POLITECNICO DI TORINO
Repository ISTITUZIONALE

Generative AI Pipeline with Model-Guided Filtering for Sim-to-Real Transfer in Surgical Imaging

Original
Generative AI Pipeline with Model-Guided Filtering for Sim-to-Real Transfer in Surgical Imaging / Leoncini, P., Marzola, F., Pescio, M., Muratore, L., Revello, L., Barontini, F., Distefano, G., Hayashi, K., Ammirati, C.A., Arezzo, A., Dagnino, G.. - In: COMPUTERIZED MEDICAL IMAGING AND GRAPHICS. - ISSN 0895-6111. - 132:(2026). [10.1016/j.compmedimag.2026.102775]

Availability:
This version is available at: 11583/3011040 since: 2026-05-19T10:11:40Z

Publisher:
Elsevier

Published
DOI:10.1016/j.compmedimag.2026.102775

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

17 September 2026

Computerized Medical Imaging and Graphics 132 (2026) 102775

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Generative AI pipeline with model-guided filtering for sim-to-real transfer in surgical imaging

Pietro Leoncini [a,c,1], Francesco Marzola [a,1,*], Matteo Pescio [a,c], Luigi Muratore [d], Lorenzo Revello [a], Federica Barontini [a], Giovanni Distefano [a], Kengo Hayashi [e], Carlo Alberto Ammirati [a], Alberto Arezzo [a], Giulio Dagnino [a,b]

[a] Department of Surgical Sciences, Università degli Studi di Torino, Corso Dogliotti 14, Turin, TO 10126, Italy
[b] Robotics and Mechatronics, University of Twente, Drienerlolaan 5, Enschede, NB 7522, Netherlands
[c] DIMEAS, Italy
[d] DET Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin, TO 10129, Italy
[e] Kanazawa University Hospital, Kanazawa, Japan

ARTICLE INFO

Keywords:
Synthetic Data Generation
Sim-to-Real Transfer
Generative Models
Data-Centric AI
Automated Filtering
Surgical Autonomy
Robotic Suturing

ABSTRACT

Automating surgical suturing requires reliable computer vision systems, yet annotated real surgical datasets remain scarce, costly, and difficult to obtain. To address this challenge, we introduce a data-centric pipeline that combines synthetic data generation, generative realism boosting, and model-guided filtering to improve sim-to-real transfer without relying on real annotated surgical footage. Synthetic images were created in Unity with both type-based and part-based instruments annotations, then enhanced using CycleGAN-TURBO for unpaired image-to-image translation and Real-ESRGAN for high-resolution restoration. A YOLO-based selector model, trained on synthetic images, assessed the quality of generatively enhanced data through Dice similarity scoring, discarding samples with distortions or misalignments. In the part-based configuration, on a real test set, the baseline model trained solely on synthetic images achieved a Dice score of 0.17, while combining synthetic with unfiltered enhanced data reached 0.24. Filtering proved decisive: accepted enhanced images combined with a synthetic (hybrid curated dataset) further boosted scores to 0.44. Fine-tuning strategies yielded only marginal gains, confirming that improvements were driven primarily by data quality rather than training variations. In the type-based setup, the hybrid curated dataset achieved a mean Dice score of 0.65, a substantial improvement over previous fully synthetic baselines (0.384) without requiring real training annotations. These results demonstrate that curation of generative outputs is critical for sim-to-real transfer in surgical vision. By uniting synthetic generation, generative realism, and automated filtering, this pipeline enables scalable, low-cost dataset creation, providing resources on GitHub and a reproducible foundation for developing reliable perception systems and advancing autonomy in surgical robotics.

1. Introduction

Automation in robotic-assisted surgery promises transformative benefits: enhanced precision, reduced operative times, and lower surgeon fatigue (Troccaz et al., 2019). Among the most critical and time-consuming tasks is suturing, a key part of minimally invasive procedures where even small ergonomic or cognitive challenges can compromise outcomes (Ostrander et al., 2024; Caballero et al., 2025). For this reason, automating suturing is widely regarded as a pathway to standardizing results and alleviating the burden on surgeons (Attanasio et al., 2021; Dagnino and Kundrat, 2024).

Central to this endeavor lies the challenge of surgical tool segmentation, a prerequisite for reliable instrument tracking, motion analysis, and ultimately surgical autonomy (Colleoni et al., 2020). Yet, building such vision systems requires large, well-annotated datasets, resources that are costly, time-consuming to obtain, and heavily dependent on expert annotators (Man and Chahl, 2022; Ahmed et al., 2024).

The difficulty of surgical data acquisition and annotation has become

* Corresponding author.
E-mail address: francesco.marzola@unito.it (F. Marzola).
[1] These authors contributed equally to this work

https://doi.org/10.1016/j.compmedimag.2026.102775
Received 30 September 2025; Received in revised form 16 February 2026; Accepted 11 May 2026
Available online 12 May 2026
0895-6111/© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

one of the major bottlenecks in Surgical Data Science (Rivoir et al., 2024; Majeed and Hwang, 2025). A single minimally invasive surgery video can span tens of thousands of frames, each requiring anonymization and expert labeling. This process not only draws away valuable surgeon time but also raises privacy and ethical challenges tied to patient data (Majeed and Hwang, 2025). As a result, most existing datasets remain small, imbalanced, and insufficiently diverse, limiting the generalizability of deep learning models across procedures and environments (Ahmed et al., 2024).

Synthetic data generation has emerged as a compelling solution to these challenges (Cao et al., 2025; Majeed and Hwang, 2025). By simulating surgical environments, researchers can create vast, diverse, and fully annotated datasets at a fraction of the cost and without privacy concerns (Rodrigues et al., 2022). In our previous work (Leoncini et al., 2025), we proposed a reproducible framework for generating synthetic surgical data and demonstrated how hybrid training, combining synthetic and a small fraction of real data, could significantly improve real-world model performance. However, this approach still relies on real data and synthetic datasets alone face the persistent limitation of the sim-to-real gap: models trained exclusively on simulated images often fail to generalize to the visual complexity of real environments (Hinterstoisser et al., 2019; Rivoir et al., 2024; Leoncini et al., 2025).

