# A Robust Multitask Model for Surgical Skill Evaluation

Orr Zohar
Department of Electrical Engineering
Stanford University
orrzohar@stanford.edu

## Abstract

*The development of advanced computer vision applications for open surgical procedures is plagued by the lack of annotated datasets and the difficulty of their creation. The first dataset of partially-annotated open surgical videos was recently curated from YouTube by E. D. Goodman* et al. *With it, the first multitask model trained for hand/tool detection and action segmentation on open surgical videos was also reported. However, the deployment of this multitask model on a new open-surgical dataset has shown drastically reduced performance, hindering the progression of further developments and clinical research. Here, we perform a root-cause analysis of this performance degradation and ultimately determine that the decrease in performance is due to resolution, zoom, and contrast shifts between the target and source data domains. We then retrain the model on the original dataset but with a more robust data augmentation procedure - improving model robustness to such changes. Finally, we evaluate the inference quality before/after our improvements on our novel dataset and have determined that our detection accuracy is sufficient to continue with clinical trials.*

## 1. Introduction

Artificial Intelligence (AI) and Computer Vision (CV) techniques can provide a scalable and quantitative method of surgical skill evaluation from videos of open surgical procedures [21]. However, the development of such techniques is plagued by the lack of accessible annotated surgical videos and the difficulty of acquiring and annotating such videos. Recently, E. D. Goodman *et al*. [6] have compiled 1997 videos from 23 surgical procedures on YouTube and partially annotated them to produce the first Annotated Videos of Open Surgery (AVOS) dataset [6]. In their work, they reported on a custom multitask model based on a Retinanet backbone [9] to perform both tool/hand detection and action segmentation. They then show that hand-crafted metrics, such as decisiveness and economy of motion, are correlated with the surgeons' skill.

Surgical complications are the third leading cause of death globally [12], with open procedures being the majority of surgical procedures performed overall [14]. Thus far, several studies have shown (unsurprisingly) that higher-skill considered surgeons have lower rates of surgical complications and surgically-related deaths [2, 15]. Therefore, quantitatively evaluating surgical skills could provide a valuable tool for the improvement of surgical techniques while reducing the large variation found in the field. As of yet, few works have demonstrated the capability of quantitatively evaluating surgical skill [6]. In collaboration with Harvard Medical School (HMS), we collected 250 videos of different surgical procedures and the resulting clinical outcome - specifically, when/if the patient's surgical site infections. However, deploying E. D. Goodman etal's model directly on this HMS dataset showed drastically decreased performance, impeding the estimation of the metrics mentioned above (decisiveness and economy of motion). In this work, we evaluate and improve the robustness of this multitask model (see figure 2) by utilizing both an image augmentation pipeline and test a novel "multiscale" convolutional layer design. We then retrain the model with and without this multiscale convolution layer on videos of open surgical procedures, producing both object detection and action segmentation.

## 2. Related Works

The development of machine learning based algorithms for the analysis of videos of open surgical procedures is hindered by the lack of annotated datasets, as well as the challenging nature of such videos with a high rate of occlusion, varying lighting, zoom, and the use of different equipment. Nonetheless, there have been a few works that reported on the evaluation of surgical skills from video data [5,6,21,23]. One of the first works [5] in this field attempted to detect both tools and tools-in-hands using an augmented YOLO architecture. They developed a "variable tissue simulator" setup to neutralize much of the variance, where different tissue-like materials were embedded into a static worksta-

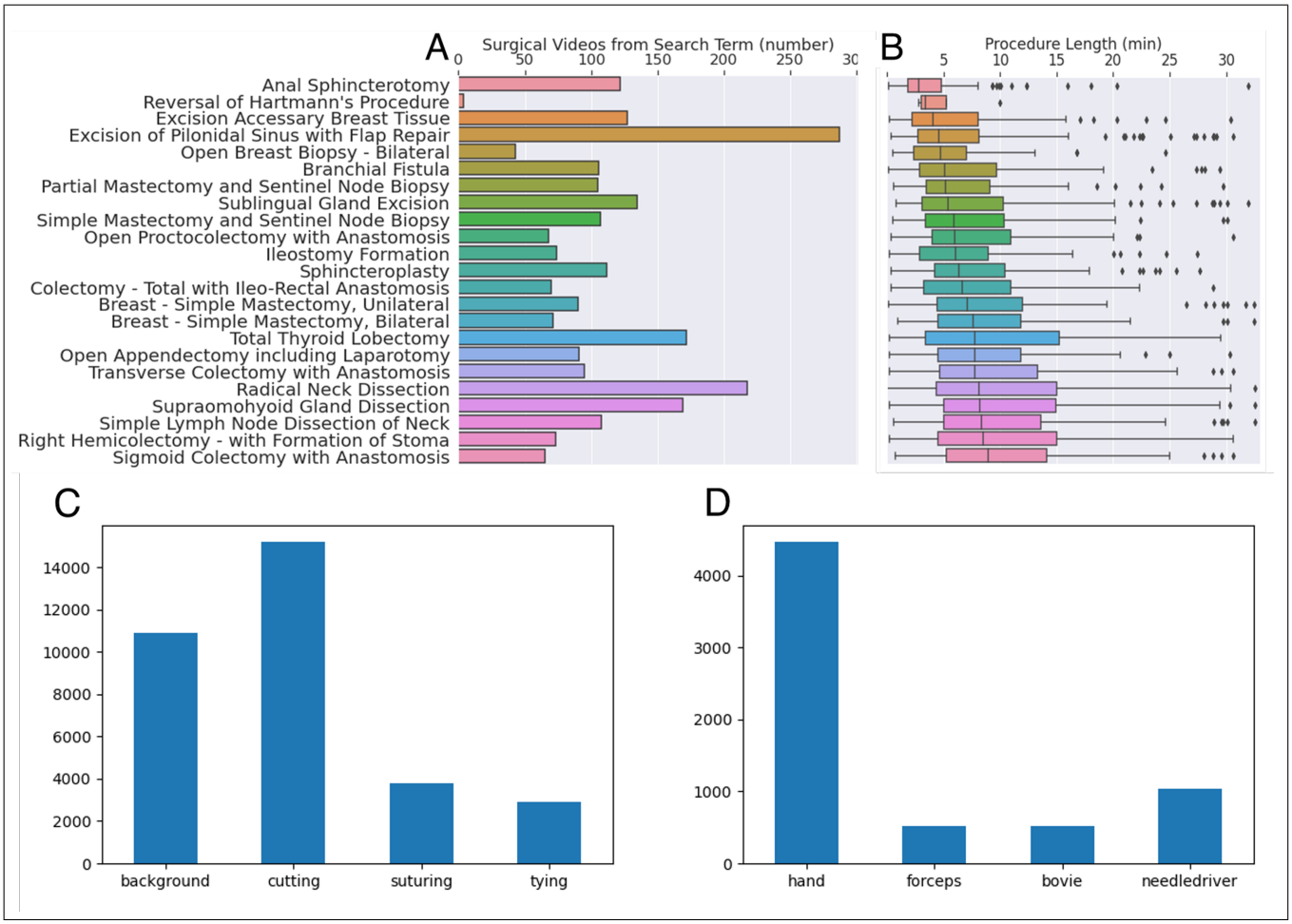


Figure 1. AVOS dataset. (A) Distribution of video counts (B) Distribution of video duration (C) Distribution of action labels (D) Distribution of object labels

tion. Using this setup, 11 medical students, one resident, and 13 attending surgeons performed four tasks collecting 100 videos of length 2-6 min. They then used the reported augmented YOLO algorithm and achieved a tool, and hand mAP@0.5 of 0.871 [5]. Unlike previous works, Zhang *et al*. collected a large dataset of videos from youtube of open surgical procedures and annotated them for training their model [21]. They then used the RetinaNet model and pre-trained it for hand detection on the EgoHands and Oxford(hands) datasets [1, 11] before performing transfer learning onto their surgical dataset. They also incorporated hand tracking. E. D. Goodman *et al*. [6] took this model one step further and attached an action detection head, which allowed them to perform also action segmentation (after adding action segmenation labeling to the AVOS dataset).

Thus far, there have been several works investigating both the evaluation [3, 7, 19, 20] and improving [4, 8, 10, 17, 18, 22] the robustness of neural network models. An interesting approach for improving model robustness is through the utilization of data augmentation [13]. In their work, Rebuffi *et al*. showed that even simple data augmentation could significantly improve model robustness - ultimately improving model generalization, resulting in increased validation set performance. In this spirit, Hendrycks *et al*. developed a new and novel image augmentation method, Augmix, where the image undergoes (randomly) different augmentations, which are averaged into the output image. Other examples of image augmentation resulting in improved model robustness come from the automotive field, where image augmentations are used to simulate different weather conditions. In their work, Michaelis *et al*. [10] used style transfer to augment their images with “winter conditions” such as snow/frost/fog. Training a model with these augmentations had improved generalization for winter scenes, which were not represented well in their source dataset. In another example, Tran *et al*. [17] had access only to daytime-driving

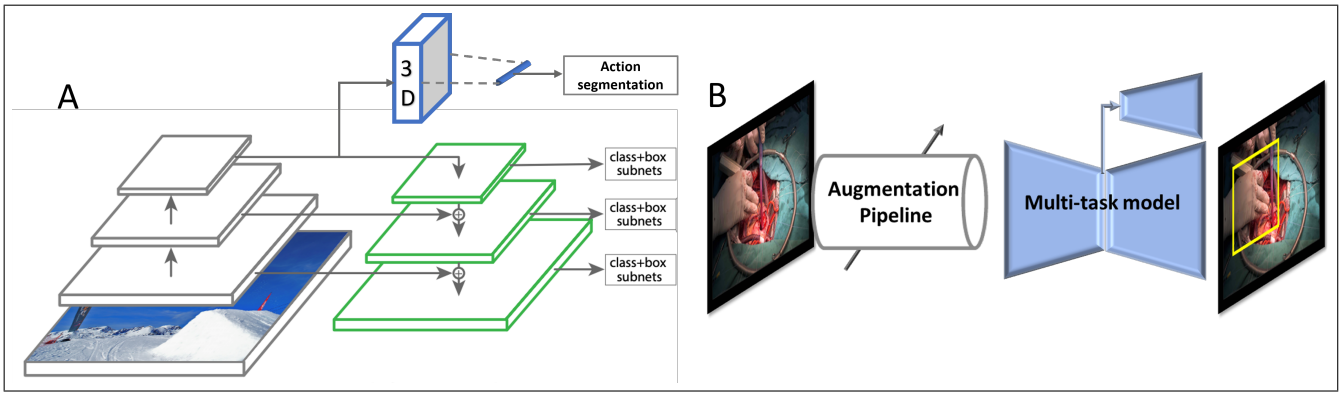


Figure 2. (A) Multitask model architecture. (B) Experimental setup for evaluating the effect of augmentations on model performance.

