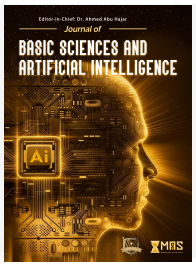
Journal of Basic Sciences and Artificial Intelligence (JBSAI)

DOI: 10.65098/jbsai.01.2026.24.36



RESEARCH ARTICLE

Visual Servoing for Robotic Manipulators: Traditional, Machine-Learning and Deep-Learning Methods, a Literature Review

Esmaeel Hilal, Abdulkader Joukhadar*

Integrated Mechatronics Laboratory, Dept. of Mechatronics Eng., University of Aleppo, Aleppo 12212, Syria
Corresponding Author Email: ajoukhadar@alepuniv.edu.sy

ARTICLE DETAILS

Article History:

Received: 02 Jul 2026
Accepted: 28 Aug 2026
Available online: 31 Aug 2026

Online Article Code

ABSTRACT

Visual servoing is the closed-loop control of robot motion directly from camera measurements. This review examines how the problem has been solved for robotic manipulators along three successive methodological lines: traditional analytic control, in which the mapping from image features to robot velocity is derived from geometry; classical machine learning, in which that mapping is estimated from motion data rather than modelled; and deep learning, in which perception, and eventually the control law itself, is replaced by trained neural networks. Twenty widely cited works are analysed in depth, namely the interaction-matrix formalism and the image-based, position-based and hybrid schemes built upon it; online Jacobian estimation and neural reinforcement learning; and deep relative-pose regression, simulation-to-reality visuomotor policies, diffusion policies and vision-language-action models; and, in the most recent period, visual servoing built on pretrained vision-transformer features and on vision foundation models. The three lines are then compared on a common set of axes: accuracy, convergence domain, model and data requirements, computational cost and verifiability. The review concludes that the eras are complementary rather than competing: analytic control still provides the only verifiable inner loop and the best terminal precision, learning provides the perception and semantic breadth that analytic methods never had, and the principal open problem is the disciplined combination of the two. The twenty works were selected purposively rather than by exhaustive database search. Only peer-reviewed work directly concerned with visual servoing for robotic manipulators was considered. A work was then used if it either established the common framework of the field or introduced a distinct methodological step, rather than an incremental refinement of an existing one; where several papers made the same step, the one most commonly cited as its standard reference was chosen. The resulting set spans 1992 to 2026 and is intended to show how the methods of the field developed, not to cover everything published in it.

KEYWORDS

Visual Servoing, Robotic Manipulators, IBVS, PBVS, Deep Learning, Literature Review

1. INTRODUCTION

A robotic manipulator that must act on objects whose position is not known in advance needs a way of converting what it sees into how it moves. Visual servoing is the family of techniques that performs this conversion inside a feedback loop: rather than measuring the scene once and then executing a planned trajectory, the controller continuously compares the current camera image with a desired one and drives the difference to zero (Chaumette & Hutchinson, 2006; Hutchinson & Hager, 1996). Closing the loop around vision is what gives the approach its practical value. A look-then-move system measures the target pose once and then moves, so every calibration error, kinematic inaccuracy, and object displacement ends up as task error; a visual servo measures the residual error at every cycle and can therefore correct for disturbances that would defeat an open-loop system entirely (Hutchinson & Hager, 1996).

The methodological history of the field falls into three phases, and this review is organised around them.

The first phase is *traditional*, or analytic. Its central object is the interaction matrix, which relates the time derivative of a chosen set of image features to the spatial velocity of the camera and is derived from projective geometry (Espiau et al., 1992). Everything follows from that derivation: the canonical proportional control law, the distinction between image-based and position-based schemes, and the stability analysis that characterises when each converges (Chaumette & Hutchinson, 2006; Hutchinson & Hager, 1996; Wilson & Hulls, 1996). The maturity of this phase is evident in the fact that its known weaknesses were established analytically, among them image singularities, local minima and unpredictable Cartesian paths, and were addressed by equally analytic remedies such as partitioned control (Corke, 2001) and hybrid 2–1/2-D servoing (Malis et al., 1999).