Recent advances in generative AI, including GANs, diffusion models, and hybrid approaches, promise to narrow this gap by enhancing the realism of synthetic data (Rivoir et al., 2024; Wang et al., 2025). Among these, CycleGAN-TURBO, a one-step image-to-image diffusion translation model adapted through adversarial learning, offers efficient, content-preserving translation even in unpaired settings (Parmar et al., 2024). While capable of generating visually compelling outputs, such models may also hallucinate artifacts or produce redundant images, which, if used uncritically, can introduce noise and reduce model generalization. This underlines a central lesson from data-centric AI: improving the quality of training data is often more beneficial than simply increasing its quantity (Bhatt et al., 2024; Joshi et al., 2024; Zhou et al., 2024). Several strategies have been explored to enhance dataset quality within the broader domain of data-centric AI, yet they face specific challenges when applied to generative sim-to-real pipelines. Data cleaning focuses primarily on detecting and filtering label noise or outliers (Côté et al., 2024), but it is generally incapable of identifying when a generative model has hallucinated structural changes in an otherwise correctly labeled image. Data pruning aims for training efficiency by removing redundant or duplicate samples (Yang et al., 2022; Yang et al., 2024); however, because these methods prioritize diversity, they may inadvertently retain unique but geometrically distorted images that degrade segmentation accuracy. Similarly, data reduction focuses on finding a minimal representative coreset (Perera-Lago et al., 2024; Chen and Zhou, 2025), assuming the underlying data is valid rather than verifying its structural integrity. In the context of automatic curation pipelines, existing methods often rely on feature clustering or entropy measures to select informative samples (Vo et al., 2024). These approaches are frequently blind to pixel-level geometric alignment; they can determine if a generated image looks like a surgical tool but cannot verify if that tool precisely matches the synthetic ground truth mask. Alternative strategies have attempted to enforce structural integrity by integrating semantic-geometric consistency losses directly into the model architecture or the training loop (Huang et al., 2025). While effective, these methods increase training complexity and may still struggle with the high-frequency distortions common in unpaired image-to-image translation. Our work introduces a distinct, simpler strategy: a post-generative, model-guided filtering mechanism based on the Dice Similarity Coefficient (DSC) computed during inference. Unlike pruning or reduction methods focused on dataset compactness or training-time constraints that modify the learning objective, our approach acts as an autonomous quality gate. It specifically targets the removal of noisy synthetic-to-real translations where geometric fidelity has been lost, ensuring that only the most reliable generative samples are used to bridge the sim-to-real gap.

This emphasis on quality over quantity reflects a broader paradigm shift in machine learning. Traditionally, research followed a model-centric approach, where performance improvements came from designing new architectures or tuning hyperparameters, with little attention to the underlying data. By contrast, the data-centric approach emphasizes preparing higher-quality datasets and has been shown to outperform purely model-focused strategies (Bhatt et al., 2024). Our work contributes to this vision by providing a practical, domain-specific pipeline for surgical data curation, generalizable to other domains.

While our earlier framework (Leoncini et al., 2025) established the generation of synthetic datasets for robotic suturing, this work advances to an autonomous, end-to-end, data-centric sim-to-real pipeline. Rather than scaling data volume, we target a key limitation of generative transfer, preserving semantic–geometric consistency, so that surgical vision models can be trained with substantially reduced reliance on real, privacy-sensitive, and labor-intensive clinical data. Moreover, this work introduces a finer-grained, part-based instrument representation for higher surgical precision.

The core contributions and implications of this work are summarized as follows:

1. Reduction of real-annotated data dependency: by combining simulated environments with a self-correcting generative pipeline, we demonstrate a pathway to train high-performance models that minimize or eliminate the reliance on scarce, privacy-sensitive real-world surgical images. This strategy is inherently generalizable, providing a blueprint for any domain where real-world annotations are hard to obtain.
2. Autonomous self-curation: our model-guided filtering serves as a post-generative quality gate, replacing labor-intensive manual data auditing with an automated mechanism. This moves the field toward fully autonomous perception pipelines that can evaluate their own training data for geometric and quality reliability.
3. Computational efficiency: the filtering mechanism ensures that training resources are not wasted on noisy or distorted samples, leading to more efficient model convergence and stronger generalization by prioritizing high-fidelity data over sheer volume.

Unlike the previous framework, which relied on raw synthetic data, this pipeline acts as a realism booster and an autonomous quality gate, automatically discarding hallucinated or noisy generative outputs that would otherwise degrade model generalization. The results confirm that careful data curation leads to consistent improvements over unfiltered approaches. More importantly, the pipeline offers an extensible and open foundation for surgical vision research. This contribution represents the next logical step: moving from synthetic data generation toward autonomous data curation, a shift essential for advancing autonomy in surgical robotics (Cao et al., 2025). To ensure reproducibility and support the research community, we extend our previously released open-source repository[1] by sharing new Unity-based virtual scenarios for synthetic data generation, the enhanced images, and annotations produced in this study. By consolidating both works into a single resource, we provide datasets, the tools, and code necessary for others to replicate, adapt, and build upon this framework.

2. Materials and methods

2.1. Data generation

2.1.1. Synthetic data generation: unity

Synthetic surgical datasets were generated using the Unity3D engine

[1] https://github.com/PietroLeoncini/Surgical-Synthetic-Data-Generation-and-Segmentation

(Unity Technologies, version 2022.3.24f1) and the Perception Package (version 1.0.0-preview.1) for automatic annotation. This process builds upon the framework presented in our previous work (Leoncini et al., 2025), where three distinct simulation scenarios were originally employed. In the present study, we refined the workflow by consolidating the simulation into a single, more versatile virtual environment. This environment was enriched with other randomizers to increase intra-scenario variability, including variations in camera position and orientation, lighting conditions, and randomized tool poses and orientations.

To enable finer-grained visual understanding, all surgical tool models were also reconfigured into part-based (PB) representations, where each instrument (e.g., Cadiere Forceps, Needle Driver) was decomposed into functionally distinct segments (e.g., Tool-Shaft, Tool-Wrist, and Tool-Clasper) as shown in Fig. 1. This configuration provides more precise instance segmentation and allows the vision system to identify tool sub-components that are critical for tasks such as suturing. In parallel, a type-based (TB) dataset was also generated to compare the results with the previous work.

Both type-based and part-based annotations were produced automatically using the Perception toolkit's ground-truth labeling. A total of 5000 high-definition synthetic images (1920 ×1080) were generated in this stage, with instance segmentation masks covering all relevant surgical objects, including instruments, tissue, needle, and thread.

