Real-Time Ultra-Fine-Grained Surgical Instrument Classification

Md. Atabuzzaman[1] Gino DiMatteo[1] Hani Alomari[1] Chiawei Tang[1] Connor Hale[2]
Adam E. Goode[2] David Ryan King[2] Chris Thomas[1]
[1]Virginia Tech, Blacksburg, Virginia, USA [2]Carilion Clinic, Roanoke, Virginia, USA
{atabuzzaman, ginod25, hani, cwtang, christhomas}@vt.edu
{cehale, aegoode, drking}@carilionclinic.org

Figure 1. Surgical instruments from Eye Vitrectomy (1st row), Major Laparotomy and Minor Laparotomy trays (2nd row). Each image shows a different instrument class, though items within each colored group (green, red, and purple) look nearly identical despite their subtle differences. Data from Hospital 'X' (Carilion Clinic, Roanoke, Virginia, USA) shows roughly 80% of trays have problems like miscounts or wrong instruments, sometimes leading to surgery cancellations when issues are not caught during pre-operative checks.

Abstract

Accurate classification of ultra-fine-grained surgical instruments can significantly reduce the rate of canceled or postponed surgical procedures and improve a hospital's overall operational efficiency. However, accurately classifying these instruments is challenging due to the vast number of surgical instruments in a hospital's Central Sterile Services Department (CSSD) and their ultra-fine-grained distinctions. To address this challenge and assist CSSD technicians, we propose a real-time ultra-fine-grained surgical instrument classification system. Our system consists of a unique open-environment image acquisition platform and multi-view CNN and transformer-based architectures to capture and classify multi-view images of instruments in real-time. We train models on images from three globally recognized surgical trays: Eye Vitrectomy, Major Laparotomy, and Minor Laparotomy, encompassing 95 distinct classes. We evaluate our system in real-time and on image-based datasets, demonstrating state-of-the-art (SoTA) performance. A user study conducted after deployment in a hospital CSSD reveals that the system significantly improves workflow efficiency, streamlining CSSD operations.

1. Introduction

In hospitals, the CSSD plays a crucial role in managing surgical instruments, which are transported in organized trays between operating rooms and the CSSD for sterilization, sorting, and reassembly. A critical responsibility of CSSD technicians is to accurately assemble these trays according to predefined "count" sheets. This is a significantly challenging task, particularly for new staff, and becomes even

more complex when dealing with specialty trays containing uncommon instruments used in procedures such as rib plating, spinal surgery, or ophthalmology. This labor-intensive process requires highly skilled technicians. However, this field struggles with severe staffing shortages. At Hospital X's CSSD, nearly 25% of the positions remain unfilled, and similar situations are reported in other urban hospitals. These staffing shortages have resulted in serious consequences: improperly sterilized instruments, insufficient tray availability, and delays in surgeries and patient care [21].

Manual instrument handling and identification pose significant challenges, largely due to the vast volume of items in the CSSD and their fine-grained to ultra-fine-grained differences (Figure 1). A study at Hospital X found that roughly 80% of trays contain assembly errors that raise patient safety concerns when essential items are missing [13, 29]. A CSSD processes over 100,000 trays and 2.6 million instruments annually [40], with each tray containing an average of 38 instrument categories [28]. To manage this volume of trays, CSSDs need an automated system to enhance accuracy and efficiency in instrument classification.

While coarse-level categorization of surgical instruments is straightforward, the real challenge lies in distinguishing between visually similar instruments that serve different functions. Examples include various types of clamps, scissors, or forceps that differ slightly in design (Figure 1). Recent advancements in computer vision and deep learning, particularly Convolutional Neural Networks (CNNs) [30], have shown significant promise in addressing these challenges. CNNs excel at image recognition tasks, including fine-grained classification of objects with subtle visual differences, making them well-suited for applications in medical imaging and surgical instrument identification [14, 15, 45, 46]. However, applying CNNs to the ultra-fine-grained surgical instrument classification in a real-time, multi-view setting remains under-explored.

Given these technological advancements, a real-time ultra-fine-grained image classification system can transform healthcare delivery by improving the tray sorting process in hospitals' CSSDs. Such a system would assist technicians in accurate instrument identification, reducing errors and enhancing efficiency. This improved accuracy ensures surgeons have precise tools for life-saving surgeries in critical situations. Additionally, it would help new hires quickly learn to recognize ultra-fine-grained, uncommon instruments and assemble trays, reducing supervised training overhead. Beyond patient care, this system could significantly lower assembly costs by enabling efficient instrument identification and organization. This could also facilitate future advances toward fully robotic tray assembly.

In this paper, we propose a system that uses CNN and transformer-based architectures on multi-view images to automatically recognize and classify ultra-fine-grained surgical instruments in real-time. Our system employs a simple yet effective addition-based feature fusion technique that combines complementary features from multiple views to enhance classification accuracy. Specifically, we use late addition feature fusion, where feature vectors extracted from the side and top views of surgical instruments are added and passed through an additional fully connected layer. This fusion strategy leverages complementary information from different views before passing the combined features to the classification layer, improving the model's discriminative power. Our system's hardware consists of a two-camera image acquisition platform for continuous capture, enabling real-time predictions. We also develop a user interface (UI) that allows users to interact with the system by viewing model predictions, retrieving the top three predictions with their confidence scores, counting instruments assembled in trays, and accessing high-resolution images of predicted instruments for verification or training purposes.