The second phase is *machine learning* in its pre-deep sense: the recognition that the visual-motor mapping need not be derived at all, but can be estimated from observed motion. Hosoda and Asada showed that a Jacobian estimated online by recursive least squares supports convergent servoing with no prior knowledge of camera or link parameters (Hosoda & Asada, 1994), and Lampe and Riedmiller showed that the control policy itself can be acquired by reinforcement learning from success and failure signals alone (Lampe & Riedmiller, 2013). Both replace a model with data, but neither requires the representational capacity that would arrive with deep networks.

The third phase is *deep learning*. It began by substituting convolutional networks for the fragile feature-extraction and pose-estimation stages of the classical pipeline while retaining a classical control law (Bateux et al., 2018; Lee & Levine, 2017), and it proceeded to dissolve the perception-control boundary altogether: policies trained end-to-end from

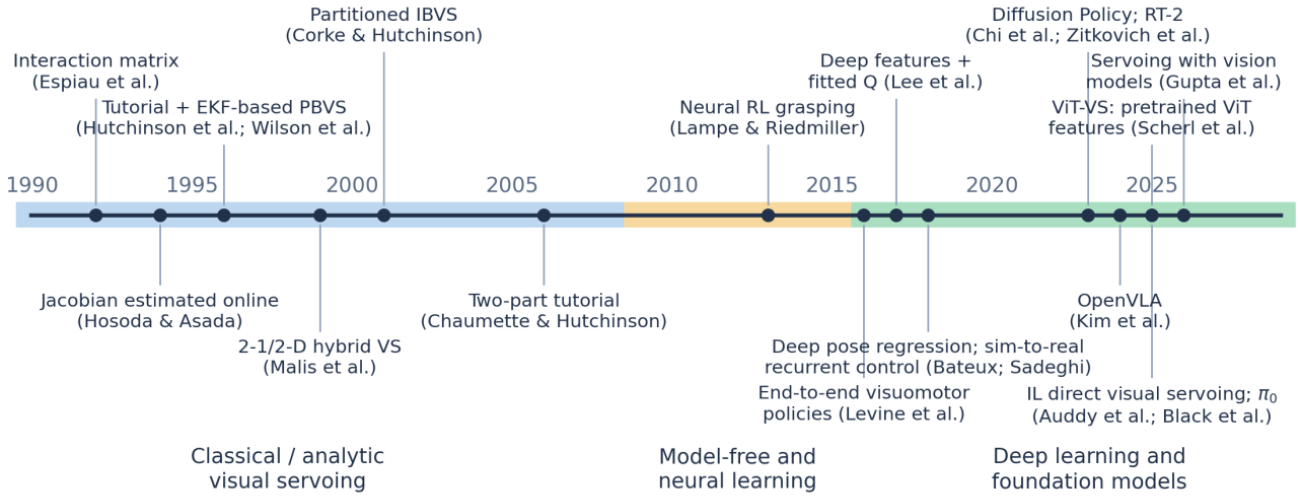


Figure 1 Timeline of Reviewed Literature Grouped by Visual Servoing Methodological Phases

