Practical Considerations in Utilization of Computer Vision

Thomas Ward, MD

Surgical AI & Innovation Fellow

Surgical AI & Innovation Laboratory

Massachusetts General Hospital

15 Parkman Street, WAC-460

Boston, MA 02114

tmward@mgh.harvard.edu

Introduction ____ *4*
Overview ____ **4**
Data acquisition and preparation ____ *5*
Existing datasets ____ **6**
New dataset creation ____ **7**
Data Security and Anonymity ____ **14**
Data Annotation ____ *16*
Modeling ____ *18*
Conclusion ____ *22*
References ____ *22*

Highlights

1. Adherence to and compliance with local laws and regulations in data capture, storage, and use is just as important as compliance with accepted ethical guidelines around research
2. True anonymization and full deidentification of data are difficult to achieve. Appropriate measures must be in place to ensure optimal protection of patient privacy.
3. To obtain a reliable result in your research, establishing a well-planned database structure for storage and access to data is as important as your machine learning model and analysis.
4.

Introduction

Computer vision is a young field. Its current era started in 2012 with Krizhevsky's application of deep convolutional neural networks,[1] and this "deep learning revolution" made accurate image and video classification a reality with a resultant explosion in computer vision research (nearly 30,000 publications in 2017). The medical community, though, has underperformed in contributions with under one thousand annual publications in artificial intelligence.[2] Computer vision medical research, particularly that focused on surgery, is therefore an even smaller slice of the pie.

With a paucity of research groups, there is no definitive roadmap to creating an effective medical computer vision group. New labs will inevitably encounter hurdles and unforeseen complications. Our group, the Surgical Artificial Intelligence and Innovation Laboratory (SAIIL) at the Massachusetts General Hospital (MGH), is one of just a handful of groups that focus on computer vision and surgery. Using our laboratory's experience, this chapter will provide a roadmap to navigating these poorly charted waters.

Overview

Prior to embarking on a computer vision study in surgery, first any researcher must determine a question to address. This question will be unique to each group's interests, with common computer vision problems including operative temporal segmentation, phase recognition, tool recognition, operative skill assessment, and clinical decision support. Once there is a question, one must then identify the appropriate technical expertise in answering the question.

As with any research, the selection of appropriate methodology is critical. Thus, the first recommendation for anyone looking to engage in computer vision research is to partner with an expert in, at least, the technical aspects of the field. Guidance from a PhD-trained expert will help you refine your question to be specific, measurable, and achievable while also preventing you from falling into common traps regarding dataset selection and model evaluation.

While the specifics for any particular research project will vary, the remaining chapter will go through the three major considerations when conducting research in surgical computer vision:

1. Data acquisition and preparation
2. Data Annotation
3. Modeling

Data acquisition and preparation

Machine learning and computer vision models live or die on their data inputs. For this image and video data to be useful for modeling, it must go through a multi-step pipeline. First, researchers must acquire data, either from a publicly available dataset or from data captured *de novo*. Second, the data must be processed to a standardized format for use with the modelling tools. Lastly, the data needs proper labels for supervised model training. While this pipeline seems straightforward, it actually presents numerous challenges. Without consideration of the various minutiae, a research group will ultimately

lose months and years in data preparation, which for such a fast-moving field as machine learning, translates to numerous research discoveries lost to other groups.

Existing datasets

As a relatively nascent field bound by numerous health privacy regulations, few publicly available medical datasets for computer vision in surgery exist. Most readily available medical datasets consist of radiographic images or clinical still images. For example, the KiTS19 Challenge Data has Computer Tomography (CT) images and clinical outcomes for 300 patients with kidney tumors.[3] Other datasets include the ChestX-ray8 (over 100,000 chest X-rays for 32,000 patients), Messidor (1,200 eye fundus images for diabetic retinopathy), and HAM10000 (10,000 dermatoscopic images).[4–6] Many more examples are centrally located on the Cancer Imaging Archive (TCIA).[7]

Although there are several medical image datasets, few have data that can answer surgical questions. Ophthalmology groups have generated some useful datasets from their procedures. These procedures lend themselves to recording as they occur in a non-moving field under a microscope. The CATARACTS dataset holds fifty cataract operations with included surgical tool labels.[8] Zisimopoulos *et al* built upon this dataset by adding labels for the surgical phases.[9] The same group then, over a select 200 frames per video, performed highly granular labeling of tools and anatomy for the CATARACTS dataset.[10]