For our system, we focus on three surgical trays (Eye Vitrectomy, Major Laparotomy, and Minor Laparotomy) containing fine-grained to ultra-fine-grained surgical instruments (Figure 1). These trays comprise 113 categories, of which 95 are unique and 18 are shared between the Major and Minor trays. We train individual models for each tray and develop a comprehensive model capable of classifying all 95 unique instrument categories. Our dataset includes 50 image pairs per category, captured with variations in orientation, position, and lighting conditions. These surgical instruments follow standardized designs consistent across healthcare facilities worldwide, making our system applicable to CSSDs in various hospitals. When deployed in Hospital X's CSSD, our system demonstrates exceptional performance with accuracy rates consistently exceeding 99.5% in real-time. A user study shows significant improvements in workflow efficiency, indicating its potential to transform CSSD operations and enhance surgical procedure safety.

In this paper, we have the following contributions:

- We propose a real-time ultra-fine-grained surgical instrument classification system with an addition-based feature fusion technique that combines complementary features from multiple views, achieving competitive or better performance than SoTA methods.
- Our system introduces a cost-effective, easy-to-clean open-environment image acquisition platform and a user-friendly interface for CSSD technicians.
- To the best of our knowledge, this is the first real-time ultra-fine-grained multi-view surgical instrument classification system that achieves exceptional accuracy exceeding 99.5% in real-time, real-world CSSD testing.
- We have released a dataset of 10 paired images per instrument class (totaling 950 pairs) to support future research[1].

[1]Datasets: https://githubhtbprolcom s.evpn.librar .nenu.edu.cn/

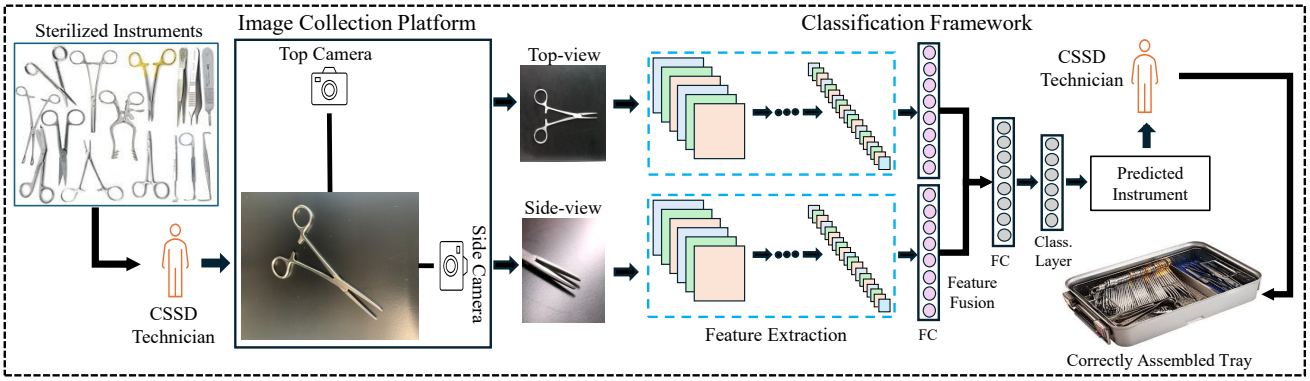

Figure 2. Overview of our proposed real-time ultra-fine-grained surgical instrument classification system. The system consists of two main components: (1) an image collection platform that continuously captures images using dual cameras, and (2) a classification framework that predicts instrument types using multi-view images. The process begins when a CSSD technician places an unknown surgical instrument on the platform, then the system automatically captures images from two different angles. Using the images, the trained model predicts the instrument class. Based on the prediction, the CSSD technician assembles the tray for further sterilization and storage.

2. Related Work

Surgical Instruments Classification

The classification of surgical instruments is crucial due to their role in successful procedures and the increasing complexity of fine-grained tools used in modern surgeries. Recent works have addressed this challenge using machine learning and computer vision techniques [25, 31, 33].

Research on surgical instrument classification has largely focused on laparoscopic procedures for semantic segmentation of video footage [1, 37, 48]. [31] employed a Support Vector Machine with Bag-of-Words features extracted from densely sampled key points in training images. Their study compared the efficacy of three keypoint descriptors: ORB [34], SIFT [27], and SURF [2]. [4] proposed an image-based surgical tool classification method using detected bounding boxes and a cascade of random forest algorithms based on multiple features, including histograms of hue and saturation, gradients, and SURF features [2].

Recent work has incorporated deep learning approaches. [1] introduced a novel neural network framework that incorporates a classification module to enhance the accuracy of instrument mask identification. [24] developed a full-resolution CNN for efficient organ and surgical instrument classification using laparoscopic image datasets.

Several studies have explored various aspects of instrument recognition. [26] developed a camera-based surgical instrument classification system using CNN architecture, employing three cameras to capture different views while processing one image at a time. [17] addressed the challenge of unbalanced data in the publicly available Cholec80 [43] laparoscopy video dataset for classification by implementing multiple data augmentation techniques and a fine-tuned CNN. [22] applied a region-based convolutional neural network (R-CNN) to recognize surgical instruments using a custom dataset generated from laparoscopic gynecological videos. [18] combined a faster R-CNN with VGG16 to detect laparoscopic surgical tools and perform operative skill assessment using the M2CAI dataset [19]. [7] explored a self-supervised method for segmenting surgical instruments in laparoscopic surgery, utilizing the robot's kinematic model as a source of information. While most prior work has focused on instrument segmentation in surgical videos or single-view classification, we propose a real-time ultra-fine-grained surgical instrument classification system using multi-view images, CNN, and transformer-based architectures.