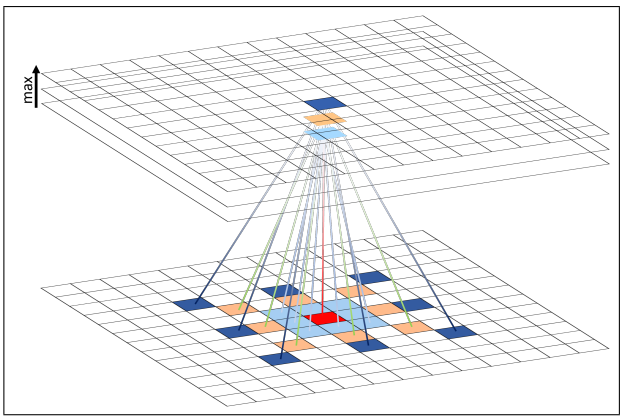


Figure 3. Multiscale convolution

labeled data. To improve "nighttime" performance, they used a novel augmentation method to effectively darken the image, more closely approximating nighttime driving. By varying (randomly) this augmentation during training, their YOLO-based model was less affected by the lighting conditions.

## 3. Data

The Annotated Videos of Open Surgery (AVOS) dataset contains 1997 manually identified videos of 23 standard and common open surgical procedures from YouTube (see figure 1 A, B). Of these 1997 videos, 343 videos were spatiotemporally annotated. Specifically, ten frames were annotated from each video with bounding boxes of hands and surgical tools (electrocautery, needle drivers, and forceps). The distribution of these labels can be seen in figure 1 D, where a clear bias towards the hand-class can be seen. Temporally, each video was labeled for action segmentation at a per-second resolution with either cutting, tying, suturing, or background labels (see figure 1 C). We split the 343 labeled videos into a test/val/train of 55/41/247. Ten frames were annotated for object detection and second-scale action labeling for each video. The pre-processing consists of only a channel-wise mean-subtraction. The frames are resized to a minimum dimension of 320 and max dimension of 1024 and down-sampled to 30 fps.

The Harvard Medical School (HMS) dataset contains 250 videos of different surgical procedures and the resulting clinical outcome. Unlike most videos in the AVOS dataset, it was collected using a static camera setup where a camera was placed directly above the surgical table. We randomly selected 57 videos from the HMS dataset and labeled ten frames from each video (a total of 570 annotated frames). As the data is for evaluating the effects of surgical skill on surgical site infections, only the closing was recorded, and therefore the "bovie" tool is unused - and no annotations of this tool exist in the dataset. Unfortunately, due to patient privacy concerns, we are not yet able to display any images from this dataset. Currently, we are undergoing a de-identification process to enable wider access to this data in the future.

## 4. Methods

The domain shift in surgical videos can largely be attributed to recording quality and type differences. Most of the videos in the AVOS dataset were captured using a head-mounted camera, while the HMS dataset was captured using a stationary overhead camera. As a result, the HMS videos are substantially more zoomed out than those found in AVOS. Furthermore, it is apparent that the lighting is quite different, with our videos appearing somewhat more "whitened out". To evaluate model robustness, we will perform different data augmentations and study their effect on mAP on the AVOS validation dataset.

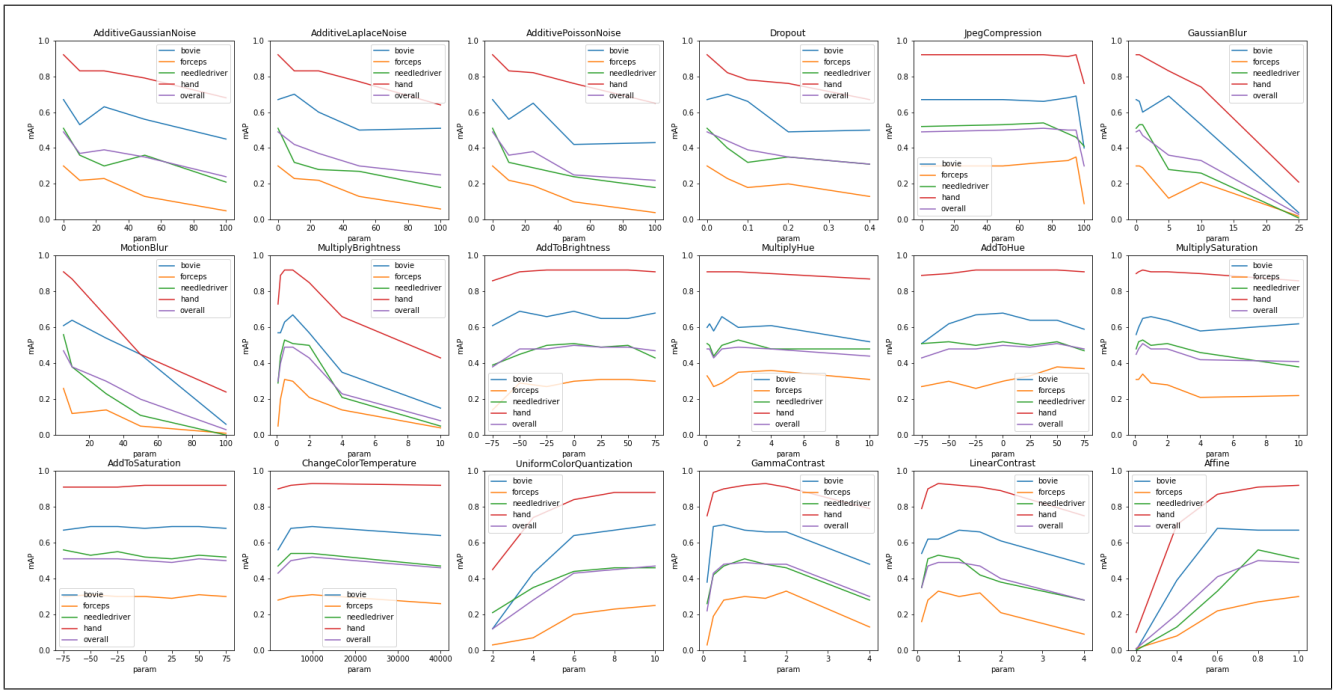


Figure 4. mAP vs. augmentation parameter for 18 different augmentations

### 4.1. Multitask Model

We used a pre-trained Multitask model [6], which utilizes a modified Retinanet architecture [9]. The model utilizes a shared Retinanet encoder connected to both a Retinanet decoder, which performs object detection, and a single 3D convolutional layer from R2PLUS1D5-18 [16] followed by one FC layer for Action Segmentation (see figure 2 A for the network architecture) - and was trained for hand and tool (bovie, needle driver, and forceps) detection and action (cutting, tying, and suturing) segmentation.

The Retinanet model is a single-stage object detection model that utilizes a focal loss function in training. It's backbone is a Resnet50 model (figure 2 A, left), and it utilizes a Feature Pyramid Network (FPN) (figure 2 A, right). Anchors are then extracted from the FPN and put through different object detection heads. We additionally attach an action segmentation head after the Resnet50 feature encoder. Specifically, we use a single 3D convolutional layer from R2PLUS1D5-18 [16] followed by 2D adaptive max pooling to a 1x1 spatial dimension and one FC layer for Action Segmentation (see figure 2 A for the network architecture). Action segmentation was trained with cross-entropy loss. As we have a single-backbone, dual-head setup - for every batch, we first perform forward propagation through the backbone to both heads, then back-propagate the loss from both heads and average the loss gradients from both heads over the backbone during the update step.

### 4.2. Multitask Model Robustness Evaluation - Augmentation Pipeline

To characterize what image variations have caused the performance degradation, we created a video augmentation pipeline that can adaptively perform any combination of image augmentations. We then used this pipeline before the Multitask model and used it to evaluate the effect different augmentations had on the model's performance (see figure 2 B). Specifically, we are interested in the degree a specific augmentation causes a reduction in object detection accuracy – or Mean Average Precision (mAP) at 0.5 intersention-over-union (@0.5).