Laparoscopic and robotic-assisted surgery similarly lends itself to video recording given the camera stability and continuous video feed inherent to minimally invasive operations. One group published a small dataset of nine partial nephrectomies that

included surgical phase annotations.[11] The 2017 Robotic Instrument Segmentation Challenge also published nephrectomy videos, though their dataset contained robotic instrument data for ten sequences of porcine abodminal operations.[12] Another robotic instrument dataset comes from the JIGSAWS study with videos of a robotic surgery system performing surgical tasks on a model. Their dataset includes annotations for instrument kinematics, surgical gestures, and operator skill.[13]

Multiple groups have also published datasets for laparoscopic cholecystectomies. The CAMMA group in Strasbourg, France, has released a series of successively larger datasets, and their most recent publicly available dataset contains eighty surgeries with annotations for both surgical phases and tools.[14] Technische Universität München released another dataset, the TUM LapChole, which has twenty surgeries with annotated surgical phases.[15]

The above examples are, for the time being, a nearly comprehensive list of surgical video datasets. Unfortunately, the vast majority of surgical computer vision questions cannot be answered with this little data. Therefore, most computer vision groups will need to pursue *de novo* data acquisition which brings its own unique set of challenges.

New dataset creation

Medical media acquisition is a process fraught with hurdles and challenges. The below section will present an overview map with acquisition of surgical video as a case example to help with navigation of this difficult process, starting from steps that precede even data acquisition and ending with labeled data suitable for computer vision modeling.

First, prior to media acquisition, one must obtain patient and institutional consent. These processes are specific to the institution's policies and its country's laws. Some institutions have media recording consent included in all procedure consents while others require specific consent forms or addenda to allow for video recording of an operation. This may vary based on each individual institution's culture regarding use of video for quality improvement, education, and research purposes. For the time being, United States regulations are simpler to comply with than data regulation laws in Europe, especially General Data Protection Regulation (GDPR). As regulations are apt to change year to year, all researchers should use the available resources of their institution to maintain compliance at all times. It is the responsibility of the investigator to ensure that all data is captured in an ethical manner that preserves patient autonomy and privacy.

Once a patient grants permission to capture media, there are then many ways to actually capture the raw data. Early standardization of the capturing process will yield the most useful data without needing to discard ultimately inutile captures. All data should be captured at the highest quality and resolution possible. Capturing devices (cameras, video recorders, etc) save the visual data in various file formats. File formats are either "lossless" or "lossy." Lossy formats, to save disk space, compress similar data together (eg various shades of blue will save to the disk as just one type of blue). There are many common examples both for images (JPEG, WMP) and videos (MPEG-4, Quicktime).[16,17] Using a lossy format may save an image that appears to the human eye identical to the original photo. To computer vision models, though, which analyze at the pixel level, this compression may negatively hamper analysis and model performance. Astronomy, with their massive

petabyte level of data input has dealt with this tradeoff extensively and still argues for highest resolution input.[18]

The importance of capture process standardization cannot be emphasized enough. Capture devices should ideally be identical or at least of similar capabilities. In the early stages of our data acquisition we had disparate video capture devices that ultimately led to significant downstream headaches with differences in file formats, captured metadata, and video quality. Even for simple photo acquisition, there are numerous variables to consider for standardization (eg image orientation, ambient light, flash versus no flash). While images can be processed after acquisition through normalization techniques, *a priori* consideration of the needs of a project will allow for more robust data capture.

Even the most perfect capture plan falls short if it fails to capture a sufficient quantity of data. Although the process of hitting a record button seems to be an easy task, asking already busy nurses and physicians to add an additional step to only certain procedures will inevitably lead to poor initial data capture. Institutions may have a "culture of capture", where video recording becomes the default rather than the exception – primarily for purposes of patient safety and quality improvement. The work performed with the OR Black Box (Chapter 10) is an excellent example of how a culture of capture can assist in improving acquisition of surgical video data.

