynamic surgical procedures [14]. This requires a reliable multi-task online detector that models fine-grained spatial-temporal relationships in surgical videos for accurate predictions. Further, the lack of publicly available multi-task real bleeding datasets remains a major obstacle to progress in surgical intelligence.

To advance research on bleeding region and point detection in surgical videos, we construct a new actual laparoscopic surgery bleeding dataset, named **SurgBlood**. Our dataset comprises a total of 5,330 video frames from 95 laparoscopic video clips, encompassing multiple types and intensities of bleeding during surgery. As displayed in Fig. 1(a), SurgBlood also provides pixel-level annotations of bleeding regions and bleeding point coordinates by hepatobiliary surgeons, supporting the joint detection of bleeding regions and points. We evaluate several task-relevant methods on SurgBlood to establish a comprehensive benchmark for laparoscopic intraoperative bleeding detection, driving further research in surgical assistance.

Existing learning-based methods have been demonstrated to be effective in bleeding region detection [1, 17]. However, most algorithms [1, 8, 17] are designed for image or keyframe analysis, lacking the ability to model temporal dependencies in surgical videos. In addition, previous methods [23, 35, 38] mainly focus on bleeding region detection, which falls short of addressing the clinical needs in locating the bleeding source. With the emergence of large vision models, Segment Anything Model 2 (SAM 2) [34] unleashed powerful visual representation capabilities for video sequence modeling. Subsequently, a series of SAM 2-based frameworks [21, 22, 32] have been proposed in the clinical medicine domain, but have not yet been unified into a multi-task paradigm. Multi-task frameworks [8, 23] can detect both regions and keypoints, but they neglect spatial-temporal modeling, leading to difficulties in stably tracking the movement of bleeding sources. Moreover, current multi-task architectures are unable to fully harness the potential of SAM 2 in cross-task joint optimization.

To meet the clinical demand of bleeding region and point detection, in this paper, we propose a dual-task online baseline model called **BlooDet**, which embraces a dual-branch bidirectional guidance scheme based on SAM 2 to synergistically optimize both tasks. As illustrated in Fig. 1(b), our framework consists of two branches: *Mask branch* and *Point branch*. In the mask branch, we embed an edge generator that performs multi-scale perception of spatial-temporal features with the wavelet Laplacian filter to generate edge prompts, mitigating the problem of blurred bleeding boundaries. Meanwhile, we incorporate bleeding points produced from the point branch as point prompts and combine them with edge prompts to facilitate bleeding region detection. For the point branch, considering the movement of bleeding points within the field of view is influenced by the relative motion of the camera, we leverage inter-frame optical flow and mask memory to estimate camera motion and viewpoint offsets, improving the location accuracy of bleeding points and reducing interference caused by surrounding blood blurring. Further, we integrate mask memory trails from the mask branch to enhance bleeding point perception. By coupling clues and co-guiding across two branches, BlooDet can exploit spatial-temporal associations between bleeding regions and points. Extensive experiments demonstrate that our approach achieves superior performance in both tasks, *e.g.*, 64.88% IoU for bleeding region detection and 83.69% PCK-10% for bleeding point detection. The main contributions are four-fold:

- We debut intraoperative bleeding region and point detection tasks in surgical videos and contribute a real-world dataset to advance the surgical intelligent assistance.
- We introduce a dual-task synergistic online detector BlooDet, which adopts a dual-branch structure and performs co-optimization by mutual prompts and bidirectional guidance for bleeding region and point detection.
- The inter-frame optical flow and mask memory are utilized in point branch for capturing movement cues and providing spatial-temporal modeling. Besides, the edge generator and point prompting strategy in the mask branch exploits multi-scale wavelet Laplacian filters to enhance edge perception, while incorporating bleeding points for mask prompt embedding.
- We establish a comprehensive benchmark and evaluate 13 task-related models on SurgBlood. Experimental results indicate that our method achieves superior performance for joint bleeding region and point detection.

2. Related Work

Bleeding Region Detection. Bleeding region detection has been explored across various medical scenarios, such as intracranial hemorrhage detection [17], capsule endoscopy bleeding recognition [1], and retinal hemorrhage identification [43]. Deep learning-based methods employ convolution and attention mechanisms to extract discriminative features for bleeding localization. Sunakawa *et al.* [38] developed a semantic segmentation model for automatically recognizing bleeding regions on the anatomical structure of the liver. Nonetheless, existing methods primarily focus on mask-level bleeding detection, overlooking the localization of the bleeding source. To bridge this gap, we introduce a unified paradigm for the synergistic detection of bleeding regions and points in surgical videos.

Keypoint Detection in Medical Domain. Keypoint detection plays a crucial role in various clinical applications, *e.g.*, pathological site identification and anatomical landmark localization [6, 41]. Existing keypoint detection methods are usually classified into three categories: 1) Context-aware

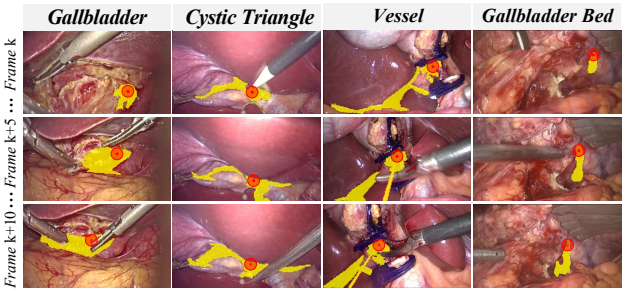

Figure 2. Illustration of bleeding types in SurgBlood.

Table 1. Summary of publicly available surgical bleeding datasets. SurgBlood provides richer real-world surgical videos with *region* and *point*-level annotations, showing holistic bleeding conditions.

Dataset	Year	Source	Type	#Videos	#Frames	Region	Point
Rabbani *et al.* [33]	2022	Human	Image	–	751	✓	–
HemoSet [25]	2024	Animal	Video	10	857	✓	–
SurgBlood (Ours)	2025	Human	Video	**95**	**5,330**	✓	✓

spatial methods. This technique exploits the stability and uniformity of keypoint spatial distributions to improve localization accuracy [5, 20, 29, 44]. 2) Multi-stage learning strategies. These architectures follow a coarse-to-fine process to refine keypoint localization through gradual optimization and integrating shallow and deep layers [7, 28, 46]. 3) Multi-task learning frameworks, which jointly optimize image segmentation and keypoint detection by constructing a union network [8, 23, 45]. To enhance mutual guidance between tasks [30], we propose a collaborative dual-branch model that couples bleeding region and point detection.