Multi-view Fusion Techniques

Multi-view fusion strategies in deep learning architectures can be categorized based on the stage of feature integration in the network pipeline. Early fusion combines convolutional feature maps from different views before deeper processing, enabling joint representation learning but potentially losing view-specific discriminative features [12]. Late fusion, which processes features independently before integration, has gained popularity due to its effectiveness in preserving view-specific information. Various late fusion methods have been proposed, including concatenation-based approaches [5, 44], pooling operations [11, 41], attention mechanisms [8, 35], and transformer-based fusion [23], though the latter can introduce significant computational overhead. Score fusion, representing the most downstream integration approach, combines predictions at the classifier output level through methods like softmax score averaging [38] and temporal window fusion [20], but may miss important feature interactions that occur at earlier stages.

Our method employs efficient element-wise addition of features from dedicated fully connected layers. Unlike transformer-based feature addition fusion [23] that requires

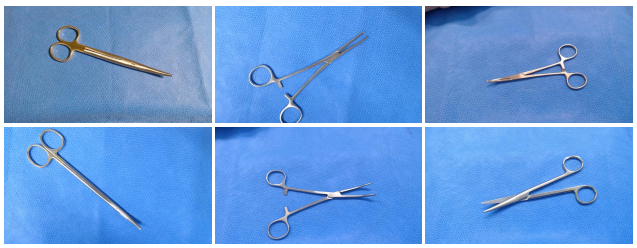

Figure 3. Initial single-view images of surgical instruments from the major laparotomy tray.

additional encoding steps, or concatenation approaches [5] that increase dimensionality, our approach maintains computational efficiency while preserving discriminative spatial information. This makes it particularly suited for fine-grained surgical instrument classification, where both efficiency and detailed feature preservation are essential.

3. Method

Our proposed real-time ultra-fine-grained surgical instrument classification system mainly consists of two components: (i) an image collection platform to continuously collect multi-view images, and (ii) a multi-view image classification framework for real-time predictions. Figure 2 illustrates our proposed system[2].

3.1. Image Collection Platform

We initially collect 35 single-view images per instrument class using a basic imaging platform. Figure 3 shows sample images from the major laparotomy tray, which contains larger instruments compared to the eye vitrectomy tray. Despite using SoTA models like EfficientNet [42] and Vision Transformer (ViT) [9], we achieve only 82% and 74% accuracy, respectively. These suboptimal results motivate us to develop a cost-effective, user-friendly image collection platform equipped with dual UVC cameras to capture multiple views of each instrument. Figure 4 depicts our designed platform, fabricated from solid aluminum extrusion through iterative CAD design and rapid prototyping. Our $18'' \times 6''$ platform features a top overhead camera positioned $14''$ above the center via joined arms and a side camera at the base corner oriented at a $45°$ angle with a slight downward tilt. We engrave a rectangular box to indicate the side-view camera focus area. The blue handler helps to consistently place the instrument's tips in the focus area. The platform accommodates one instrument at a time, with camera positions allowing clear visualization of instruments from both small Eye Vitrectomy tray and larger Major and Minor laparotomy trays (Figure 1). For image classification, instruments are laid flat on the platform base. A Luxonis OAK-D camera with fixed focus and 4K resolution captures an overhead view of the entire instrument. An oblique side-angle profile of the instrument is captured using a 4K USB camera with a fixed short focal length to acquire ultra-fine-grained details of the instrument tips (e.g., curvature, surface texture, and tooth pattern).

[2]We design a UI that could be found in the supplementary material.

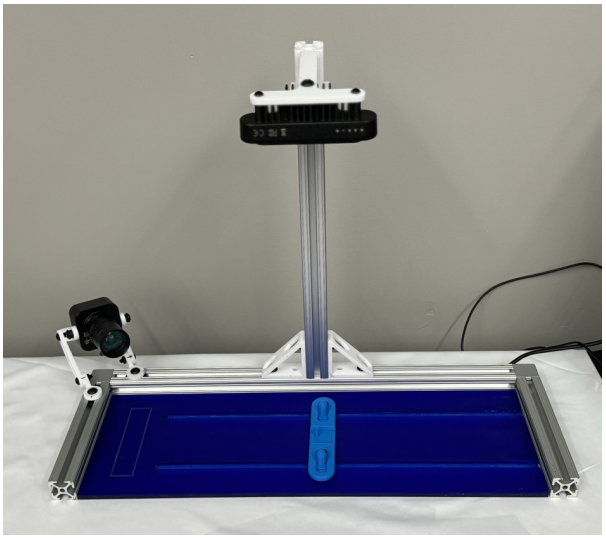

Figure 4. Our proposed image collection platform for real-time ultra-fine-grained surgical instruments classification.

3.2. Multi-view Image Classification Framework

Our proposed ultra-fine-grained surgical instrument classification system uses multi-view CNN architectures (Figure 2) inspired by [3, 10, 36, 41, 44], along with a multi-view ViT architecture [6, 47]. Let I_s and I_t be a multi-view image pair of our dataset D $=\{(I_{si}, I_{ti}), c_i\}_{i=1}^{N}$ of size N, comprising side-view (I_s), top-view (I_t) images, and category label c. We extract the features (f_s and f_t) of I_s and I_t separately using identical architectures ($M_{cnn/vit}$), including EfficientNet [42], ResNet50 [14], VGG16 [39], and ViT [9] (Eq. 1 & 2). Each model processes the input image pair independently, extracting features (f_s and f_t) using the CNN or ViT architecture. These features pass through respective fully connected (L_{FC}) layers within each model branch (Eq. 3 & 4). To fuse the feature vectors (f_{sv} and f_{tv}) from both views, we employ late feature fusion techniques, specifically concatenation as motivated by [3, 5, 44]. Besides this feature concatenation fusion, we employ a feature addition fusion technique (Eq. 5) for our proposed system inspired by [23]. In concatenation fusion, feature vectors are concatenated end-to-end into a single larger vector. This concatenated vector is then passed through a fully connected layer (L_{fusion}). In addition fusion, feature vectors are first summed element-wise into a same-size vector before passing through the L_{fusion} layer. This late-stage feature fusion leverages complementary information from different views, enhancing the discriminative power of the