All personnel in the operating room should ideally be trained to use the data capture devices so that recording video does not become dependent on one individual (e.g. most often the circulating nurse). For institutions that do not already routinely record surgical video, there may be significant inertia to changing workflows to accommodate video

recording, even when the change in workflow involves just pressing the “record” button on a pre-existing device. Communicating clearly the value of the recordings in potentially improving patient care can help with acceptance. For example, the research video database could be used not only for computer vision projects but also as a repository for surgeons to compare their techniques to other surgeons, review cases during morbidity and mortality conferences, and for instruct residents. Do not be discouraged with initially low video yields: as more and more people record and the culture changes, increasing media capture will happen.

Once surgical media is captured reliably, the data must be centralized and processed. Ideally, capture systems, such as operative video recorders, will connect to the hospital’s network and allow for easy transfer of data to a central, secure repository. If not, regular weekly data pulls should occur to prevent build-up of data and possible lost footage. Having various media files scattered across different laptops, flash drives, and computers can lead to data loss and risk patient privacy. Rapid centralization prevents inadvertent data loss and also increases the informational security of the group. This centralized data storage must also undergo regular secure backups. The American Society of Media Photographers proposes the 3-2-1 rule: all files have 3 copies, on 2 different types of media (hard drives or tapes), with at least 1 copy offsite.[19] Although the 3-2-1 rule seems cumbersome, data redundancy ensures no data loss.

Be cognizant that media take up significant disk space, especially surgical videos: one minute of high-definition surgical video has twenty-five times the amount of data in a high-resolution computed tomography scan.[20] Data storage costs are expensive, so these

must factor into a laboratory's budget. For example, storing 500 operative videos will take nearly half a terabyte of storage which translates to a minimum of $1,000 annually for the cheapest of offline backups (as of end of 2019).[21]

Once the data is in centrally located and reliably backed up, local (and possibly remote) researchers will need access. Multiple factors will affect data access. First, decide where the data should reside: either on the local network or at a remote off-site server. Local network will guarantee higher data transfer speeds and data security while remote servers will make collaborations easier (collaborators will not need to do the training and accreditation necessary to access the local hospital or university network). Compliance with local and institutional requirements will likely determine the ultimate storage location. Second, there needs to be guaranteed data availability. Many remote storage options provide a website to browse files, but any file manipulation and modeling then requires the user to download the files locally, which with hundreds of videos, can quickly outstrip any local hardwares storage capabilities. We prefer solutions that allow researchers to browse files on their computer like regular folders, which requires using systems such as Samba or the Network File System.[22,23] This allows for easy, user-transparent, file-browsing and data manipulation over remote links. Lastly, consider the network speed link. Many networks still use slower speeds (100 megabits per second (Mbps)) rather than 1,000 Mbps. This ten times speed difference significantly affects workload times. For example, transferring 100 surgical videos of a 45 minute procedure across the network would take an estimated nine hours with the 100 Mbps link, while the faster link allows the files to transfer in under an hour.

At this point in our example process, video data is now centrally located, replicated, and quickly accessible. Next, it must undergo meticulous tracking in a database. You should evaluate all the variables you are recording for each patient. An example initial database structure for operative videos could store records for file name, medical record number (or other identifier), and type of operation. Depending on the research question, you may also want to track other associated data, such as the surgeon's prior experience, presence of assistants, etc. Be flexible and open to changing the database structure early on as you start to scale. Additionally, you may wish to plan for the unknown future and consider tracking more information than may initially seem necessary (if approved by your Institutional Review Board). Collecting data prospectively will always be more accurate than retrospectively filling in missing data variables.

There are various programs suited to database creation. Using a spreadsheet (for example, Microsoft Excel®), is an easy initial choice with a low barrier to entry. We have found though, that with a burgeoning dataset size, spreadsheets rapidly become messy and are prone to errors. We suggest using a formal database system (e.g. PostgreSQL) due to its considerable benefits.[24] Formal databases allow for concurrent editing and for setting information sanity checks to ensure clean data. For example, our data enterers can only input certain surgeons names and certain procedures rather than having free text input fields. The database also checks that valid times are entered (e.g. procedure start times must be numbers that occur before the recorded end times). These simple sanity checks can bolster confidence in the data quality during informal assessments of the captured data. Though deploying a formal database seems daunting, check your institution's resources. You may find that your institution can offer you these services for free or at a

low cost. For example, we leveraged an institutional database server to offload the mundane, though highly important, tasks of database administration such as backups, tuning, and security, giving us the benefits of a formal information database without the administrative headaches.