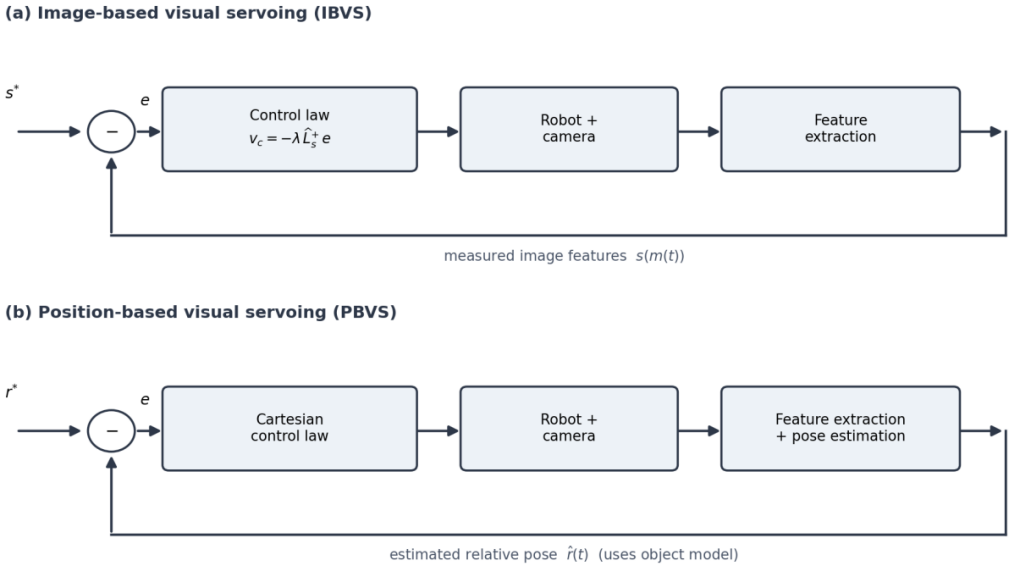


Figure 2 Closed-Loop Architectures of IBVS and PBVS (Adapted from Chaumette & Hutchinson, 2006)

pixels to torques (Levine et al., 2016), recurrent controllers transferred from simulation (Sadeghi et al., 2018), generative policies that model whole action sequences (Chi et al., 2023), and vision-language-action models that accept a natural-language instruction as the task specification (Zitkovich et al., 2023).

The review has three aims. First, to give an accurate and self-contained account of each phase, grounded in the results the primary sources actually report. Second, to make the phases comparable by analysing them along the same axes of accuracy, convergence domain, model and data requirements, computational cost and verifiability. Third, to identify what remains unsolved, which is less a matter of any single method being inadequate than of the field lacking a principled way to combine the certifiability of analytic control with the generality of learned perception. A final section brings the review up to date with work published between 2024 and 2026. Figure 1 places the twenty works reviewed here on a common timeline.

1.1 Review Methodology

Because the account that follows is a narrative review rather than a systematic one, the procedure behind it is set out here so that its coverage can be judged. The scope was fixed first. The subject is visual servoing for robotic manipulators, that is, closed-loop control of manipulator motion from camera measurements; visual navigation for mobile robots and aerial vehicles was therefore excluded except where a method was demonstrated on a manipulator. Candidates were identified from the tutorial and survey literature (Chaumette & Hutchinson, 2006; Chaumette & Hutchinson, 2007; Hutchinson & Hager, 1996) and the primary works it cites, which covers the analytic period, and from searches of IEEE Xplore and arXiv for the learning-based and recent periods. Searching stopped when further queries ceased to return methodologically distinct work.

Each candidate was then assessed against the criteria stated in the abstract: publication in a peer-reviewed venue, a subject directly concerned with visual servoing for manipulators, and a contribution that either established the common framework of the field or introduced a distinct methodological step rather than an incremental refinement of an existing one. Where several papers made the same step, the one most commonly cited as its standard reference was chosen. For the recent additions, candidates were rejected for three recurring reasons: validation in simulation only, no peer-reviewed venue, or refinement of an existing method rather than a new one. The outcome is the sample of twenty works that this review analyses; the benchmarks and datasets cited in Section 9, together with the further works cited in support of individual points, are supporting references rather than reviewed methods. The review does not claim to have surveyed the literature exhaustively; the intention is instead that the methodological trajectory be represented faithfully.

