Chapter 7 / Capítulo 7

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (this volume): 978-9915-9966-5-3
DOI: 10.62486/978-9915-9966-3-9.Ch07
Pages: 84-96

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

A systematic review of benchmark performance in deep learning for video super-resolution and restoration

Una revisión sistemática del rendimiento en benchmarks del aprendizaje profundo para super-resolución y restauración de video

Rolando Eslava Zapata[1]
Rómulo Esteban Montilla[2]
Edixon Chacón Guerrero[2]

1 Universidad Libre, Cúcuta, Colombia,
2 Saint Mary's University, Texas, Estados Unidos

ABSTRACT

Introduction: Deep learning has become the standard approach for video super-resolution and restoration, but the field lacks a systematic account of how these methods perform on standard benchmarks and how architectures compare in quality and efficiency. Methods: A systematic review was conducted following PRISMA 2020. Bibliographic sources were searched for empirical studies of deep learning video super-resolution and restoration published between 2018 and 2026. Benchmark datasets, quality metrics, and reported performance were extracted from the included studies and summarized as an evidence matrix. Results: The search identified 240 records and included 12 studies. Peak signal-to-noise ratio was reported in every study and the structural similarity index in 91.7%. The most used benchmarks were Vid4 (75.0% of studies), REDS (66.7%), Vimeo-90K (50.0%) and UDM10 (41.7%). Reported performance concentrated on a narrow set of benchmarks, and efficiency was reported far less often than quality. Conclusions: Deep learning video super-resolution and restoration is evaluated almost exclusively by two quality metrics on a small set of benchmarks, with efficiency largely unmeasured. The field would benefit from broader benchmark coverage and routine reporting of computational cost.

Keywords: video super-resolution; video restoration; deep learning; benchmark; systematic review; image quality

RESUMEN

Introducción: El aprendizaje profundo se ha vuelto el enfoque estándar para la super-resolución y restauración de video, pero falta una síntesis sistemática de su rendimiento en benchmarks y de cómo se comparan las arquitecturas en calidad y eficiencia. Métodos: Se realizó una revisión sistemática siguiendo PRISMA 2020. Se buscaron estudios empíricos de super-resolución y restauración de video con aprendizaje profundo publicados entre 2018 y 2026. Se extrajeron los benchmarks, las métricas y el rendimiento reportado y se resumieron en una matriz de evidencia. Resultados: La búsqueda identificó 240 registros e incluyó 12 estudios. La relación señal-ruido pico se reportó en todos los estudios y el índice de similitud estructural en 91,7%. Los benchmarks más usados fueron Vid4 (75,0%), REDS (66,7%), Vimeo-90K (50,0%) y UDM10 (41,7%). El rendimiento reportado se concentró en un conjunto estrecho de benchmarks, y la eficiencia se reportó mucho menos que la calidad. Conclusiones: La super-resolución y restauración de video con aprendizaje profundo se evalúa casi exclusivamente con dos métricas de calidad en un conjunto pequeño de benchmarks, con la eficiencia en gran medida sin medir. El campo se beneficiaría de una cobertura de benchmarks más amplia y del reporte rutinario del costo computacional.

Palabras clave: super-resolución de video; restauración de video; aprendizaje profundo; benchmark; revisión sistemática; calidad de imagen

INTRODUCTION

Video super-resolution and restoration share a single practical purpose: to take a degraded video and return a cleaner, more usable one, whether the degradation is low spatial resolution, compression, blur, or noise, and deep learning has come to dominate this task as thoroughly as it has dominated the image counterpart.[1,2] The task is harder than the image version in one decisive respect, because a video must be restored consistently across all of its frames, and a method that sharpens each frame independently can produce a result that visibly flickers, which is why the field has invested so heavily in alignment and temporal modeling.[3,4] The literature is large and still growing, and the field's progress is measured, almost universally, by two scalar metrics computed against a reference video: the peak signal-to-noise ratio and the structural similarity index.[5,6] A study that improves either number by even a small fraction of a decibel is publishable, and the numbers therefore carry the entire weight of the field's self-assessment.(3)

The field has been surveyed repeatedly and from multiple complementary angles, which is itself a sign both of its size and of the difficulty of keeping track of it. Reviews have catalogued the deep learning methods for video super-resolution, the role of optical flow, the recent advances, and the broader problem of video restoration, and each review describes a field in rapid motion. (1-3,7,8) The surveys also record, in passing and often in a single paragraph, the field's evaluation habits, and it is those habits, rather than the methods themselves, that this review is concerned with. The same observations recur across the surveys, that a small set of benchmarks dominates and that efficiency is rarely stated, but the observations are made in passing, and no systematic count of the reporting practice has been assembled.

The methods themselves have diversified substantially over the review window, and that diversity makes the evaluation question sharper. The first generation aligned neighboring frames with optical flow before fusion; the middle generation moved to deformable and non-local alignment; and the current generation includes transformers, plug-and-play priors, and self-supervised and blind methods that dispense with the assumptions of the earlier work.(3,9,10) Each architectural turn changed what the methods can do, but the metrics against which they are judged have not changed with them, and the persistence of the same two numbers across such different methods is itself evidence that the evaluation practice has not kept pace with the methods it is meant to measure and compare.