2.1.2. Real data acquisition and labeling

To provide external test sets for evaluating generalization, two small real datasets of 40 images each were constructed. These images were extracted from 2 slightly different videos acquired using the da Vinci robotic surgical system's endoscopic camera. Frames were sampled at approximately 1.3-second intervals to ensure temporal independence and representative coverage of the surgical scene.

The selected frames (resolution 1920 ×1080) depicted a representative suturing setup: a porcine colon with a 3 cm horizontal cut placed centrally, flanked by a Needle Driver and a Cadiere Forceps, with a surgical needle positioned within the operative field. Manual annotations were created using the Roboflow platform (Dwyer et al., 2024). Each frame was labeled with segmentation masks for all instruments, tissue, needle, and thread, both in type-based and part-based formats (Fig. 2). These datasets serve exclusively as evaluation sets and were not included in the training pipeline.

2.2. Image-to-image translation and super-resolution

The synthetic datasets were enhanced through a two-step pipeline combining image-to-image translation and super-resolution upscaling.

2.2.1. Generative translation with CycleGAN-TURBO

To reduce the visual gap between simulated and real surgical images, we applied CycleGAN-TURBO (Parmar et al., 2024), a one-step diffusion model adapted with adversarial training. A key feature of this model is its unpaired training strategy, which enables translation between domains without requiring one-to-one correspondence between synthetic and real images. This property is particularly valuable in surgical contexts, as it removes the constraint of creating synthetic images that exactly replicate real ones and instead allows broader exploration of variability during simulation.

Equally important, CycleGAN-TURBO is designed to preserve the input structure of the original image while transferring realistic textures and color distributions. For this study, this structural fidelity was essential because annotations generated in Unity rely on precise spatial correspondence. Distortion or misalignment introduced by the generative model would render those annotations unusable. TURBO's ability to maintain spatial geometry while enhancing realism, therefore ensured that the automatically generated segmentation masks could still be reliably exploited in the subsequent pipeline.

Due to computational constraints, images were first downscaled from 1920 × 1080–320 × 184. This resolution was selected empirically, as TURBO exhibited more stable and consistent translations at this scale for our domain. The model was trained using unpaired sets of 5000 synthetic and real surgical images and then applied to the synthetic dataset to produce 5000 generatively enhanced images. Fig. 3 shows a comparison of the images in the three different domains.

2.2.2. Super resolution with real-ESRGAN

After generative enhancement, the TURBO outputs (320 ×184) were restored to full resolution using Real-ESRGAN (Wang et al., 2021). To adapt the model to surgical imagery, it was trained with synthetic image pairs consisting of high-resolution frames (1920 ×1080) and their downscaled counterparts.

The upscaling was performed in stages. TURBO-enhanced images were first upscaled to 640 × 368, then doubled again to 1280 × 736 using Real-ESRGAN, and finally resized with bicubic interpolation to reach the native resolution of 1920 × 1080. This progressive strategy reduced artifacts that typically arise from single large-scale upscaling operations.

The outcome of this stage was a set of 5000 high-resolution, realism-enhanced images, each aligned with the original Unity-automatically-generated annotations. Despite the improvements, it is important to note that both CycleGAN-TURBO and Real-ESRGAN occasionally introduced artifacts, distortions, or subtle annotation misalignments, motivating the need for a subsequent curation step.

2.3. Self-curation pipeline

To ensure only high-quality enhanced images are included in model training, we designed a model-guided self-curation pipeline. This pipeline integrates a closed-loop filtering mechanism in which a YOLO-based "selector model" evaluates the quality of enhanced images and automatically rejects samples likely to degrade model performance.

2.3.1. Selector model

The selector model is a YOLOv11-segmentation network (Glenn and Jing, 2024) trained exclusively on 1000 synthetic images. Its role is not task deployment but rather quality assessment: by running inference on enhanced images, it provides segmentation predictions that can be compared against the known ground-truth masks generated in Unity.

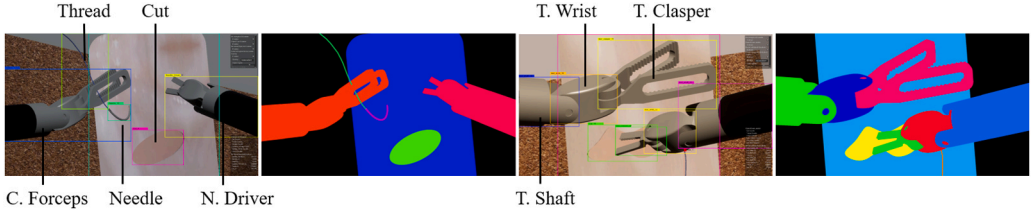


Fig. 1. Example of type-based (a) and part-based (c) synthetic images generated in Unity and their corresponding instance segmentation masks (b, d). Type-based classes: Cut, Tissue, Needle, Cadiere-Forceps, Needle-Driver, Thread. Part-based classes: Cut, Tissue, Needle, Tool-Shaft, Tool-Wrist, Tool-Clasper, Thread.

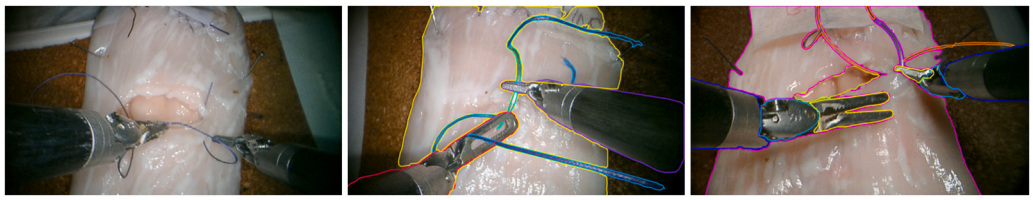


Fig. 2. (a) Example of a real RGB image obtained by sampling a video taken from the Da Vinci endoscope. (b) Example of manual annotation for Type-based configuration. (c) Example of manual annotation for Part-based configuration.

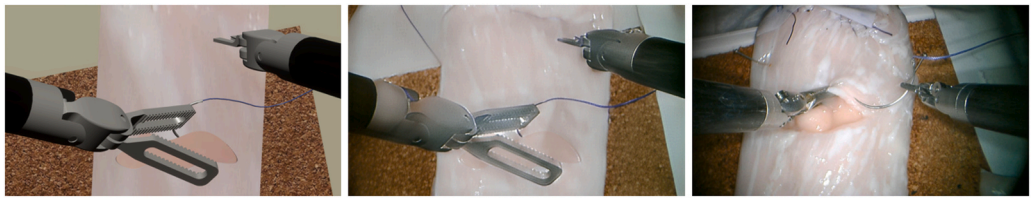


Fig. 3. (a) Synthetic RGB image generated in Unity and its corresponding TURBO-enhanced image (b). Real RGB image obtained by sampling a video taken from the Da Vinci endoscope.