### 4.3. Multitask Model Robust Retraining

After evaluating which augmentations caused the most severe performance degradation, as well as which augmentations are actually (visually) plausible, we will select several augmentations to use during retraining. We will then retrain the multitask model while randomly augmenting the data during training. As we are training on video data, we cannot simply augment the videos and create a larger database; but rather, each time we load a new batch, we will randomly select some augmentations to apply to each sample in the batch. I chose to use the Adam optimizer with a learning rate schedualer of "ReduceLROnPlateau". For images, a batch size of 8 was used and for videos SGD was used with 8 frames due to GPU constraints. L1 whieght reg-

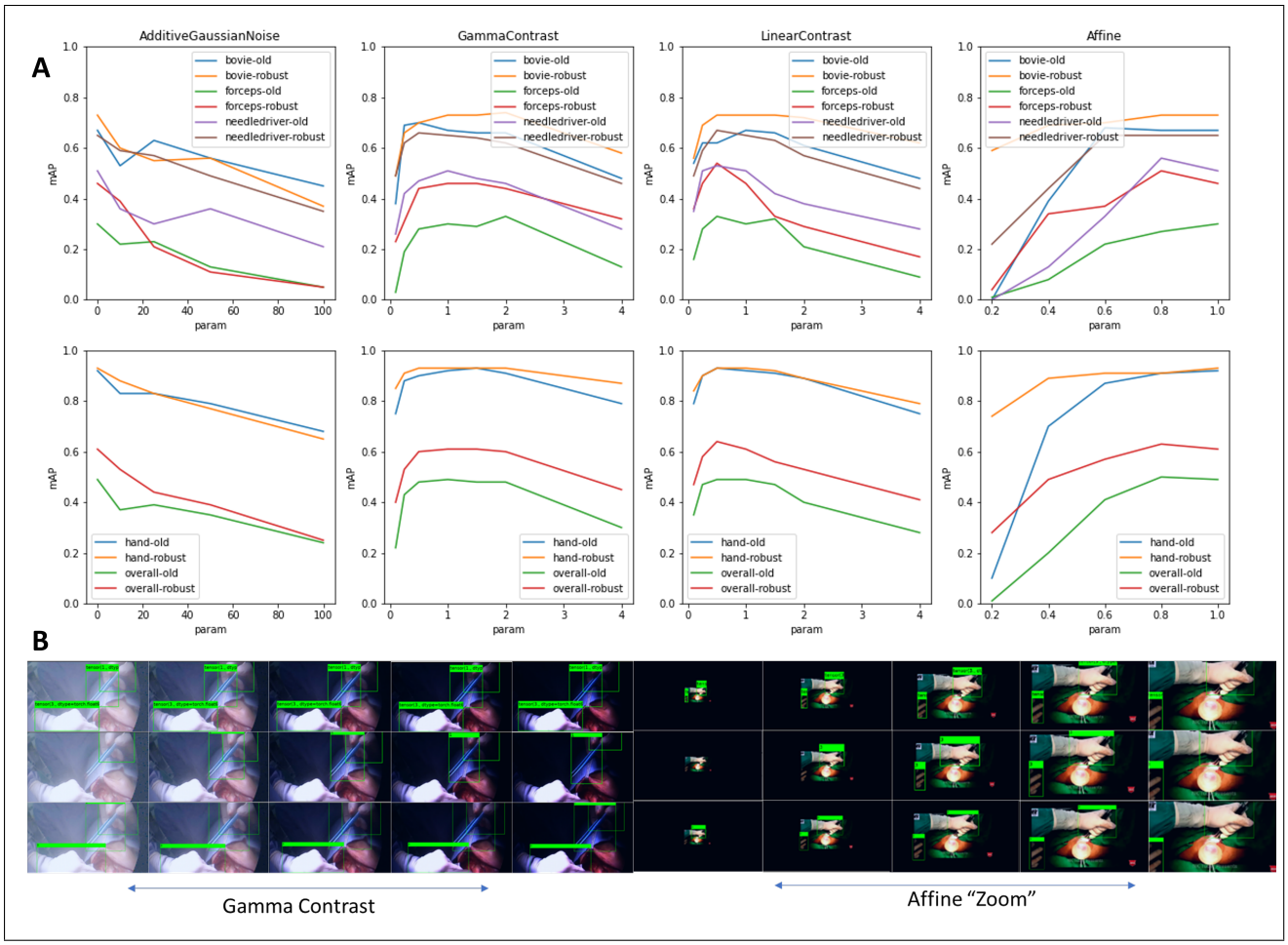


Figure 5. (A) mAP vs. augmentation strength of the robust original model for different augmentations. (B) Representative example of the effect of Gamma Contrast (left) and Affine (right) augmentations. Top row is ground truth, middle row is the predictions of the original multitask model, and the bottom row is the predictions of the robust multitask model.

ularization was also used to reduce overfitting on the target dataset.

### 4.4. Multiscale Convolution for Zoom Robustness

After viewing the results from the augmentations pipeline, it became apparent that some of the issues arise from scale (zoom) differences between the datasets. Specifically, the HMS dataset seemed significantly more zoomed out. Upon further investigation, I discovered that scaling is a common issue with object detection neural networks. In order to combat this issue, I theorized that perhaps using dilated convolutions might serve as a type of solution. Specifically, I designed a custom "multiscale" convolutional layer where each convolutional filter is applied three times with a dilation of 1, 2, and 3 (in each case, the image is padded to produce the same output size). I then take the maximum activation across these three filters per pixel. Specifically, at any pixel (i,j) - see figure 3 - I convolve the input image with the *same* filter at different dilations and take the maximum value as my output. Using this method adds no additional parameters to the model as I use the same weights for all the dilated convolutions. This method does, however, increase the number of computations needed for the forward pass, decreasing the maximum fps. I then replaced the first (7x7) convolutional layer in the multitask model with this "multiscale" convolutional layer to see if the model's robustness to scale differences is improved. While it would also be interesting to replace all convolutions with this formulation, it would also slow down the forward run significantly (as we would need to perform 3x the number of computations per convolutional layer), and therefore it would make sense to apply this model only on a few layers in the network.

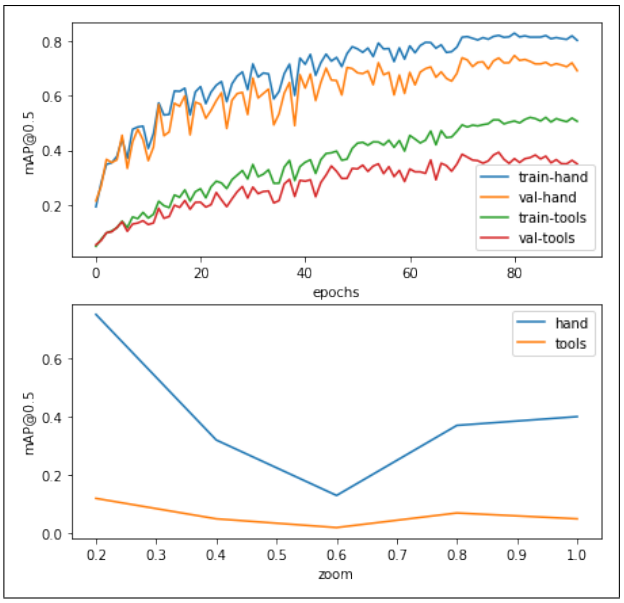


Figure 6. (top) training curves depicting mAP of the hands/tools vs Epoch on the val/train dataset. (bottom) trained multiscale convolution mAP@0.5 vs zoom

### 4.5. Additions to Existing Codebase

I will be using the original multitask model codebase, published by E. D. Goodman *et al*. [6], and have added the following: (1) Robustness evaluation script. (2) Augmentation pipeline. (3) Editing data loader to enable control of image resolution. (4) Multiscale convolution layer new model implementation. (5) labeling HMS dataset. (7) Model deployment and evaluation on HMS dataset.

## 5. Results

### 5.1. Model Robustness Evaluation on AVOS

We begin by evaluating the original multitask model on the AVOS dataset with a custom data augmentation pipeline (see figures 2 B and 4). We have selected 18 possible augmentations and tested the model's object detection accuracy on the validation dataset while sweeping each of the 18 augmentation parameters[1]. After studying the graphs seen in fig. 4 and the corresponding augmented images; we determined that the Gaussian Noise, Affine (Zoom), and Gamma Contrast augmentations were the most substantial (and reasonable) augmentations (see figure 5 B for examples of Gamma Contrast and Affine augmentations).

### 5.2. Robust Retraining

We then retrained the Multitask model with the robust augmentation pipeline, where we employed more extreme image augmentations. Using our augmentation pipeline, our model seemed to generalize better than that of E. D. Goodman *et al*., resulting in improved tool and hand detection. We considered "Model Robustness" to be the extent to which a model is unaffected by our augmentations. Therefore, we would hope to see that our retrained model maintains high mAP values while undergoing a wide range of augmentations and augmentation intensities. By studying figure 5 A, it is apparent that our model manages to produce more accurate predictions even when undergoing extreme augmentation. For example, for an Affine augmentation (zoom) of 0.2, our model succeeded in producing an overall detection mAP@0.5 of 0.28 while the original model's mAP@0.5 dropped to 0.01.

Furthermore, our model's relative drop in overall mAP@0.5 (mAP@0.5 drop at 0.2 affine augmentation vs. no augmentation) was smaller than that of the original model - with a drop of 0.35 vs. 0.49, respectively. The improvement of model robustness is even more apparent when studying the effect of affine (zoom) augmentation on hand detection - where the retrained model retained high accuracy (0.74 mAP@0.5) while the original model's accuracy dropped to near zero (0.1 mAP@0.5) - see figure 5 A, bottom left. When studying the other augmentations, which can be seen in figure 5, the same observations are apparent, although not as dominantly as with affine (zoom) augmentation. In figure 5 B, two example images were used to demonstrate the effect of Gamma Contrast and affine (zoom) augmentation on the image and the resulting detections – with the top row being ground truth and the middle and bottom row being the inference results of the original and robust pipelines respectively. When studying the effect of both Gamma Contrast and Affine augmentations, it is clear that the robust multitask model (figure 5 B bottom) is more resilient to these augmentations compared to the original multitask model(figure 5 B middle), where there is a higher rate of missed detection.

### 5.3. Multiscale Convolution Model

Next, we tried to train our multiscale convolutional model. We only replaced the first convolutional layer with our proposed "multiscale" convolution. As can be seen in figure 6 (top), the model had convergence issues and converged to a lower setpoint of 0.7 vs. 0.89 mAP@0.5, which was reached in previous experiments. Furthermore, we found that the model generalized much worse to our test set, with drastically decreased performance and higher sensitivity to zoom. This can be seen in figure 6 (bottom), where a mAP@0.5 of 0.3-0.5 for hands and a near-zero mAP for tools was reached. We, therefore, concluded not to use multiscale convolutions on the HMS dataset.

