ISSN: 3066-3679

Research Article

Journal of Applied Engineering Education

Computational Imaging from Acquisition to Action: A Review of Efficiency, Evidence, and Reliability

Nguyn Gia Huy[1]* and Valentina Rodríguez[2]

[1]*Saigon University, Vietnam*

[2]*Escuela Colombiana de Ingenieria Julio Garavito, Colombia*

***Corresponding Author**
Nguyn Gia Huy, Saigon University, Vietnam.

Submitted: 2026, Oct 12; **Accepted:** 2026, Nov 16; **Published:** 2026, Nov 25

Citation: Huy, N. G., Rodríguez, V. (2026). Computational Imaging from Acquisition to Action: A Review of Efficiency, Evidence, and Reliability. *J of App Eng Education, 3*(2), 01-13.

Abstract

Computational imaging combines physical measurements with numerical models to produce images and task-specific estimates. Its practical performance depends on how acquisition, reconstruction, inference, and data transfer share limited resources. Improvements within one stage can shift costs elsewhere or change the evidence available for a decision. This narrative review examines those interactions across tomography, interventional guidance, endoscopy, and digital pathology. A measurement-to-decision framework relates task loss to acquisition effort, computational demand, communication load, energy use, and response deadlines. Within this framework, the review considers sparse sampling, accelerated optimization, learned reconstruction, compact neural networks, adaptive compression, and distributed processing. Three distinctions guide the synthesis: acquired information versus inferred content, algorithmic complexity versus measured implementation cost, and processing throughput versus the age of a usable output. The analysis connects these distinctions to evaluation practices, including matched task comparisons, resource–performance curves, explicit timing boundaries, and tests of sensitivity to uncertain or restored regions. It also identifies challenges in coordinating adaptive stages, validating intermediate results, and retaining information for future tasks. The resulting reporting framework supports assessment of complete imaging workflows under stated operating conditions. Reliable efficiency gains require evidence that resource savings preserve the information, timing, and failure behavior needed by the intended application.

Keywords: Computational Imaging, Acquisition Design, Inverse Problems, Efficient Inference, Task-Aware Compression, End-To-End Latency, Uncertainty, Digital Pathology

1. Introduction: Connecting Imaging Efficiency to Usable Evidence

Imaging workflows transform observations into decisions through a sequence of physical and computational operations. Acquisition determines which aspects of an object are observed; reconstruction and inference convert those observations into interpretable outputs; storage and transmission determine where and when the outputs can be used. A useful assessment of efficiency must follow this sequence. Reduced acquisition time has limited value if it creates ambiguity that the required task cannot tolerate. Faster reconstruction matters when its result remains accurate enough for interpretation or guidance. Smaller files are useful when their contents still support the analyses for which they are retained.

Several established research areas explain different parts of this problem. Compressed sensing links recoverability to measurement design and assumptions about signal structure [1,2]. Analysis of neuralnetwork implementation connects computational work with memory access and hardware behavior [3]. Task-oriented communication relates representation to the purpose for which information is transmitted [4]. Visual servoing connects estimation accuracy and timing to the response of a controlled system [5]. Considered together, these perspectives suggest that imaging methods should be compared through their effects on a complete

workflow as well as through their individual objectives.

This review asks how measurement effort, computation, and communication can be coordinated to deliver an output that remains useful at the point of decision. Tomography, interventional imaging, endoscopy, and digital pathology provide contrasting settings in which the required evidence, available resources, and consequences of delay differ. The review uses these examples to examine common design questions while keeping reconstruction metrics, downstream task performance, and clinical outcomes distinct. Findings demonstrated under one set of acquisition or deployment conditions require further evaluation before they support a broader application claim.

The synthesis develops three connected arguments. The first concerns evidence: reducing measurements or generating missing content changes the balance between direct observation and model assumptions. The second concerns implementation: arithmetic, memory traffic, communication, and scheduling jointly determine the resources consumed by an output. The third concerns evaluation: an efficiency claim needs a defined task, a stated operating range, and a test of whether the output arrives with adequate information and reliability. A constrained systems model organizes these arguments, and a reporting framework translates them into comparison criteria. Both are conceptual syntheses of existing research; no new algorithm or experimental benchmark is introduced.

2. Scope, Terminology, and A Measurement-To-Decision Framework

2.1. Scope and Evidence Boundaries

The review is a selective narrative synthesis of inverse problems, efficient inference, communication, and feedback, interpreted through applications in tomography, guidance, endoscopy, and pathology. Foundational methods establish the relevant measurement and computational principles; application studies illustrate how those principles interact with deployment requirements. The purpose is to connect design choices and evaluation criteria across stages. The review does not use a reproducible systematic-search protocol, estimate pooled effects, or claim exhaustive coverage. Quantitative comparisons are restricted to publications whose methods and results could be checked, and examples from abstracts or incompletely documented reports are treated according to their available evidence.

The breadth of the subject also requires care with terminology. Acquisition effort can mean exposure time, sampled locations, views, or frame count. Computation can mean operations, memory traffic, energy, or elapsed processing time. Communication can mean stored bytes, transmitted bytes, or a guaranteed delivery interval. Response time begins at an application-defined physical event and ends when a usable decision or action is available. These quantities interact, and none is a universal substitute for the others.

2.2. A Measurement-To-Decision Model

Let x denote the underlying scene or object, a an acquisition configuration, and y the measured data. A general observation model is

y = A_a(x) + epsilon. (1)

The operator A_a may include projection geometry, sampling, blur, or modality-specific physics; epsilon represents noise and modeling error. An encoder E_b produces a transmitted or stored representation z at coding setting b. A reconstruction or inference module F_theta then forms an image, feature, or decision representation. A task module G converts that representation into an output d:

z = E_b(y), x_hat = F_theta(z; a), d = G(x_hat). (2)

The intermediate x_hat need not be a conventional image. It may be instrument pose, a segmented vessel, or a slide-level embedding. Reconstruction remains essential when the task requires a human-interpretable image or several future analyses. Direct task inference can be appropriate when the output requirements are narrower, provided its validation matches those requirements.

A conceptual design objective is minimize E[L_task(d, d_star)] over a, b, theta, and deployment policy pi, subject to M(a) <= M_max, C(theta, pi) <= C_max, B(b, pi) <= B_max,

E_energy(a, b, theta, pi) <= E_max, Pr[T_end-to-end > D] <= delta. (3)

Here M is measurement effort, C computational resource use, B stored or transmitted information, D a decision deadline, and delta an allowed deadline-miss probability. Hard real-time requirements need a stronger guarantee under explicitly defined workload and hardware assumptions. Equation (3) is an organizing device: the admissible losses, constraints, and uncertainty models must be chosen for the application. It does not imply that every quantity is differentiable or that a single weighted loss resolves all trade-offs.

2.3. Coupling and Bottlenecks

Reducing measured views changes both the inverse problem and the amount of data transferred. Moving inference to the sensor can reduce link traffic while increasing local power consumption. A denoiser can improve reconstruction but also delay every subsequent stage. These dependencies make the four resource categories distinguishable but coupled. Additional costs arise from training, annotation, calibration, maintenance, and validation.