2.3.2. *Dice-based quality assessment*

Quality assessment of the enhanced images was carried out using the Dice similarity coefficient (DSC), computed on a per-class basis (needle, tissue, tool parts, etc.). The DSC is defined as:

$$Dice = \frac{2TP}{2TP + FP + FN} = \frac{2 * P * R}{P + R} \quad (1)$$

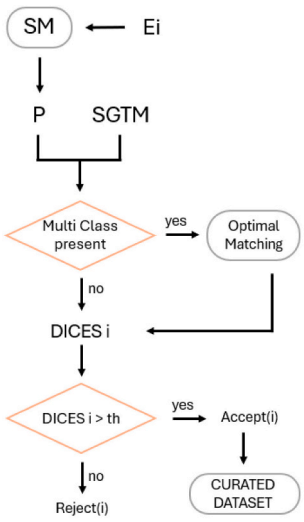
Where TP denotes true positives, FP false positives, FN false negatives, P precision, and R recall. The coefficient ranges from 0 (no overlap) to 1 (perfect overlap) and quantifies how closely predicted segmentation masks align with the ground-truth annotations.

This metric was chosen because it provides a direct measure of spatial consistency, making it particularly suitable for detecting structural distortions or misalignments that generative models may introduce. By comparing selector model predictions against the Unity-generated annotations, Dice scoring allowed us to identify images in which realism enhancements had degraded object boundaries, shifted contours, or otherwise compromised the usability of annotations for training.

2.3.3. *Dice-based model-guided filtering*

The filtering process is summarized in the pseudo-code and illustrated in the accompanying flowchart (Table 1), which together outline the sequential evaluation of all 4000 enhanced images (1000 enhanced images derived from their synthetic counterparts, used for the selector model training, were kept outside the filtering process). Each image was first processed by the selector model, which generated segmentation predictions that were compared against the original Unity ground-truth annotations. To address multi-instance cases, where multiple objects of the same class were present in the same frame, an optimal matching algorithm was applied to maximize the probability that each predicted mask was paired with its correct ground-truth counterpart. This ensured that Dice scores were computed object by object in the right way,

Table 1
Model-guided data filtering pipeline.

Algorithm 1	Flowchart
Input: Selector Model (SM), 4000 Enhanced Images (E), Classes threshold (th), Synthetic Ground Truth Masks (SGTM) **Output**: Curated (D_{cur}) and Hybrid Datasets (D_{hyb}) **for** each image *I* in *E*: *Predicted Mask (PM)* ← *SM.predict(I)* $SGTM_I$ ← Synthetic Annotations of I **if** present Multi Instances of the same class Matched Pairs *(P)* ← *OptimalMatching(PM, SGTM$_I$)* **for** each class *c* in *P*: $Dice_c$ ← *Dice(P_c, $SGTM_{I,c}$)* **if** $Dice_c \geq th_c$ for all classes *c* in *I*: $D_{cur} \leftarrow D_{cur} \cup \{I\}$ $D_{hyb} \leftarrow D_{cur} \cup$ original synthetic dataset Return D_{hyb} , D_{cur}	

avoiding mismatches that could otherwise underestimate the true segmentation quality. Once scores were assigned, images were retained only if every predicted Dice-class present in the frame exceeded its class-specific threshold, with values empirically determined through qualitative and quantitative inspection. This strict criterion acted as a gatekeeping mechanism: only structurally reliable, high-quality images were accepted, while those containing distortions, boundary shifts, or annotation misalignments introduced during generative enhancement were discarded.

2.3.4. Datasets construction

For this study, the focus was placed on the part-based configuration, as it provides the finer level of detail required for advancing surgical autonomy. The type-based setup was included only as a benchmark: rather than exploring multiple dataset variants, we trained it exclusively on the best-performing dataset configuration identified in the part-based experiments. This allowed us to directly compare the new model's performance with our earlier work, which was conducted solely in the type-based format.

Following the filtering stage, multiple dataset configurations were constructed to train different models and systematically evaluate the effect of different data sources on them. The first configuration consisted of the original 1000 purely synthetic images, serving as a stable baseline for all comparisons. To test whether the sheer volume of generative outputs could improve training, a second dataset combined the 1000 synthetic images with the entire pool of 4000 enhanced samples. The core of our approach was the curated configuration, where only the enhanced images that passed Dice-based filtering thresholds were retained. This produced two additional datasets: one composed exclusively of the filtered subset and another hybrid dataset combining these curated enhanced images with the 1000 synthetic baseline. These two sets were considered the most representative of our data-centric philosophy, balancing realism with structural fidelity.

To further investigate transfer strategies, the baseline YOLO-filter model trained on 1000 synthetic images was fine-tuned on the curated enhanced dataset in two modes: (1) offline fine-tuning on the entire curated set at once and (2) online iterative fine-tuning, where the model was progressively updated as new batches of 200 filtered images were accepted. All configurations are summarized in Table 2, which recaps the datasets used for training and the corresponding YOLO model variants.

2.4. Pipeline summary

The proposed end-to-end pipeline is structured into four functional stages, as illustrated in Fig. 4:

1. Scene preparation (Blue): the process begins with the 3D design of surgical instruments and their integration into a Unity-based simulated environment. This virtual scene replicates porcine suturing procedures, incorporating extensive domain randomization, including variable lighting, camera angles, and textures, to maximize data diversity.
2. Synthetic generation (Green): automated scripts in Unity generate high-fidelity synthetic images and their corresponding pixel-perfect ground truth masks (type-based and part-based), providing a scalable foundation for training without manual labeling.
3. Realism enhancement (Orange): a generative enhancement step utilizes real and synthetic image pairs to fine-tune the CycleGAN-TURBO model for unpaired image-to-image translation. This stage outputs realism-enhanced images, which are then processed via Real-ESRGAN for super-resolution restoration, while strictly maintaining the original synthetic annotations.
4. Self-curation and training (Pink): the final stage implements the YOLO-based self-curation mechanism (detailed in Table 1). This quality gate filters out generative outputs that exhibit geometric distortions or semantic misalignments. Only the resulting hybrid dataset, consisting of curated enhanced images and raw synthetic data, is used to train the final YOLO model for real-time inference on surgical footage.