3. SurgBlood Dataset

Currently, there are few publicly available datasets for surgical bleeding detection, as summarized in Table 1. HemoSet [25] provides bleeding samples based on live animal robotic surgery, including 857 labeled frames for bleeding areas without bleeding points. Rabbani *et al.* [33] released a dataset of gynecologic laparoscopic surgeries, containing 751 frames of active bleeding rather than video clips. To this end, we construct a brand-new dataset, SurgBlood, specifically for bleeding region and point detection in laparoscopic surgery. We provide an overview of our dataset from the following three aspects: data collection, data annotation, and data analysis.

3.1. Data Collection

To ensure high-quality and representative data, we invited four hepatobiliary surgeons from partner hospitals to carefully select 95 video clips from 42 cholecystectomy surgical cases. Each clip covers the entire bleeding process while retaining the non-bleeding scene for approximately 3 seconds before and after the bleeding event. We collect a total of

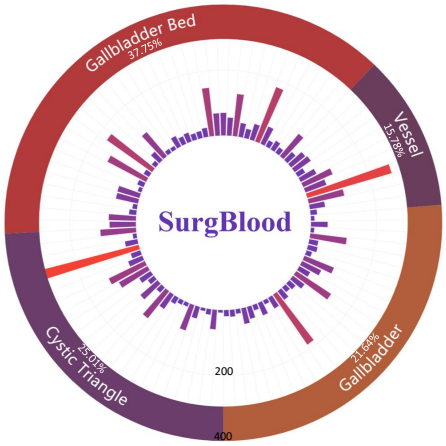

Figure 3. Statistical distribution of video clips in SurgBlood.

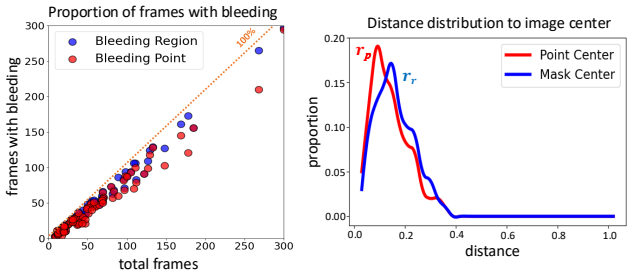

Figure 4. **Bleeding distribution**. Left: proportion of frames with bleeding region and point; Right: Distance of bleeding region center and bleeding point to image center.

5,330 video frames with a resolution of 1280×720 from all clips using a sampling rate of 2 fps. Notably, we focus exclusively on dynamic bleeding regions within the surgical action field, as this is the critical location that directly interferes with the surgeon and contains key bleeding points. As shown in Fig. 2, there are four bleeding types by tissue location: gallbladder, cystic triangle, vessel, and gallbladder bed. The proportion of bleeding types corresponds to the frequency distribution observed in actual surgery.

3.2. Data Annotation

To ensure the annotation quality of SurgBlood, the invited hepatobiliary surgeons meticulously annotate and review each video clip. During the labeling process, the surgeon uses both static frames and dynamic video sequences to label bleeding regions and bleeding point coordinates for each frame. Annotation is guided by the following principles: 1) For bleeding regions, pooled blood and inactive sparse bloodstains are not annotated. 2) For bleeding points, if the bleeding point is not obscured by tissue or instruments, its coordinates are labeled; if coordinates are difficult to confirm, refer to surrounding anatomy or previous frames to label it. To ensure annotation consistency, we adopt the cross-validation strategy: each clip is initially annotated by four

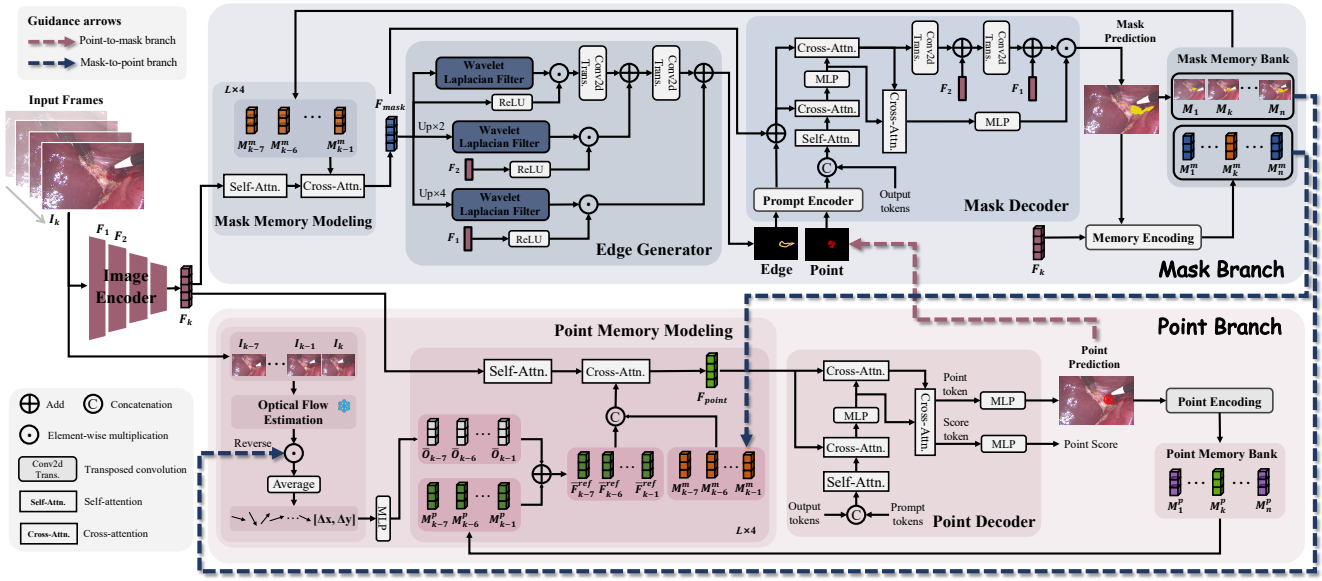

Figure 5. Overview of the proposed BlooDet. Our framework comprises a mask branch and a point branch to jointly detect bleeding regions and bleeding points. Cross-branch guidance and adaptive prompt embedding allow our framework to reach a co-optimized state.

surgeons, followed by a review and refinement process conducted by two additional surgeons. Fig. 2 presents examples of annotations for various bleeding situations.

3.3. Data Analysis

- ***Clip Distribution***: SurgBlood includes 5,330 frames extracted from 95 video clips. As shown in Fig. 3, each clip contains an average of 56 frames, with the longest containing 300 frames and the shortest containing 8 frames. We also counted the distribution of bleeding types: gallbladder (21.64%), cystic triangle (25.01%), vessel (15.78%), and gallbladder bed (37.75%).
- ***Bleeding Ratio***: We calculate the proportion of frames containing bleeding regions and points. As shown on the left of Fig. 4, both bleeding regions and points have a high frame rate, where the slightly higher rate with bleeding regions is due to the partial occlusion of bleeding points.
- ***Space Statistics***: We analyze spatial distribution of bleeding regions and points across all samples in SurgBlood. The right of Fig. 4 provides statistical insights into the distances of bleeding region centers and bleeding points from the image center. We observed that bleeding primarily occurred in the central zone of surgical manipulation.