[1] see supporting information for examples of all the augmentations

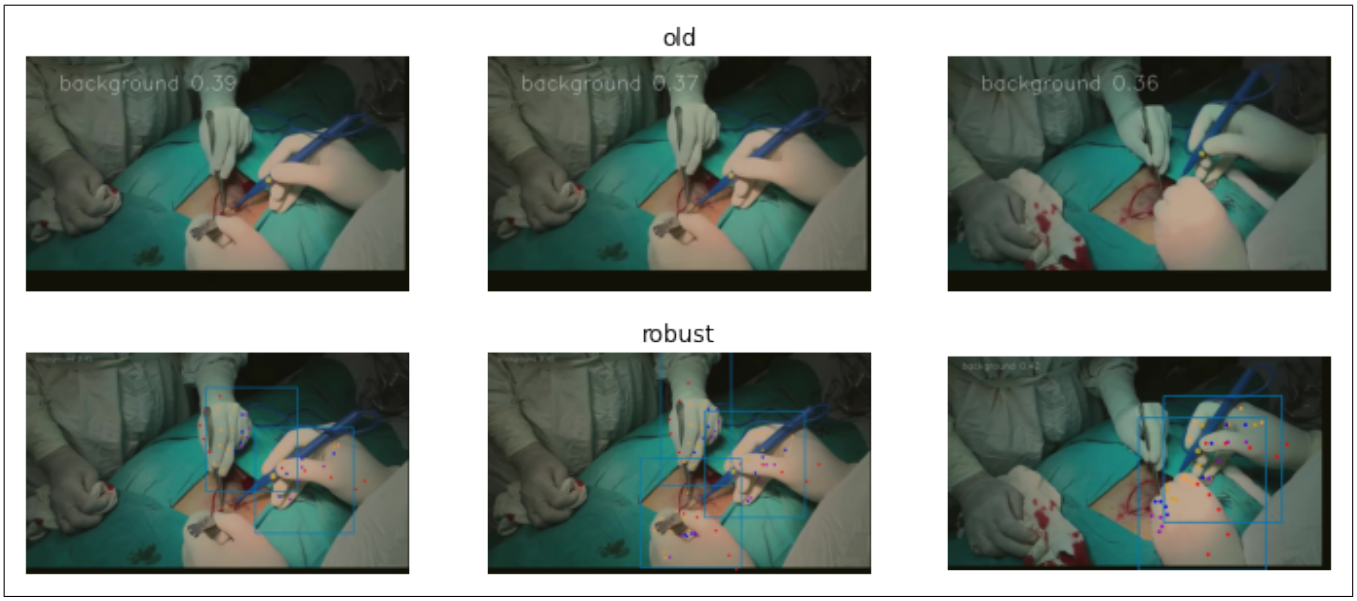


Figure 7. Robust vs. old multitask model on data that is visually simular to HMS dataset

| mAP | bovie | forceps | n. driver | hand | overall |
|---|---|---|---|---|---|
| Original | - | 0.11 | 0.06 | 0.65 | 0.085 |
| Robust | - | 0.25 | 0.30 | 0.88 | 0.285 |

Table 1. mAP@0.5 on the HMS dataset

### 5.4. Results on HMS Dataset

We ran the robust model on the HMS dataset and saw significant improvement in tool and hand detection compared to the original model (Table 1). Hand mAP@0.5 performance improved from 0.65 to 0.88, while the tool mAP@0.5 increased from 0.085 to 0.285. These increases in accuracy are quite drastic and result in a much-improved capability of object tracking – ultimately allowing us to calculate the needed metrics. While we cannot show images from the HMS dataset, we did select a visually similar video from AVOS and compared the detection results of the robust and original models.

To qualitatively compare the two models, we selected a single video from AVOS that was visually similar to the HMS dataset videos. We then ran inference of both the old and robust multitask models and compared the detection results (see figure 7). We found that the robust model significantly outperformed the old multitask model qualitatively as well. However, it is not great at detecting all the objects in the frame. For example, there is almost no difference between the left and right frames; however, for some reason, the robust model did not identify the bottom hand in the left frame. Furthermore, the left hand was not identified at all throughout the video.

## 6. Conclusion/Future Work

As can be seen above, the proposed "multiscale convolution" failed to make the model more scale-invariant. This could perhaps be due to the selection of only the first convolutional layer for this model. It is entirely possible the deeper layers would benefit more from this approach and improve model robustness to zoom. If I had additional compute power, I would test how the robustness is affected by replacing convolutional of different depths with the proposed multiscale convolutional model.

Our robust model had significantly improved performance on the HMS dataset, allowing for the continuation of the proposed clinical trial. However, there is still much room for improvement. One major issue with the AVOS dataset is that most of the videos are very zoomed in – with the hands mostly out of frame, with only a close-up of the surgical site. However, we need to see the entire hand to analyze its motion to analyze surgical skills, requiring a more zoomed-out video. Unfortunately, as we can only train on AVOS, we have issues generalizing into the HMS dataset. Therefore, future work will be two-pronged. Firstly, we will focus on the clinical side of the research. With the improved detections, detection-based multi-object tracking methods will benefit greatly, allowing us to continue with the proposed clinical study. Specifically, we will extract and analyze different surgical skill metrics, such as economy of motion, and see if it correlates to surgical outcomes – in this

case, surgical site infections. The second prong of future research will focus on incorporating low-data regime techniques such as unsupervised pre-training, semi-supervised training, and active learning.

## 7. Contributions Acknowledgements

I would like to thank MARVL lab for use of their GPC credits. This allowed me to ceate my own virtual enviorment with 30 Gb of GPU. I did make use of the (not yet public) codebase: https://github.com/yeung-lab/spatiotemporal-open-surgery. The codebase will be made available upon publication of the article (currently under review). For any questions, feel free to reach out to me.

## References

[1] Sven Bambach, Stefan Lee, David J. Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. December 2015. 2

[2] John D. Birkmeyer, Jonathan F. Finks, Amanda O'Reilly, Mary Oerline, Arthur M. Carlin, Andre R. Nunn, Justin Dimick, Mousumi Banerjee, and Nancy J.O. Birkmeyer. Surgical skill and complication rates after bariatric surgery. *New England Journal of Medicine*, 369(15):1434–1442, 2013. PMID: 24106936. 1

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. 2016. 2

[4] Xiang Gao, Ripon K. Saha, Mukul R. Prasad, and Abhik Roychoudhury. Fuzz testing based data augmentation to improve robustness of deep neural networks. page 1147–1158, 2020. 2

[5] Adam Goldbraikh, Anne-Lise D'Angelo, Carla M. Pugh, and Shlomi Laufer. Video-based fully automatic assessment of open surgery suturing skills. *International Journal of Computer Assisted Radiology and Surgery*, 17(3):437–448, Mar 2022. 1, 2

[6] Emmett D. Goodman, Krishna K. Patel, Yilun Zhang, William Locke, Chris J. Kennedy, Rohan Mehrotra, Stephen Ren, Melody Y. Guan, Maren Downing, Hao Wei Chen, Jevin Z. Clark, Gabriel A. Brat, and Serena Yeung. A real-time spatiotemporal AI model analyzes skill in open surgical videos. *CoRR*, abs/2112.07219, 2021. 1, 2, 4, 6

[7] Keren Gu, Brandon Yang, Jiquan Ngiam, Quoc Le, and Jonathon Shlens. Using videos to evaluate image model robustness. 2019. 2

[8] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. 2019. 2

[9] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. 1, 4

[10] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. 2019. 2

[11] Arpit Mittal, Andrew Zisserman, and Philip H. S. Torr. Hand detection using multiple proposals. 2011. 2

[12] Dmitri Nepogodiev, Janet Martin, and Bruce Biccard et al. Global burden of postoperative death. *The Lancet*, 393(10170):401, 2019. 1

[13] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. 34:29935–29948, 2021. 2

[14] Morgan K. Richards, Jarod P. McAteer, F. Thurston Drake, Adam B. Goldin, Saurabh Khandelwal, and Kenneth W. Gow. A National Review of the Frequency of Minimally Invasive Surgery Among General Surgery Residents: Assessment of ACGME Case Logs During 2 Decades of General Surgery Resident Training. *JAMA Surgery*, 150(2):169–172, 02 2015. 1

[15] Greg D. Sacks, Elise H. Lawson, Aaron J. Dawes, Marcia M. Russell, Melinda Maggard-Gibbons, David S. Zingmond, and Clifford Y. Ko. Relationship Between Hospital Performance on a Patient Satisfaction Survey and Surgical Quality. *JAMA Surgery*, 150(9):858–864, 09 2015. 1

[16] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. *CoRR*, abs/1711.11248, 2017. 4

[17] Le-Anh Tran, Truong-Dong Do, Dong-Chul Park, and My-Ha Le. Enhancement of robustness in object detection module for advanced driver assistance systems. pages 158–163, 2021. 2

[18] Georg Volk, Stefan Müller, Alexander von Bernuth, Dennis Hospach, and Oliver Bringmann. Towards robust cnn-based object detection through augmentation with synthetic rain variations. pages 285–292, 2019. 2

[19] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. 2018. 2

[20] Fuxun Yu, Zhuwei Qin, Chenchen Liu, Liang Zhao, Yanzhi Wang, and Xiang Chen. Interpreting and evaluating neural network robustness. 2019. 2

[21] Michael Zhang, Xiaotian Cheng, Daniel Copeland, Arjun D. Desai, Melody Y. Guan, Gabriel A. Brat, and Serena Yeung. Using computer vision to automate hand detection and tracking of surgeon movements in videos of open surgery. *CoRR*, abs/2012.06948, 2020. 1, 2

[22] Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. 2016. 2

[23] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L. Sarin, Thomas Ploetz, Mark A. Clements, and Irfan Essa. Automated video-based assessment of surgical skills for training and evaluation in medical schools. *International Journal of Computer Assisted Radiology and Surgery*, 11(9):1623–1636, Sep 2016. 1