For a sequential workflow, latency can be approximated by a sum of acquisition, readout, transfer, queueing, processing, and output delays. With overlap or streaming, latency follows the critical path instead. Steady-state throughput is yet another quantity, often limited by the slowest stage. A system can deliver many outputs per second while individual outputs remain too old for control. Evaluation must identify which of these properties an optimization changes.

Intervention	Intended benefit	Cost or failure mechanism	Necessary comparison
Fewer acquired views	Shorter acquisition or reduced exposure opportunities	Ambiguity, noise, geometrydependent loss	Task quality versus actual measured views and acquisition settings
Stronger regularization or prior	Stable reconstruction from incomplete data	Bias and unsupported structure	Data consistency, uncertainty, and rarestructure sensitivity
Accelerated iteration	Less time to a usable estimate	Extra work per update or unstable behavior	Wall time and operator calls at matched quality
Learned reconstruction	Reusable inference after training	Training cost and distribution dependence	Deployment runtime, external validation, and failure cases
Adaptive compression	Lower storage or transfer load	Task-dependent information loss	Bytes, codec time, and downstream performance
Edge placement	Smaller remote payload or shorter transfer path	Local memory, energy, and scheduling demands	Complete workflow latency under realistic link conditions

Table 1: Common Interventions and the Evidence Needed to Judge Them

3. Acquisition Design and the Limits of Observable Information

3.1. Sampling Reduction Changes the Evidence

Compressed sensing links sparse structure with recovery from suitably chosen incomplete measurements [1,2]. Sparse MRI translates this principle into an acquisition setting where sampling duration matters directly [6]. The implication is not that any object can be reconstructed from arbitrarily few measurements. Recovery depends on the signal model, sampling operator, noise, and reconstruction procedure. The computational cost relative to a conventional method must also be established experimentally rather than assumed to increase in every implementation.

Sparse-view tomography and limited-angle tomography require separate treatment. Sparse-view acquisition reduces angular sampling density over a range; limited-angle acquisition leaves directions unobserved. Both can be difficult, but their ambiguities differ. Analysis of incomplete tomography and limited-angle artifacts explains why certain structures are poorly determined or unstable [7,8]. In continuous idealized settings, uniqueness and stability also depend on assumptions about the object and data. The practical conclusion is that additional computation cannot replace independent measurement evidence without introducing assumptions.

Total-variation regularization is effective when piecewise-smooth structure is a useful approximation. Its constrained use in cone-beam reconstruction provides a foundation for recovering images from insufficient projections [9]. The associated smoothing can nevertheless remove fine texture or bias boundaries. Earlier reports on few-view reconstruction and total-variation CT document related directions [10,11]. These reports provide historical context rather than a basis for quantitative performance comparisons.

3.2. Structural Priors and Synthesized Projections

Prior-based enhanced ASD-POCS integrates structural information into sparse-view CBCT reconstruction [12]. Deep-prior limited-angle tomography combines learned prior information with a reconstruction procedure designed for deficient angular coverage [13]. These works illustrate a useful strategy: retain the physical forward model while introducing information that stabilizes uncertain components. They do not establish that every recovered detail is supported by the measured projections, or that convergence and clinical reliability follow automatically from prior integration.

Projection interpolation acts at a different stage. An early interpolation report and a later journal study investigate combining acquired and interpolated projections with reconstruction [14,15]. This can improve an estimator by supplying a smooth, structured completion of the projection sequence. However, interpolated views are derived from measured views. They do not constitute additional independent observations and should not be counted as if the scanner had acquired them. Evaluation should report the numbers of real and synthesized projections separately and include interpolation time in the total runtime.

The appropriate question is therefore not simply whether a reconstruction looks better after completion. It is whether the completed data improve the required task across the range of angular gaps, noise levels, motion, and object structures expected in use. Holding out genuinely acquired projections provides one way to test projection prediction. Independent task annotations provide another. Both are more informative than assuming a smoother sinogram is necessarily a more faithful observation.

3.3. Observability Beyond Tomography

The same distinction appears in geometric vision. Two-view reconstruction without known scale yields a scene and motion estimate up to scale under standard assumptions [16]. Trajectory factorization and modern structure-from-motion pipelines improve estimation but do not remove every ambiguity or degenerate configuration [17,18]. Forward camera motion can provide weak triangulation near the viewing axis; the Bappy and Rahman thesis examines structure detection in that setting [19].

Photometric assumptions introduce another limit. SIFT, ORB, evaluations of affine detectors, and SuperPoint represent different approaches to local feature extraction [20-23]. Their usefulness depends on image formation and the structures visible in the scene.

Specular highlights can violate the correspondence assumptions used by tracking, although specularity can itself convey shape information under an appropriate model [24,25]. Work on specular-region detection addresses this reliability problem in feature extraction [26]. Transparent X-ray sequences raise a related issue because superimposed structures can move differently; an early motion-vector report identifies this as a distinct estimation problem [27]. Reducing acquisition before understanding these conditions can amplify an existing modeling error.

4. Reconstruction and Restoration Under Computational Constraints

4.1. Comparing Solver Work Fairly

Algebraic reconstruction, simultaneous iterative updates, SART, and analytic cone-beam reconstruction embody different compromises between geometry handling, update structure, and computation [28-31]. Their relative efficiency cannot be inferred from iteration counts alone. An iteration may process a row, a subset, or the complete measurement set; regularization and memory access may dominate outside the projector. A meaningful comparison counts forward and adjoint applications, processed rays or subsets, auxiliary operations, peak memory, and measured elapsed time at matched output quality.

Software and hardware determine how these operations become runtime. ASTRA, TIGRE, TomoPy, and RTK provide reconstruction infrastructure with different supported geometries and implementation choices [32-35]. GPU reconstruction and FPGA acceleration demonstrate the value of parallel execution and specialized data paths [36,37]. Results obtained on different resolutions, precision settings, or devices should not be combined into a single speed ranking. Implementation details are part of the evidence for an efficiency claim.

4.2. Acceleration Changes More than the Iteration Count

Ordered subsets can produce substantial early progress by updating from portions of the data [38]. Their speedup need not scale linearly with subset count, and convergence behavior depends on the algorithm and relaxation strategy. FISTA provides an accelerated objective convergence rate under its convex assumptions this guarantee does not establish an equally improved rate for task accuracy or for arbitrary learned updates [39]. Combining subsets and momentum applies these ideas to CT while making the update schedule part of the method [40].

Preconditioning aims to improve conditioning, but its application can add computation and storage [41]. Randomized Kaczmarz theory and stochastic primal-dual analysis show how sampling and assumptions can support convergence guarantees [42,43]. Continuation solves a sequence of related problems and can improve practical progress when intermediate solutions provide effective starting points [44]. The relevant comparison remains time to an acceptable estimate, including the cost of constructing or applying the acceleration mechanism.

Modified simultaneous iterative reconstruction provides a specific example of changing the update rule to improve the speed-quality compromise [45]. An early CGTV conference report concerns the broader effort to integrate reconstruction with regularization [46]. These examples motivate an implementation-aware comparison, rather than a universal claim that any named method reduces runtime under every setting.