2. FOUNDATIONS

2.1 Problem Formulation

Every visual servoing scheme regulates to zero an error defined between a vector of visual features measured in the current image and the value

that vector takes in the desired configuration (Chaumette & Hutchinson, 2006). Let $\mathbf{s}(\mathbf{m}(t), \mathbf{a}) \in \mathbb{R}^k$ denote the feature vector computed from image measurements $\mathbf{m}(t)$, such as pixel coordinates, image moments, or, in the learning-based methods of Section 5, the activations of a trained network, and using whatever prior knowledge $\mathbf{a}$ is available, such as camera intrinsics or an object model. The task is encoded by

$$\mathbf{e}(t) = \mathbf{s}(\mathbf{m}(t), \mathbf{a}) - \mathbf{s}^* \tag{1}$$

For a camera moving with spatial velocity $\mathbf{v}_c \in \mathbb{R}^6$, the feature velocity is related to the camera velocity by the *interaction matrix* $\mathbf{L}_\mathbf{s} \in \mathbb{R}^{k\times 6}$ through $\dot{\mathbf{s}} = \mathbf{L}_\mathbf{s}\mathbf{v}_c$ (Espiau et al., 1992). Requiring the error to decrease exponentially, $\dot{\mathbf{e}} = -\lambda\mathbf{e}$, yields the control law that underlies the entire classical literature,

$$\mathbf{v}_c = -\lambda\, \hat{\mathbf{L}}_\mathbf{s}^{+}\, \mathbf{e} \tag{2}$$

in which $\hat{\mathbf{L}}_\mathbf{s}^{+}$ is the Moore-Penrose pseudo-inverse of an *estimate* of the interaction matrix (Chaumette & Hutchinson, 2006); the resulting camera velocity is mapped to joint velocities through the robot Jacobian and the hand-eye transformation. For a single image point with normalised coordinates (x, y) at depth Z, the interaction matrix is

$$\mathbf{L}_x = \begin{bmatrix} -1/Z & 0 & x/Z & xy & -(1+x^2) & y \\ 0 & -1/Z & y/Z & 1+y^2 & -xy & -x \end{bmatrix} \tag{3}$$

which exposes the two structural difficulties that recur throughout this review (Chaumette & Hutchinson, 2006). The matrix depends on the depth Z, which a single camera does not measure, so the control law is always executed with an approximation $\hat{\mathbf{L}}_\mathbf{s} \neq \mathbf{L}_\mathbf{s}$; and it couples translational and rotational velocity nonlinearly, so the image trajectory and the Cartesian trajectory cannot both be shaped by a single proportional gain.

Read as a template, Eqs. (1)–(2) organise the whole field: a method is a choice of what $\mathbf{s}$ is, how $\hat{\mathbf{L}}_\mathbf{s}$ is obtained, and what replaces the proportional regulator. Traditional methods choose geometric features and derive $\hat{\mathbf{L}}_\mathbf{s}$ in closed form (Section 3); machine-learning methods keep the structure but estimate $\hat{\mathbf{L}}_\mathbf{s}$ from data (Section 4); deep-learning methods learn $\mathbf{s}$, and in the limit discard Eq. (2) altogether (Section 5).

2.2 Taxonomy and Camera Configuration

Two independent classifications are used throughout, both introduced in the foundational tutorial literature (Hutchinson & Hager, 1996).

The first concerns the space in which the error is defined. In image-based visual servoing (IBVS) the features are image quantities and no pose is reconstructed, which makes the loop largely insensitive to calibration error but leaves the Cartesian path uncontrolled. In position-based visual servoing (PBVS), the image measurements are first converted into an estimate of the relative camera-object pose, and the error is defined in Cartesian space; the trajectory is then predictable, but the accuracy of the result is bounded by the accuracy of the camera calibration and the object model (Chaumette & Hutchinson, 2006; Wilson & Hulls, 1996). *Hybrid* schemes partition the error between the two spaces (Malis et al., 1999). Figure 2 contrasts the IBVS and PBVS loop structures. The learning-based methods reviewed later inherit this taxonomy rather than escaping it: networks that regress a relative pose and feed a Cartesian control law are structurally PBVS (Bateux et al., 2018), methods that servo on learned image features are structurally IBVS (Lee & Levine, 2017), and end-to-end policies drop the distinction altogether (Levine et al., 2016; Zitkovich et al., 2023).