Once the data is appropriately captured and stored, it must undergo processing from its raw state to a state amenable to analysis and modeling. Bennett *et al* presented a framework for optimal workflow for radiology datasets which shares many similarities with our own pipeline (Figure 1).[25] Think of data processing as a series of finer-and-finer filters: each filter removes extraneous details and standardizes the data until it reaches a usable final state. Take videos: videos for one operation often span multiple smaller video files that will need to be combined, trimmed to contain only operative footage, stripped of audio (if necessary), de-identified, and finally converted to an appropriate format for long-term storage and model creation. Performing each step by-hand would take hours per video, and quickly, the captured data volume (once a culture of capture has been established at your institution!) will consume the lab member's time and prevent actual research from occurring. Fortunately for videos and images, there is ffmpeg and ImageMagick, both powerful open-source tools that can perform all this manipulation and more in an automated fashion.[26,27]

Our group created a program to expedite our data pipeline, and now all videos are rapidly processed, uploaded to the central server, and cataloged in the database, with just a single command, freeing up members for more important tasks. Thus, automating as much of the capture, processing, and storage of the database will be important to preserve your

research team's time for actual research tasks. However, we stress that the initial, appropriate capture and catalogue of data is just as critical a step as the modeling and analysis.

Data Security and Anonymity

A comprehensive guide on data security and safe protected health information (PHI) handling is beyond the scope of this chapter. We do, however, want to underline the importance of data safety and anonymity. You should create, for each research project, a thoughtful data security plan.

Data security occurs at three phases: rest, motion, and in use. Data at rest refers to stored data on a computer, server, or portable hard drive. Storage, especially if the data travels off the main research campus, must be encrypted. Similarly, during data transfer between remote storage and data-processing servers, all communication should be encrypted. Encryption while the data is actively in use is a yet unsolved problem: in order for the data to make sense to any program using it, it has to be decrypted. Raskar's group at the Massachusetts Institute of Technology (MIT) have proposed an interesting work-around: split learning.[28] Split learning allows neural networks to train across multiple data sources, so that data never needs to leave a local facility and risk compromise (Figure 2). There is no large scale application of this technology as of yet, but implementations may become more common as data privacy laws such as the GDPR continue to be enacted.

The hassle that comes with encryption and enhanced data safety may seem unnecessary, especially if surgical media was de-identified. True "de-identification"

however, is a goal that is increasingly more difficult, if not impossible, to achieve. Sweeney, an expert in the "re-identification" field, could exactly identify health records for almost 50% of people who had newspaper stories written on their hospital visits from "de-identified" health record databases.[29] Similarly, researchers showed that with increasing number of demographic attributes in a database, the likelihood of re-identification increases. With just fifteen attributes, 99.98% of "de-identified" patients could be re-identified.[30] Therefore, since all data is essentially PHI, it should be treated as such.

Efforts should still be made to minimize the amount of PHI contained in the data, as each variable removed makes the cost function of re-identification more expensive. As an example, when the laparoscope is extra-corporeal for intermittent lens cleaning during an operation, the camera often incidentally captures images from the operating room itself, such as the clock on the side of the wall or a whiteboard on which the date and surgeon are written. As an example of how reidentification can occur, we have actually been able to use this information leakage to help identify videos to particular surgeries when a glitch in the video recorder stopped it from recording medical record numbers. To eliminate this information leakage, we use ffmpeg, the software previously mentioned, which has a filter to blur images during certain time intervals. There are few published guides to systematic media de-identification, though one excellent example came from the TCIA group who describe thorough procedures for de-identification of radiology images.[31] Ultimately, applying extra effort during initial media processing will help keep videos more secure, prevent regulatory punishment in the future, and more importantly, keep patient information safe.

Data Annotation

The arduous journey towards usable data has one last peak to overcome: data annotation (also referred to as labeling). Much of the work thus far in surgical computer vision has utilized supervised learning, where models train on human-labeled data. Sadly there is a scarcity of pre-existing labeled data. This pre-labeled data usually is of little utility for more than a narrow question. For example, even though a dataset may have extensive labeling of surgical tools, if the model hopes to identify correct phases, each video will need to be re-watched and re-labeled, similar to the what Zisimopoulos's group did for their DeepPhase paper.[9] A key component of productive computer vision research is minimization of data preparation costs, so below we will describe one work-flow to maximize annotation efficiency.