4. Proposed Method

BlooDet is a dual-task collaborative detector for simultaneous bleeding region and point detection in surgical videos. The region detection and point localization tasks are tightly coupled: point localization enforces fine-grained geometric consistency within each predicted region (suppressing loose or false detections), while region detection supplies instance-level spatial context that constrains and stabilizes keypoint regression in cluttered scenes. This mutual dependency implies that the two tasks constrain and optimize each other rather than being solved in isolation. We formalize this interdependence with a coupled objective in which parameters of one branch impact the optimization of the other:

$$\{\boldsymbol{\theta}^*, \boldsymbol{\vartheta}^*\} = \arg\min_{\boldsymbol{\theta},\,\boldsymbol{\vartheta}} \Big[\mathcal{L}_{\mathrm{m}}\big(\boldsymbol{\theta}(\boldsymbol{\vartheta})\big) + \mathcal{L}_{\mathrm{p}}\big(\boldsymbol{\vartheta}(\boldsymbol{\theta})\big)\Big], \qquad (1)$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ denote the model parameters of the region detection and the point localization networks, respectively. The objectives $\mathcal{L}_{\mathrm{m}}$ and $\mathcal{L}_{\mathrm{p}}$ are the training losses associated with these two tasks.

4.1. Overall Architecture

As shown in Fig. 5, our framework empowers edge clues to detect bleeding regions and incorporates a point branch to localize bleeding points based on SAM 2 [34]. The whole model consists of the following processes: image encoder, mask/point memory modeling, Edge generator, mask/point decoder, and mask/point memory bank.

Image Encoding. Given a set of N video frames, including the current frame $I_k \in \mathbb{R}^{H\times W\times 3}$ and the previous N-1 frames $X = \{I_i\}_{i=k-N+1}^{k-1}$, we first flatten frames and feed them into the image encoder inherited from SAM 2 to produce multi-scale spatial features $F \in \mathbb{R}^{s\times c\times N}$, where s denotes the length of the feature sequence and c is the feature dimension. Then, the output sequential frame features $F_{k-N}, ..., F_k$ are fed into the mask and point branches, respectively, for memory modeling.

Point Branch. As shown in the bottom of Fig. 5, the point branch comprises three parts: point memory modeling, point decoder, and point memory bank. The point memory modeling module embeds optical flow estimation to predict the displacement field between consecutive frames $[I_{k-7}, ..., I_k]$, enabling the inference of laparoscopic camera motion and viewpoint offset during surgery, thereby identifying the movement direction of current bleeding points. Further, we integrate previous mask memory features $\{M_q^m\}_{q=k-7}^{k-1}$ from mask branch with the corresponding point features to enhance location and temporal perception as well as to narrow the search space for bleeding point coordinates. We describe this process in detail in Sec. 4.2.

Afterward, the memory-enhanced point feature F_{point} passes through the point decoder to predict the bleeding point. Different from the upsampling fusion in the mask decoder, we employ learnable output tokens and prompt tokens that interact with F_{point} via self-attention and cross-attention [3], followed by MLP layers to predict point coordinates and confidence scores. The point memory is stored in the point memory bank for temporal modeling.

Mask Branch. The top of Fig. 5 illustrates the pipeline of the mask branch, which produces bleeding regions by mask memory modeling and coupling edge and point prompts. The current frame features F_k are first fed into mask memory modeling to perform self- and cross-attention interaction [40] with mask memory features from previous frames, producing the spatial-temporal feature F_{mask}. After that, we introduce an edge generator that adopts multi-scale wavelet Laplacian filters to F_{mask} for edge refinement. Then, we incorporate high-resolution features from the image encoder to obtain edge maps (detailed description in Sec. 4.3). Unlike the manual intervention prompts [15, 34], we form adaptive prompt embeddings by combining the edge map E_m from the edge generator with the point map P_m from point branch. Then, the prompt encoder is utilized to yield prompt features E_p and P_p:

$$E_p, P_p = \mathcal{P}[E_m, P_m], \tag{2}$$

$$E_p = \texttt{Conv}\big(\texttt{LN}(\mathbf{G}(\texttt{Conv}(\texttt{LN}(\mathbf{G}(\texttt{Conv}(E_m)))))\big), \tag{3}$$

$$P_p = \mathcal{C}[sin(2\pi(\texttt{Po}(P_m))), cos(2\pi(\texttt{Po}(P_m)))] + \texttt{Le}, \tag{4}$$

where $\mathcal{P}$ represents prompt encoding, $\mathbf{G}$ denotes the GeLU function, $\texttt{Conv}$ refers to 2×2 convolutions, and $\texttt{LN}$ is layer normalization. Also, $\texttt{Po}$ denotes positional encoding, $\texttt{Le}$ stands for learned embeddings, and $\mathcal{C}[,]$ is the concatenation operation. We input prompt features along with F_{mask} into the mask decoder and attain the bleeding mask by upsampling and integrating with high-resolution features. Then, we employ memory encoding to achieve mask memory feature M_k^m and store it in the mask memory bank. The mask maps are also updated in the memory bank to provide spatial guidance for bleeding point detection.

Cross-branch Guidance. Our framework embraces bidirectional collaborative guidance between masks and point branches, enabling simultaneous optimization of bleeding region and bleeding point predictions. In mask decoder, we exploit the point map produced by the point decoder as an automatic prompt input. This helps guide the decoder to focus on the target bleeding region while mitigating the interference from residual blood in the surrounding area. In point memory modeling, the predicted mask maps from previous frames provide temporal and directional cues that can improve the accuracy of bleeding point localization. Besides, mask memory features are merged with point memory features to induce the point decoder to concentrate on the most likely bleeding areas while mitigating the impact of low-contrast background. The alternating optimization strategy for cross-branch guidance is found in Sec. 4.4.

4.2. Point Memory Modeling in Point Branch

To detect bleeding points in consecutive frames effectively, we embed the point memory modeling module in the point branch to develop temporal clues for point features. As illustrated in Fig. 5, point memory modeling is divided into two steps: 1) combining the optical flow between consecutive frames with region maps to compensate for the relative displacement of the bleeding point caused by camera viewpoint offset; 2) interacting the average camera displacement of previous frames with mask memory features from the mask branch to obtain point memory features.