## 8. Supporting Information


Figure 1. Affine(ZOOM)

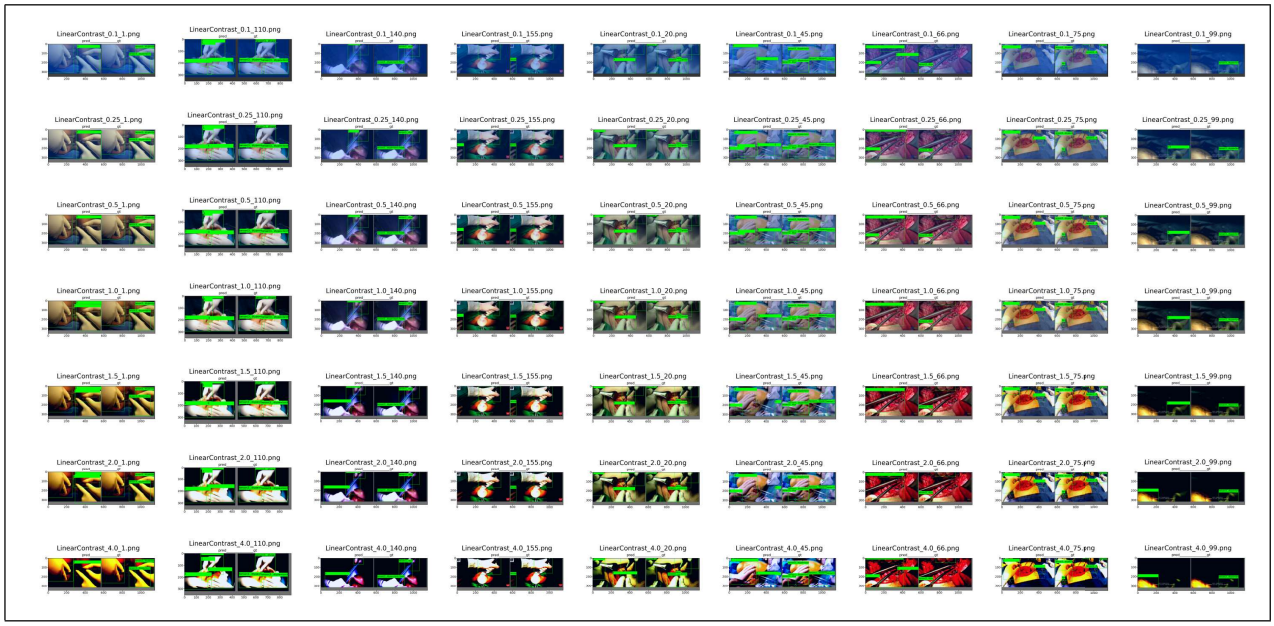


Figure 2. Linear Constast

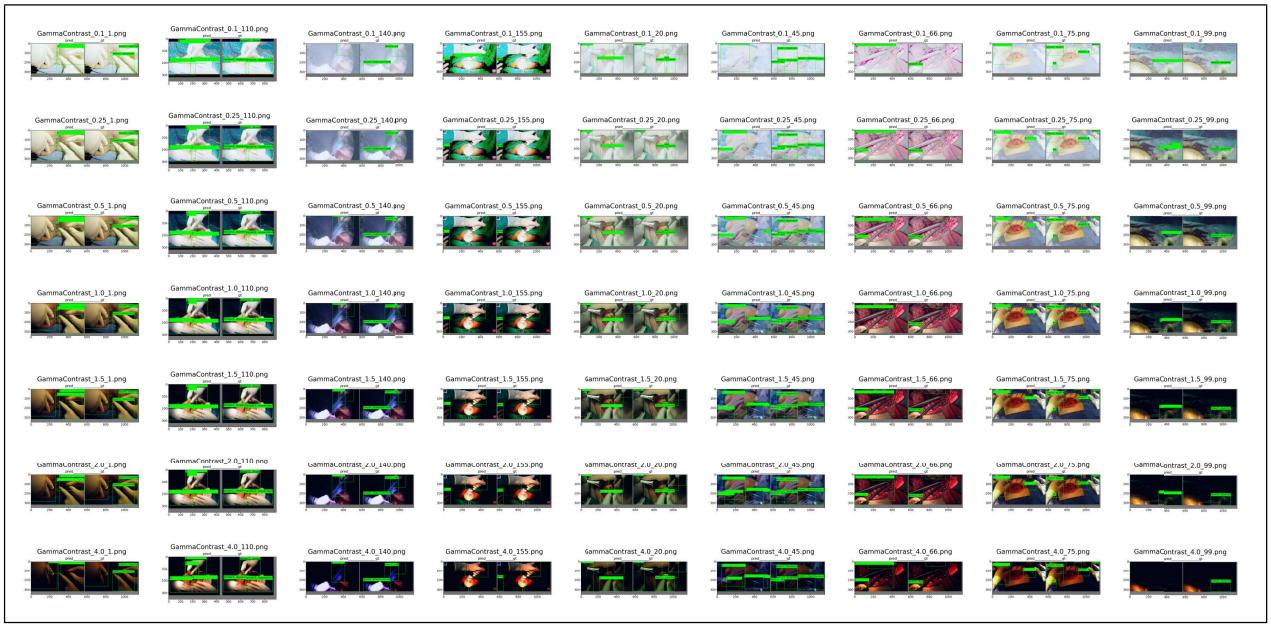


Figure 3. Gamma Constast

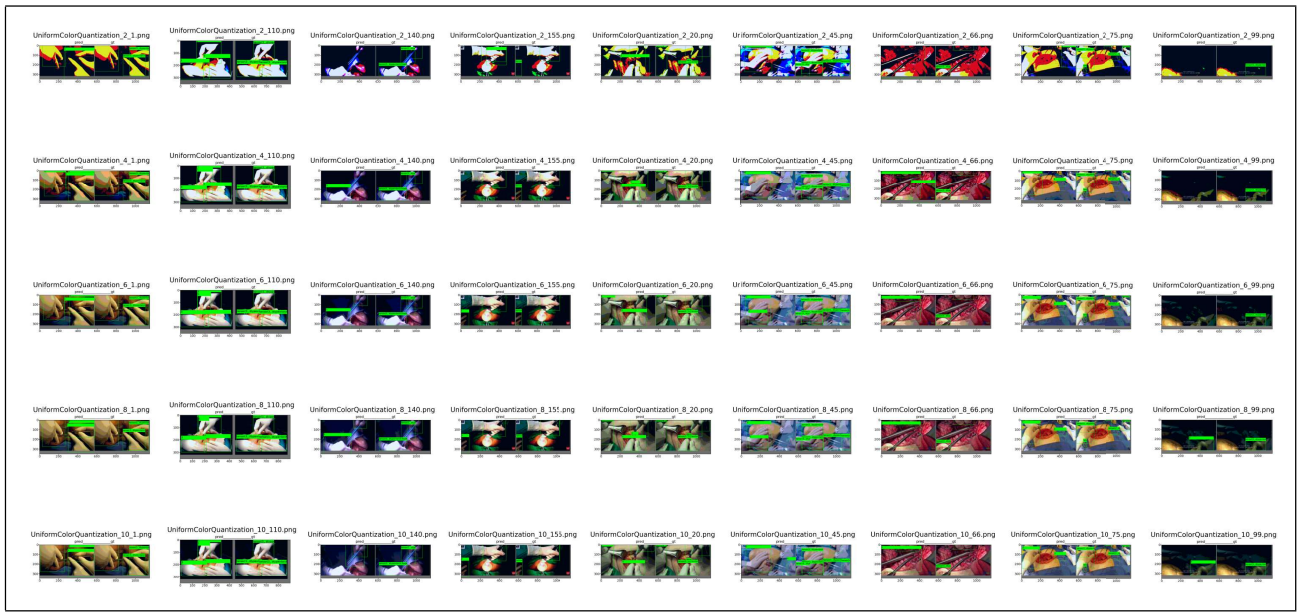


Figure 4. Uniform Color Quantization

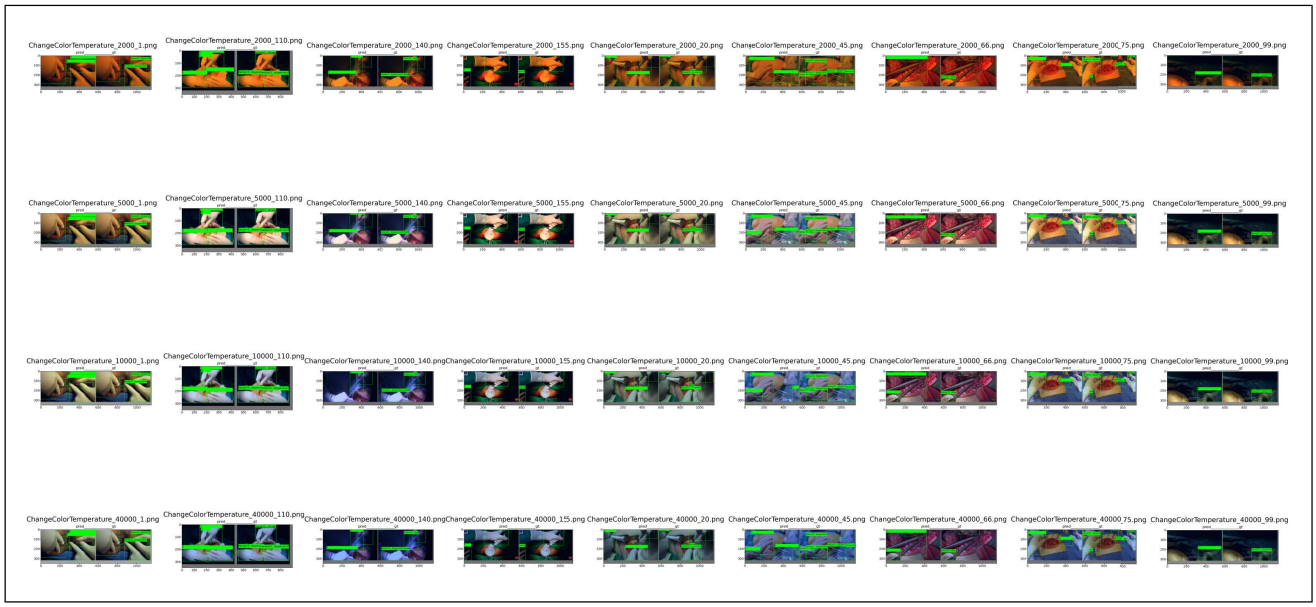