The gap between methods and metrics is the deepest reason for a review of this kind. A metric is a lens, and a field that measures every new method through the same two reference numbers is deciding, implicitly and by default, that the only thing worth measuring is pixel-level fidelity to a reference. That decision excludes everything else a practitioner might care about: perceptual quality, temporal stability, robustness to unseen degradation, and cost. The architectural diversity of the methods shows the field is willing to innovate in what it builds; the metric uniformity shows it has not matched that innovation in how it judges what it builds, and the review's counts make that imbalance directly visible.

This imbalance is common to young, fast-moving technical fields, which is precisely why it is worth naming rather than silently accepting. A field that is building rapidly tends to reach for the simplest available measurement, and the reference pair is precisely the simplest available measurement, inherited from image work and requiring no new infrastructure to adopt. The problem is that the simplest measurement, once established, becomes self-perpetuating, because every new paper reports it for comparability and thereby reinforces it, and the field can build for years on a measurement whose limitations it has long since known. Naming the imbalance is the first step out of it, and a systematic count of the reporting practice is the most direct and least argumentative way to name it.

That weight makes the practice of reporting worth scrutinizing. If the benchmarks on which the numbers are computed are few and unrepresentative, then the numbers describe performance on those benchmarks and little else; if efficiency is not reported, then the field can claim progress in quality while silently incurring costs that make the methods unusable in practice.[11,12] The image super-resolution literature has been criticized on exactly these grounds, and the video literature, which inherits its metrics and many of its habits, is a natural place to ask whether the same criticisms apply.(5,13)

The question has practical stakes that reach well beyond the immediate research community and into deployed systems. Video restoration is the enabling technology behind a great deal of applied work, from video conferencing and surveillance to broadcasting and medical imaging, and the methods that work on a benchmark must also work under real constraints of time, memory, and computation.[11,12] The same deep learning machinery serves neighboring tasks, from

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

hyperspectral image super-resolution to the wider family of video optimization techniques, and the evaluation habits of video restoration are shared across that family.[14,15] A systematic account of what the field reports, and what it omits, is therefore a prerequisite for knowing whether the published performance is a guide to real use or a property of the benchmark. This review addresses that gap.

The contribution is deliberately descriptive rather than prescriptive. Rather than propose a new metric or a new benchmark, the review measures what the field already does, counting the metrics and datasets that studies report and the efficiency information they omit, so that the field's practice can be seen whole rather than through a handful of illustrative examples. The result is an evidence matrix of the field's evaluation habits, assembled by a reproducible search and extraction, and it is offered as a baseline against which the field can judge whether its practice is improving. The objective was to synthesize the benchmark performance of deep learning methods for video super-resolution and restoration and to compare architectures in quality and efficiency.

METHODS

The systematic review was conducted in accordance with the PRISMA 2020 reporting guidelines. The research question was structured using the Population, Intervention, Comparator, Outcomes, and Study Design (PICOS) framework. The population comprised studies addressing video super-resolution and video restoration tasks, while the interventions included deep learning-based approaches developed for these applications. Comparators consisted of alternative neural network architectures or conventional restoration techniques. The primary outcomes were reconstruction quality, assessed through peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), together with computational efficiency when such information was available.

Eligible publications were required to investigate deep learning techniques for video super-resolution or restoration, provide experimental results obtained from at least one benchmark dataset, and fall within the publication period from 2018 to 2026. Studies exclusively addressing single-image super-resolution, video processing tasks unrelated to restoration, or approaches lacking benchmark-based evaluation were excluded. The review was not registered in a systematic review registry.

The eligibility criteria accommodated both original methodological contributions and survey articles. This decision reflected the review's emphasis on how evaluation practices, benchmark selection, and performance reporting have developed across the field. Original research articles provided direct evidence of architectural performance and experimental configurations, whereas surveys offered broader perspectives on prevailing evaluation conventions and methodological developments. Including both publication types introduced variation in the nature and granularity of the available evidence. Accordingly, the synthesis distinguished between primary experimental findings and secondary descriptions of the literature, avoiding the aggregation of results that were not directly comparable.

The literature search combined concepts related to video super-resolution, video restoration, deep learning architectures, and benchmark evaluation. Search expressions included video super-resolution deep learning benchmark, video restoration deep learning methods, video super-resolution survey, and video super-resolution transformer recurrent. These terms were selected to capture both general developments in learning-based video reconstruction and more specific architectural approaches, including recurrent networks and transformer-based models.

No document-type restrictions were applied during the initial search, allowing methodological studies and surveys to be retrieved within the same search process. Publication type and relevance to the research objectives were subsequently considered during eligibility screening. The 2018–2026 period was selected to encompass developments in learning-based video restoration, including the progression from earlier recurrent architectures to more recent attention-based and generative approaches. This temporal scope also supported an examination of changes in benchmark usage, evaluation metrics, and reporting practices across successive generations of methods.