For the viewpoint offset of the camera, we first utilize the frozen PWC-Net [36] for optical flow estimation. Given N frames $\{I_i\}_{i=k-N+1}^{k}$, the optical flow $O_i(x, y) \in \mathbb{R}^{H\times W\times 2}$ between two consecutive frames can be expressed as $O_i(x, y) = \texttt{PWC-Net}(I_{i-1}, I_i)$. Considering the instability of the optical flow in the rapidly changing bleeding region, we reverse the mask map M_i from mask branch for each frame and combine with $O_i(x, y)$ to obtain the average viewpoint offset $\bar{O}_i(\Delta x, \Delta y)$:

$$\bar{O}_i(\Delta x, \Delta y) = \frac{1}{H \times W} \sum_{X=1}^{H} \sum_{Y=1}^{W} (1 - M_i) \cdot O_i(x, y). \tag{5}$$

Then, the global offset coordinates $\bar{O}_i \in \mathbb{R}^2$ of previous frames can be produced via an MLP layer.

After that, we aggregate point memory features M_i^p of previous frames in the point memory bank with $\bar{O}_i$ and concatenate with mask memory features M_i^m to obtain the mask-guided corrected point features $\bar{F}_i^{ref}$. Lastly, we perform self-attention on F_k and cross-attention with $\bar{F}_i^{ref}$ to the memory-enhanced point feature F_{point}. Through the

optical flow estimation as well as mask guidance, we model the effective memory of point traits with camera offset.

4.3. Edge Generator in Mask Branch

To address low contrast and high noise in surgical fields, we embed an edge generator in the mask branch that generates edge map prompts by combining multi-scale wavelet Laplace filters [16] with high-resolution features containing low-level texture cues. Concretely, we first input the spatial-temporal features F_{mask} into the Gabor wavelet Laplacian filter to enhance edge structures in the spatial domain. The Gabor wavelet operation on position (x, y) is calculated as

$$\mathcal{G}(x, y; \lambda, \theta, \psi, \sigma, \gamma) = exp(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}) exp(i(\frac{2\pi}{\lambda} x' + \psi)), \tag{6}$$

where $x' = xcos\theta + ysin\theta$, $y' = -xsin\theta + ycos\theta$, λ denotes the wavelength, θ is the orientation angle of the Gabor kernel, ψ is the phase offset, σ is the standard deviation of Gaussian function, and γ stands for the aspect ratio. Thus, Laplacian filtering based on the Gabor wavelet is defined as

$$\mathbf{L_g}(x, y) = \Delta f(x, y) \cdot \mathcal{G}(x, y), \tag{7}$$

$$\Delta f(x, y) = \frac{\partial^2 f}{x^2} + \frac{\partial^2 f}{y^2}, \tag{8}$$

where $\Delta f(x, y)$ represents the Laplacian operator in 2D space. Then, we perform an activation operation on F_{mask} and interact with the filtered features to suppress low-confidence signals and preserve refined edge features. The whole process of the edge generator can be described as

$$F^{'}_{mask} = (\texttt{ReLU}(F_{mask})) \odot (\mathbf{L_g}(x, y) * F_{mask}), \tag{9}$$

where $\odot$ and $*$ denote the element-wise multiplication and convolution operation. The feature $F^{'}_{mask}$ represents the output from the edge generator. As illustrated in Fig. 5, we parallel upsample F_{mask} twice and separately pass through the wavelet Laplacian filter, and then interact with high-resolution features (F_1 and F_2) to refine bleeding edges. Finally, the generated edge map is fed into the mask decoder as prompts for bleeding region detection.

4.4. Objective Function

Synergizing two tasks within a unified model to solve coupled objectives is nontrivial, as updating one set of parameters immediately changes the optimal solution for the other. To this end, we adopt an alternating optimization scheme: fixing one branch during iteration while iteratively updating the other. At iteration t, we perform two coordinated steps:

1) Update region detection given the current point branch. We first fix $\boldsymbol{\vartheta}^{(t)}$ and update $\boldsymbol{\theta}$ by minimizing the region detection loss, which is conditioned on (and regularized by) the current point branch:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta_\theta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathtt{m}}\Big(\boldsymbol{\theta}^{(t)}\big(\boldsymbol{\vartheta}^{(t)}\big)\Big). \tag{10}$$

2) Update point localization given the updated detector. We then fix $\boldsymbol{\theta}^{(t+1)}$ and update $\boldsymbol{\vartheta}$ by minimizing the localization loss under the latest detector context:

$$\boldsymbol{\vartheta}^{(t+1)} = \boldsymbol{\vartheta}^{(t)} - \eta_\vartheta \nabla_{\boldsymbol{\vartheta}} \mathcal{L}_{\mathtt{p}}\Big(\boldsymbol{\vartheta}^{(t)}\big(\boldsymbol{\theta}^{(t+1)}\big)\Big). \tag{11}$$

This alternating strategy enforces mutual adaptation: the detector branch $\boldsymbol{\theta}$ is optimized under fine-grained geometric constraints from the current localization branch $\boldsymbol{\vartheta}^{(t)}$, while the localization branch $\boldsymbol{\vartheta}$ is optimized within the updated spatial context provided by $\boldsymbol{\theta}^{(t+1)}$. Repeating Eq. (10) and Eq. (11) drives both sets of parameters toward a joint fixed point that is consistent across region- and point-level predictions, thus approximating the coupled optimum in Eq. (1).

The loss $\mathcal{L}_{\mathtt{m}}$ of mask branch consists of the region loss $\mathcal{L}_{\mathtt{r}}$ and edge loss $\mathcal{L}_{\mathtt{e}}$:

$$\mathcal{L}_{\mathtt{m}} = \lambda_{\mathtt{r}}\mathcal{L}_{\mathtt{r}} + \lambda_{\mathtt{e}}\mathcal{L}_{\mathtt{e}}. \tag{12}$$

$\mathcal{L}_{\mathtt{r}}$ and $\mathcal{L}_{\mathtt{e}}$ are computed as a combination of Focal loss [19] and Dice loss [26]. In addition, the point branch loss $\mathcal{L}_{\mathtt{p}}$ consists of the point loss $\mathcal{L}_{\mathcal{P}}$ and the scoring loss $\mathcal{L}_{\mathtt{s}}$:

$$\mathcal{L}_{\mathtt{p}} = \lambda_{\mathcal{P}}\mathcal{L}_{\mathcal{P}} + \lambda_{\mathtt{s}}\mathcal{L}_{\mathtt{s}}. \tag{13}$$

$\mathcal{L}_{\mathcal{P}}$ employs the smooth L1 loss for point-level supervision. $\mathcal{L}_{\mathtt{s}}$ is the binary cross-entropy loss for point existence. $\lambda_{\mathtt{r}}$, $\lambda_{\mathtt{e}}$, $\lambda_{\mathtt{s}}$, and $\lambda_{\mathcal{P}}$ are set to 1, 1, 1, and 0.5.