Figure 5. Change Color Temperature

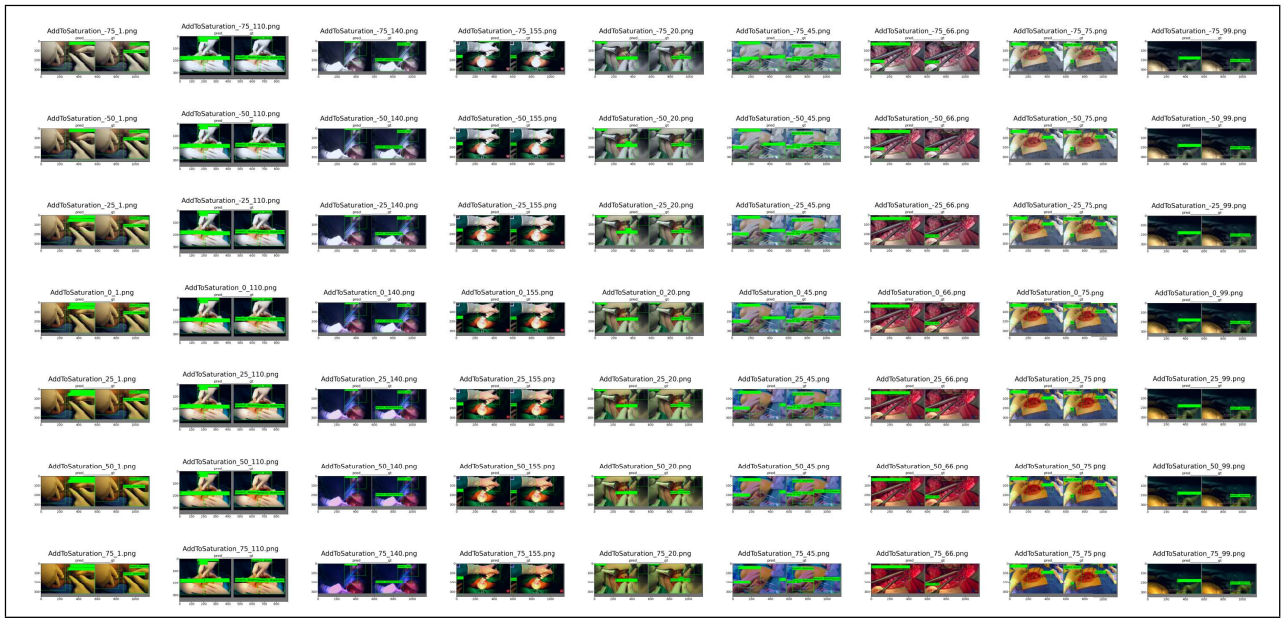


Figure 6. Add To Saturation

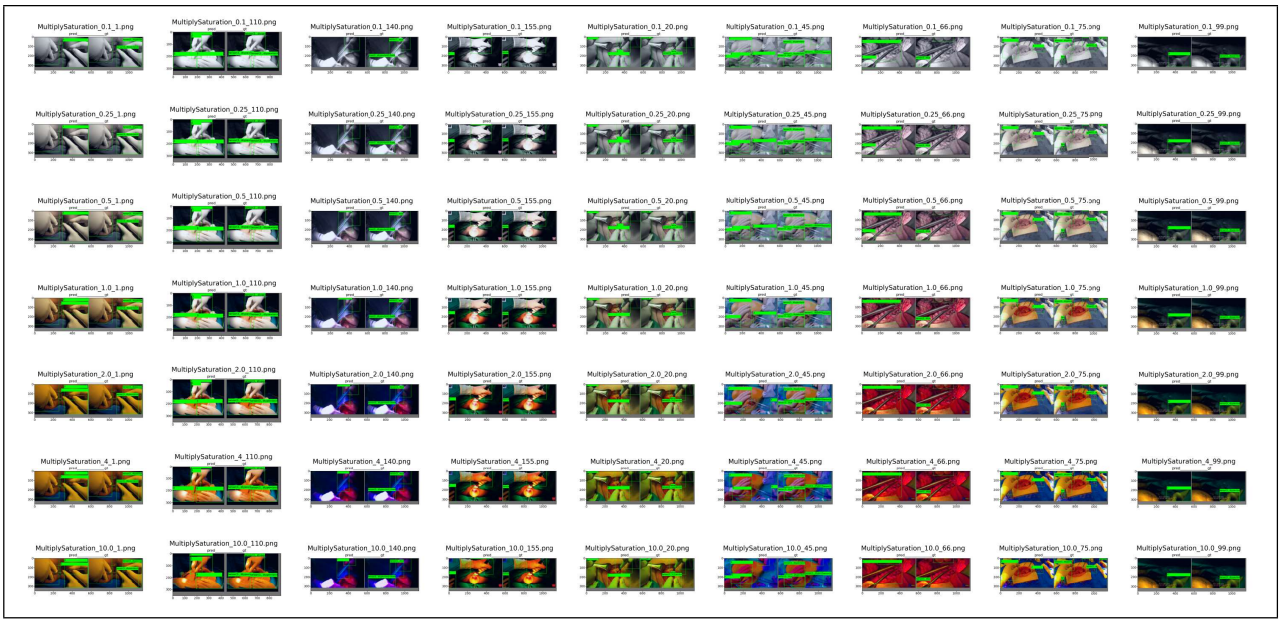


Figure 7. Multiply Saturation

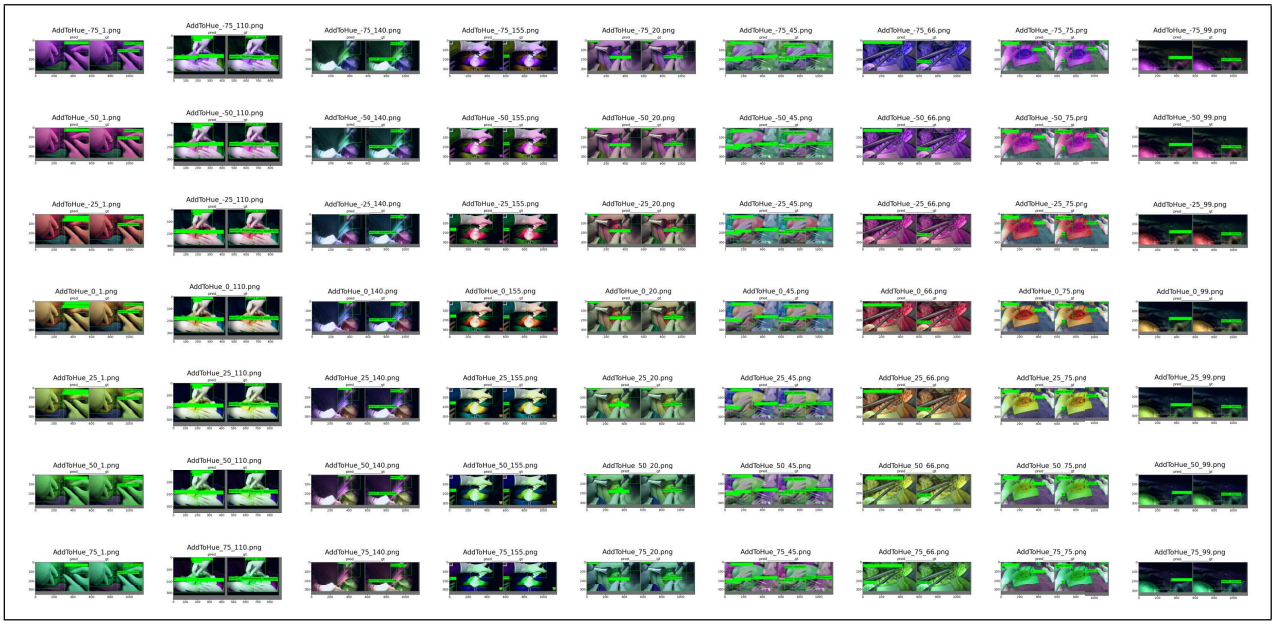


Figure 8. Add to Hue

Figure 9. Multiply Hue