This closed-loop design transforms a large set of raw generative samples into a compact, high-quality dataset, ensuring that downstream training relies on quality-controlled data that preserves beneficial variability while mitigating generative noise.

Table 2
Training-data part based configuration (S: synthetic, T-E: Turbo Enhanced, T: Training, FT: Fine-Tuning).

	# S	# T-E Filtered	# T-E	Mode
Model 1	1000	0	0	T
Model 2	1000	0	4000	T
Model 3	0	538	0	T
Model 4	1000	538	0	T
Model 5	1000	538	0	Online FT
Model 6	1000	538	0	Offline FT

2.5. Framework and tools

All experiments were implemented in PyTorch (version 2.4.0) with GPU acceleration provided by CUDA (version 12.4).

The CycleGAN-TURBO and Real-ESRGAN were trained on a workstation equipped with an RTX A6000 GPU. For segmentation, we employed the YOLOv11-m architecture (version 11.0.0), selected for its balance between inference speed and accuracy, properties that are essential for real-time deployment in surgical contexts (Hai-Binh et al., 2023; Pan et al., 2024; Leoncini et al., 2025).

While the pipeline integrates multiple heavyweight architectures, a key distinction must be made regarding computational complexity: the generative enhancement and model-guided filtering stages operate strictly offline. This one-time computational cost, focused on autonomous curation, replaces the manual effort of human annotation and data cleaning. Consequently, there is no complexity tradeoff during deployment; the system yields a highly optimized, curated training set that results in a lightweight YOLOv11 model capable of real-time inference on standard hardware.

Training and evaluation were carried out on an NVIDIA GeForce RTX 3070 Ti. The Unity environments, 3D models, and annotated datasets used in this work are openly accessible through our GitHub[1].

3. Results

3.1. Dataset filtering and acceptance rate

The CycleGAN-TURBO and Real-ESRGAN pipeline produced a total of 4000 realism-enhanced surgical images at 1920 × 1080 resolution. However, as shown in Fig. 5, not all enhanced images contributed equally to model performance. Using the YOLO-based selector model, guided by Dice similarity scores, 538 images met the predefined quality thresholds across all classes and were retained in the part-based configuration, corresponding to an acceptance rate of roughly 13.5%. In the type-based configuration, where the task is simpler and less sensitive to fine boundary errors, 641 images were accepted, corresponding to an acceptance rate of approximately 16%. These results (Table 3) emphasize both the variability of generative outputs and the critical role of automated filtering in establishing a higher-quality dataset.

3.2. Models comparison on real test data

A series of models trained under different data configurations were evaluated in the part-based setup, spanning purely synthetic, purely enhanced, and curated datasets. The baseline YOLO-filter trained on

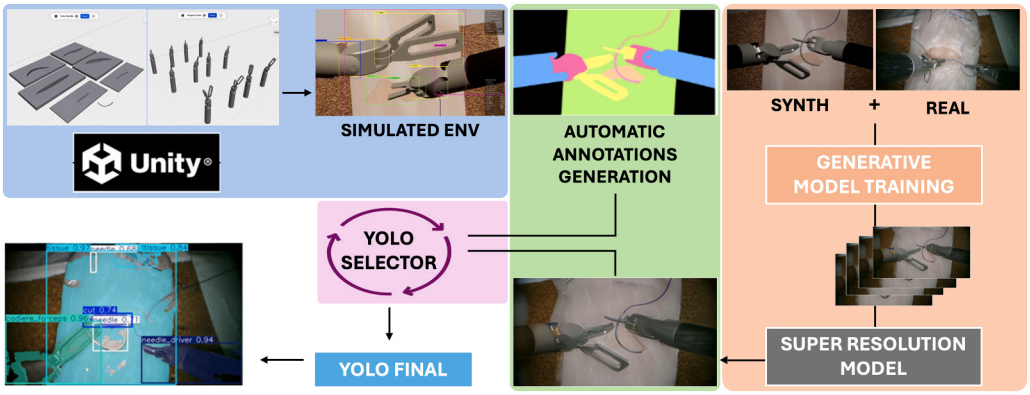


Fig. 4. End-to-end pipeline from 3D model design and synthetic data generation in Unity to realism enhancement with CycleGAN-TURBO, super-resolution restoration to 1920 × 1080, and filtering with a YOLO-based selector. The curated hybrid dataset (synthetic + filtered enhanced images) was then used to train YOLO-final, shown here with an example of type-based predictions on real data.

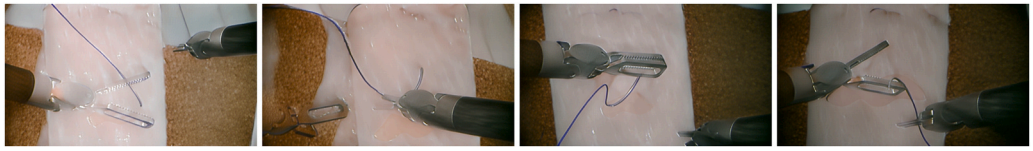


Fig. 5. (a, b) Example of TURBO-enhanced images rejected by the filtering pipeline. (c, d) Example of TURBO-enhanced images accepted by the filtering pipeline and added to the curated dataset.

Table 3
Results of the filtering pipeline.

	# Tot Images	# Accepted Images	% Accepted Images	Filtering Time
Type-Based	4000	641	16%	10 min
Part-Based	4000	538	13.5%	14 min

1000 synthetic images (M1) achieved a Dice score of 0.17, while combining synthetic with all enhanced images (M2) plateaued at 0.24. Filtering provided the most consistent benefits: training on the 538 filtered images (M3) yielded a Dice score of 0.36, and combining the filtered set with 1000 synthetic images (M4) reached 0.44. Fine-tuning the baseline YOLO-filter model on the curated enhanced subsets (M5) yielded only modest improvements (0.25), and iterative fine-tuning with incremental batches of 200 accepted images (M6) did not further enhance performance, plateauing after the first update. These findings (Table 4) suggest that the improvements were primarily driven by the quality of the curated data, while variations in training strategy had only a minor impact.

For the type-based setup, we restricted the analysis to the best-performing configuration identified in the part-based experiments, namely the hybrid dataset combining 1000 synthetic images with the curated enhanced subset. This choice was made to ensure comparability with our previous type-based framework while avoiding redundant configurations. In this case, the curated type-based model, trained on 1000 synthetic images and 641 filtered enhanced images, achieved a mean Dice score of 0.65 on the real test set.