Many tools exist for data annotation. Unfortunately, the majority of them focus on data labeling for a particular problem, such as tool tracking, gesture tracking, or labeling of objects in a video frame.[32] When searching for a suitable software, ensure that it is easy to use, works on all operating systems in the lab, can do the labeling needed for the project, and can output to a file format that your modeling tools can read. We in the past have used two programs: Anvil and VIA. Anvil is an annotation tool originally meant for labeling dialogue and gestures but has convenient ways to label phases of an operation. It also outputs the labels into a file format that is easy to process (XML).[33] We also use VIA, an annotation software that runs straight in the browser and allows for similar temporal segmentation in addition to direct labeling and outlining of objects in the video. It also outputs to easily readable plain text formats (JSON and CSV).[34] These are not the only

available annotation tools, and Table 1 contains a (far from comprehensive) list of known annotation tools that are available.

Now with the tools to label data, an annotation consensus and standard must be reached. No community-wide surgical standard for media annotation exists at this point so each group will need to create their own annotation consensus. Take, as an example, phase segmentation of operation videos. The group must agree upon the phases to label, their exact start and stop times, and whether or not they can repeat. Truly drill down to a granularity that will prevent disagreement between multiple annotators. With an initial draft, then annotate a select few media with multiple annotators. Afterwards, compare the annotators' notes and closely review their areas of disagreements to then create a revised annotation guide.

You will likely be surprised by how many unplanned variations occur in the annotations. Take dissection of a structure: does the phase start when the instrument is in view or when it touches the tissue? Does it end every time it leaves the tissue or just when the next phase start? What should happen when the laparoscope leaves the body? This refinement process will take time and is nicely encapsulated by the Japanese word "Kaizen" which represents an iterative testing and refinement process.[35] Put in the time up front because inconsistent labeling will only lead to poor modeling performance. The model is only as good as its data inputs. Poor model performance from inconsistent annotations will lead to a serious time-sink that a little up-front effort will avoid.

Understandably, manually labeling large numbers of images and videos is quite costly from a time-perspective, particularly since it requires highly-educated labelers (for

many labels only trained surgeons with suffice). Research groups must plan for an annotation budget to help accrue multiple annotators. Additionally, a training program for data annotation should be created (that is, to train new annotators). Lastly, annotators should cultivate a continued "Kaizen" culture with regular annotator audits and re-training to ensure high-quality data labeling.

Recent studies have tried to reduce the labor-intensive labeling process. If computer vision models will need hundreds of videos to train, then the data-labeling task will rapidly outpace the ability of surgeons to annotate. Some groups have looked at pre-training models on unlabelled data to then reduce the amount of actual labelled data needed to perform accurate predictions.[36] Others have trained a neural network on a small proportion of labeled videos then use this trained network to generate annotations on other videos. Using both synthetically generated annotations and a small number of human-generated annotations, Yu *et al* could almost halve the prediction accuracy gap between a model trained solely on 20 annotations versus a model trained on 80 annotations.[37] With continued advances, models will one day generate their own annotations and only require a trained-human input for those on which the model is uncertain (i.e. semi- or weakly supervised learning). Until then, we will continue to have to dedicate large amounts of time towards annotations.

Modeling

With good data, now the modeling and predictions can begin. Just like data collection, modeling requires a broad toolset. First, the lab will need to pick a programming

language in which to design the models. The most commonly used languages include Python, C++, Julia, and R. We use Python given its readability, ease to learn, and extensive machine learning package availability. In fact, the top ten packages used for machine learning on GitHub are only compatible with Python programs.[38] Next a deep learning framework should be selected for easy neural network development and to optimize model training at any scale. Our lab uses PyTorch though there are other popular alternatives, including TensorFlow.[39,40]

Once a model is programmed, it will need appropriate hardware for training and testing. Machine learning and computer vision is quite computer resource intensive. Typically models leverage computer Graphical Processing Units (GPUs) to perform the myriad calculations required. These GPUs were developed originally for computer graphical tasks, such as video games but were found to be well-suited towards machine learning computations with orders of magnitude improvements over traditional Computer Processing Units (CPUs).[41] Machine learning specific chips also have been created, such as the Tensor Processing Unit (TPU) from Google, which runs with improved speed and energy costs.[42] GPUs can be accessed either locally using a computer or server within your institution or through third-party services such as those offered by Amazon, Microsoft, and Google.