4.3. Restoration as Part of the Reconstruction Pipeline

Regularization balances stability with fidelity to observations. Tikhonov regularization and L-curve analysis formalize aspects of that balance, while total variation offers a prominent edge-preserving model [47-49]. Half-quadratic methods, split Bregman, and primal-dual optimization provide different ways to handle the resulting objectives [50-52]. Their costs include auxiliary variables, inner solves, and parameter selection. Faster reduction of the optimization objective is useful only when the resulting estimate meets the imaging task.

Variable penalty strategies are represented by Bappy and Jeon's early restoration report [53]. Their hybrid median and total-variation journal method combines corruption handling with variational restoration earlier denoising and L1-minimization reports are also retained [54-56]. Adaptive median filtering and detector-based treatment of impulsive corruption provide context for this combination [57,58]. The evidence supports selecting a noise model and solver jointly. It does not support assuming that a low-cost filter is sufficient for every corruption distribution.

Non-local means, collaborative 3-D filtering, and dictionary learning offer richer image models with different computational requirements [59-61]. CNN denoising shifts much of model construction into training [62]. Plug-and-play methods make the distinction between the forward model and a denoising prior particularly explicit [63]. Their flexibility is valuable, but arbitrary denoiser substitution does not automatically supply a proximal interpretation, convergence, or measurement fidelity. Restoration should be assessed both independently and through its effect on subsequent reconstruction or decision stages.

4.4. Moving Optimization into Training

Learned sparse-coding approximations and algorithm unrolling replace some repeated optimization with learned finite computation [64,65]. Learned primal-dual reconstruction and CNN treatment of inverse problems illustrate model-based and image-domain routes to fast inference [66,67]. Learning an optimizer generalizes the idea of sharing effort across related problems [68]. Deep image prior occupies a different position, using network structure while optimizing parameters for the particular observation [69]. These approaches exchange training requirements, per-instance optimization, and prior assumptions in different ways.

A fixed number of network layers or unrolled stages defines a bounded computational graph, but it does not itself guarantee a wall-clock deadline. Data transfer, accelerator contention, synchronization, and thermal behavior remain relevant. Conversely, an iterative solver can be stopped after a fixed number of iterations

when a predictable work limit is required. The practical choice concerns the quality and reliability of the estimate available at that limit, including whether a partially executed method produces a meaningful output.

5. Efficient Inference Across Hardware and Data Domains

5.1. Arithmetic, Memory, and Measured Execution

MobileNets, MobileNetV2, and ShuffleNet reduce neural computation through architectural changes [70-72]. SqueezeNet emphasizes a compact parameter representation [73]. EfficientNet coordinates depth, width, and input resolution [74]. These developments offer useful design options for imaging, but small lesions, fine boundaries, and large fields of view make resolution a task variable as well as a cost variable. A model that is efficient for coarse classification may be inappropriate for a small-target localization task.

Hardware-aware search makes deployment measurements part of architecture selection. MnasNet, ProxylessNAS, and once-for-all networks illustrate related approaches to choosing or extracting models for particular constraints [75-77]. The general lesson is to evaluate the operators and tensor shapes that the target platform actually executes. An architecture with fewer multiply-accumulate operations can still have longer latency because of memory access, kernel overhead, or limited operator support [3].

5.2. Compression of the Model and Adaptive Execution

Pruning changes parameter or channel structure. Unstructured sparsity can reduce storage, while realized speed depends on sparse execution support; structured filter pruning more directly changes dense tensor dimensions [78,79]. The lottery-ticket hypothesis concerns trainable sparse subnetworks and should not be interpreted as a deployment speed guarantee [80]. Energy-aware pruning further shows why minimizing parameter count is different from minimizing energy [81].

Integer quantization changes numerical representation and can reduce model and activation traffic [82]. Replacing a 32-bit value with an 8-bit value yields a fourfold reduction for that tensor representation; complete model size and latency depend on metadata, mixed precision, kernels, and hardware. Binary networks explore a more aggressive approximation [83]. For quantitative imaging, the effects on lowcontrast signals, calibration, and rare structures require direct testing rather than inference from classification benchmarks.

Knowledge distillation aims to improve a smaller or otherwise constrained model using a teacher [84]. Self-distillation can support learning across network components [85]. Early-exit architectures make inference depth conditional on a stopping policy [86]. Such a policy changes the distribution of work; a lower mean does not guarantee a lower maximum. Efficient attention addresses another scaling constraint in high-resolution representations [87]. Each mechanism should be evaluated on a curve linking task loss with the resource it is intended to save.

5.3. Annotation, Domain Shift, and Synthetic Content

Learning also depends on what data can be labeled and which distributions those labels represent. Earlier image texture synthesis work provides a mechanism for duplicating content [88], but synthetic duplication alone does not establish representative disease or defect variation. The independence of training and evaluation data must survive any such transformation. Gains from augmentation should be measured on independently acquired cases.

Frozen-section pathology gives a concrete example of distribution dependence. Joint feature alignment and self-knowledge distillation were studied for tumor detection across histopathology domains [89]. This work connects model adaptation with practical data limitations; it is evidence about domain adaptation in the studied setting, not proof that compression or reduced resolution preserves diagnostic information. A published conference abstract on weakly supervised gastric-cancer molecular subtype prediction illustrates another downstream objective for whole-slide representations [90]. Its abstract status limits the methodological and performance conclusions that can be drawn from it.

6. Task-Aware Representation, Communication, and Processing Placement

6.1. The Representation Must Serve its Intended Use

Communication theory establishes the relationship between rate and distortion under specified models and objectives [91,92]. JPEG, JPEG 2000, and HEVC provide widely used mechanisms for representing image and video data [93-95]. Their rate-control decisions involve engineering approximations and quality objectives; they should not be described as all optimizing the same perceptual loss. For scientific or diagnostic imaging, the key issue is whether the representation preserves the information needed for the intended interpretation.

Pixel error and structural similarity can help characterize image changes [96]. However, classifier behavior under image degradation shows why image quality and machine-task performance should be measured separately [97]. A small global distortion can coexist with the loss of a small but consequential region.

Conversely, visible changes in irrelevant background may have little effect on a specific detector. The relationship is empirical and depends on the task, data, and model. Learned image compression provides a way to optimize a rate-distortion objective jointly with an encoder and decoder [98,99]. Video coding for machines, learned image coding for machines, and task-oriented communication broaden the objective to include the consumer of the representation [100,101,4]. These developments support joint design, while also exposing a limitation: a representation tailored to one task may discard evidence needed by another. Archival images require evaluation beyond the model used to train the codec.

6.2. Digital Pathology as A Test of Spatially Adaptive Coding

Whole-slide pathology images have large spatial extent and uneven diagnostic content. AdaSlide combines regional compression decisions with image enhancement, providing an example of adapting storage expenditure to image content [102]. Its reported downstream evaluation makes the study directly relevant to a systems account of imaging efficiency. The wider design implication is that regional bit allocation, enhancement, and task validation belong in the same comparison. Perceptual acceptability alone does not demonstrate preservation of every future diagnostic use.