5. Experiments

5.1. Datasets and Evaluation Metrics

Datasets. Since the tasks of the laparoscopic bleeding region and bleeding point cooperative detection are proposed for the first time, we adopt the SurgBlood dataset to train and test our method and related comparative methods. We randomly split a total of 95 video clips into two sets: 75 for training and the remaining 20 for testing. For further evaluation, we assess the region-level detection performance on HemoSet [25] and provide comparative results on the Rabbani *et al.* [33] dataset in the supplementary materials.

Evaluation Metrics. Following previous studies [8, 34], we adopt the Intersection over Union (IoU) and Dice Coefficient (Dice) metrics to evaluate bleeding region detection performance. For bleeding points, the Percentage of Correct Keypoints (PCK) metric is used to measure localization accuracy. Unlike the 10% to 40% threshold range applied in [8, 23] for anatomical structure centroids, we adopt a narrower threshold range of 2%–10% to ensure greater assessment, *i.e.*, PCK-2%, PCK-5%, and PCK-10%. This is due to the requirement for higher precision and lower tolerance of bleeding point detection in laparoscopic surgery.

Table 2. Overall comparison with task-related methods on SurgBlood test set and HemoSet [25] dataset. Scores are reported as mean ± standard deviation. † and ∗ denote region- and point-level models with an additional point/region prediction head for dual-task learning.

Methods	Volumes	Bleeding Region Detection				Bleeding Point Detection			Params ↓
		SurgBlood		HemoSet [25]		SurgBlood			
		IoU ↑	Dice ↑	IoU ↑	Dice ↑	PCK-2% ↑	PCK-5% ↑	PCK-10% ↑	
Swin-UNet† [2]	ECCV'22	34.83 ± 0.15	51.67 ± 0.20	29.86 ± 0.12	45.99 ± 0.17	1.65 ± 0.12	11.70 ± 0.18	37.23 ± 0.22	27.2M
SAM† [15]	ICCV'23	37.94 ± 0.17	55.01 ± 0.23	38.81 ± 0.18	55.91 ± 0.21	4.61 ± 0.14	25.86 ± 0.19	60.79 ± 0.21	93.9M
MemSAM† [9]	CVPR'24	52.84 ± 0.16	69.14 ± 0.22	46.89 ± 0.17	63.84 ± 0.24	5.27 ± 0.13	31.80 ± 0.24	64.91 ± 0.18	133.5M
STDDNet† [4]	ICCV'25	47.49 ± 0.18	64.39 ± 0.19	43.98 ± 0.16	61.10 ± 0.26	1.48 ± 0.14	13.67 ± 0.17	45.80 ± 0.20	38.2M
SAM 2† [34]	ICLR'25	50.93 ± 0.12	67.49 ± 0.14	56.53 ± 0.19	72.02 ± 0.24	12.35 ± 0.17	41.68 ± 0.21	71.99 ± 0.26	81.0M
HRNet∗ [37]	CVPR'19	48.90 ± 0.12	65.68 ± 0.17	48.07 ± 0.18	64.93 ± 0.26	1.98 ± 0.15	15.98 ± 0.18	47.28 ± 0.23	63.6M
SimCC∗ [18]	ECCV'22	51.02 ± 0.17	67.57 ± 0.20	51.62 ± 0.19	68.09 ± 0.23	3.62 ± 0.13	17.30 ± 0.20	48.76 ± 0.16	66.3M
GTPT∗ [42]	ECCV'24	41.10 ± 0.14	58.26 ± 0.18	50.27 ± 0.20	66.91 ± 0.22	3.29 ± 0.11	15.98 ± 0.25	38.55 ± 0.18	16.7M
D-CeLR∗ [7]	ECCV'24	51.30 ± 0.19	67.82 ± 0.19	50.85 ± 0.21	67.42 ± 0.25	2.97 ± 0.12	24.22 ± 0.24	63.92 ± 0.17	53.4M
CalibratedSL∗ [11]	TMI'25	36.96 ± 0.16	53.97 ± 0.16	44.59 ± 0.16	61.68 ± 0.19	3.46 ± 0.14	15.82 ± 0.22	39.70 ± 0.13	6.6M
PAINet [8]	MICCAI'23	44.14 ± 0.14	61.24 ± 0.15	53.78 ± 0.17	69.87 ± 0.21	2.47 ± 0.19	15.48 ± 0.26	48.43 ± 0.20	13.6M
PitSurgRT [23]	IJCARS'24	30.48 ± 0.15	46.72 ± 0.17	27.11 ± 0.18	42.65 ± 0.24	2.47 ± 0.18	13.84 ± 0.23	41.68 ± 0.18	67.3M
ConsisTNet [24]	IJCARS'25	40.43 ± 0.16	57.59 ± 0.18	49.76 ± 0.22	66.45 ± 0.26	7.09 ± 0.15	32.83 ± 0.19	68.15 ± 0.21	143.5M
BlooDet (Ours)	-	**64.88** ± 0.09	**78.70** ± 0.09	**59.62** ± 0.15	**74.70** ± 0.20	**18.62** ± 0.13	**55.85** ± 0.18	**83.69** ± 0.23	91.6M

Table 3. Comparison with region-specific methods on SurgBlood.

Methods	Volumes	IoU ↑	Dice ↑
Swin-UNet [2]	ECCV'22	41.31 ± 0.18	58.47 ± 0.17
SAM [15]	ICCV'23	40.43 ± 0.13	57.49 ± 0.15
MemSAM [9]	CVPR'24	55.34 ± 0.23	71.28 ± 0.13
STDDNet [4]	ICCV'25	50.42 ± 0.17	67.04 ± 0.15
SAM 2 [34]	ICLR'25	63.51 ± 0.11	77.68 ± 0.19
BlooDet (Ours)	-	**64.88** ± 0.09	**78.70** ± 0.09

Table 4. Comparison with point-specific methods on SurgBlood.