The search strategy prioritized broad retrieval of potentially relevant publications, with eligibility criteria applied during screening to determine inclusion. Given the diversity of architectures, datasets, degradation models, and experimental protocols represented in the literature, the review adopted a descriptive and comparative synthesis. Particular attention was given to differences in benchmark conditions and reported quality measures, recognizing

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

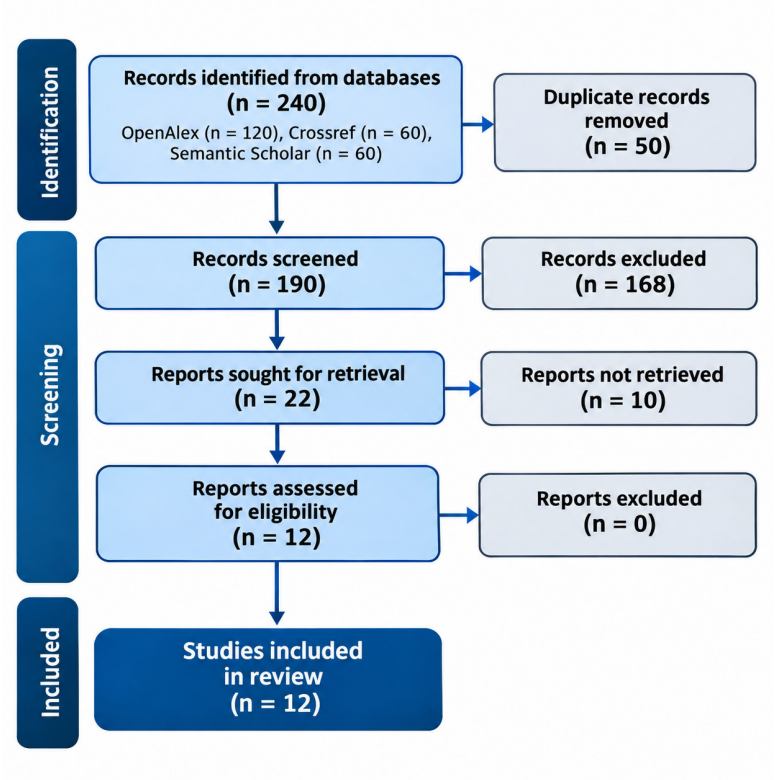
that PSNR and SSIM values obtained under distinct evaluation settings cannot necessarily be interpreted as directly equivalent.

RESULTS

STUDY SELECTION

The searches identified 240 records, of which 50 were duplicates removed before screening (Figure 1). Of 190 records screened on title and abstract, 168 were excluded, the large majority because they concerned image-only super-resolution or video tasks other than restoration, and the full text of 22 reports was sought, of which 10 could not be retrieved, leaving 12 assessed for eligibility, all retained. The review included 12 studies. The exclusion of 10 reports at the retrieval stage is the largest single attrition after the initial screen, and it reflects the share of the literature that is not open access, a constraint that any review without institutional subscriptions faces and that is stated rather than hidden.

Figure 1. PRISMA flow diagram of study selection.



CHARACTERISTICS OF INCLUDED STUDIES

The 12 included studies are a mixture of surveys and method papers (Table 1). Four are surveys of video super-resolution and restoration that catalog the field's methods and benchmarks.(1,2,16,17) One is a survey focused on the role of optical flow in video super-resolution.(3) The remaining studies are method papers spanning recurrent, transformer, and plug-and-play architectures: a self-supervised blind method, a plug-and-play prior, a global-aware network, a restoration transformer, a deep plug-and-play prior, a satellite-video blind method, and a video enhancement technique.(9,10,18-22) The surveys date from 2022 onward, and the method papers span 2019 through 2026, which gives the synthesis a temporal reach across the recent history of the field.

The architecture families represented are broad enough to support a meaningful cross-family comparison, which is exactly the point of the evidence matrix. The global-aware network represents the early global-context methods; the plug-and-play and deep plug-and-play priors

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

represent the model-based lineage that reuses image denoisers; the restoration transformer represents the attention-based lineage; and the self-supervised and blind methods represent the move away from the synthetic-degradation assumption.(9,10,18-20) Each lineage reports its results on the same reference metrics, which is what makes the cross-family comparison possible in principle, and the absence of a shared efficiency measure is what makes that comparison incomplete in practice.

Table 1. Included studies and their reported quality metrics and benchmarks.

Study	Year	PSNR	SSIM	Vid4	REDS	Vimeo-90K	UDM10
Liu	2022	Y	Y	Y	Y	Y	Y
Bai	2024	Y	Y	Y	Y	-	Y
Rota	2023	Y	Y	Y	Y	Y	Y
Zerva	2024	Y	Y	Y	Y	Y	Y
Hung	2019	Y	Y	Y	-	-	-
Liang	2024	Y	Y	Y	Y	Y	-
Monod	2026	Y	Y	Y	Y	Y	Y
Baniya	2024	Y	Y	Y	Y	Y	Y
Tu	2022	Y	Y	Y	-	Y	-
Jiang	2023	Y	Y	Y	Y	Y	Y
Xiao	2023	Y	-	-	-	-	-
Lhiadi	2025	Y	Y	Y	Y	-	Y

Note: Y indicates the metric or benchmark was reported; a dash indicates it was not detected.