Training and running models is a costly endeavor with both resource and expertise-requiring challenges. The monetary cost of model development at first seems reasonable: one model averages 120 hours to train with a cloud compute cost of around $100. However, effective models will need parameter tuning to each dataset which estimates to

nearly 2,880 training hours with cloud compute costs of up to $4,000. Beyond training hours and direct cost for cloud computing power, these models generate large carbon dioxide emissions. Training one model, including tuning, generates almost 80,000 pounds of carbon dioxide, which is over double that produced by an American citizen each year.[43]

Gaining the expertise to appropriately select a neural network and optimize its parameters also presents a significant challenge. It is not a field quickly learned: small changes to a model's hyperparameters can lead to an inability to train on a dataset. Even modifying the random seed for a model with otherwise identical network and hyperparameters can lead to learning curves with zero overlap of distribution.[44] Machine learning and computer vision additionally are rapidly evolving fields. Few medical practitioners will have the expertise for thoughtful and productive work machine learning analysis. Typically medical groups will overcome the knowledge gap through partnerships with groups who have the necessary computer vision expertise, as we did with MIT's Computer Science and Artificial Intelligence Laboratory. Their collaboration was critical for our success at tackling surgical computer vision problems.

If resources such as a world-leading institution are not available, there is an alternative: automated machine learning (AutoML). AutoML seeks to provide off-the-shelf solutions to non-experts without requiring machine learning knowledge. AutoML solutions will select a model, optimize hyperparameters, interpret the results, and analyse data.[45] Currently available options include the Google Cloud AutoML and Microsoft AzureML.[46,47] Faes *et al* performed a proof-of-concept study where they had two physicians with no previous programming nor deep learning experience use AutoML to create models for

image classification on public medical datasets. They then compared the AutoML models' predictions against the previously published human-designed models and found nearly comparable performance on smaller datasets.[48]

This AutoML approach, however, still has short-comings. The environmental cost is significant: AutoML uses a method called neural architecture search to automatically select networks and hyperparameters. [49] Training a neural architecture search can increase the carbon dioxide emissions of a model over 3,000-fold. This increase translates to a carbon emission that five cars would produce over their lifetime.[43] Despite this drastic increase in resource usage, AutoML unacceptably does not reach performance parity with hand-designed models. In the physician proof-of-concept paper from Faes *et al*, AutoML performed similar to traditional models only in binary classification (abnormal versus normal) with poor performance on labeling images. For example, it classified a melanoma as a nevus 28.6% of the time. When the AutoML models were used on an external test dataset, their near perfect accuracy then dropped to less than 50%.[48] Clearly further development is necessary before any medical practitioner can become a "machine-learning" researcher in a day with an AutoML approach. Rather than replacing machine learning experts, AutoML's future may involve model development facilitation through initial guidance on hyperparameter choices and neural architecture design that then trained machine learning experts can build upon to improve performance.[50]

Conclusion

Although it is a young field, computer vision and its applications in the medical arena promise an exciting future. Computer vision work first starts with a defined question. Second, it requires meticulous data acquisition and processing. Third, experts label the data following a reproducible annotation consensus. Lastly, with the prepared data, researchers can train deep learning models to answer their question, whatever it may be. The above chapter presented the MGH Saiil experience and provided a roadmap to avoid the numerous hurdles any researcher is destined to hit when utilizing computer vision. An AI-augmented future is exciting, and we cannot wait to experience it with you.

References

1. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*.; 2012:1097-1105.

2. Shoham Y, Perrault, Raymond, Brynjolfsson, Erik, et al. *2018 Annual Report*. Artificial Intelligence Index https://aiindex.org. Accessed September 13, 2019.

3. Heller N, Sathianathen N, Kalapara A, et al. The KiTS19 Challenge Data: 300 Kidney Tumor Cases with Clinical Context, CT Semantic Segmentations, and Surgical Outcomes. *arXiv:190400445 [cs, q-bio, stat]*. March 2019. http://arxiv.org/abs/1904.00445. Accessed September 11, 2019.

4. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: 2017:2097-2106. http://openaccess.thecvf.com/content_cvpr_2017/html/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.html. Accessed September 11, 2019.