Table 4
Mean Dice on the real test set of different models (Mi) in Part-based configuration.

Part-based	M1	M2	M3	M4	M5	M6
Dices	0.17	0.24	0.36	0.44	0.25	0.27

To further validate the pipeline and demonstrate its architectural independence, we extended the part-based evaluation on the real test set to SegFormer and U-Net. For both models, training solely on synthetic images established a baseline (SegFormer Dice: 0.47, U-Net Dice: 0.39). Combining synthetic images with the full, unfiltered realism-enhanced set showed modest improvements (SegFormer Dice: 0.51, U-Net: 0.46). Conversely, training on our curated hybrid set led to the highest results for both models (SegFormer Dice: 0.67, U-Net Dice: 0.54), confirming that generative noise and hallucinations universally degrade performance regardless of architecture. Full results for this analysis are available in Supplementary Table 1.

3.3. *Threshold sensitivity and quality-quantity trade off*

To further investigate the balance between dataset size and quality, we systematically lowered the Dice thresholds used in the filtering process. In the part-based configuration (Table 6), reducing the thresholds (th2) nearly doubled the acceptance rate, from 13,5% (538 images) to close to 25% (around 993 images). Despite this increase in dataset size, performance declined sharply: the best-performing model (M4), trained with 1000 synthetic images combined with the curated enhanced subset, dropped from a Dice score of 0.44–0.35. The additional data introduced noise in the form of visually distorted or misaligned samples, reducing the model's ability to generalize to real test images.

A similar trend was observed in the type-based configuration (Table 5). Here, the initial acceptance rate was higher, with 1240 images retained under the stricter threshold. Loosening the criteria raised the acceptance rate above 31%, but performance again declined: the Dice score of the hybrid model (M4 data configuration) decreased from 0.65

Table 5
Type-based threshold sensitivity and M4 mean Dice.

TB	#Acc images	%Acc images	Dice
Th1	641	16%	0.65
Th2	1240	31%	0.53

Table 6
Part-based threshold sensitivity and M4 mean dice.

PB	#Acc images	%Acc images	Dice
Th1	538	13.5%	0.44
Th2	993	25%	0.35

to 0.53. Although the absolute values were higher than in the part-based case, the relative drop in performance confirmed that enlarging the dataset without controlling for quality consistently undermines generalization.

Together, these findings demonstrate that stricter filtering yields smaller but more reliable training sets, and that model improvements are driven by the quality of generative data rather than its quantity.

4. Discussion

The central finding of this work is that data quality outweighs data volume in surgical AI. Even when using modest subsets, carefully curated enhanced images consistently outperformed larger, unfiltered collections, underscoring that poorly controlled generative outputs can actively harm sim-to-real transfer. By embedding a YOLO-based selector model within the pipeline, we introduced a model-guided gatekeeper that filters out low-quality generative samples, ensuring that only reliable images contribute to training. To the best of our knowledge, this use of model inference results to automatically filter out geometrically distorted generative noise is a novel approach in surgical imaging. This self-curation strategy illustrates a practical and effective way of operationalizing the data-centric AI paradigm: rather than focusing on increasingly complex architectures, meaningful performance gains arise from systematically improving the training data itself. Anyway, to rigorously validate the effectiveness of this novel curation strategy and demonstrate that it is architecture-agnostic, we extended the study to include SegFormer and U-Net. Both models mirrored the performance trends observed with YOLO, peaking significantly when trained on the curated hybrid set and confirming that data integrity is more critical than the specific model choice. While SegFormer and U-Net achieved higher absolute Dice scores (0.67 and 0.54 respectively), we focused on YOLOv11-seg, which provides the optimal trade-off between the precision and the low-latency inference required for safe intraoperative robotic assistance. Furthermore, we compared our approach against a filtering strategy based on the MANIQA (Yang et al., 2022) metric, which prioritizes perceptual image realism. Our results demonstrated that images appearing realistic to a quality metric are insufficient for training if they lack post-generative geometric accuracy, as our inference-based filtering consistently led to superior downstream models' performance, as shown in the supplementary Table 2.

A key advance of this work lies in combining realism-enhancing generative models with reliability-focused curation, representing a significant technological leap over our previous work (Leoncini et al., 2025). This pipeline transitions from simple data generation to an autonomous, self-correcting system by integrating high-fidelity enhancement with a novel curation layer. While CycleGAN-TURBO and Real-ESRGAN enriched synthetic images with realistic textures and high-resolution details, it was the subsequent filtering step that determined whether these enhancements translated into performance gains. Our results show that naive use of unfiltered generative outputs can degrade model performance, a critical risk that should be (has not been sufficiently) emphasized in (prior) sim-to-real approaches. In contrast, the curated subsets consistently outperformed both purely synthetic and unfiltered enhanced datasets, highlighting the value of combining generation and curation into a single, self-correcting pipeline.

The introduction of part-based annotations represents another important contribution, offering a granularity leap over type-based labelling. By decomposing instruments into functionally distinct components, such as Tool-Shaft, Tool-Wrist, and Tool-Clasper, the vision system acquires a richer representation of surgical tools, which is essential for enabling fine-grained tasks such as needle handling and suture placement. This refinement marks a step forward from our previous work, where only type-based segmentation was considered.

Type-based models were also evaluated under the best-performing configuration found in the part-based experiments. The curated type-based model achieved a Dice score of 0.65, a substantial improvement compared to the 0.384 Dice reported in our previous work for models trained solely on synthetic images. While this still falls short of the 0.92 Dice obtained in the earlier hybrid setting, where synthetic data was combined with real annotated images, a costly and resource-intensive strategy, the present results are notable because no real images were used for training. Moreover, this new type-based dataset included the surgical thread, one of the most difficult classes to segment, alongside the needle, further lowering the mean Dice compared to the earlier study. Encouragingly, when analyzed at the class level, the TURBO-enhanced pipeline achieved Dice scores of 0.71 and 0.77 for the two instruments (Cadiere Forceps and Needle Driver), demonstrating that generative enhancement combined with filtering can deliver strong results for critical objects without relying on real annotations.