Artifact restoration introduces a related distinction between image completion and observed evidence. Patch-to-Global, published online in September 2026 in the MICCAI 2026 proceedings, uses diffusionbased inpainting to improve global consistency in whole-slide artifact restoration [103]. Such completion may make downstream analysis more useful, but restored tissue remains inferred content. A trustworthy workflow should preserve the acquired image, record restored regions, and evaluate whether decisions depend on those regions. This requirement parallels the handling of missing-angle structure in tomography.

The pathology examples reveal a broader interaction. A codec can suppress subtle evidence, and a generative enhancer can add plausible structure. Their combination may produce a visually coherent image while changing the basis of a decision. Testing each module separately is insufficient; the composed workflow must be assessed against the original acquisition and task annotations. Reader studies, downstream models, rare-finding analysis, and cross-site data provide complementary evidence.

6.3. Where Processing Should Run

Edge computing and collaborative inference provide mechanisms for placing computation nearer the sensor or dividing a network between local and remote resources [104,105]. The best partition depends on intermediate tensor sizes, local processing time, link conditions, and the deadline. A feature tensor is not necessarily smaller than the input image. Partition selection should therefore use measured payloads and include encoding, decoding, and synchronization.

Teleradiology demonstrates why nominal network capacity is insufficient: transport behavior and roundtrip delay affect effective transfer performance [106]. At the physical layer, multicarrier communication and an OFDM channel study concern transmission behavior [107,108]. These are enabling mechanisms rather than direct demonstrations of improved imaging decisions. At the network level, tail reliability and deterministic delivery become relevant when outputs must arrive within a prescribed interval [109,110].

Microservices can make pipeline components independently deployable and scalable, as illustrated by work on AI-IoT architecture [111]. Service decomposition also introduces serialization, network hops, and scheduling costs. Sensor-network intrusion detection provides a separate perspective on integrity [112]. Authentication and provenance can help establish where data came from, but they do not make a noisy or miscalibrated measurement physically correct. Both computational overhead and the consequences of corrupted inputs belong in system evaluation.

7. Timing and Reliability in Interactive Imaging

7.1. The Age of an Estimate Matters

When an image or pose estimate drives action, its acquisition time is part of its meaning. Visual feedback control and time-delay systems theory explain why delayed information can affect stability and tracking performance [5,113]. The effect depends on the plant, controller, sampling, and compensation. A missed deadline can be unacceptable in a hard real-time loop, whereas a human-facing application may degrade gradually as results become older. These distinctions should be specified rather than collapsed into a universal failure rule.

Classic remote manipulation studies describe behavioral changes under delay, and bilateral control theory provides mechanisms for maintaining stability under defined conditions [114,115]. Transatlantic telesurgery demonstrates a demanding networked application, while simulator studies examine latencyrelated changes in surgical performance [116,117]. The latter do not establish a single acceptable delay for every intervention. A system claim must connect timing measurements to the task and control conditions actually studied.

Separating fast tracking from slower mapping is one architectural response to unequal stage requirements [118]. Correlation-filter tracking supplies another example of emphasizing rapid per-frame estimates [119]. Real-time MRI and learned ultrafast ultrasound reconstruction show how modality and reconstruction design can respond to temporal requirements [120,121]. Their frame or reconstruction rates should still be distinguished from the response time of a complete guidance system.

7.2. Guidance and Operator Feedback

MRI-guided robotic catheter work combines imaging, shape estimation, and control [122]. Biplane X-ray microrobot posture recognition requires position and orientation estimates useful for electromagnetic actuation [123]. Automated vessel reconstruction from biplane angiography addresses a related geometric output for guidance [124]. These applications demonstrate that the appropriate image-processing endpoint can be a calibrated geometric quantity. Evaluation must then include pose error, registration, temporal alignment, and the effect of uncertainty on subsequent action.

Human feedback creates a different loop. A prospective randomized study of automatic polyp detection evaluates a clinically meaningful outcome rather than image quality alone [125]. Automatic endoscopy feedback and specular-region analysis address the timeliness and reliability of information shown during acquisition [126,26]. A quality gate can help identify unusable frames, with no-reference image quality analysis providing relevant tools [127]. However, rejection may also suppress difficult but valuable frames.

Reporting the rejection fraction and the findings in rejected cases is necessary to interpret any gain in average performance.

7.3. Scheduling and Partial Results

Predictability requires examining queueing and resource sharing. S3DNN, CPU/GPU scheduling for multiple networks, and Clockwork represent approaches to scheduling or controlling inference execution [128-130]. Their guarantees or service objectives rely on particular assumptions. They do not certify every imaging network on every accelerator. A static computational graph can still experience variable runtime, while a dynamic method can sometimes be bounded under a restricted workload.

Streaming tomography illustrates the possibility of producing useful intermediate outputs during acquisition [131]. Early-exit inference offers a related mechanism at the network level [86]. Neither approach automatically guarantees that every intermediate output is valid or that task quality improves monotonically with time. A useful partial result needs a defined interpretation, an uncertainty estimate or acceptance rule, and a fallback when it cannot support the required decision. These properties should be evaluated at the stopping times that deployment is likely to impose.

Application	Required output	Consequence of excess delay or information loss	Evaluation emphasis
Sparse-view CT	Reconstructed structure or task measurement	Artifacts, missed structures, delayed interpretation	Real-view count, dose-related settings, task errors, reconstruction time
Instrument guidance	Calibrated pose or geometry	Stale location, tracking error, inappropriate action	Spatial error, temporal alignment, age of estimates, control response
Endoscopy assistance	Timely feedback on the current field	An alert after the relevant view has passed	Detection performance, false alerts, rejected frames, feedback delay
Digital pathology	Preserved morphology and task evidence	Storage burden or loss of small diagnostic regions	Bytes per slide, region-level fidelity, downstream tasks, provenance

Table 2: Application-Dependent Definitions of an Acceptable Output

8. Evaluating complete Workflows and Defining Research Priorities

8.1. From Isolated Benchmarks to Operating Ranges

A method should be evaluated over the resource range in which it is intended to operate. For reconstruction, this includes measurement count, angular coverage, noise, and stopping time. For coding, it includes stored or transmitted bytes, codec overhead, and task loss. For inference, it includes precision, resolution, memory use, and concurrent workload. An operating curve reveals whether a method is useful only near one carefully selected setting or remains reliable as conditions change.

Fairness also requires matched conditions. Competing methods should use comparable input evidence, output dimensions, and task definitions. Hyperparameter selection and training access should be disclosed. Runtime should include preprocessing and postprocessing when these stages are necessary for the reported output. Warm-up and synchronization should be described, and offline throughput should not be substituted for single-case response time. MLPerf's scenario-based inference benchmarking offers a relevant model for separating deployment conditions [132].

Latency reporting should include a distribution and workload description. Median, upper quantiles, deadline-miss rate, and the number of measured cases provide more context than an average alone. The largest observed delay is an empirical observation, not a proof of worst-case execution time. Claims of a hard bound require an analysis tied to scheduling, resources, and the permitted workload. For streamed data, timestamps and the age of the output are needed to connect software timing with the physical event.