Methods	Volumes	PCK-2% ↑	PCK-5% ↑	PCK-10% ↑
HRNet [37]	CVPR'19	3.13 ± 0.17	15.98 ± 0.24	44.31 ± 0.19
SimCC [18]	ECCV'22	2.14 ± 0.18	14.99 ± 0.22	46.95 ± 0.11
GTPT [42]	ECCV'24	2.80 ± 0.15	13.01 ± 0.20	38.38 ± 0.26
D-CeLR [7]	ECCV'24	5.10 ± 0.16	27.67 ± 0.28	60.13 ± 0.27
CalibratedSL [11]	TMI'25	3.13 ± 0.15	15.82 ± 0.23	35.42 ± 0.24
BlooDet (Ours)	-	**18.62** ± 0.13	**55.85** ± 0.18	**83.69** ± 0.23

5.2. Implementation Details

Our framework is implemented on a single RTX 4090 GPU. During training, we input eight consecutive frames in an online manner with resolution resized to 512×512 pixels. The image encoder of BlooDet is initialized with pre-trained weights from SAM 2_base [34]. Additionally, we utilize a frozen PWC-Net [36] to compute inter-frame optical flow. No data augmentation is applied during data loading. The maximum learning rate for the image encoder is set to 5e-6, while other parts are trained with 5e-4. We employ the Adam optimizer with a warm-up strategy and linear decay, training for 20 epochs. During inference, we perform frame-by-frame inference in line with SAM 2.

5.3. Performance Comparison

We evaluate 13 task-related methods for bleeding region and point detection to build a comprehensive benchmark on SurgBlood, including multi-task detection models [8, 23, 24], region-level object segmentation methods [2, 4, 9, 15, 34], and point-level detection methods [7, 11, 18, 37, 42]. Besides, we conduct zero-shot testing on the HemoSet [25] dataset using pre-trained models from SurgBlood.

Quantitative Evaluation. We display the performance of our framework and comparison methods for bleeding region and point detection in Table 2. Region-level and point-level models respectively append point/region prediction heads for dual-task learning. We can see that our framework outperforms competitors across both tasks on SurgBlood and HemoSet [25] datasets. Benefiting from edge-enhanced prompts and bleeding point guidance, BlooDet achieves notable improvements in bleeding region detection. For bleeding point localization, BlooDet significantly surpasses other methods thanks to mutual coupling guidance. Moreover, BlooDet obtains superior accuracy with proper parameters.

We also compare with task-specific models on individual tasks. As shown in Table 3 and Table 4, BlooDet maintains state-of-the-art performance in both bleeding region detection and point localization. Comparing the results in Table 2 further reveals that task-specific models impact the performance of the original task during dual-task learning, illustrating the effectiveness of our cross-branch guidance and alternating optimization strategy in multi-task learning.

Qualitative Evaluation. Fig. 6 shows a visual comparison of our approach with multi-task methods. BlooDet provides greater stability and consistency in detecting bleeding regions and points. In surgical environments with low contrast, competitors tend to be disturbed by surrounding noise. In contrast, our method ensures robust detection across consecutive frames. More visual comparisons can refer to supplementary materials.

5.4. Ablation Analysis

Contributions of Key Component. Table 5 illustrates the contribution of key components in BlooDet. The experimental results show that point memory modeling contributes significantly to detecting bleeding points, *e.g.*, improving PCK-5% by 13.02%. Moreover, the edge generator offers effective edge prompt embedding, enhancing the accuracy of bleeding region detection. In short, each compo-

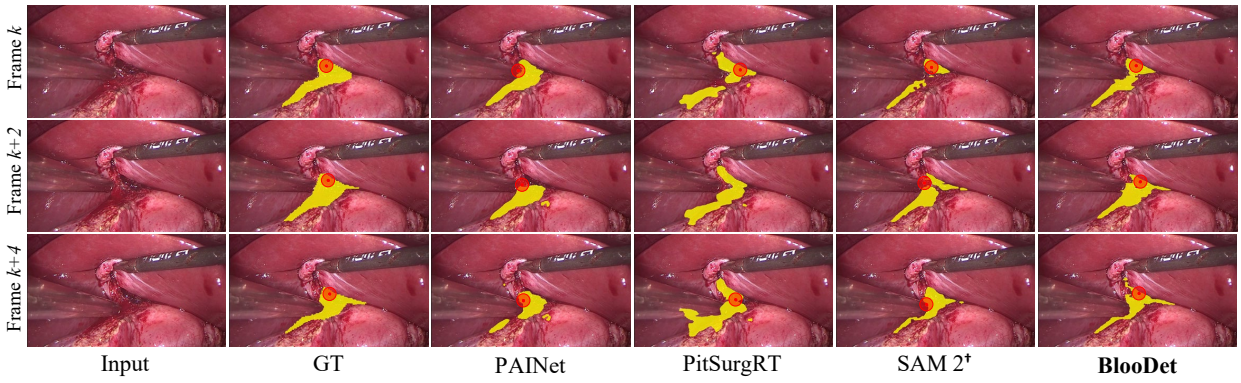

Figure 6. Visual comparison of bleeding region and point detection on SurgBlood test set.

Table 5. Ablations for key components of BlooDet for bleeding region and point detection on SurgBlood test set. EG and PMM denote edge generator and point memory modeling modules.

EG	PMM	IoU ↑	Dice ↑	PCK-2% ↑	PCK-5% ↑	PCK-10% ↑
✗	✗	56.22	71.98	13.18	45.14	75.78
✓	✗	64.38	78.36	12.52	42.83	75.94
✗	✓	61.20	75.93	14.33	51.57	80.89
✓	✓	**64.88**	**78.70**	**18.62**	**55.85**	**83.69**

Table 6. Influence of edge generator for bleeding region detection.

Configs	IoU ↑	Dice ↑	# Params (M)
w/o Edge generator	61.20	75.93	**90.99**
w/o Laplacian Filter	62.41	76.85	91.58
w/o F_1&F_2	64.74	78.60	91.55
Edge generator	**64.88**	**78.70**	91.58

Table 7. Ablations for optical flow operation via mask maps.

Mask Map	PCK-2% ↑	PCK-5% ↑	PCK-10% ↑
w/o optical flow	14.17	45.47	77.10
Global	15.49	49.59	82.00
Foreground	12.03	41.02	71.99
Background	**18.62**	**55.85**	**83.69**

nent makes a positive contribution to model performance.

Ablations for Edge Generator. We investigate the effect of edge generator designs in the mask branch. As exhibited in Table 6, the wavelet Laplacian filter shows a strong response to edge clues, mitigating interference from complex backgrounds. Meanwhile, integrating high-resolution features further enhances the quality of edge maps.

Optical Flow Operation Design. In point memory modeling, we adopt reversed mask maps in conjunction with optical flow maps to estimate the average camera displacement. Thus, we ablate the impact of focusing on different regions in mask maps for bleeding point localization. Table 7 indicates that using the foreground region leads to inferior performance. It may stem from the instability of optical flow in the rapidly changing bleeding areas. In contrast, utilizing the background enables stable motion modeling.