Another central outcome is the value of hybrid datasets. Models trained on a combination of synthetic and curated enhanced images consistently outperformed those trained exclusively on either modality. This suggests that hybrid approaches balance the structural fidelity and diversity of synthetic data with the realism introduced by generative enhancement, producing models with stronger generalization capabilities. Importantly, this was achieved without incorporating real images into training; real data was used solely for evaluation, demonstrating that our pipeline can deliver measurable improvements in real-world applicability while remaining fully independent of annotated surgical footage (real data annotated only for evaluation).

While the results are promising, several limitations deserve attention. Although the Unity-based simulation was quite variable, it was restricted to a single scenario, which constrained the variability of tool–tissue interactions, lighting conditions, and surgical contexts available for training. This constrained diversity in the input domain likely influenced the behavior of CycleGAN-TURBO, which, although content-preserving, tended to produce outputs that were realistic but not highly diverse. In practice, TURBO was translating a narrow distribution of inputs, which may explain why its outputs often lacked variability and why the overall performance gains, though consistent, plateaued at a modest level. This highlights the need for broader simulation diversity in tandem with generative models capable of leveraging that variability to achieve stronger generalization. In addition, thresholds for Dice-based filtering were static and set empirically, and although already class-specific, future work could explore adaptive calibration strategies to optimize selection dynamically. Finally, relying solely on Dice as the quality metric also presents limitations; future work should include complementary measures such as perceptual similarity or uncertainty-based indicators to enrich the filtering process.

5. Conclusion

This study underscores that progress in surgical computer vision depends less on designing increasingly complex architectures and more on systematically improving the quality of training data. By integrating generative realism boosting with automated, model-guided filtering, we demonstrated how synthetic data can be transformed into a reliable resource for sim-to-real transfer. The pipeline also extends the field by introducing part-based annotations, which elevate tool recognition from coarse type-level identification to a more precise understanding of functional components, essential for tasks such as Tool-Wrist and Tool-Clasper. Beyond methodological refinements our framework enables scalable, low-cost training of computer vision models without reliance on real data annotations, supporting both cost savings and progress

toward autonomy. These resources not only help overcome the scarcity of task-specific datasets but also provide researchers with reproducible and accessible tools to further explore this field. By consolidating Unity environments, enhanced datasets, and filtering tools in our open-source repository, we contribute a foundation upon which others can build, adapt, and extend. Looking forward, we see this approach as a step toward self-supervised model improvement and as a pathway to developing vision systems capable of supporting reliable, fine-grained autonomy in surgical robotics.

CRediT authorship contribution statement

Giulio Dagnino: Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Conceptualization. **Lorenzo Revello:** Formal analysis, Methodology, Writing – review & editing. **Kengo Hayashi:** Writing – review & editing, Writing – original draft, Validation, Supervision. **Giovanni Distefano:** Writing – review & editing, Writing – original draft, Validation, Supervision. **Alberto Arezzo:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Carlo Alberto Ammirati:** Writing – review & editing, Writing – original draft, Validation, Supervision. **Matteo Pescio:** Writing – review & editing, Writing – original draft, Validation, Data curation. **Francesco Marzola:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Federica Barontini:** Writing – review & editing, Writing – original draft, Validation, Supervision, Conceptualization. **Luigi Muratore:** Writing – review & editing, Visualization, Data curation. **Pietro Leoncini:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to refine the English phrasing, improve the overall textual clarity, and enhance the readability and flow of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding statement

This work was supported by the European Research Council (ERC) under the Horizon Europe programme (Grant EndoTheranostics, Grant Agreement No. 101118626, DOI: 10.3030/101118626), by the CLASSICA project (Grant Agreement No. 101057321, DOI: 10.3030/101057321) and the PALPABLE project (Grant Agreement No. 101092518, DOI: 10.3030/101092518). The work has also been partially supported by the Italian Ministry of University and Research (MUR) under the PRIN 2022 programme “Towards Intelligent ROBOTIC ENDOSCOPIC DISSECTION (TI-RED)” Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union, the European Research Council Executive Agency, or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.compmedimag.2026.102775.

Data availability

Data and code are open access. The GitHub link to the research data and code has been included in the manuscript.

References

Ahmed, F.A., Yousef, M., Ahmed, M.A., Ali, H.O., Mahboob, A., Ali, H., Shah, Z., Aboumarzouk, O., Al Ansari, A., Balakrishnan, S., 2024. Deep learning for surgical instrument recognition and segmentation in robotic-assisted surgeries: A systematic review. Artif. Intell. Rev. 58 (1), 1. https://doi.org/10.1007/s10462-024-10979-w.

Attanasio, A., Scaglioni, B., De Momi, E., Fiorini, P., Valdastri, P., 2021. Autonomy in surgical robotics. Annu. Rev. Control Robot. Auton. Syst. 4 (1), 651–679. https://doi.org/10.1146/annurev-control-062420-090543.

Bhatt, N., Bhatt, N., Prajapati, P., Sorathiya, V., Alshathri, S., El-Shafai, W., 2024. A Data-Centric Approach to improve performance of deep learning models. Sci. Rep. 14. https://doi.org/10.1038/s41598-024-73643-x.

Caballero, D., Sánchez-Margallo, J.A., Pérez-Salazar, M.J., Sánchez-Margallo, F.M., 2025. Applications of artificial intelligence in minimally invasive surgery training: A scoping review (Article). Surgeries 6 (1), 7. https://doi.org/10.3390/surgeries6010007.

Cao, Y., Zhang, J., Li, H., Ren, B., Zhao, Y., Gao, C., Zhou, Y., 2025. A review of critical deep learning-based image guidance technologies for surgical robots. IEEE/ASME Trans. Mechatron. 1–22. https://doi.org/10.1109/TMECH.2025.3584339.

Chen, F., Zhou, W., 2025. Quality over quantity: An effective large-scale data reduction strategy based on pointwise v-information (Article). Electronics 14 (15), 3092. https://doi.org/10.3390/electronics14153092.

Colleoni, Emanuele, Edwards, Philip, Stoyanov, Danail, 2020. Synthetic and real inputs for tool segmentation in robotic surgery. Int. Conf. Med. Image Comput. Comput. -Assist. Interv. 700–710. https://doi.org/10.1007/978-3-030-59716-0_67.

Côté, P.-O., Nikanjam, A., Ahmed, N., Humeniuk, D., Khomh, F., 2024. Data cleaning and machine learning: A systematic literature review. Autom. Softw. Eng. 31 (2), 54. https://doi.org/10.1007/s10515-024-00453-w.