Energy should be reported at a stated measurement boundary. Device-only inference energy, full-system energy, training expenditure, and energy per completed decision answer different questions. Memory traffic and energy-aware pruning explain why arithmetic is an incomplete proxy [3,81]. Green AI motivates reporting efficiency as part of the research result [133]. Training costs may be amortized across many deployments, but the number of uses and any retraining requirements should be explicit.

Category	Required information
Task and evidence	Output definition, intended user, task loss, independent test data, uncertainty and failure criteria
Acquisition	Actual measurements, synthetic measurements, geometry, exposure or sampling settings, acquisition duration
Category	Required information
Category	Required information
Computation	Algorithm settings, operator calls, precision, hardware, software, peak memory, training versus inference costs
Representation and transfer	Codec settings, actual bytes, intermediate tensor sizes, transfer conditions, encoding and decoding time
Timing	Timing boundaries, concurrency, warm-up, sample count, quantiles, deadline misses, output age
Robustness	External conditions, small or rare findings, missing data, artifacts, abstentions, fallback behavior
Reproducibility	Source code or configuration, data access conditions, preprocessing, selection procedures, measurement scripts

Table 3: Proposed Minimum Reporting Set for Efficiency Claims

8.2. Coordinated Adaptation with Explicit Failure Conditions

An adaptive pipeline could jointly choose whether to acquire another view, spend another reconstruction step, preserve more bits in a region, or move a task to another device. Equation (3) describes the shared objective, but achieving it requires a state estimate for both the scene and the computing environment. Independent adaptation policies may interact badly: stronger compression can increase inference uncertainty, which can trigger more computation and erase the anticipated latency saving. Research should compare coordinated policies with strong stage-wise baselines under changing conditions.

The most useful target is often a Pareto set of feasible operating points rather than a single weighted optimum. A clinician, operator, or system controller can then choose among settings that meet the current constraints. The choice must include an explicit condition under which no available setting supports a reliable result. Escalation, additional acquisition, or abstention may then be preferable to an apparently complete output based mostly on prior assumptions.

8.3. Evidence-Aware Reconstruction and Restoration

Uncertainty should distinguish measurement support, model mismatch, and distribution shift. In limitedangle CT, the relevant question is which structures remain unstable under the acquisition geometry. In generative pathology restoration, it is which regions contain inferred tissue rather than directly observed morphology. In both settings, high visual fidelity can obscure the distinction. Future systems should retain provenance and test decision sensitivity to uncertain regions rather than treating every output pixel as equally supported.

Anytime processing deserves the same scrutiny. An objective may decrease while a small lesion becomes less visible; a network confidence score may increase while calibration worsens. Partial-result evaluation should therefore measure the task, calibration, and failure rate across actual interruption points. Useful guarantees will likely depend on restricted application conditions and explicit acceptance tests rather than a universal promise of monotonically improving images.

8.4. Reusable Data and Changing Downstream Tasks

Task-aware coding is most straightforward when the task is fixed. Archival imaging and foundation-model development have changing objectives. A representation optimized for current labels may remove evidence needed by future analyses. Multitask evaluation, preserved high-fidelity regions, residual coding, and retention of original data are possible responses, each with its own resource cost. Recent whole-slide compression, artifact restoration, domain adaptation, and weakly supervised prediction provide complementary case studies [89,90,102,103]. They should be compared through shared data and task protocols before their benefits are assumed to compose.

9. Conclusions and Implications for Imaging-System Design

Computational imaging efficiency is a property of the path from measurement to use. Acquisition, optimization, learned inference, representation, and processing placement determine both the cost of an output and the evidence on which it rests. Their interactions explain why an improvement in one module may have a different effect on the complete application. Selecting a method therefore requires a task definition and operating conditions against which its savings and failure modes can be assessed.

Across the reviewed settings, three distinctions remain central. Inferred structure must be distinguished from independently acquired evidence. Computational proxies must be related to measured memory, energy, and elapsed time on the intended implementation. Output rate must be evaluated alongside response latency and the age of information used for action. These

distinctions connect sparse-view reconstruction with generative restoration, compact inference with hardware behavior, and distributed processing with feedback requirements.

Future comparisons should report how task performance changes over realistic acquisition, coding, computation, and timing budgets. They should also establish when uncertainty, missing measurements, or delayed results require additional acquisition, fallback processing, or abstention. The reporting framework assembled here provides a basis for those comparisons. Its practical aim is to support imaging workflows whose resource savings remain compatible with reliable interpretation and timely action.

References

1. Candès, E. J., Romberg, J., & Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory, 52*(2), 489-509.
2. Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory, 52*(4), 1289-1306.
3. Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12), 2295-2329.
4. Gündüz, D., Qin, Z., Aguerri, I. E., Dhillon, H. S., & Yang, Z., et al. (2022). Beyond transmitting bits: Context, semantics, and task-oriented communications. IEEE Journal on Selected Areas in Communications, 41(1), 5-41.
5. Hutchinson, S., Hager, G. D., & Corke, P. I. (1996). A tutorial on visual servo control. IEEE transactions on robotics and automation, 12(5), 651-670.
6. Lustig, M., Donoho, D., & Pauly, J. M. (2007). Sparse MRI: The application of compressed sensing for rapid MR imaging. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 58(6), 1182-1195.
7. Louis, A. K. (1986). Incomplete data problems in x-ray computerized tomography: I. Singular value decomposition of the limited angle transform. Numerische Mathematik, 48(3), 251-262.
8. Frikel, J., & Quinto, E. T. (2013). Characterization and reduction of artifacts in limited angle tomography. Inverse Problems, 29(12), 125007.
9. Sidky, E. Y., & Pan, X. (2008). Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. Physics in Medicine & Biology, 53(17), 4777-4807.
10. Bappy, D. M., & Jeon, I. (2015). Few View X-ray CT Image Reconstruction. In Proceedings of the Korean Society of Mechanical Engineers Spring and Autumn Conference (pp. 3123–3124).
11. Bappy, D. M., & Jeon, I. (n.d.). X-ray CT image reconstruction using total-variation.
12. Bappy, D. M., Kang, D., Lee, J., & Baek, H. (2024, April). Prior-based enhanced asd-pocs for artifact suppression and structural preservation in sparse-view cbct. In Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing (pp. 496-505).
13. Bappy, D. M., Kang, D., Lee, J., Lee, Y., & Baek, H. (2025). Deep prior based limited-angle tomography. In International Conference on Pattern Recognition (pp. 79-95). Springer, Cham.
14. Bappy, D. M., & Jeon, I. (2013). X-ray image interpolation for high quality CT reconstruction.
15. Bappy, D. M., & Jeon, I. (2019). High-quality X-ray computed tomography reconstruction using projected and interpolated images. IET Image Processing, 13(7), 1074-1080.
16. Longuet-Higgins, H. C. (1981). A computer algorithm for reconstructing a scene from two projections. Nature, 293(5828), 133-135.
17. Tomasi, C., & Kanade, T. (1992). Shape and motion from image streams under orthography: a factorization method. International journal of computer vision, 9(2), 137-154.
18. Schonberger, J. L., & Frahm, J. M. (2016). Structure-from-motion revisited. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4104-4113).
19. Bappy, D. M., & RAHMAN, M. H. (2012). A Study in 3D Structure Detection Implementing Forward Camera Motion.
20. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.
21. Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011, November). ORB: An efficient alternative to SIFT or SURF. In 2011 International conference on computer vision (pp. 2564-2571). Ieee.
22. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., & Matas, J., et al. (2005). A Comparison of Affine Region Detectors: Mikolajczyk et al. International journal of computer vision, 65(1), 43-72.
23. DeTone, D., Malisiewicz, T., & Rabinovich, A. (2018, June). Superpoint: Self-supervised interest point detection and description. In 2018 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW) (pp. 337-33712). IEEE.
24. Shafer, S. A. (1985). Using color to separate reflection components. Color Research & Application, 10(4), 210-218.
25. Tan, R. T., & Ikeuchi, K. (2005). Separating reflection components of textured surfaces using a single image. IEEE transactions on pattern analysis and machine intelligence, 27(2), 178-193.
26. Bappy, D. M., Kang, D., Lee, J., Lee, Y., & Koo, M., et al. (2024, December). Specular Region Detection and Covariant Feature Extraction. In International Conference on Pattern Recognition (pp. 170-186). Cham: Springer Nature Switzerland.
27. Bappy, D. M., & Jeon, I. (2015). Motion Vector Estimation for Transparent Sequences. In
28. Proceedings of the Korean Society of Mechanical Engineers Spring and Autumn Conference (pp. 171– 172). Korean Society of Mechanical Engineers.
29. Gordon, R., Bender, R., & Herman, G. T. (1970). Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. *Journal of*