combined feature representation [3, 32, 36, 41]. The fused feature vector f_v feeds into a classification layer L_{CLS}, which produces logits P_{logits} (Eq. 6). These logits are converted into class probabilities P_{class} using the softmax activation, enabling the prediction of the instrument class (Eq. 7). Mathematically, we can express these as follows:

$$f_s = M_{\text{cnn/vit}}(I_s) \tag{1}$$
$$f_t = M_{\text{cnn/vit}}(I_t) \tag{2}$$
$$f_{sv} = L_{FC}(f_s) \tag{3}$$
$$f_{tv} = L_{FC}(f_t) \tag{4}$$
$$f_v = L_{\text{fusion}}(f_{sv} + f_{tv}) \tag{5}$$
$$P_{\text{logits}} = L_{CLS}(f_v) \tag{6}$$
$$P_{\text{class}} = \text{softmax}(P_{\text{logits}}) \tag{7}$$

The cross-entropy loss $\mathcal{L}_{CE}$ between the predicted class probabilities P_{class} and the true label y is defined as:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{C} y_c \log(P_{class,c}) \tag{8}$$

where C is the total number of classes, and y_c is the binary indicator (0 or 1) for class c. The predicted class for the instrument pair can then be obtained by selecting the class $\hat{c}$ with the highest probability:

$$\hat{c} = \arg\max_{c} P_{class,c} \tag{9}$$

We deploy the trained models for real-time prediction in the CSSD of Hospital X.

4. Experiments and Evaluations

We evaluate our proposed ultra-fine-grained surgical instrument classification system on three surgical instrument trays containing instruments ranging from small to large with varying degrees of granularity (Figure 1).

4.1. Datasets Collection

We collect the dataset by systematically acquiring images using our image acquisition platform and three surgical trays: Eye Vitrectomy, Major Laparotomy, and Minor Laparotomy. The Eye Vitrectomy tray contains small, ultra-fine-grained surgical instruments (Figure 1, 1st row), while the Major and Minor trays contain relatively larger, fine-grained to ultra-fine-grained surgical instruments (Figure 1, 2nd row). Table 1 presents the dataset statistics and train-test distribution used in our experiments [3].

To ensure system robustness across varying lighting conditions and address the challenges of controlling environmental lighting, we conduct image acquisition in an open environment. For each of the 95 unique surgical instrument categories, we collect 50 image pairs, totaling 4,749 pairs (one category in the Eye Vitrectomy tray has 49 pairs). When multiple instances of the same instrument category are present in a tray, we include all within their respective 50 image-pair set. Our platform's dedicated focus area ensures consistent and optimal instrument positioning, while we introduce controlled variations by slightly adjusting instrument placement across images. We further enhance dataset diversity by capturing images under different lighting conditions, including shadows induced by a large board and additional illumination from a flashlight. This systematic approach to data collection, combining controlled positioning with environmental variations, enables robust model training while maintaining high recognition accuracy for ultra-fine-grained surgical instruments.

Tray Type	# Categories	# Instruments	Training	Testing	Total
Eye	41	58	1536	513	2049
Major	34	91	1275	425	1700
Minor	38	89	1425	475	1900
# Total	113	238	-	-	5649
# Unique	95	238	3561	1188	4749

Table 1. Dataset statistics for the Eye, Major, and Minor Laparotomy Trays. Each instrument category contains 50 image pairs, with one exception in the Eye tray. The "# Instruments" column indicates the total count of distinct instruments per tray across all categories, including manufacturer variations. The "Training" and "Testing" columns show the distribution of image pairs for model development and evaluation. While the total "# Categories" is 113, the Major and Minor Laparotomy trays share 18 common categories, resulting an overall of 95 unique categories.

4.2. Experimental Setups

To validate our ultra-fine-grained surgical instrument classification system, we conduct experiments using side-view, top-view, and multi-view settings at different image resolutions. For single-view settings (side-view and top-view), we use single-view models (e.g., traditional EfficientNet, ResNet50, etc.). For multi-view settings, we apply these models with feature fusion from separate views. For simplicity, we retain the single-view model names (e.g., EfficientNet, ResNet50, etc.) to refer to our different multi-view instrument classification architectures.

We utilize pre-trained weights from the torchvision library[4] for three CNN-based architectures (EfficientNet, ResNet50, VGG16) and from the timm library[5] for the transformer-based ViT-B/16. We train all models using SGD optimizer with a learning rate of 0.001 and Cross Entropy Loss. With 50 image pairs per class, models reach optimal performance within 15 epochs of the total 20 epochs. For reduced datasets (10 or 20 pairs per class), while EfficientNet and ViT-B/16 converge quickly, ResNet50 requires 60 epochs. We train the models on an NVIDIA A40 GPU,

[3] Surgical instruments' names are in the supplementary material.