Effect of Mutual Guidance. To validate the effect of our cross-branch mutual guidance, Fig. 8 ablates the impact of point prompt from point branch as well as mask map and mask memory from mask branch. The results indicate that the point prompt contributes to bleeding region detection. For the point branch, the mask maps from previous frames assist with optical flow operation, while mask memory fosters point memory modeling to improve the accuracy of point localization. The attention maps visualized in Fig. 7 also verify the effect of mutual guidance.

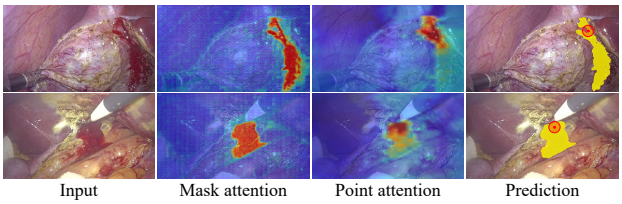

Figure 7. Attention maps in mask and point branches of BlooDet.

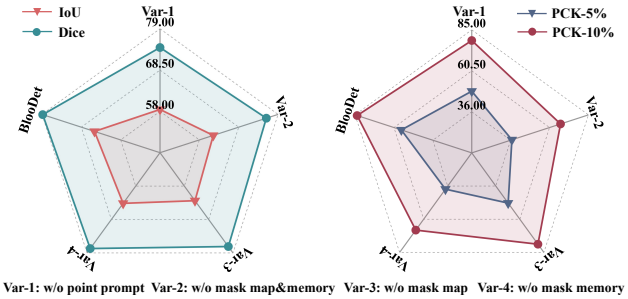

Figure 8. Effect of mutual guidance between two branches.

6. Conclusion

This work advances the intelligent bleeding region and point detection in laparoscopic surgery. We contribute an open-source dataset for bleeding detection in real-world surgical videos to facilitate benchmark construction. Accordingly, we design a dual-task synergistic online framework called BlooDet, which assembles mask and point branches in a bidirectional guidance structure, and exploits an edge generator and point memory modeling to enhance the adaptive prompting scheme. Extensive experimental results demonstrate that our method outperforms existing models for both bleeding tasks. We believe this study can facilitate research in intelligent surgical assistance, enhancing intraoperative decision-making and clinical outcomes.

Acknowledgements

This work was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project T45-401/22-N and in part by the Innovative Research Group Project of Hubei Province under Grants 2024AFA017.

References

[1] N Bourbakis, Sokratis Makrogiannis, and Despina Kavraki. A neural network-based detection of bleeding in sequences of wce images. In *Fifth IEEE Symposium on Bioinformatics and Bioengineering*, pages 324–327, 2005. 2

[2] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *ECCV*, 2022. 7

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 5

[4] Guilian Chen, Huisi Wu, and Jing Qin. Stddnet: Harnessing mamba for video polyp segmentation via spatial-aligned temporal modeling and discriminative dynamic representation learning. In *IEEE ICCV*, pages 21364–21373, 2025. 7

[5] Ruize Cui, Jiaan Zhang, Jialun Pei, Kai Wang, Pheng-Ann Heng, and Jing Qin. Topology-constrained learning for efficient laparoscopic liver landmark detection. In *MICCAI*, pages 585–594, 2025. 3

[6] Ruize Cui, Weixin Si, Zhixi Li, Kai Wang, Jialun Pei, Pheng-Ann Heng, and Jing Qin. Depth-induced prompt learning for laparoscopic liver landmark detection. *Medical Image Analysis*, page 103940, 2026. 2

[7] Chao Dai, Yang Wang, Chaolin Huang, Jiakai Zhou, Qilin Xu, and Minpeng Xu. A cephalometric landmark regression method based on dual-encoder for high-resolution x-ray image. In *ECCV*, pages 93–109, 2024. 3, 7

[8] Adrito Das, Danyal Z Khan, Simon C Williams, John G Hanrahan, Anouk Borg, Neil L Dorward, Sophia Bano, Hani J Marcus, and Danail Stoyanov. A multi-task network for anatomy identification in endoscopic pituitary surgery. In *MICCAI*, pages 472–482, 2023. 2, 3, 6, 7

[9] Xiaolong Deng, Huisi Wu, Runhao Zeng, and Jing Qin. Memsam: taming segment anything model for echocardiography video segmentation. In *IEEE CVPR*, pages 9622–9631, 2024. 7

[10] Daniel J Deziel, Keith W Millikan, Steven G Economou, Alexander Doolas, Sung-Tao Ko, and Mohan C Airan. Complications of laparoscopic cholecystectomy: a national survey of 4,292 hospitals and an analysis of 77,604 cases. *The American Journal of Surgery*, 165(1):9–14, 1993. 1

[11] Yong Feng, Jinzhu Yang, Lingzhi Tang, Song Sun, and Yonghuai Wang. Uncertainty quantification and quality control for heatmap-based landmark detection models. *IEEE TMI*, 2025. 7

[12] Jared R Gallaher and Anthony Charles. Acute cholecystitis: a review. *Jama*, 327(10):965–975, 2022. 1

[13] Diandian Guo, Weixin Si, Zhixi Li, Jialun Pei, and Pheng-Ann Heng. Surgical workflow recognition and blocking effectiveness detection in laparoscopic liver resection with pringle maneuver. In *AAAI*, pages 3220–3228, 2025. 1

[14] Yuichiro Hirai, Ai Fujimoto, Naomi Matsutani, Soichiro Murakami, Yuki Nakajima, Ryoichi Miyanaga, Yoshihiro Nakazato, Kazuyo Watanabe, Masahiro Kikuchi, and Naohisa Yahagi. Evaluation of the visibility of bleeding points using red dichromatic imaging in endoscopic hemostasis for acute gi bleeding (with video). *Gastrointestinal Endoscopy*, 95(4):692–700, 2022. 2

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *IEEE ICCV*, pages 4015–4026, 2023. 5, 7

[16] Tai Sing Lee. Image representation using 2d gabor wavelets. *IEEE TPAMI*, 18(10):959–971, 1996. 6

[17] Lu Li, Meng Wei, BO Liu, Kunakorn Atchaneeyasakul, Fugen Zhou, Zehao Pan, Shimran A Kumar, Jason Y Zhang, Yuehua Pu, David S Liebeskind, et al. Deep learning for hemorrhagic lesion detection and segmentation on brain ct images. *IEEE JBHI*, 25(5):1646–1659, 2020. 2

[18] Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. Simcc: A simple coordinate classification perspective for human pose estimation. In *ECCV*, pages 89–106, 2022. 7

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE ICCV*, pages 2980–2988, 2017. 6