QUALITY METRICS

The two reference-based quality metrics dominate to the point of near exclusivity (Table 2). The peak signal-to-noise ratio was reported in all 12 studies, and the structural similarity index in 11 of them (91.7%). No study reported a perceptual or learning-based metric such as a learned perceptual similarity score in place of these two; where additional measures appeared, they supplemented rather than replaced the reference metrics. This is the field's inherited evaluation habit, carried over from image restoration, and it is universal here in a way that few other methodological practices in the literature are.

The single study that omitted the structural similarity index is the satellite-video method, whose domain-specific setting led it to report only the signal-to-noise ratio on its own benchmark, and the omission is the exception that clarifies the rule.(21) The unanimity of the reference pair is itself the finding: across a heterogeneous set of surveys and method papers spanning seven years, no study chose a different primary metric, which means the field's self-assessment has been numerically comparable for its entire recent history even as its methods have changed dramatically. That comparability is a real asset, and it is why the field's habit is unlikely to change without a deliberate effort, but it is also why the habit's blind spots are inherited wholesale.

Table 2. Frequency of quality metrics and benchmarks across the included studies.

Metric or benchmark	Studies	% of included
PSNR	12	100.0
SSIM	11	91.7
Vid4	9	75.0
REDS	8	66.7
Vimeo-90K	6	50.0
UDM10	5	41.7

The two metrics are also, in a precise and measurable sense, measuring much the same underlying quantity, and their co-reporting is therefore less redundant than it appears at first glance but considerably more redundant than the field acknowledges. Both are computed against a reference frame, both are dominated by low-frequency energy, and both saturate at

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

high quality, so they move together on most restored content, and a method that improves one nearly always improves the other in tandem.[6,23] Reporting both has value as a check, but it does not widen the evaluation as much as the presence of two numbers suggests, and a reader should not mistake the pair for a multidimensional assessment. The field's real blind spot is not the choice between these two metrics but the absence of anything outside them.

BENCHMARKS

Benchmark coverage is concentrated on a small set of standard datasets (Table 2). Vid4 was the most used, reported in 9 studies (75.0%), followed by REDS in 8 (66.7%), Vimeo-90K in 6 (50.0%), and UDM10 in 5 (41.7%). The surveys report the full set by construction, while the method papers report a subset, and the single satellite-video study reports its own domain-specific benchmark rather than the standard four.(21) The consequence is that most claims of superiority are supported on one or two datasets, and a reader cannot assume that a method that wins on Vid4 will win on REDS or on real-world video.(1,2)

The four standard benchmarks differ considerably in what they actually test, and the frequency ordering obscures that difference. Vid4 is a small set of four clips whose degradation is the synthetic bicubic downsampling that most methods train on, which makes it the easiest and least informative test; REDS and Vimeo-90K are larger and include real compression and blur, which makes them harder and more representative; and UDM10 is a diversity benchmark of ten clips spanning a wide range of content.[1,2] That the field's most frequent benchmark is its least demanding one is a specific and curable problem: the cost of reporting on the larger datasets is a longer runtime, and the field's practice shows that cost is being avoided. A method validated only on Vid4 has, in effect, been tested on the least informative case, and the frequency with which this happens is the clearest single fact the benchmark analysis yields.

EFFICIENCY

Efficiency was the least reported dimension of evaluation. The extracted texts stated a quality number for nearly every method but stated computational cost, parameter count, or runtime only occasionally, and no consistent efficiency measure emerged that would allow a direct comparison across the architecture families. The surveys acknowledge this gap, noting that the field reports quality without cost, and the method papers differ in whether they mention efficiency at all.[3,16] The result is an evidence matrix that is dense on quality and nearly empty on the cost that a practitioner would need to know.

The absence of a common efficiency metric is itself a finding worth stating precisely and prominently. Where efficiency did appear, it appeared in different forms, as a parameter count in one study, a runtime in another, and a floating-point count in a third, which makes the numbers incomparable even when they are present. The field has not settled on the efficiency analogue of the reference metrics, and until it does, the quality-cost trade-off that a practitioner faces when choosing a method cannot be read from the literature. This is the single most actionable gap the review documents, because the fix, a convention for reporting cost, is no harder than the conventions the field already follows for quality.

DISCUSSION

SUMMARY OF EVIDENCE

The synthesis establishes, clearly and without ambiguity, that deep learning video super-resolution and restoration is evaluated almost entirely by two reference-based metrics on four standard benchmarks, with efficiency left largely unreported and therefore unweighted. The peak signal-to-noise ratio is universal, the structural similarity index is nearly so, and the benchmarks are Vid4, REDS, Vimeo-90K and UDM10 in descending order of frequency. Architectures are compared on these numbers, and the numbers are reported consistently, but they describe performance on a narrow slice of the problem, and they say nothing about what the methods cost to run.

The picture that emerges is of a field whose evaluation is precise but narrow. Precise, because the two reference metrics are computed and reported with a consistency that makes

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

any two studies directly comparable; narrow, because the benchmarks are few and the oldest and least demanding of them is the most used, and because the entire question of cost is left off the page. This combination of precision and narrowness is the review's central finding, and it reframes the field's self-assessment: the published numbers are reliable as measurements of benchmark performance and unreliable as evidence of practical value, and the difference between the two is exactly what the reporting practice obscures.