The second classification concerns where the loop is closed. In a *dynamic look-and-move* architecture, the vision system supplies setpoints to an existing joint-level controller running at a much higher rate, whereas a *direct visual servo* computes joint torques from the image alone (Hutchinson & Hager, 1996). Almost all systems, classical and learned, are of the first kind, which matters for Section 6: the vision loop inherits a stable inner loop and may treat the manipulator as an ideal velocity source. The design of that inner loop is a subject in its own right, adaptive computed-torque control being one of the standard approaches to joint-level manipulator control (Alhaddad et al., 2019).

A third choice applies to both classifications: where the camera is mounted. An eye-in-hand camera gives resolution that improves as the target is approached but a field of view that changes with every motion; a fixed eye-to-hand camera gives a stable view of the workspace at lower resolution and can be occluded by the manipulator (Hutchinson & Hager, 1996). The choice affects which methods apply: recursive pose estimation was first made practical for eye-to-hand PBVS (Wilson & Hulls, 1996), while learned policies that seek viewpoint invariance are trying not to commit to either arrangement at all (Sadeghi et al., 2018).

3. TRADITIONAL VISUAL SERVOING

3.1 The Analytic Framework

The modern form of the field dates from the task-function formulation of Espiau, Chaumette and Rives, who showed that a servoing task can be specified as a function of the sensor signal that must be regulated to zero, and derived the interaction matrix for generic geometric primitives rather than for points alone (Espiau et al., 1992). Two consequences made this the reference formulation. It is constructive: given a primitive, the interaction matrix follows from projective geometry, so a control law can be written for any feature set the designer chooses. And it is analysable: because Eq. (2) is derived rather than tuned, the closed loop can be examined with Lyapunov methods, and the consequences of using an approximate $\hat{\mathbf{L}}_\mathbf{s}$, which is inevitable since Z is unknown, can be bounded rather than merely observed (Chaumette & Hutchinson, 2006; Espiau et al., 1992).

Hutchinson, Hager and Corke consolidated the framework into the tutorial that defined the vocabulary of the field, setting out the taxonomy of Section 2.2 and surveying the feature tracking on which any implementation depends (Hutchinson & Hager, 1996). A decade later Chaumette and Hutchinson revisited the subject in two parts: the first treats the basic IBVS and PBVS schemes and their stability (Chaumette & Hutchinson, 2006), the second the advanced material the intervening decade produced, including feature selection to improve the conditioning of $\mathbf{L}_\mathbf{s}$, partitioned and hybrid schemes, and second-order minimisation and path planning to enlarge the domain of convergence (Chaumette & Hutchinson, 2007). These three references form the analytic backbone against which all later work is most usefully read.

3.2 Image-Based Visual Servoing

IBVS defines the error directly in the image (Figure 2 (a)), which is its principal strength and the source of its principal difficulty. Because no pose is reconstructed, errors in the camera intrinsics, the hand-eye transformation, and the depth estimate perturb the *direction* of the commanded velocity but do not displace the equilibrium: the loop converges to the configuration in which the features coincide with their desired values, whatever the calibration happens to be (Chaumette & Hutchinson, 2006). This robustness is why IBVS is the default choice when the camera is imperfectly calibrated.