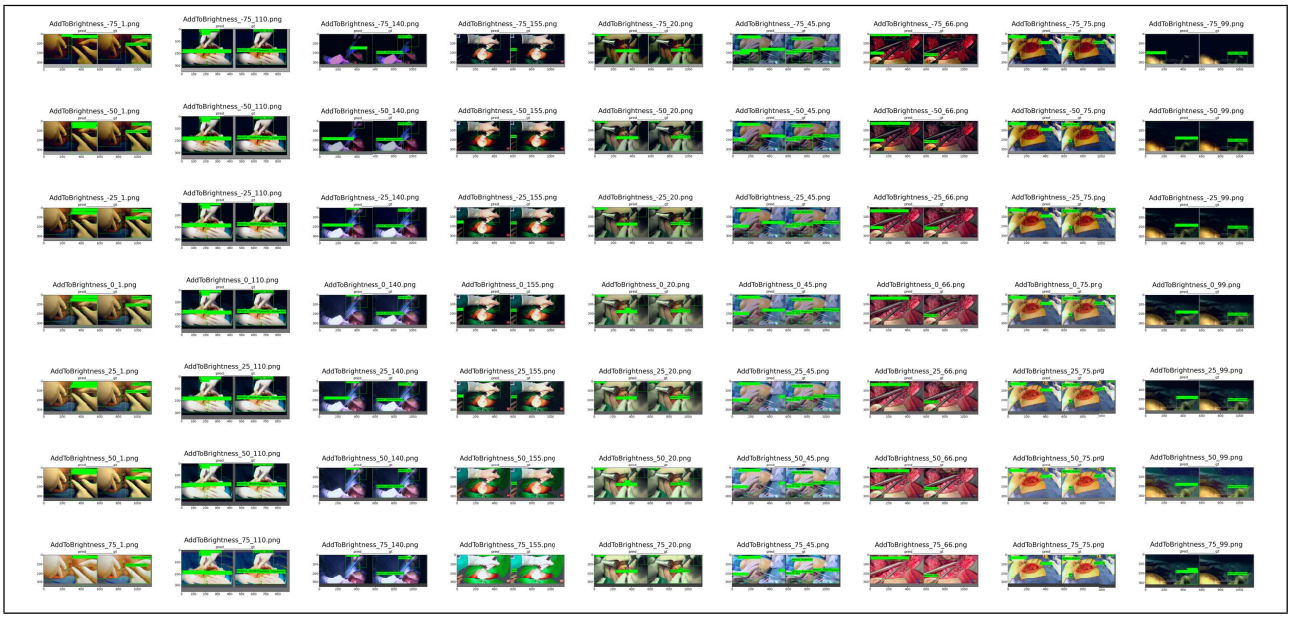

Figure 10. Add to Brightness

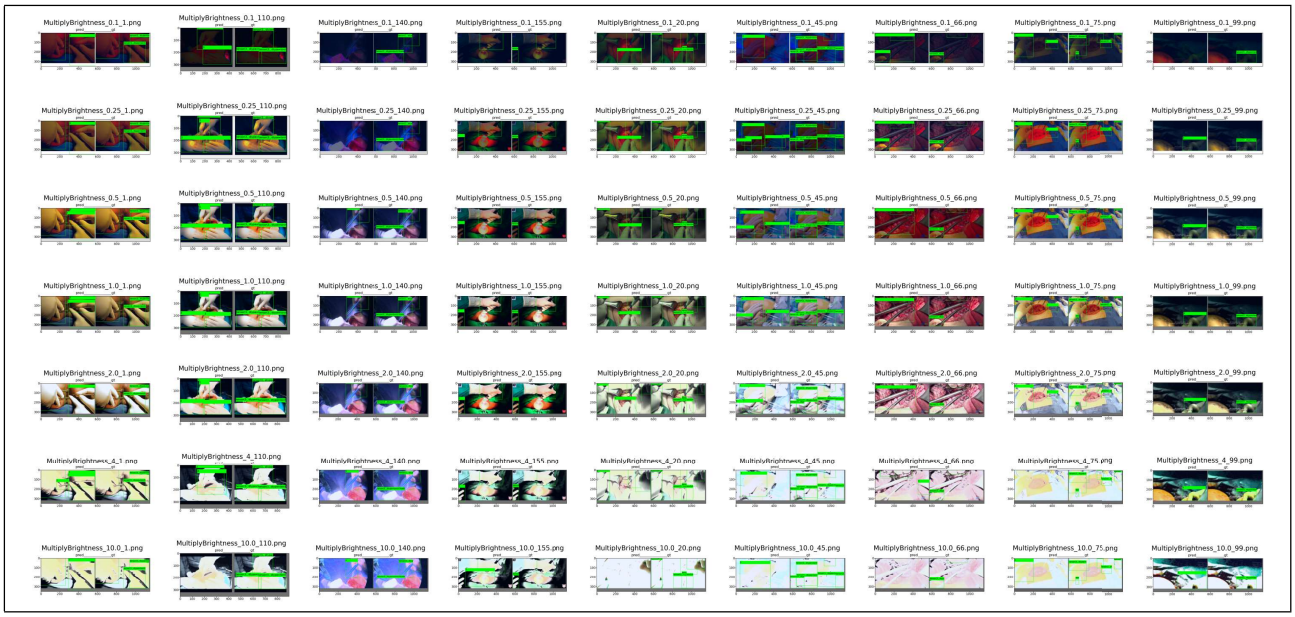


Figure 11. Multiply Brightness

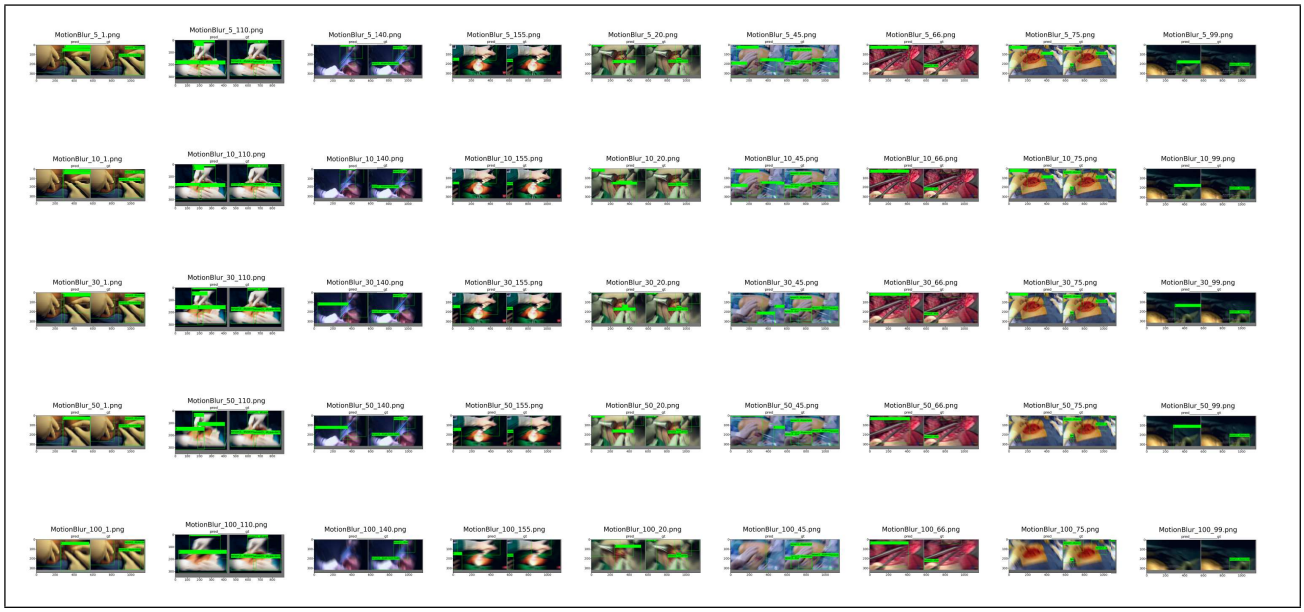


Figure 12. Motion Blur

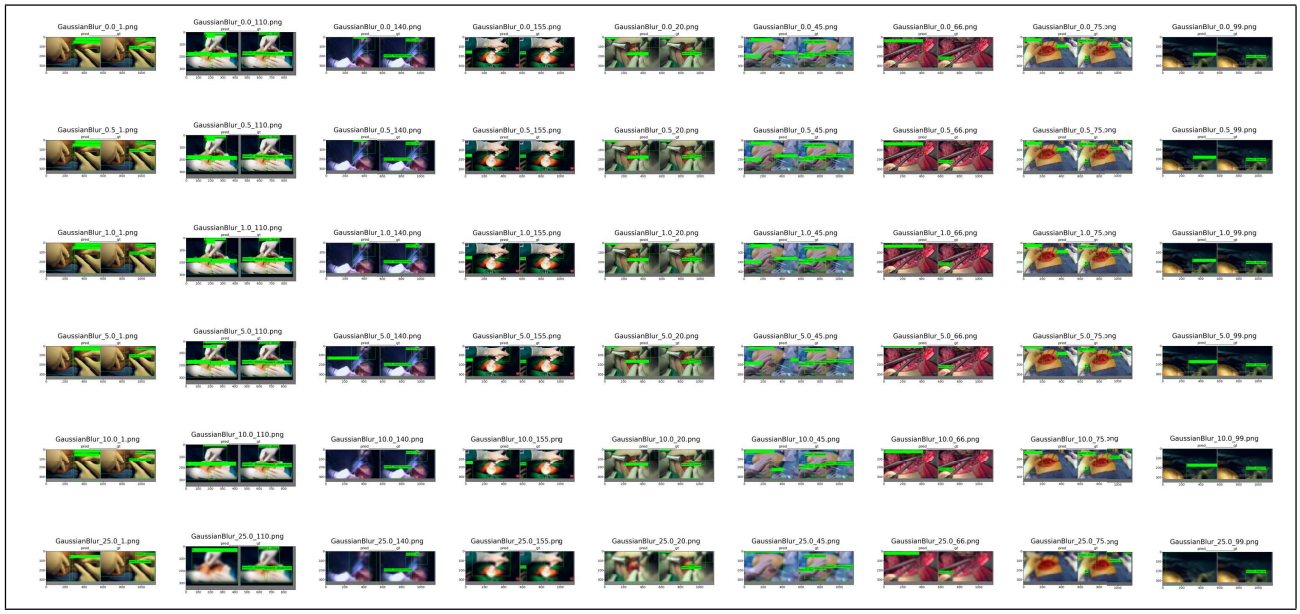


Figure 13. Gaussian Blur

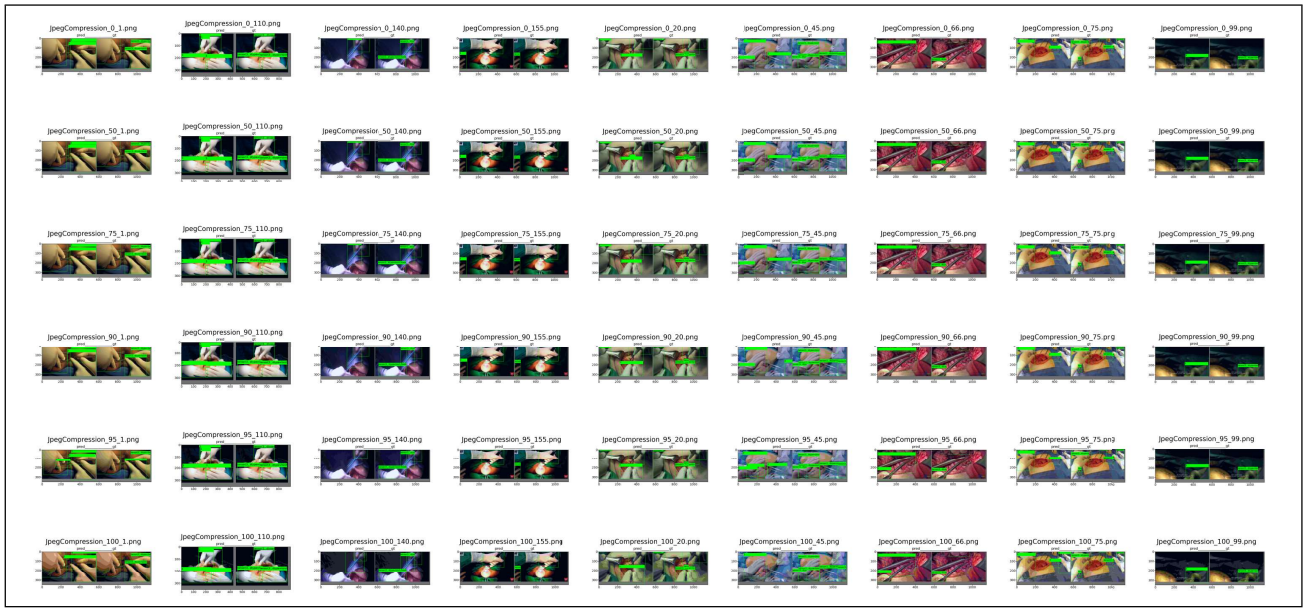