[20] Chuanbin Liu, Hongtao Xie, Sicheng Zhang, Jingyuan Xu, Jun Sun, and Yongdong Zhang. Misshapen pelvis landmark detection by spatial local correlation mining for diagnosing developmental dysplasia of the hip. In *MICCAI*, pages 441–449, 2019. 3

[21] Haofeng Liu, Erli Zhang, Junde Wu, Mingxuan Hong, and Yueming Jin. Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. In *NeurIPS Workshop*, 2024. 2

[22] Haofeng Liu, Mingqi Gao, Xuxiao Luo, Ziyue Wang, Guanyi Qin, Junde Wu, and Yueming Jin. Resurgsam2: Referring segment anything in surgical video via credible long-term tracking. In *MICCAI*, pages 435–445, 2025. 2

[23] Zhehua Mao, Adrito Das, Mobarakol Islam, Danyal Z Khan, Simon C Williams, John G Hanrahan, Anouk Borg, Neil L Dorward, Matthew J Clarkson, Danail Stoyanov, et al. Pitsurgrt: real-time localization of critical anatomical structures in endoscopic pituitary surgery. *IJCARS*, pages 1–8, 2024. 2, 3, 6, 7

[24] Zhehua Mao, Adrito Das, Danyal Z Khan, Simon C Williams, John G Hanrahan, Danail Stoyanov, Hani J Marcus, and Sophia Bano. Consistnet: a spatio-temporal approach for consistent anatomical localization in endoscopic pituitary surgery. *IJCARS*, pages 1–10, 2025. 7

[25] Albert J Miao, Shan Lin, Jingpei Lu, Florian Richter, Benjamin Ostrander, Emily K Funk, Ryan K Orosco, and Michael C Yip. Hemoset: The first blood segmentation dataset for automation of hemostasis management. In *ISMR*, pages 1–7, 2024. 3, 6, 7

[26] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, pages 565–571, 2016. 6

[27] Yosuke Mori, Taro Iwatsubo, Akitoshi Hakoda, Shin Kameishi, Kazuki Takayama, Shun Sasaki, Ryoji Koshiba, Shinya Nishida, Satoshi Harada, Hironori Tanaka, et al. Red dichromatic imaging improves the recognition of bleeding points during endoscopic submucosal dissection. *Digestive Diseases and Sciences*, 69(1):216–227, 2024. 1

[28] Jialun Pei, Tianyang Cheng, Deng-Ping Fan, He Tang, Chuanbo Chen, and Luc Van Gool. Osformer: One-stage camouflaged instance segmentation with transformers. In *European conference on computer vision*, pages 19–37. Springer, 2022. 3

[29] Jialun Pei, Ruize Cui, Yaoqian Li, Weixin Si, Jing Qin, and Pheng-Ann Heng. Depth-driven geometric prompt learning for laparoscopic liver landmark detection. In *MICCAI*, pages 154–164, 2024. 3

[30] Jialun Pei, Tao Jiang, He Tang, Nian Liu, Yueming Jin, Deng-Ping Fan, and Pheng-Ann Heng. Calibnet: Dual-branch cross-modal calibration for rgb-d salient instance segmentation. *IEEE TIP*, 33:4348–4362, 2024. 3

[31] Jialun Pei, Jiaan Zhang, Guanyi Qin, Kai Wang, Yueming Jin, and Pheng-Ann Heng. Instrument-tissue-guided surgical action triplet detection via textual-temporal trail exploration. *IEEE TMI*, 2025. 1

[32] Jialun Pei, Zhangjun Zhou, and Tiantian Zhang. Evaluation study on sam 2 for class-agnostic instance-level segmentation. *CAAI Artificial Intelligence Research*, 2025. 2

[33] Navid Rabbani, Callyane Sève-d'Erceville, Nicolas Bourdel, and Adrien Bartoli. Video-based computer-aided laparoscopic bleeding management: a space-time memory neural network with positional encoding and adversarial domain adaptation. In *MIDL*, pages 961–974, 2022. 3, 6

[34] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *ICLR*, 2025. 2, 4, 5, 6, 7

[35] Ruisheng Su, Matthijs van der Sluijs, Sandra AP Cornelissen, Geert Lycklama, Jeannette Hofmeijer, Charles BLM Majoie, Pieter Jan van Doormaal, Adriaan CGM Van Es, Danny Ruijters, Wiro J Niessen, et al. Spatio-temporal deep learning for automatic detection of intracranial vessel perforation in digital subtraction angiography during endovascular thrombectomy. *Medical Image Analysis*, 77:102377, 2022. 2

[36] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE CVPR*, pages 8934–8943, 2018. 5, 7

[37] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE CVPR*, pages 5693–5703, 2019. 7

[38] Taiki Sunakawa, Daichi Kitaguchi, Shin Kobayashi, Keishiro Aoki, Manabu Kujiraoka, Kimimasa Sasaki, Lena Azuma, Atsushi Yamada, Masashi Kudo, Motokazu Sugimoto, et al. Deep learning-based automatic bleeding recognition during liver resection in laparoscopic hepatectomy. *Surgical Endoscopy*, pages 1–7, 2024. 1, 2

[39] Chris Varghese, Ewen M Harrison, Greg O'Grady, and Eric J Topol. Artificial intelligence in surgery. *Nature Medicine*, pages 1–12, 2024. 1

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 5

[41] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30 (10):2924–2935, 2024. 2

[42] Haonan Wang, Jie Liu, Jie Tang, Gangshan Wu, Bo Xu, Yanbing Chou, and Yong Wang. Gtpt: Group-based token pruning transformer for efficient human pose estimation. In *ECCV*, pages 213–230, 2024. 7

[43] Renkai Wu, Pengchen Liang, Yiqi Huang, Qing Chang, and Huiping Yao. Automatic segmentation of hemorrhages in the ultra-wide field retina: multi-scale attention subtraction networks and an ultra-wide field retinal hemorrhage dataset. *IEEE JBHI*, 2024. 2

[44] Dongqing Zhang, Jianing Wang, Jack H Noble, and Benoit M Dawant. Headlocnet: Deep convolutional neural networks for accurate classification and multi-landmark localization of head cts. *Medical Image Analysis*, 61:101659, 2020. 3

[45] Jun Zhang, Mingxia Liu, Li Wang, Si Chen, Peng Yuan, Jianfu Li, Steve Guo-Fang Shen, Zhen Tang, Ken-Chung Chen, James J Xia, et al. Context-guided fully convolutional networks for joint craniomaxillofacial bone segmentation and landmark digitization. *Medical Image Analysis*, 60:101621, 2020. 3

[46] Zhusi Zhong, Jie Li, Zhenxi Zhang, Zhicheng Jiao, and Xinbo Gao. An attention-guided deep regression model for landmark detection in cephalograms. In *MICCAI*, pages 540–548, 2019. 3