5. Decencière E, Zhang X, Cazuguel G, et al. FEEDBACK ON A PUBLICLY DISTRIBUTED IMAGE DATABASE: THE MESSIDOR DATABASE. *Image Analysis & Stereology*. 2014;33(3):231-234. doi:10.5566/ias.1155

6. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*. 2018;5:180161. doi:10.1038/sdata.2018.161

7. Prior FW, Clark K, Commean P, et al. TCIA: An information resource to enable open science. In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*.; 2013:1282-1285. doi:10.1109/EMBC.2013.6609742

8. Al Hajj H, Lamard M, Conze P-H, et al. CATARACTS: Challenge on automatic tool annotation for cataRACT surgery. *Medical Image Analysis*. 2019;52:24-41. doi:10.1016/j.media.2018.11.008

9. Zisimopoulos O, Flouty E, Luengo I, et al. DeepPhase: Surgical Phase Recognition in CATARACTS Videos. *arXiv:180710565 [cs, stat]*. July 2018. http://arxiv.org/abs/1807.10565. Accessed September 14, 2019.

10. Flouty E, Kadkhodamohammadi A, Luengo I, et al. CaDIS: Cataract Dataset for Image Segmentation. *arXiv:190611586 [cs]*. June 2019. http://arxiv.org/abs/1906.11586. Accessed September 11, 2019.

11. Nakawala H. Nephrec9. November 2017. doi:10.5281/zenodo.1066831

12. Allan M, Shvets A, Kurmann T, et al. 2017 Robotic Instrument Segmentation Challenge. *arXiv:190206426 [cs]*. February 2019. http://arxiv.org/abs/1902.06426. Accessed September 14, 2019.

13. Narges Ahmidi, Tao L, Sefati S, et al. A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery. *IEEE transactions on bio-medical engineering*. 2017;64(9):2025-2041. doi:10.1109/TBME.2016.2647680

14. Twinanda AP, Shehata S, Mutter D, Marescaux J, Mathelin M de, Padoy N. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *arXiv:160203012 [cs]*. February 2016. http://arxiv.org/abs/1602.03012. Accessed September 11, 2019.

15. Stauder R, Ostler D, Kranzfelder M, Koller S, Feußner H, Navab N. The TUM LapChole dataset for the M2CAI 2016 workflow challenge. *arXiv:161009278 [cs]*. October 2016. http://arxiv.org/abs/1610.09278. Accessed September 11, 2019.

16. Format Descriptions for Still Image. https://www.loc.gov/preservation/digital/formats/fdd/still_fdd.shtml. Accessed September 12, 2019.

17. Format Descriptions for Moving Images. https://www.loc.gov/preservation/digital/formats/fdd/video_fdd.shtml. Accessed September 12, 2019.

18. Vohl D, Fluke CJ, Vernardos G. Data Compression in the Petascale Astronomy Era: A GERLUMPH case study. *Astronomy and Computing*. 2015;12:200-211. doi:10.1016/j.ascom.2015.05.003

19. Backup Overview dpBestflow. https://dpbestflow.org/node/262. Accessed September 12, 2019.

20. Natarajan P, Frenzel JC, Smaltz DH. *Demystifying Big Data and Machine Learning for Healthcare*. Boca Raton: CRC Press, Taylor & Francis Group; 2017.

21. Cloud Storage Pricing Comparison: Amazon S3 vs Azure vs B2. https://www.backblaze.com/b2/cloud-storage-pricing.html. Accessed September 15, 2019.

22. Samba - opening windows to a wider world. https://www.samba.org/. Accessed September 12, 2019.

23. Staubach P, Pawlowski B, Callaghan B. NFS Version 3 Protocol Specification. https://tools.ietf.org/html/rfc1813. Accessed September 12, 2019.

24. PostgreSQL: The world's most advanced open source database. https://www.postgresql.org/. Accessed September 15, 2019.

25. Bennett W, Smith K, Jarosz Q, Nolan T, Bosch W. Reengineering Workflow for Curation of DICOM Datasets. *Journal of Digital Imaging*. 2018;31(6):783-791. doi:10.1007/s10278-018-0097-4

26. FFmpeg. https://www.ffmpeg.org/. Accessed September 12, 2019.

27. LLC IS. ImageMagick. *ImageMagick*. https://imagemagick.org/. Accessed September 17, 2019.

28. Gupta O, Raskar R. Distributed learning of deep neural network over multiple agents. *arXiv:181006060 [cs, stat]*. October 2018. http://arxiv.org/abs/1810.06060. Accessed September 17, 2019.