[4] https://pytorch.org/vision/stable/models.html

[5] https://pypi.org/project/timm/

Tray Type	Model	Side View	Top View	Add Fusion	Cat Fusion
Eye	EfficientNet [42]	93.18	91.81	94.93	91.62
	ResNet50 [14]	94.93	94.35	**96.10**	90.45
	VGG16 [39]	90.06	**94.54**	88.69	**93.57**
	ViT-B/16 [9]	**95.52**	93.96	**96.10**	92.98
Minor	EfficientNet [42]	83.58	90.45	94.11	96.21
	ResNet50 [14]	**92.21**	**97.26**	**99.79**	**98.95**
	VGG16 [39]	88.42	96.42	96.42	95.79
	ViT-B/16 [9]	89.26	95.58	**98.32**	98.32
Major	EfficientNet [42]	92.00	94.82	97.41	97.88
	ResNet50 [14]	93.41	**99.29**	**99.76**	**99.53**
	VGG16 [39]	95.06	97.88	98.82	97.88
	ViT-B/16 [9]	**97.18**	97.65	99.53	**99.53**

Table 2. Performance comparison across surgical instrument trays. The evaluation includes side-view only, top-view only, and multi-view settings, with multi-view using either addition-based (Add) or concatenation-based (Cat) feature fusion. All experiments use 224×224 image resolution. **Bold** values indicate the best performance within each setting for each tray, while highlighted cells show the best performance across settings, demonstrating the effectiveness of our employed Add-feature fusion technique.

and at inference, they require approximately 5 GB of GPU memory for a 1000×1000 image pair. Since our dataset is balanced across all classes, we use accuracy as our primary evaluation metric to assess model performance.

4.3. Results and Discussion

We evaluate the models using three configurations: side-view images only, top-view images only, and multi-view (combined views). For multi-view evaluation, we employ two feature fusion techniques: late concatenation (Cat) and our introduced late addition (Add).

4.3.1. Models' Performance across Different Settings

Table 2 presents the performance of our used four models across three surgical tray images at 224×224 resolution.

Eye Vitrectomy Tray. ResNet50 and ViT-B/16 achieve the highest accuracy of 96.10% with Add fusion. While VGG16 excels in top-view (94.54%), it shows lower performance in side-view (90.06%) and Add fusion (88.69%). EfficientNet demonstrates consistent performance across settings, peaking at 94.93% with Add fusion. Notably, Cat fusion shows lower performance compared to Add fusion, with ResNet50's accuracy decreasing by 5.65% to 90.45%.

Minor and Major Laparotomy Trays. ResNet50 maintains superior performance, achieving the highest accuracies with Add fusion in both Minor (99.79%) and Major (99.76%) Laparotomy trays. ViT-B/16 closely follows in the Major tray, reaching 99.53% in both fusion settings. While EfficientNet shows lower side-view accuracies (83.58% and 92.00%), it improves significantly with Add fusion (94.11% and 97.88%).

Overall Findings. The results demonstrate ResNet50's superior performance across all trays and fusion techniques, particularly with Add fusion at 224×224 resolution. ViT-B/16 shows promising performance, while EfficientNet and VGG16, though competent, generally lag behind. As shown in the highlighted cells of Table 2, our employed Add fusion consistently outperforms Cat fusion [5, 41] across all trays, establishing its effectiveness for feature fusion.

Tray Type	Model	Image resolution		
		224x224	384x384	512x512
Eye	EfficientNet	94.93	97.08	**99.81**
	ResNet50	96.10	98.05	95.13
	VGG16	88.69	96.10	92.79
	ViT-B/16	96.10	96.10	97.08
	MV-HFMD [3]	**96.88**	**98.64**	98.05
Minor	EfficientNet	94.11	98.32	**99.58**
	ResNet50	**99.79**	99.16	98.74
	VGG16	96.42	97.05	96.84
	ViT-B/16	98.32	99.16	**99.58**
	MV-HFMD [3]	96.84	**99.37**	98.95
Major	EfficientNet	97.41	99.53	99.76
	ResNet50	**99.76**	99.76	99.76
	VGG16	98.82	99.53	99.29
	ViT-B/16	99.53	**100.00**	99.76
	MV-HFMD [3]	99.53	100.00	**100.00**
All-trays	EfficientNet	94.95	96.89	98.82
	ResNet50	95.37	97.05	97.22
	VGG16	92.46	92.26	93.52
	ViT-B/16	96.89	**99.24**	**99.16**
	MV-HFMD [3]	**97.81**	**99.24**	97.64

Table 3. Accuracy comparison of different models with varying image sizes on Eye, Minor, Major trays, and all trays together. All models employed the late addition feature fusion technique, utilizing both side-view and top-view (multi-view) images. The results show performance improvements as the image size increased from 224×224 to 384×384. However, at 512×512, EfficientNet constantly achieves better accuracy than at the 384×384 image size. The best results are marked in **bold**.

4.3.2. Model Performance across Image Resolutions

While the models perform efficiently in image-based testing, they struggle to correctly classify ultra-fine-grained instruments in real-time open-environment settings with 224×224 image resolution due to insufficient instrument detail. To address this limitation, we conduct experiments using higher image resolutions of 384×384 and 512×512. Table 3 presents the performance comparison of four models across different image resolutions using late addition feature fusion for the trays. Higher resolutions generally improve model performance due to better instrument detail capture, crucial for ultra-fine-grained classification.