theoretical Biology, 29(3), 471-481.
30. Gilbert, P. (1972). Iterative methods for the three-dimensional reconstruction of an object from projections. Journal of theoretical biology, 36(1), 105-117.
31. Andersen, A. H., & Kak, A. C. (1984). Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm. Ultrasonic imaging, 6(1), 81-94.
32. Feldkamp, L. A., Davis, L. C., & Kress, J. W. (1984). Practical cone-beam algorithm. Journal of the Optical Society of America A, 1(6), 612-619.
33. Van Aarle, W., Palenstijn, W. J., Cant, J., Janssens, E., & Bleichrodt, F., et al. (2016). Fast and flexible X-ray tomography using the ASTRA toolbox. Optics express, 24(22), 25129-25147.
34. Biguri, A., Dosanjh, M., Hancock, S., & Soleimani, M. (2016). TIGRE: a MATLAB-GPU toolbox for CBCT image reconstruction. Biomedical Physics & Engineering Express, 2(5), 055010.
35. Gürsoy, D., De Carlo, F., Xiao, X., & Jacobsen, C. (2014). TomoPy: a framework for the analysis of synchrotron tomographic data. Synchrotron Radiation, 21(5), 1188-1193.
36. Rit, S., Vila Oliva, M., Brousmiche, S., Labarbe, R., & Sarrut, D., et al. (2014, March). The Reconstruction Toolkit (RTK), an open-source cone-beam CT reconstruction toolkit based on the Insight Toolkit (ITK). In Journal of Physics: Conference Series (Vol. 489, No. 1, p. 012079).
37. Xu, F., & Mueller, K. (2007). Real-time 3D computed tomographic reconstruction using commodity graphics hardware. Physics in Medicine & Biology, 52(12), 3405-3419.
38. Zhang, W., Qiao, L., Hsu, W., Cui, Y., & Jiang, M., et al. (2020). Fpga acceleration for 3-d low-dose tomographic reconstruction. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 40(4), 666-679.
39. Hudson, H. M., & Larkin, R. S. (1994). Accelerated image reconstruction using ordered subsets of projection data. IEEE transactions on medical imaging, 13(4), 601-609.
40. Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences, 2(1), 183-202.
41. Kim, D., Ramani, S., & Fessler, J. A. (2014). Combining ordered subsets and momentum for accelerated X-ray CT image reconstruction. IEEE transactions on medical imaging, 34(1), 167-178.
42. Fessler, J. A., & Booth, S. D. (1999). Conjugate-gradient preconditioning methods for shift-variant PET image reconstruction. IEEE transactions on image processing, 8(5), 688-699.
43. Strohmer, T., & Vershynin, R. (2009). A randomized Kaczmarz algorithm with exponential convergence. Journal of Fourier Analysis and Applications, 15(2), 262-278.
44. Chambolle, A., Ehrhardt, M. J., Richtárik, P., & Schonlieb, C. B. (2018). Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. SIAM Journal on Optimization, 28(4), 2783-2808.
45. Hale, E. T., Yin, W., & Zhang, Y. (2008). Fixed-point continuation for \ell_1-minimization: Methodology and convergence. SIAM Journal on Optimization, 19(3), 1107-1130.
46. Bappy, D. M., & Jeon, I. (2017). Modified simultaneous iterative reconstruction technique for fast, high-quality CT reconstruction. IET Image Processing, 11(9), 701-708.
47. Bappy, D. M., & Jeon, I. (2016). A Modified CGTV Algorithm for X-ray CT reconstruction. In Proceedings of the Korean Society of Mechanical Engineers Spring and Autumn Conference (p. 37). Korean Society of Mechanical Engineers.
48. Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. Sov Dok, 4, 1035-1038.
49. Hansen, P. C. (1992). Analysis of discrete ill-posed problems by means of the L-curve. SIAM review, 34(4), 561-580.
50. Rudin, L. I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena, 60(1-4), 259-268.
51. Geman, D., & Yang, C. (1995). Nonlinear image recovery with half-quadratic regularization. IEEE transactions on Image Processing, 4(7), 932-946.
52. Goldstein, T., & Osher, S. (2009). The split Bregman method for L1-regularized problems. SIAM journal on imaging sciences, 2(2), 323-343.
53. Chambolle, A., & Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of mathematical imaging and vision, 40(1), 120-145.
54. Bappy, D. M., & Jeon, I. (2014). Fast Medical Image Restoration Using Variable Half-quadratic Penalty Parameter. In Proceedings of the Korean Society of Mechanical Engineers Spring and Autumn Conference (pp. 2595–2597). Korean Society of Mechanical Engineers.
55. Bappy, D. M., & Jeon, I. (2016). Combination of hybrid median filter and total variation minimisation for medical X-ray image restoration. IET Image Processing, 10(4), 261-271.
56. Bappy, D. M., & Jeon, I. (2013). Total Variation Denoising Using X-Ray Images. In Proceedings of the Korean Society of Mechanical Engineers Spring and Autumn Conference (pp. 3861–3863). Korean Society of Mechanical Engineers.
57. Bappy, D. M., & Jeon, I. (2015). Medical Image Restoration Using L1-Minimization. In Proceedings of the Korean Society of Mechanical Engineers Spring and Autumn Conference (pp. 3–4). Korean Society of Mechanical Engineers.
58. Hwang, H., & Haddad, R. A. (1995). Adaptive median filters: new algorithms and results. IEEE Transactions on image processing, 4(4), 499-502.
59. Chan, R. H., Ho, C. W., & Nikolova, M. (2005). Salt-and-pepper noise removal by median-type noise detectors and detail-preserving regularization. IEEE Transactions on image processing, 14(10), 1479-1485.
60. Buades, A., Coll, B., & Morel, J. M. (2005, June). A non-local algorithm for image denoising. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 2, pp. 60-65). Ieee.
61. Dabov, K., Foi, A., Katkovnik, V., & Egiazarian, K. (2007). Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Transactions on image processing, 16(8),