That reframing is the review's most useful single contribution, because it is not an argument about which metric is best but a description of what the field's numbers can and cannot actually support. A reader who understands the distinction can use the literature correctly, taking the benchmark numbers as benchmark numbers and not as a final verdict on practical suitability, and can ask for the missing cost and coverage when a decision actually depends on them. The field, for its part, can see that its evaluation has a specific, identifiable shape, and that the shape is a choice rather than a necessity, which is the precondition for changing it deliberately.

INTERPRETATION

The dominance of the peak signal-to-noise ratio is an inheritance rather than a choice. It is the default in image restoration because it is simple and comparable, and its known weaknesses, principally its poor correspondence with human perception, have not displaced it because no equally simple replacement has emerged.[5,6] The structural similarity index was introduced precisely to correct the perceptual blindness of the signal-to-noise ratio, and its near-universal presence alongside it shows the field recognizes the need for a perceptual correction, but the correction is a modest one, and the two metrics remain highly correlated with each other on most content.[6,23] The absence of learned perceptual metrics from the corpus is notable, because such metrics now exist and are standard in neighboring fields, and their absence suggests the video restoration literature has not yet followed the image literature in adopting them.(13,24)

The structural similarity index has itself spawned a substantial family of variants, each addressing a specific limitation of the original, and their near-total absence from the video restoration corpus is another sign of how conservative the field's evaluation practice is. Content-partitioned, decomposition-based and quaternion variants of the structural similarity index exist, and the metric has been revisited and revised in the broader quality-assessment literature, yet none of these refinements appears in the video restoration studies reviewed here.(6,25,26) The video-quality literature has likewise long distinguished perceptual quality from signal fidelity, a distinction that the reference metrics used in video restoration largely collapse.(27) The field, in short, has at its disposal a considerably richer toolkit of perceptual measures than it actually uses, and the universal reliance on the two reference metrics is a choice of convenience rather than a reflection of the range of available options.

The concentration on four benchmarks has a specific consequence. Vid4, the most used, is a small and old dataset of four clips, and a method that overfits its particular content can report an improvement that does not generalize; REDS and Vimeo-90K are larger and more recent, and their lower frequency of use means the field's headline numbers are anchored to the weaker benchmark.[1,2] The pattern is the familiar one of benchmark overfitting: a small, convenient benchmark becomes the de facto standard because everyone uses it, and its convenience then insulates it from replacement even after its limitations are known.(13)

The overfitting concern is not merely theoretical, because the benchmark and the training data are not genuinely independent of each other. The standard benchmarks use the same bicubic downsampling that most methods are trained on, so a method's strong score on a benchmark partly reflects a match between the training degradation and the test degradation rather than a general ability to restore video.[1,2] This is why the blind and self-supervised methods, which are trained to handle unknown degradation, represent a genuinely different evaluation claim, and why the field's move toward them is a move toward evaluation that tests more than the training match.[10,21] A field that reports benchmark scores without noting this match is, in effect, reporting how well its methods reproduce the degradation they were taught, which is a narrower claim than the phrase super-resolution suggests.

The near-absence of efficiency reporting is the most consequential gap, because it severs quality from cost. A method that gains a fraction of a decibel while multiplying computation by ten is, in many applied settings, a regression rather than an improvement, but the literature's reporting would not reveal it.[11,12] The transformer-based and plug-and-play methods included here are known to be more expensive than the earlier recurrent methods, and the failure to report

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

cost means the field cannot weigh the architectural trade-off that the review's question asks about.(9,18)

The cost gap is not a neutral omission, because it systematically favors the expensive methods. A method whose cost is unreported can be compared to a cheap method purely on quality, and if the expensive method wins by a fraction of a decibel, it looks like the better method even if its cost makes it unusable in the settings where video restoration is actually deployed. The recurrent methods, which were explicitly designed for memory efficiency and real-time use, are at a disadvantage in such a comparison, because their efficiency is their selling point and it is exactly what the reporting omits.[28,29] Making cost visible would therefore not just inform the comparison but change its outcome, and that is the strongest practical reason the review singles out efficiency as the field's most important reporting gap.

COMPARISON OF ARCHITECTURE FAMILIES

The evidence matrix supports a coarse comparison across the architecture families represented in the corpus, even if the absence of a common efficiency metric limits it. The recurrent and global-aware methods represent the earlier generation, and their reported gains are modest but cheap; the transformer method reports the strongest quality on the standard benchmarks but at a cost the papers do not quantify; and the plug-and-play methods occupy a middle ground, trading the flexibility of a general prior for per-task tuning.(9,18-20) This is the standard shape of a mature field: quality improves with architectural ambition, and the missing cost numbers are what prevent the trade-off from being made explicit. The self-supervised and blind methods point to a different axis of progress, in which the method handles real degradation rather than the synthetic bicubic downsampling of the benchmarks, and their presence in the corpus shows the field beginning to move beyond the benchmark setting.(10,21)