29. Sweeney L. Only You, Your Doctor, and Many Others May Know. *Technology Science*. September 2015. /a/2015092903/. Accessed September 12, 2019.

30. Rocher L, Hendrickx JM, Montjoye Y-A de. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*. 2019;10(1):1-9. doi:10.1038/s41467-019-10933-3

31. Freymann JB, Kirby JS, Perry JH, Clunie DA, Jaffe CC. Image Data Sharing for Biomedical Research—Meeting HIPAA Requirements for De-identification. *Journal of Digital Imaging*. 2012;25(1):14-24. doi:10.1007/s10278-011-9422-x

32. Gaur E, Saxena V, Singh SK. Video annotation tools: A Review. In: *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*.; 2018:911-914. doi:10.1109/ICACCCN.2018.8748669

33. Kipp M. *ANVIL - A Generic Annotation Tool for Multimodal Dialogue*.; 2001.

34. Dutta A, Zisserman A. The VIA Annotation Software for Images, Audio and Video. *arXiv:190410699 [cs]*. April 2019. doi:10.1145/3343031.3350535

35. Imai M. *Kaizen (Ky'zen): The Key to Japan's Competitive Success*. New York: McGraw-Hill; 1986.

36. Funke I, Jenke A, Mees ST, Weitz J, Speidel S, Bodenstedt S. Temporal Coherence-based Self-supervised Learning for Laparoscopic Workflow Analysis. In: Stoyanov D, Taylor Z, Sarikaya D, et al., eds. *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Lecture Notes in Computer Science. Springer International Publishing; 2018:85-93.

37. Yu T, Mutter D, Marescaux J, Padoy N. Learning from a tiny dataset of manual annotations: A teacher/student approach for surgical phase recognition. *arXiv:181200033 [cs, stat]*. November 2018. http://arxiv.org/abs/1812.00033. Accessed September 18, 2019.

38. The State of the Octoverse: Machine learning. *The GitHub Blog*. January 2019. https://github.blog/2019-01-24-the-state-of-the-octoverse-machine-learning/. Accessed September 4, 2019.

39. Abadi M, Barham P, Chen J, et al. TensorFlow: A System for Large-Scale Machine Learning. In: 2016:265-283. https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi. Accessed September 20, 2019.

40. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch. October 2017. https://openreview.net/forum?id=BJJsrmfCZ. Accessed September 20, 2019.

41. Mittal S, Vaishay S. A survey of techniques for optimizing deep learning on GPUs. *Journal of Systems Architecture*. 2019;99.

42. Jouppi N, Young C, Patil N, Patterson D. Motivation for and Evaluation of the First Tensor Processing Unit. *IEEE Micro*. 2018;38(3):10-19. doi:10.1109/MM.2018.032271057

43. Strubell E, Ganesh A, McCallum A. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:190602243 [cs]*. June 2019. http://arxiv.org/abs/1906.02243. Accessed September 4, 2019.

44. Henderson P, Islam R, Bachman P, Pineau J, Precup D, Meger D. Deep Reinforcement Learning that Matters. *arXiv:170906560 [cs, stat]*. September 2017. http://arxiv.org/abs/1709.06560. Accessed September 10, 2019.

45. Truong A, Walters A, Goodsitt J, Hines K, Bruss CB, Farivar R. Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools. *arXiv:190805557 [cs, stat]*. August 2019. http://arxiv.org/abs/1908.05557. Accessed September 12, 2019.

46. Cloud AutoML - Custom Machine Learning Models AutoML. *Google Cloud*. https://cloud.google.com/automl/. Accessed September 20, 2019.

47. AutoML. *Microsoft Research*. https://www.microsoft.com/en-us/research/project/automl/. Accessed September 20, 2019.

48. Faes L, Wagner SK, Fu DJ, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: A feasibility study. *The Lancet Digital Health*. 2019;1(5):e232-e242. doi:10.1016/S2589-7500(19)30108-6

49. Zoph B, Le QV. Neural Architecture Search with Reinforcement Learning. *arXiv:161101578 [cs]*. November 2016. http://arxiv.org/abs/1611.01578. Accessed September 20, 2019.

50. Wang D, Weisz JD, Muller M, et al. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *arXiv:190902309 [cs]*. September 2019. doi:10.1145/3359313