2080-2095.
62. Aharon, M., Elad, M., & Bruckstein, A. (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on signal processing, 54(11), 4311-4322.
63. Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE transactions on image processing, 26(7), 3142-3155.
64. Venkatakrishnan, S. V., Bouman, C. A., & Wohlberg, B. (2013, December). Plug-and-play priors for model based reconstruction. In 2013 IEEE global conference on signal and information processing (pp. 945-948). IEEE.
65. Gregor, K., & LeCun, Y. (2010, June). Learning fast approximations of sparse coding. In Proceedings of the 27th international conference on international conference on machine learning (pp. 399-406).
66. Monga, V., Li, Y., & Eldar, Y. C. (2021). Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. IEEE Signal Processing Magazine, 38(2), 18-44.
67. Adler, J., & Öktem, O. (2018). Learned primal-dual reconstruction. IEEE transactions on medical imaging, 37(6), 1322-1332.
68. Jin, K. H., McCann, M. T., Froustey, E., & Unser, M. (2017). Deep convolutional neural network for inverse problems in imaging. IEEE transactions on image processing, 26(9), 4509-4522.
69. Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., & Pfau, D., et al. (2016). Learning to learn by gradient descent by gradient descent. Advances in neural information processing systems, 29.
70. Dmitry, U., Vedaldi, A., & Victor, L. (2020). Deep image prior. International Journal of Computer Vision, 128(7), 1867-1888.
71. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., & Wang, W., et al. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
72. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018, June). Mobilenetv2: Inverted residuals and linear bottlenecks. In 2018 IEEE/CVF conference on computer vision and pattern recognition (pp. 4510-4520). Ieee.
73. Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018, June). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In 2018 IEEE/CVF conference on computer vision and pattern recognition (pp. 6848-6856). IEEE.
74. Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., & Dally, W. J., et al. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360.
75. Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PmLR.
76. Tan, M., Chen, B., Pang, R., Vasudevan, V., & Sandler, M., et al. (2019, June). Mnasnet: Platform-aware neural architecture search for mobile. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2815-2823). IEEE.
77. Cai, H., Zhu, L., & Han, S. (2018). Proxylessnas: Direct neural architecture search on target task and hardware. arXiv preprint arXiv:1812.00332.
78. Cai, H., Gan, C., Wang, T., Zhang, Z., & Han, S. (2019). Once-for-all: Train one network and specialize it for efficient deployment. arXiv preprint arXiv:1908.09791.
79. Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for efficient neural network. Advances in neural information processing systems, 28.
80. Li, H., Kadav, A., Durdanovic, I., Samet, H., & Graf, H. P. (2017). Pruning filters for efficient ConvNets. In International Conference on Learning Representations.
81. Frankle, J., & Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. In International Conference on Learning Representations.
82. Yang, T. J., Chen, Y. H., & Sze, V. (2017, July). Designing energy-efficient convolutional neural networks using energy-aware pruning. In 2017 IEEE conference on computer vision and pattern recognition (CVPR) (pp. 6071-6079). IEEE.
83. Jacob, B., Kligys, S., Chen, B., Zhu, M., & Tang, M., et al. (2018, June). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In 2018 IEEE/CVF conference on computer vision and pattern recognition (pp. 2704-2713). IEEE.
84. Rastegari, M., Ordonez, V., Redmon, J., & Farhadi, A. (2016, September). Xnor-net: Imagenet classification using binary convolutional neural networks. In European conference on computer vision (pp. 525-542). Cham: Springer International Publishing.
85. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.
86. Zhang, L., Song, J., Gao, A., Chen, J., & Bao, C., et al. (2019, October). Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 3712-3721). IEEE.
87. Teerapittayanon, S., McDanel, B., & Kung, H. T. (2016, December). Branchynet: Fast inference via early exiting from deep neural networks. In 2016 23rd international conference on pattern recognition (ICPR) (pp. 2464-2469). IEEE.
88. Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020, November). Transformers are rnns: Fast autoregressive transformers with linear attention. In International conference on machine learning (pp. 5156-5165). PMLR.
89. Rahman, M. H., Yusuf, M. S. U., Bappy, D. M., Islam, M. Z., & Dey, A. K. (2012). Object Duplication in an Image using Texture Synthesis Technique. International Journal of Engineering and Technology, 2(4).
90. Kim, E., Lee, Y. J., Lee, J., Lee, H., & Bappy, D. M., et al. (2025). Fusing Domains for Tumor Detection in Frozen Section Histopathology Using Joint Feature Alignment and Self-Knowledge Distillation. IEEE Access, 13, 190745-190753.