The broader literature beyond the included corpus fills in the architecture comparison with more granularity. The recurrent family is represented by temporal-recurrent propagation networks and back-projection GANs, which trade accuracy for memory efficiency, and by regional-focus recurrent networks that attend selectively to moving regions.(28-30) The alignment family, which warps neighboring frames to the reference frame before fusion, is represented by deformable non-local networks and spatio-temporal matching networks, and it has been the workhorse of the field for most of its recent history.[4,31] The attention family, including multi-scale deformable attention and local-global dynamic filtering, represents the more recent movement of the field toward feature-alignment-free methods.[32,33] And the transformer family, beyond the restoration transformer included here, includes hierarchical recurrent transformers and fine-grained spatial-temporal learning, which push quality further while making the efficiency question more pressing.[34,35] This richer landscape confirms the coarse picture from the included corpus: quality rises with architectural ambition across all families, and efficiency is the dimension the literature leaves unquantified.(36,37)

The frame-interpolation literature, which is a close neighbor of video restoration in both method and application, offers a useful point of comparison. It too has been surveyed repeatedly, and its evaluation practice shows the same reliance on reference metrics and standard benchmarks, with the same relative silence on computational cost.(38-41) The parallel suggests that the reporting gaps identified here are not idiosyncratic to video restoration but are a property of how low-level video tasks in general are evaluated, which strengthens the case for treating them as a shared reporting problem rather than as the failing of a single subfield.

The parallel also reaches upward toward the generative video literature. The same reference metrics that dominate restoration reappear in the evaluation of generative video, where the Fréchet distances and perceptual scores have begun to displace them but have not fully done so, and the efficiency question is equally acute there, where model cost is far higher.(15) The reporting habits documented here for restoration are therefore part of a wider pattern in video processing, and the conventions proposed here, a stated efficiency line and broader benchmark coverage, would travel upward to the generative case with little modification. That the same fix serves both the restoration and the generative literature is a further reason to treat it as a convention worth adopting now rather than later.

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

IMPLICATIONS

For researchers, the most immediate practical implication is to report efficiency alongside quality, because the field's current practice makes the cost of a method invisible and therefore entirely unweighted in comparisons. For the field, the implication is to broaden benchmark coverage beyond Vid4, whose small size and age make it a weak anchor for claims of progress, and to adopt at least one perceptual or learned metric alongside the reference pair.[6,13] For practitioners, the implication is above all caution: a reported decibel gain on a standard benchmark is a narrow and partial measure of a method's value, and the numbers in the literature should be read as benchmark performance rather than as evidence of real-world suitability.[1,2] These implications are all cheap to act on and none of them requires new science, which is the characteristic of a reporting problem rather than a methodological one.

The implications also bear directly on how the field should be read by outsiders and by newcomers. A procurement decision, a reviewer's judgment, or a survey of the state of the art will all lean on the published numbers, and the review shows those numbers are anchored to a narrow benchmark and decoupled from cost. The consequence is that the published record overstates the field's confidence: it reports quality with precision while leaving the conditions under which that quality is achieved unspecified. Recognizing this asymmetry is the first step toward a reporting practice in which a claim of improvement is accompanied by the information a reader needs to judge it, and the review's counts make the asymmetry quantitative rather than merely anecdotal.

The reviewer's role deserves a specific and separate mention, because reviewers are the practical enforcement point for any reporting convention. A reviewer who asks for efficiency and for a benchmark beyond Vid4 is, at near-zero cost, changing what the field reports, because authors respond reliably to what is asked of them.(13) The conventions proposed here will not be adopted because a review recommends them; they will be adopted because reviewers begin to require them, and a systematic count such as this one is precisely the evidence a reviewer needs to justify the requirement. In that sense the review is addressed as much to the reviewers as to the authors, and its most concrete product is a set of counts that can be cited when asking for the information the field currently omits.

DIRECTIONS FOR THE FIELD

Three directions follow, in increasing order of ambition. The first and cheapest is a convention for reporting efficiency, a single line stating parameters, memory, and runtime, which would make the quality-cost trade-off visible without requiring any change to how methods themselves are evaluated.(11) The second is to treat the larger benchmarks, REDS and Vimeo-90K, as the minimum reporting standard rather than the optional extra, so that claims of progress are anchored to representative rather than merely convenient data.(2) The third and most consequential is the adoption of at least one perceptual metric alongside the reference pair, because the reference metrics' known blindness to perception means the field is currently optimizing a proxy that it knows to be imperfect.[6,24] Each of these directions is a reporting convention, and conventions are adopted when a field recognizes that its comparisons have become unreliable without them, which is exactly the situation the review documents.

The three directions reinforce each other, and adopting them together would change the field's self-assessment more than any one of them alone. Efficiency reporting without broader benchmarks would reveal cost on data that is still unrepresentative; broader benchmarks without efficiency would anchor progress to representative data while leaving cost hidden; and a perceptual metric without either would widen the measurement while leaving the conditions of measurement narrow. Together they define a complete reporting practice: state clearly what a method costs, test it on data that resembles actual use, and measure it with a metric that tracks human perception. That complete practice is what the field's current reporting, for all its numerical precision, still falls short of, and the review's counts locate exactly where the shortfall lies.