EfficientNet demonstrates consistent improvement with

Tray Type	Model	# Image-pair/class			
		10	20	35	50
Eye	EfficientNet	**81.55**	**96.10**	**95.82**	97.08
	ResNet50	72.82	92.68	93.87	**98.05**
	ViT-B/16	72.82	95.12	95.12	96.10
Minor	EfficientNet	**85.26**	94.21	98.20	98.32
	ResNet50	77.89	88.42	98.50	**99.16**
	ViT-B/16	84.21	**96.32**	**99.70**	**99.16**
Major	EfficientNet	78.82	94.12	96.98	99.53
	ResNet50	75.29	97.06	97.65	99.76
	ViT-B/16	**85.88**	**98.24**	**99.66**	**100.00**

Table 4. Results demonstrate the effect of varying the number of image pairs per category on models' performance. For these experiments, we used an image resolution of 384×384, as most models demonstrated optimal performance at this resolution.

increased resolution, achieving peak accuracies of 99.81% and 99.58% at 512×512 for Eye and Minor trays, respectively. ResNet50 and ViT-B/16 show optimal performance at 384×384, with slight degradation at 512×512. VGG16 performs best at 384×384 before declining at higher resolutions, indicating sensitivity to larger image sizes. For all 95 unique categories (All-trays), ViT-B/16 achieves the highest accuracy of 99.24% at 384×384 resolution, closely followed by EfficientNet reaching 98.82% at 512×512. Compared to the current SoTA MV-HFMD [3], which utilizes hybrid fusion techniques (concatenation and score fusions), our system demonstrates superior performance with high-resolution images across most settings. The only exception is in Major laparotomy classification, where MV-HFMD achieves 100.00% accuracy at 512×512 resolution compared to our system's 99.76%.

Therefore, we train EfficientNet with a higher image resolution of 1000×1000 and with our employed addition fusion technique. We perform real-time testing of our system with the trained EfficientNet and find that with higher-resolution (1000×1000) images, EfficientNet accurately classifies all the ultra-fine-grained surgical instruments in the open-environment with an inference rate of 7 image pairs/second. Based on these results, we deploy our system for real-time ultra-fine-grained surgical instrument classification using EfficientNet with late addition fusion.

4.3.3. Impact of Training Data Volume

While our system uses 50 image pairs per class, capturing this volume of data is labor-intensive and challenging, given the thousands of instruments in a hospital setting. To evaluate model performance with varying data volumes, we conduct experiments using different numbers of image pairs per class. Table 4 presents these results, showing a clear correlation between accuracy and training data volume.

With only 10 image pairs per class, all models demonstrate lower performance across the surgical trays: EfficientNet achieves 81.55%, 85.26%, and 78.82% accuracy for Eye, Minor, and Major trays, respectively. Increasing to 50 pairs per class significantly improves performance, with ViT-B/16 achieving 100% accuracy for the Major tray. **While these high accuracies might suggest the task is straightforward, it's important to note that they are attributed to our controlled image acquisition protocol (Section 3.1).** Our image collection platform constrains instruments to specific focus areas, significantly reducing variations in position and angle that typically challenge fine-grained classification tasks. This standardized capture process, while enabling reliable instrument identification, also means our models may require adaptation for more variable real-world scenarios where instruments might appear in diverse orientations and environments. These results highlight both the critical role of training data volume in ultra-fine-grained surgical instrument classification and the models' limited generalization capability with reduced data.

5. Error Analysis

We present the misclassified cases in Figure 5 to analyze where the models misclassified instruments. The first and second rows show results when considering all 95 categories for ultra-fine-grained classification, while the last row shows results when considering the major and minor trays separately. Figure 5 highlights that the models encountered the most difficulty in classifying small instruments from the Eye Vitrectomy tray in ultra-fine-grained classification. Some misclassifications occurred when one view of the instrument was not visible to the model, while others were due to similar image pairs in the training set for the predicted classes—a challenge inherent to the ultra-fine-grained nature of the dataset. Notably, all models correctly classified instruments at a coarse or fine-grained level but misclassified them at the ultra-fine-grained level.

5.1. User Study

We evaluated our system's integration into CSSD technicians' workflow through a study with six participants from Hospital X, having experience ranging from 6 months to 10 years. After a brief introduction, participants assembled a mock tray containing Eye Vitrectomy and Laparotomy instruments while incorporating our system as they deemed appropriate. They verbalized their thoughts during the recorded assembly process. Afterward, participants completed a System Usability Scale (SUS) [16] questionnaire and engaged in discussions with our team of human factors and biomedical engineers. The discussions explored potential adoption, workspace compatibility, and use cases, focusing on whether technicians would use the system for all instruments or only unfamiliar ones.