Figure 14. Jpeg Compression


Figure 15. Dropout

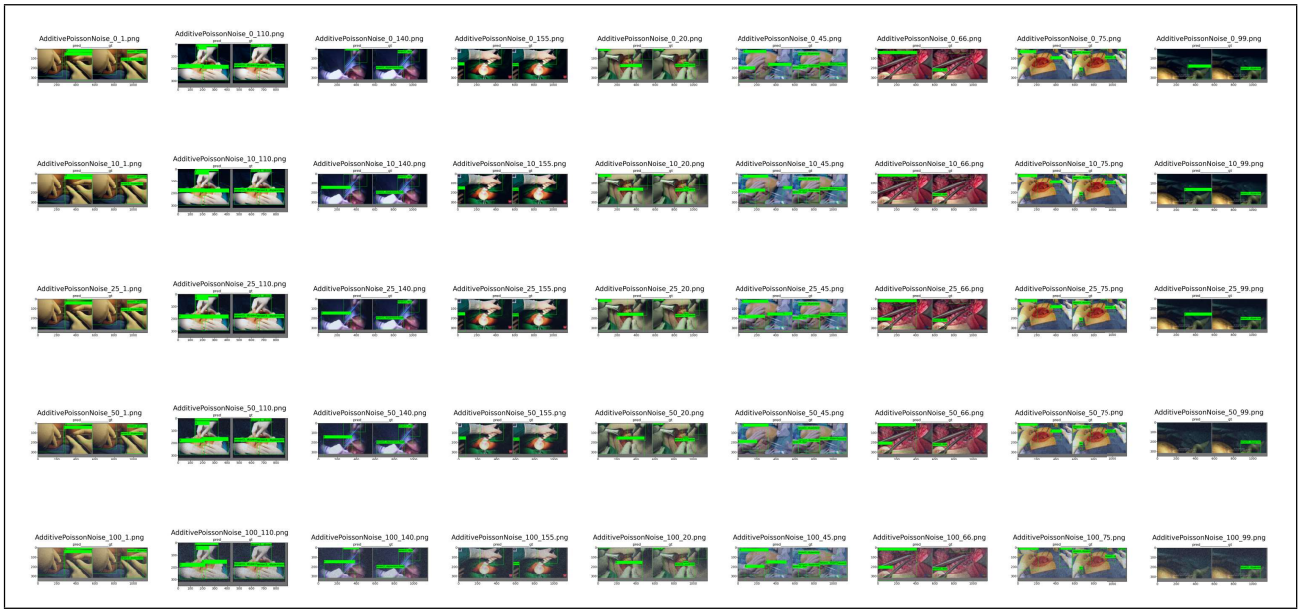


Figure 16. Additive Poisson Noise

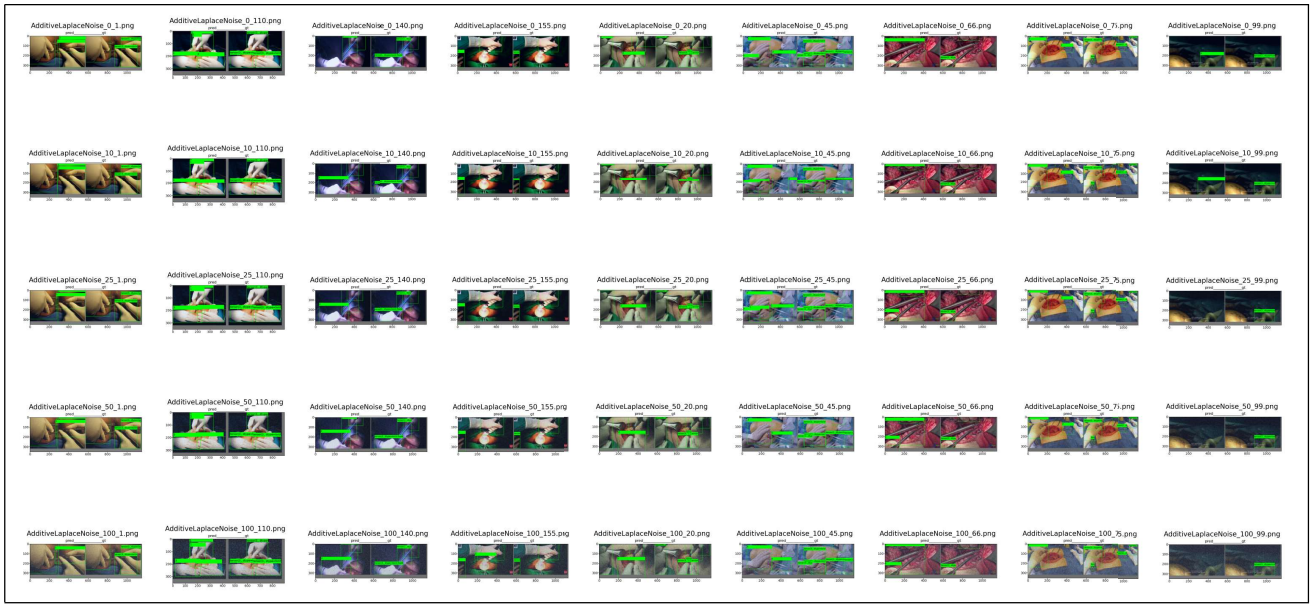


Figure 17. Additive Laplace Noise

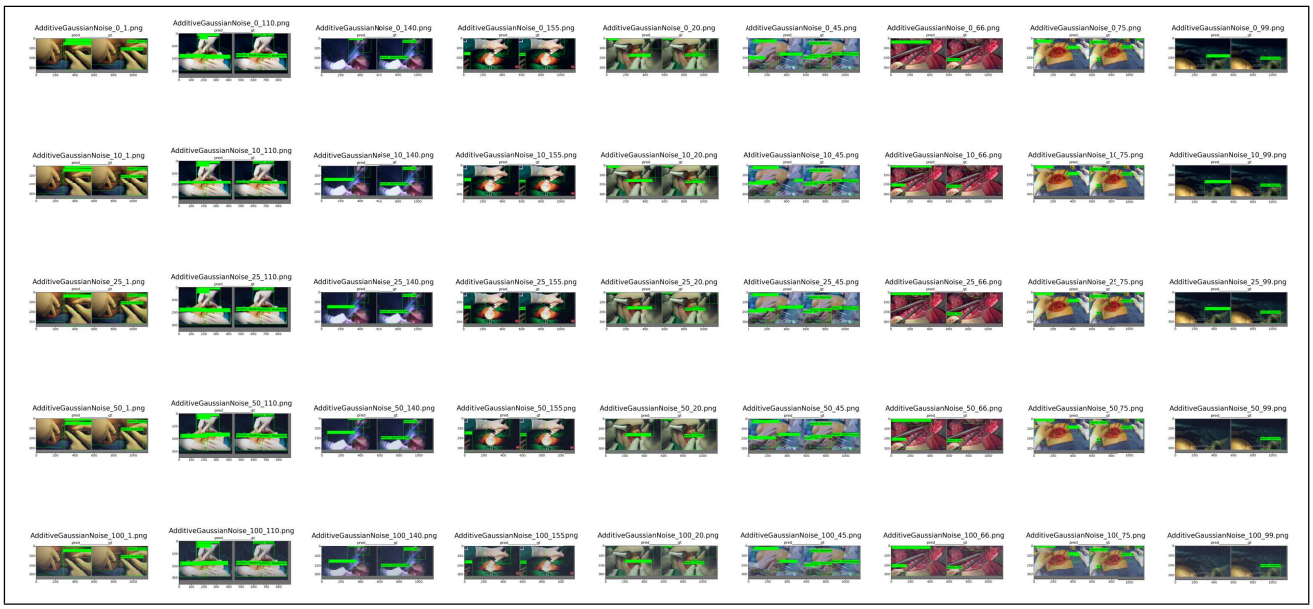


Figure 18. Additive Gaussian Noise