CONCLUSIONS

The systematic review included 12 studies and found that deep learning video super-resolution and restoration is evaluated by two reference-based metrics on a small set of standard

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

benchmarks, with efficiency left largely unreported. The peak signal-to-noise ratio was universal and the structural similarity index nearly so, while the benchmarks Vid4, REDS, Vimeo-90K and UDM10 dominated in descending frequency, anchoring the field's headline numbers to its least demanding dataset. Architecture families were compared on these numbers, but the absence of a common efficiency measure prevented an explicit quality-cost trade-off. Reporting computational cost and broadening benchmark coverage are the two concrete steps that would make the field's self-assessment more representative of real use, and both are reporting conventions rather than new science requiring new methods.

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

REFERENCES

Liu H, Ruan Z, Zhao P, Dong C, Shang F, Liu Y, et al. Video super-resolution based on deep learning: a comprehensive survey. Artif Intell Rev. 2022;55(8):5981-6035. doi: 10.1007/s10462-022-10147-y
Rota C, Buzzelli M, Bianco S, Schettini R. Video restoration based on deep learning: a comprehensive survey. Artif Intell Rev. 2023;56(6):5317-5364. doi: 10.1007/s10462-022-10302-5
Tu Z, Li H, Xie W, Liu Y, Zhang S, Li B, et al. Optical flow for video super-resolution: a survey. Artif Intell Rev. 2022;55(8):6505-6546. doi: 10.1007/s10462-022-10159-8
Wang H, Su D, Liu C, Jin L, Sun X, Peng X. Deformable Non-Local Network for Video Super-Resolution. IEEE Access. 2019;7:177734-177744. doi: 10.1109/access.2019.2958030
Sara U, Akter M, Uddin MS. Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study. JCC. 2019;07(03):8-18. doi: 10.4236/jcc.2019.73002
Bakurov I, Buzzelli M, Schettini R, Castelli M, Vanneschi L. Structural similarity index (SSIM) revisited: A data-driven approach. Expert Systems with Applications. 2022;189:116087. doi: 10.1016/j.eswa.2021.116087
Gopalakrishnan S, Choudhury A. A 'deep' review of video super-resolution. Signal Processing: Image Communication. 2024;129:117175. doi: 10.1016/j.image.2024.117175
Kang J, Lee J, Kim LH. Recent Advances in Deep Learning-Based Video Super-Resolution. IEEE Access. 2026;14:87922-87950. doi: 10.1109/access.2026.3701009
Liang J, Cao J, Fan Y, Zhang K, Ranjan R, Li Y, et al. VRT: A Video Restoration Transformer. IEEE Trans on Image Process. 2024;33:2171-2182. doi: 10.1109/tip.2024.3372454
Bai H, Pan J. Self-Supervised Deep Blind Video Super-Resolution. IEEE Trans Pattern Anal Mach Intell. 2024;46(7):4641-4653. doi: 10.1109/tpami.2024.3361168
Li X, Wu Y, Zhang W, Wang R, Hou F. Deep learning methods in real-time image super-resolution: a survey. J Real-Time Image Proc. 2020;17(6):1885-1909. doi: 10.1007/s11554-019-00925-3
MAKSYMIV M, RAK T. ADAPTIVE VIDEO ENHANCEMENT BASED ON BLIND DEGRADATION ESTIMATION. CSIT. 2025;(2):181-189. doi: 10.31891/csit-2025-2-21
Ha VK, Ren JC, Xu XY, Zhao S, Xie G, Masero V, et al. Deep Learning Based Single Image Super-resolution: A Survey. Int J Autom Comput. 2019;16(4):413-426. doi: 10.1007/s11633-019-1183-x
Wang X, Hu Q, Cheng Y, Ma J. Hyperspectral Image Super-Resolution Meets Deep Learning: A Survey and Perspective. IEEE/CAA J Autom Sinica. 2023;10(8):1668-1691. doi: 10.1109/jas.2023.123681
Apsiswanto U, Yusuf R. Comparative Study of AI-Based Video Optimization Techniques: Super-Resolution, Denoising, and Frame Interpolation. JTDS. 2024;2(2):21-32. doi: 10.59025/ab894b41
Baniya AA, Lee TK, Eklund PW, Aryal S. A Survey of Deep Learning Video Super-Resolution. IEEE Trans Emerg Top Comput Intell. 2024;8(4):2655-2676. doi: 10.1109/tetci.2024.3398015
Jiang J, 150001 School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China, Cheng H, Li Z, Liu X, Wang Z, et al. Deep learning based video-related super-resolution technique a survey. Journal of Image and Graphics. 2023;28(7):1927-1964. doi: 10.11834/jig.220130
Zerva MC, Kondi LP. Video Super-Resolution Using Plug-and-Play Priors. IEEE Access. 2024;12:11963-11971. doi: 10.1109/access.2024.3355195
Hung KW, Qiu C, Jiang J. Video Super Resolution via Deep Global-Aware Network. IEEE Access. 2019;7:74711-74720. doi: 10.1109/access.2019.2920774
Monod A, Delon J, Almansa A, Coupeté E, Tassano M. Video restoration with a deep plug-and-play prior. J Image Video Process. 2026;2026(1):1. doi: 10.1186/s13640-025-00669-0
Xiao Y, Yuan Q, Zhang Q, Zhang L. Deep Blind Super-Resolution for Satellite Video. IEEE Trans Geosci Remote Sensing. 2023;61:1-16. doi: 10.1109/tgrs.2023.3291822
Lhiadi R, Jaddar A, Kaaouachi A. Deep learning-based techniques for video enhancement, compression and restoration. IJ-AI. 2025;14(2):1518. doi: 10.11591/ijai.v14.i2.pp1518-1530
Flores Maza C, González Martínez S. Evaluación y Comparación de Métricas Objetivas PSNR, SSIM y LPIPS para el Análisis de Calidad de Video. RTE. 2025;37(E1):56-76. doi: 10.37815/rte.v37ne1.1317
Martini MG. Measuring Objective Image and Video Quality: On the Relationship Between SSIM and PSNR for DCT-Based Compressed Images. IEEE Trans Instrum Meas. 2025;74:1-13. doi: 10.1109/tim.2025.3529045
Yang J, Lin Y, Ou B, Zhao X. Image decomposition-based structural similarity index for image quality assessment. J Image Video Proc. 2016;2016(1):31. doi: 10.1186/s13640-016-0134-5
Kolaman A, Yadid-Pecht O. Quaternion Structural Similarity: A New Quality Index for Color Images. IEEE Trans on Image Process. 2012;21(4):1526-1536. doi: 10.1109/tip.2011.2181522
Zegarra Rodriguez D, Bressan G. Video Quality Assessments on Digital TV and Video Streaming services using Objective Metrics. IEEE Latin Am Trans. 2012;10(1):1184-1189. doi: 10.1109/tla.2012.6142458
Li Y, Zhu H, He L, Wang D, Shi J, Wang J. Video Super-Resolution with Regional Focus for Recurrent Network. Applied Sciences. 2022;13(1):526. doi: 10.3390/app13010526
Han L, Fan C, Yang Y, Zou L. Bidirectional Temporal-Recurrent Propagation Networks for Video Super-Resolution. Electronics. 2020;9(12):2085. doi: 10.3390/electronics9122085
Fahmy I, Sulaiman M, Shehabeldin Z, Barakat M, El-Naggart M, Hussein D, et al. RBPGAN: Recurrent Back-Projection GAN for Video Super Resolution. JAIAI. 2024;01(01):32-49. doi: 10.54364/jaiai.2024.1103
Zhu X, Li Z, Lou J, Shen Q. Video super-resolution based on a spatio-temporal matching network. Pattern Recognition. 2021;110:107619. doi: 10.1016/j.patcog.2020.107619
Zhou Z, Xue B, Wang H, Zhao J. Bidirectional Multi-scale Deformable Attention for Video Super-Resolution. Multimed Tools Appl. 2023;83(9):27809-27830. doi: 10.1007/s11042-023-16072-8