The difficulty is that the equilibrium is only guaranteed to be *locally* asymptotically stable. Because six degrees of freedom are typically controlled with more than six feature coordinates, $\hat{\mathbf{L}}_\mathbf{s}^{+}$ has a non-trivial null space, and configurations exist for which the commanded velocity lies in that null space, producing a local minimum with non-zero pose error; configurations also exist at which $\mathbf{L}_\mathbf{s}$ loses rank (Chaumette & Hutchinson, 2006). The most cited concrete failure is the *camera retreat* induced by a large rotation about the optical axis: the image-plane error is minimised by a trajectory that translates the camera backwards, in the limit to infinity for a rotation of 180°, because the coupling terms in Eq. (3) tie optical-axis rotation to optical-axis translation (Corke, 2001).

Corke and Hutchinson answered this structurally rather than heuristically. Their partitioned scheme removes the rotation and translation about the optical axis from the interaction matrix and controls them with dedicated features, leaving a reduced, better-behaved subsystem for the remaining four degrees of freedom; a repulsive potential function additionally keeps feature points away from the image boundary, so the field-of-view constraint is enforced by the control law rather than hoped for (Corke,

2001). This is how the traditional phase worked in general: analysis identifies a failure, traces it to a particular term in a derived matrix, and removes it by redesigning that term.

3.3 Position-Based Visual Servoing

PBVS moves the burden from control to perception (Figure 2 (b)). Image measurements are converted into an estimate of the relative pose, and the error is formed between the current and desired poses. Because the error is Cartesian, translation and rotation decouple, the end-effector follows a straight-line path, and the domain of convergence in pose space is essentially global (Chaumette & Hutchinson, 2006). The liabilities are the mirror image of the IBVS case: calibration and object-model errors now bias the estimated pose and therefore appear directly as steady-state task error, and because nothing in the formulation refers to the image, no mechanism prevents the target from leaving the field of view during the motion.

The practical obstacle to PBVS is that pose recovered from a single view is noisy. Wilson, Williams, Hulls, and Bell established the architecture that made it usable, casting relative end-effector control as a recursive state-estimation problem in which an extended Kalman filter integrates successive image measurements with a motion model to produce a filtered pose estimate for the Cartesian controller (Wilson & Hulls, 1996). This is a design pattern rather than a single algorithm, and it is worth noting for the comparison in Section 6: the estimator is explicitly separate from the controller, it carries a covariance that quantifies its own uncertainty, and it can be monitored independently. The same recursive-filtering pattern has since been carried into other uses of visual measurement, for instance an unscented Kalman filter applied to image filtering and three-dimensional reconstruction (Joukhadar et al., 2020). Its structural weakness is the dependence on a known object model, which confines classical PBVS to environments where the geometry of every manipulated object is available in advance.

3.4 Hybrid Visual Servoing

Since IBVS and PBVS fail in complementary ways, the natural response is to use each where it is strong. The 2–1/2-D scheme of Malis, Chaumette and Boudet estimates the partial camera displacement between the current and desired views at each iteration, obtains the rotation directly from it, and regulates translation through extended image coordinates: the image coordinates of a reference point on the target augmented by a supplementary normalised coordinate encoding depth (Malis et al., 1999). Three properties follow. The resulting interaction matrix is upper-triangular, decoupling the two subsystems, and has no singularity anywhere in the task space, so the positioning task converges from any initial camera position when the intrinsics are exact. No three-dimensional model of the object is required, since the displacement is recovered from image correspondences alone. And the robustness analysis is complete: necessary and sufficient conditions for local asymptotic stability under calibration error are derived, and the simple structure of the system also yields sufficient conditions for global asymptotic stability (Malis et al., 1999). Visibility, however, is not free. The authors note explicitly that positiveness of the stability matrix does not by itself guarantee that the reference point stays visible, and they add an adaptive control law in a later section to ensure that the target remains in the field of view (Malis et al., 1999). The residual cost is the sensitivity of the displacement estimation to image noise. Figure 3 shows the structure.