91. Jeong, Y., Bappy, D. M., Ahn, S., & Lee, S. H. (2026). Predicting TCGA molecular subtypes of gastric cancer from H&E whole-slide images using a weakly supervised transMIL-attention framework. Cancer Research, 86(7_Supplement), 1445-1445.
92. Shannon, C. E. (1948). A mathematical theory of communication. The Bell system technical journal, 27(3), 379-423.
93. Berger, T., & Gibson, J. D. (1998). Lossy source coding. IEEE Transactions on Information Theory, 44(6), 2693-2723.
94. Wallace, G. K. (1992). The JPEG still picture compression standard. IEEE transactions on consumer electronics, 38(1), xviii-xxxiv.
95. Skodras, A., Christopoulos, C., & Ebrahimi, T. (2001). The JPEG 2000 still image compression standard. IEEE Signal processing magazine, 18(5), 36-58.
96. Sullivan, G. J., Ohm, J. R., Han, W. J., & Wiegand, T. (2012). Overview of the high efficiency video coding (HEVC) standard. IEEE Transactions on circuits and systems for video technology, 22(12), 1649-1668.
97. Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.
98. Dodge, S., & Karam, L. (2016, June). Understanding how image quality affects deep neural networks. In 2016 eighth international conference on quality of multimedia experience (QoMEX) (pp. 1-6). IEEE.
99. Ballé, J., Laparra, V., & Simoncelli, E. P. (2016). End-to-end optimized image compression. arXiv preprint arXiv:1611.01704.
100. Ballé, J., Minnen, D., Singh, S., Hwang, S. J., & Johnston, N. (2018). Variational image compression with a scale hyperprior. arXiv preprint arXiv:1802.01436.
101. Duan, L., Liu, J., Yang, W., Huang, T., & Gao, W. (2020). Video coding for machines: A paradigm of collaborative compression and intelligent analytics. IEEE Transactions on Image Processing, 29, 8680-8695.
102. Le, N., Zhang, H., Cricri, F., Ghaznavi-Youvalari, R., & Rahtu, E. (2021, June). Image coding for machines: an end-to-end learned approach. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1590-1594). IEEE.
103. Lee, J., Takemaru, L., Bappy, D. M., Jeong, Y. S., & Jeong, W. K., et al. (2025). Adaptive compression framework for gigapixel whole slide images. Nature communications, 17(1), 207.
104. Lee, H., Kim, E., Bappy, D. M., Kim, H. H., & Lee, Y., et al. (2026, September). Patch-to-Global: Random Patch Diffusion for Globally Consistent Megapixel Artifact Inpainting in Whole Slide Images. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 353-363). Cham: Springer Nature Switzerland.
105. Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE internet of things journal, 3(5), 637-646.
106. Kang, Y., Hauswald, J., Gao, C., Rovinski, A., & Mudge, T., et al. (2017). Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. ACM SIGARCH Computer Architecture News, 45(1), 615-629.
107. Langer, S. G., French, T., & Segovis, C. (2011). TCP/IP optimization over wide area networks: implications for teleradiology. Journal of Digital Imaging, 24(2), 314-321.
108. Weinstein, S., & Ebert, P. (1971). Data transmission by frequency-division multiplexing using the discrete Fourier transform. IEEE transactions on Communication Technology, 19(5), 628-634.
109. Bappy, D. M., Dey, A. K., Saha, S., Saha, A., & Ghosh, S. (2010). OFDM System analysis for reduction of inter symbol interference using the AWGN channel platform. International Journal of Advanced Computer Science and Applications, 1(5).
110. Bennis, M., Debbah, M., & Poor, H. V. (2018). Ultrareliable and low-latency wireless communication: Tail, risk, and scale. Proceedings of the IEEE, 106(10), 1834-1853.
111. Nasrallah, A., Thyagaturu, A. S., Alharbi, Z., Wang, C., & Shao, X., et al. (2018). Ultra-low latency (ULL) networks: The IEEE TSN and IETF DetNet standards and related 5G ULL research. IEEE Communications Surveys & Tutorials, 21(1), 88-145.
112. Kim, Y., BAPPY, D., & Kim, J. (2021). Artificial intelligence IoT architecture based on microservices. The Journal of Contents Computing, 3(1), 289-295.
113. Islam, M. S., Khan, R. H., & Bappy, D. M. (2010). A hierarchical intrusion detection system in wireless sensor networks. Computer Science and Network Security, 10(8), 21-26.
114. Richard, J. P. (2003). Time-delay systems: an overview of some recent advances and open problems. automatica, 39(10), 1667-1694.
115. Sheridan, T. B., & Ferrell, W. R. (1963). Remote manipulative control with transmission delay. IEEE Transactions on Human Factors in Electronics, (1), 25-29.
116. Anderson, R. J., & Spong, M. W. (1988, August). Bilateral control of teleoperators with time delay. In Proceedings of the 1988 IEEE International Conference on Systems, Man, and Cybernetics (Vol. 1, pp. 131-138). IEEE.
117. Marescaux, J., Leroy, J., Gagner, M., Rubino, F., & Mutter, D., et al. (2001). Transatlantic robot-assisted telesurgery. Nature, 413(6854), 379-380.
118. Xu, S., Perez, M., Yang, K., Perrenot, C., & Felblinger, J., et al. (2014). Determination of the latency effects on surgical performance and the acceptable latency levels in telesurgery using the dV-Trainer® simulator. Surgical endoscopy, 28(9), 2569-2576.
119. Klein, G., & Murray, D. (2007, November). Parallel tracking and mapping for small AR workspaces. In 2007 6th IEEE and ACM international symposium on mixed and augmented reality (pp. 225-234). IEEE.
120. Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2014). High-speed tracking with kernelized correlation filters. IEEE transactions on pattern analysis and machine intelligence, 37(3), 583-596.

121. Uecker, M., Zhang, S., Voit, D., Karaus, A., & Merboldt, K. D., et al. (2010). Real-time MRI at a resolution of 20 ms. NMR in Biomedicine, 23(8), 986-994.
122. Perdios, D., Vonlanthen, M., Martinez, F., Arditi, M., & Thiran, J. P. (2021). CNN-based image reconstruction method for ultrafast ultrasound imaging. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, 69(4), 1154-1168.
123. Dong, Z., Wang, X., Fang, G., He, Z., & Ho, J. D. L., et al. (2022). Shape tracking and feedback control of cardiac catheter using MRI-guided robotic platform—Validation with pulmonary vein isolation simulator in MRI. IEEE Transactions on Robotics, 38(5), 2781-2798.
124. Nguyen, P. B., Kang, B., Bappy, D. M., Choi, E., & Park, S., et al. (2018). Real-time microrobot posture recognition via biplane X-ray imaging system for external electromagnetic actuation. International journal of computer assisted radiology and surgery, 13(11), 1843-1852.
125. Bappy, D. M., Hong, A., Choi, E., Park, J. O., & Kim, C. S. (2021). Automated three-dimensional vessel reconstruction based on deep segmentation and bi-plane angiographic projections. Computerized Medical Imaging and Graphics, 92, 101956.
126. Wang, P., Berzin, T. M., Glissen Brown, J. R., Bharadwaj, S., & Becq, A., et al. (2019). Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut, 68(10), 1813-1819.
127. Bappy, D. M., Kang, D., Lee, J., Lee, Y., & Koo, M., et al. (2025). Advanced Endoscopy Imaging with Automatic Feedback. In International Conference on Pattern Recognition (pp. 62-78). Springer, Cham.
128. Mittal, A., Soundararajan, R., & Bovik, A. C. (2012). Making a "completely blind" image quality analyzer. IEEE Signal processing letters, 20(3), 209-212.
129. Zhou, H., Bateni, S., & Liu, C. (2018, April). S^ 3dnn: Supervised streaming and scheduling for gpu-accelerated real-time dnn workloads. In 2018 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS) (pp. 190-201). IEEE.
130. Xiang, Y., & Kim, H. (2019, December). Pipelined data-parallel CPU/GPU scheduling for multi-DNN real-time inference. In 2019 IEEE Real-Time Systems Symposium (RTSS) (pp. 392-405). IEEE.
131. Gujarati, A., Karimi, R., Alzayat, S., Hao, W., & Kaufmann, A., et al. Vigfusson, Y., & Mace, J. (2020). Serving {DNNs} like clockwork: Performance predictability from the bottom up. In 14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20) (pp. 443-462).
132. Buurlage, J. W., Kohr, H., Jan Palenstijn, W., & Joost Batenburg, K. (2018). Real-time quasi-3D tomographic reconstruction. Measurement Science and Technology, 29(6), 064005.
133. Reddi, V. J., Cheng, C., Kanter, D., Mattson, P., & Schmuelling, G., et al. (2020, May). Mlperf inference benchmark. In 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA) (pp. 446-459). IEEE.
134. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green ai. Communications of the ACM, 63(12), 54-63.

Copyright: *©2026 Nguyn Gia Huy, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*