On the standard scale of 0 to 100, our system received

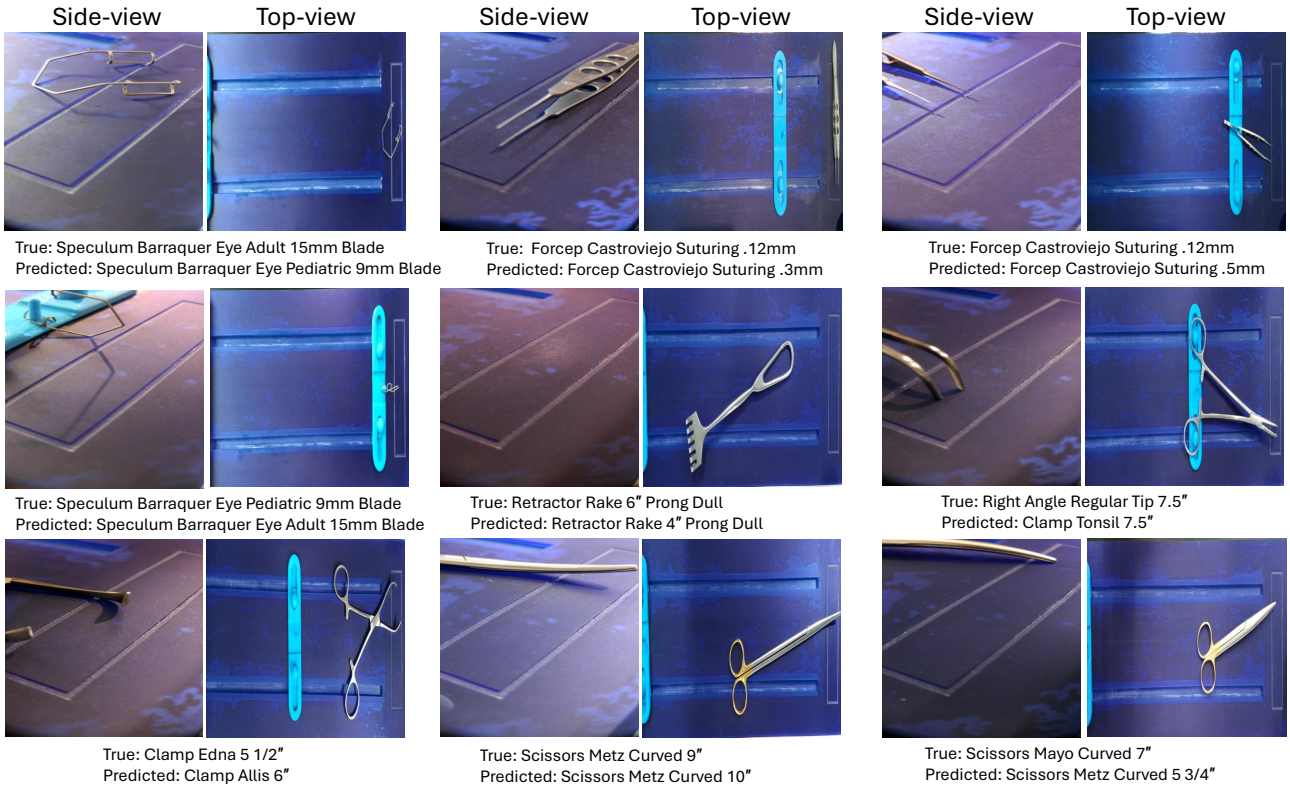

Figure 5. Illustration of misclassified surgical instruments when analyzing all 95 categories of instruments simultaneously (All-tray setting). The first two rows show examples from this complete analysis, while the last row shows results when major and minor trays are analyzed separately. The first two rows demonstrate that the models particularly struggle with distinguishing between ultra-fine-grained surgical instruments from the eye tray.

a SUS score of 81.67 (SD = 14.29). This score verifies that our system is feasible and would not be disruptive to technicians' current processes. Rather, the system would likely help accelerate and improve the accuracy of their work. While most participants expressed willingness to use this device for all instruments, some indicated they would primarily use it only when they were uncertain about the instrument's identity. We anticipate that implementing a more professional user interface will encourage broader adoption across all surgical trays, as the system proves more efficient than manual identification and assembly.

6. Conclusion

In this paper, we propose a real-time ultra-fine-grained surgical instrument classification system for deployment in the CSSD of hospitals. To enable this ultra-fine-grained classification, we employ an effective addition-based feature fusion technique that outperforms SoTA methods. We collected a comprehensive dataset of 4,749 multi-view image pairs representing 95 categories of surgical instruments. Using multi-view EfficientNet architecture with high-resolution image pairs, our system achieved near-perfect accuracy (>99.5%) in real-time testing within the CSSD of Hospital X. Our system assists CSSD technicians in accurately identifying and assembling surgical instruments, reducing errors and enhancing efficiency. This ensures that surgeons have the precise tools needed for life-saving surgeries. Beyond improving healthcare delivery, our system can significantly reduce tray assembly costs through quicker instrument identification. Our system also facilitates future advancements toward fully robotic tray assembly and robotic surgery.

7. Acknowledgments

This project was supported in part by the Virginia Tech College of Science Academy of Data Science Discovery Fund (ADSDF) (Award: 238695) and the Research Acceleration Program (RAP) grant at Carilion Clinic, Roanoke, Virginia, USA. We extend our sincere gratitude to the anonymous reviewers and the CSSD team at Carilion Clinic, Roanoke, for their valuable feedback and constructive suggestions.

References

[1] Britty Baby, Daksh Thapar, Mustafa Chasmai, Tamajit Banerjee, Kunal Dargan, Ashish Suri, Subhashis Banerjee, and Chetan Arora. From forks to forceps: A new framework for instance segmentation of surgical instruments. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6191–6201, 2023. 3

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 404–417. Springer, 2006. 3

[3] Samuel Black and Richard Souvenir. Multi-view classification using hybrid fusion and mutual distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 270–280, 2024. 4, 5, 6, 7

[4] Sebastian Bodenstedt, Antonia Ohnemus, Darko Katic, Anna-Laura Wekerle, Martin Wagner, Hannes Kenngott, Beat Müller-Stich, Rüdiger Dillmann, and Stefanie Speidel. Real-time image-based instrument classification for laparoscopic surgery. *arXiv preprint arXiv:1808.00178*, 2018. 3

[5] Gustavo Carneiro, Jacinto Nascimento, and Andrew P Bradley. Deep learning models for classifying mammogram exams containing unregistered multi-view images and segmentation maps of lesions. *Deep learning for medical image analysis*, pages 321–339, 2017. 3, 4, 6

[6] Shuo Chen, Tan Yu, and Ping Li. Mvt: Multi-view vision transformer for 3d object recognition, 2021. 4

[7] Cristian da Costa Rocha, Nicolas Padoy, and Benoit Rosa. Self-supervised surgical tool segmentation using kinematic information. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8720–8726. IEEE, 2019. 3