Dagnino, G., Kundrat, D., 2024. Robot-assistive minimally invasive surgery: Trends and future directions. Int. J. Intell. Robot. Appl. 8 (4), 812–826. https://doi.org/10.1007/s41315-024-00341-2.

Dwyer, B., Nelson, J., & Hansen, T. (2024). *Roboflow (Version 1.0) [Software], computer vision*. ⟨https://roboflow.com⟩.

Glenn J. & Jing Q. (2024). *Ultralytics YOLO11, version 11.0.0*. ⟨https://github.com/ultralytics/ultralytics⟩.

Hai-Binh, Le, Dinh Kim, Thai, Ha, Manh-Hung, Tran, Anh Long Quang, Nguyen, Duy-Thuc, Xuan-Minh, Dinh, 2023. Robust Surgical Tool Detection in Laparoscopic Surgery using YOLOv8 Model. 2023 International Conference System Science Engineering (ICSSE) 537–542. https://doi.org/10.1109/ICSSE58758.2023.10227217.

Hinterstoisser, S., Pauly, O., Heibel, H., Marek, M., Bokeloh, M., 2019. *An Annotation Saved is an Annotation Earned: Using Fully Synthetic Training for Object Instance Detection* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1902.09967.

Huang, P., Zhang, K., Ma, M., Mei, S., Wang, J., 2025. Semantic-Geometric Consistency-Enforcing With Mamba-Augmented Network for Remote Sensing Image Segmentation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 1–14. https://doi.org/10.1109/JSTARS.2025.3624209.

Joshi, S., Jain, A., Payani, A., Mirzasoleiman, B., 2024. Data-efficient contrastive language-image pretraining: Prioritizing data quality over quantity. Int. Conf. Artif. Intell. Stat. ⟨https://api.semanticscholar.org/CorpusID:268532161⟩.

Leoncini, P., Marzola, F., Pescio, M., Casadio, M., Arezzo, A., Dagnino, G., 2025. A reproducible framework for synthetic data generation and instance segmentation in robotic suturing. Int. J. Comput. Assist. Radiol. Surg. 20 (8), 1567–1576. https://doi.org/10.1007/s11548-025-03460-8.

Majeed, A., Hwang, S.O., 2025. Synthetic data: A new frontier for democratizing artificial intelligence and data access. Computer 58 (2), 106–114. https://doi.org/10.1109/MC.2024.3515412.

Man, K., Chahl, J., 2022. A review of synthetic image data and its use in computer vision. J. Imaging 8 (11), 310.

Ostrander, B.T., Massillon, D., Meller, L., Chiu, Z.-Y., Yip, M., Orosco, R.K., 2024. The current state of autonomous suturing: A systematic review. Surg. Endosc. 38 (5), 2383–2397. https://doi.org/10.1007/s00464-024-10788-w.

Pan et al. (2024). *DBH-YOLO: a surgical instrument detection method based on feature separation in laparoscopic surgery*.

Parmar, G., Park, T., Narasimhan, S., Zhu, J.-Y., 2024. One-step Image Transl. Text. -to-Image Models. *ArXiv*, *abs/2403.12036*. ⟨https://api.semanticscholar.org/CorpusID:268532317⟩.

Perera-Lago, J., Toscano-Duran, V., Paluzo-Hidalgo, E., Gonzalez-Diaz, R., Gutiérrez-Naranjo, M., Rucco, M., 2024. An in-depth analysis of data reduction methods for sustainable deep learning [version 2; peer review: 2 approved]. Open Res. Eur. 4 (101). https://doi.org/10.12688/openreseurope.17554.2.

Rivoir, D., Wagner, M., Bodenstedt, S., März, K., Kolbinger, F., Maier-Hein, L., Seidlitz, S., Brandenburg, J., Müller-Stich, B.P., Distler, M., Weitz, J., Speidel, S., 2024. Importance of the data in the surgical environment. In: Karcz, In.K., Nawrat, Z., Gumbs, A.A. (Eds.), Artificial intelligence and the perspective of autonomous surgery. Springer Nature Switzerland, pp. 29–43. https://doi.org/10.1007/978-3-031-68574-3_2.

Rodrigues, Mark, Mayo, Michael, Patros, Panos, 2022. Surgical tool datasets for machine learning research: A survey. Int. J. Comput. Vis. 130 (9), 2222–2248. https://doi.org/10.1007/s11263-022-01640-6.

Troccaz, J., Dagnino, G., Yang, G.-Z., 2019. Frontiers of medical robotics: From concept to systems to clinical translation. Annu. Rev. Biomed. Eng. 21 (1), 193–218. https://doi.org/10.1146/annurev-bioeng-060418-052502.

Vo, H.V., Khalidov, V., Darcet, T., Moutakanni, T., Smetanin, N., Szafraniec, M., Touvron, H., Couprie, C., Oquab, M., Joulin, A., J'egou, H., Labatut, P., Bojanowski, P., 2024. Autom. data Curation self-supervised Learn. A Clust. -Based Approach *2024*. ⟨https://api.semanticscholar.org/CorpusID:270045124⟩.

Wang, W., Xia, J., Luo, G., Dong, S., Li, X., Wen, J., Li, S., 2025. Diffusion model for medical image denoising, reconstruction and translation. Comput. Med. Imaging Graph. 124, 102593. https://doi.org/10.1016/j.compmedimag.2025.102593.

Wang, X., Xie, L., Dong, C., Shan, Y., 2021. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. 2021 IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW) 1905–1914. https://doi.org/10.1109/ICCVW54120.2021.00217.

Yang, S., Xie, Z., Peng, H., Xu, M., Sun, M., Li, P., 2022. Dataset pruning Reducing Train. data examining Gen. Influ. *ArXiv*, *abs/2205.09329*. ⟨https://api.semanticscholar.org/CorpusID:248887235⟩.

Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y., 2022. *MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2204.08958.

Yang, Z., Yang, H., Majumder, S., Cardoso, J., Gallego, G., 2024. Data pruning can do more: A comprehensive data pruning approach for object re-identification. Trans. Mach. Learn. Res. ⟨https://openreview.net/forum?id=vxxi7xzzn7⟩.

Zhou, Y., Tu, F., Sha, K., Ding, J., Chen, H., 2024. A survey on data quality dimensions and tools for machine learning invited paper. 2024 IEEE Int. Conf. Artif. Intell. Test. (Aitest) 120–131. https://doi.org/10.1109/AITest62860.2024.00023.