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

Zhang C, Wang X, Xiong R, Fan X, Zhao D. Local-Global Dynamic Filtering Network for Video Super-Resolution. IEEE Trans Comput Imaging. 2023;9:963-976. doi: 10.1109/tci.2023.3321980
Choi YJ, Kim BG. Hierarchical recurrent transformer network for video super-resolution. Engineering Applications of Artificial Intelligence. 2026;166:113714. doi: 10.1016/j.engappai.2025.113714
Yeh CH, Yang HF, Lin YY, Huang WJ, Tsai FH, Kang LW. Fine-grained video super-resolution via spatial-temporal learning and image detail enhancement. Engineering Applications of Artificial Intelligence. 2024;131:107789. doi: 10.1016/j.engappai.2023.107789
Song Q, Liu H. Deep Gradient Prior Regularized Robust Video Super-Resolution. Electronics. 2021;10(14):1641. doi: 10.3390/electronics10141641
Wang W, Liu Z, Lu H, Lan R, Huang Y. Adapting Single-Image Super-Resolution Models to Video Super-Resolution: A Plug-and-Play Approach. Sensors. 2023;23(11):5030. doi: 10.3390/s23115030
Dong J, Ota K, Dong M. Video Frame Interpolation: A Comprehensive Survey. ACM Trans Multimedia Comput Commun Appl. 2023;19(2s):1-31. doi: 10.1145/3556544
Parihar AS, Varshney D, Pandya K, Aggarwal A. A comprehensive survey on video frame interpolation techniques. Vis Comput. 2022;38(1):295-319. doi: 10.1007/s00371-020-02016-y
Kye D, Roh C, Ko S, Eom C, Oh J. AceVFI: A Comprehensive Survey of Advances in Video Frame Interpolation. IEEE Trans Circuits Syst Video Technol. 2026;36(7):9477-9501. doi: 10.1109/tcsvt.2026.3672288
Huang X, Wang S, Xu T, Feng Z, Yang X. A systematic survey on video frame interpolation: Advances, challenges, and future directions. Expert Systems with Applications. 2026;304:130660. doi: 10.1016/j.eswa.2025.130660

https://doi.org/10.62486/978-9915-9966-3-9.Ch07