[8] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3560–3569, 2021. 3

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 6

[10] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016. 4

[11] Y Feng, Z Zhang, X Zhao, R Ji, Y Gao, and Gvcnn. Group-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 264–272, 2018. 3

[12] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd international conference on information fusion (FUSION)*, pages 1–6. IEEE, 2020. 3

[13] Annetje CP Guédon, Linda SGL Wauben, Anne C van der Eijk, Alex SN Vernooij, Frédérique C Meeuwsen, Maarten van der Elst, Vivian Hoeijmans, Jenny Dankelman, and John J van den Dobbelsteen. Where are my instruments? hazards in delivery of surgical instruments. *Surgical endoscopy*, 30:2728–2735, 2016. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 4, 6

[15] Xiangteng He and Yuxin Peng. Fine-grained image classification via combining vision and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5994–6002, 2017. 2

[16] Maciej Hyzy, Raymond Bond, Maurice Mulvenna, Lu Bai, Alan Dix, Simon Leigh, Sophie Hunt, et al. System usability scale benchmarking for digital health apps: meta-analysis. *JMIR mHealth and uHealth*, 10(8):e37290, 2022. 7

[17] Jaafar Jaafari, Samira Douzi, Khadija Douzi, and Badr Hssina. Towards more efficient cnn-based surgical tools classification using transfer learning. *Journal of Big Data*, 8(1): 115, 2021. 3

[18] Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 691–699. IEEE, 2018. 3

[19] Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 691–699. IEEE, 2018. 3

[20] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5492–5501, 2019. 3

[21] Elise Keseler. The hospital logistics of designing the central sterile processing department. Master's thesis, NTNU, 2019. 2

[22] Sabrina Kletz, Klaus Schoeffmann, Jenny Benois-Pineau, and Heinrich Husslein. Identifying surgical instruments in laparoscopy using deep learning instance segmentation. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2019. 3

[23] Jannik Koch, Stefan Wolf, and Jürgen Beyerer. A transformer-based late-fusion mechanism for fine-grained object recognition in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 100–109, 2023. 3, 4

[24] Praveen SR Konduri and G Siva Nageswara Rao. Full resolution convolutional neural network based organ and surgical instrument classification on laparoscopic image data. *Biomedical Signal Processing and Control*, 87:105533, 2024. 3

[25] Thomas Kurmann, Pablo Márquez-Neila, Max Allan, Sebastian Wolf, and Raphael Sznitman. Mask then classify: multi-instance segmentation for surgical instruments. *International journal of computer assisted radiology and surgery*, 16(7): 1227–1236, 2021. 3

[26] Jan Lehr, Kathrin Kelterborn, Clemens Briese, Marian Schlueter, Ole Kroeger, and Joerg Krueger. Image-based recognition of surgical instruments by means of convolutional neural networks. *International journal of computer assisted radiology and surgery*, 18(11):2043–2049, 2023. 3

[27] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 3

[28] Julie M Mhlaba, Emily W Stockert, Martin Coronel, and Alexander J Langerman. Surgical instrumentation: the true cost of instrument trays and a potential strategy for optimization. *J Hosp Adm*, 4(6):82–88, 2015. 2

[29] Peter F Nichol and Mark J Saari. Risk modeling of errors in the surgical instrument cycle, insights into solutions for an expensive and persistent problem. *Perioperative Care and Operating Room Management*, 32:100333, 2023. 2

[30] K O'Shea. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015. 2

[31] Manfred Jürgen Primus, Klaus Schoeffmann, and Laszlo Böszörmenyi. Instrument classification in laparoscopic videos. In *2015 13th international workshop on content-based multimedia indexing (CBMI)*, pages 1–6. IEEE, 2015. 3

[32] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2016. 5

[33] Mark Rodrigues, Michael Mayo, and Panos Patros. Surgical tool datasets for machine learning research: a survey. *International Journal of Computer Vision*, 130(9):2222–2248, 2022. 3

[34] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 3

[35] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Cross-modal hierarchical modelling for fine-grained sketch based image retrieval. *arXiv preprint arXiv:2007.15103*, 2020. 3

[36] Marco Seeland and Patrick Mäder. Multi-view classification with convolutional neural networks. *Plos one*, 16(1): e0245230, 2021. 4, 5

[37] Yuyang Sheng, Sophia Bano, Matthew J Clarkson, and Mobarakol Islam. Surgical-desam: decoupling sam for instrument segmentation in robotic surgery. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–5, 2024. 3

[38] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 3

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4, 6

[40] Emily Walker Stockert and Alexander Langerman. Assessing the magnitude and costs of intraoperative inefficiencies attributable to surgical instrument trays. *Journal of the American College of Surgeons*, 219(4):646–655, 2014. 2

[41] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 3, 4, 5, 6

[42] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 4, 6

[43] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36 (1):86–97, 2016. 3

[44] Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen Cham. Mmss: Multi-modal sharable and specific feature learning for rgb-d object recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1125–1133, 2015. 3, 4

[45] Yafei Wang and Zepeng Wang. A survey of recent work on fine-grained image classification techniques. *Journal of Visual Communication and Image Representation*, 59:210–214, 2019. 2

[46] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8927–8948, 2021. 2

[47] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3333–3343, 2022. 4

[48] Zijian Zhou, Oluwatosin Alabi, Meng Wei, Tom Vercauteren, and Miaojing Shi. Text promptable surgical instrument segmentation with vision-language models. *Advances in Neural Information Processing Systems*, 36:28611–28623, 2023. 3