3.5 What the Traditional Phase Established, and What It Did Not

Table 1 summarises the three families. They are not steps in a progression; each trades one strength for another. IBVS maximises tolerance of modelling error and gives up control of the Cartesian path; PBVS inverts that exchange; hybrid servoing recovers much of both at the price of a noise-sensitive geometric estimation step. What all three share is a property no learned method has yet matched: the relationship between feature error and commanded velocity is written down explicitly, so stability can be argued rather than measured.

They also share three limitations, which define the agenda of the remaining sections. All of them presuppose that a designated feature set can be extracted and tracked reliably, and in practice tracking failure, not control design, is the dominant cause of failure (Chaumette & Hutchinson, 2007; Hutchinson & Hager, 1996). The task must be specified by showing the robot the desired image or pose, so nothing generalises to a new object or instruction. And the models on which PBVS and, through depth, IBVS depend must be supplied by the engineer. Estimating those models from data is the subject of Section 4; replacing the features with learned representations is the subject of Section 5.

4. MACHINE LEARNING APPROACHES

Between the analytic era and the deep-learning era lies a body of work whose premise is that the mappings appearing in Eq. (2) need not be derived at all. They can be estimated: either the interaction matrix, from the motion the robot itself produces, or the control policy, from the outcomes the robot achieves. The same idea has been applied to the manipulator itself, where inverse kinematics is solved by constrained optimisation and the solutions are intended to train a faster learned approximation (Laalou et al., 2024). Neither approach requires a large network, and both predate deep learning by two decades; together they establish the two learning targets, perception and policy, that Section 5 inherits.

4.1 Estimating the Visual-Motor Jacobian Online

The analytic construction of $\hat{\mathbf{L}}_{\mathbf{s}}$ requires the camera intrinsics, the hand-eye transformation, and the feature depths. Hosoda and Asada observed that the composite map from joint velocity to feature velocity, known as the visual-motor or image Jacobian, is simply a matrix that can be identified from the data the servo loop already generates (Hosoda & Asada, 1994). At every step the robot moves, the resulting feature displacement is observed, and the pair constrains the Jacobian; a weighted recursive least-squares update with a forgetting factor maintains the estimate, the forgetting factor being necessary because the true Jacobian is configuration-dependent and therefore drifts as the manipulator moves. Two properties of this contribution deserve emphasis. It requires no prior knowledge of the kinematic structure or of the camera and link parameters. And convergence of the image features to their desired trajectories is established by Lyapunov analysis, so the estimated model does not cost the scheme its stability argument (Hosoda & Asada, 1994).

The method trades a global model for a local one. The estimate is valid only near the current configuration and must be continuously refreshed; it degrades if the excitation is poor, and, because least squares weights all residuals quadratically, a single mis-tracked feature can corrupt it. Even so, this is the earliest clear statement of the idea that dominates everything that follows, namely that a mapping which is hard to model can be learned from interaction, and, uniquely in this review, it is learned without giving up the stability proof.

4.2 Learning the Policy: Neural Reinforcement Learning

The complementary target is the control law. Lampe and Riedmiller trained a manipulator to reach and grasp using neural reinforcement learning, with control realised in a visual feedback loop (Lampe & Riedmiller, 2013). The distinguishing feature of the work is how little is supplied: the controller is learned from scratch from success or failure signals, with no information about the solution of the task built in, and every major component of the system, from perception to value function to policy, is implemented by a neural network. Because the learned behaviour is executed as a closed visual loop rather than as a planned trajectory, it retains the reactivity and noise tolerance that motivate visual servoing in the first place (Lampe & Riedmiller, 2013).

This work also shows, on a small scale, the properties that characterise the entire reinforcement-learning branch of the field. Reward is far weaker supervision than a desired image or pose, so substantially more interaction is needed; the resulting controller is a network rather than an equation, so the stability argument available to (Hosoda & Asada, 1994) is not available here; and the behaviour is specific to the task on which it was trained. The subsequent history of learned servoing is largely the history of attacking the first two of these, sample efficiency and transfer, while the third, and the loss of verifiability, remain open.

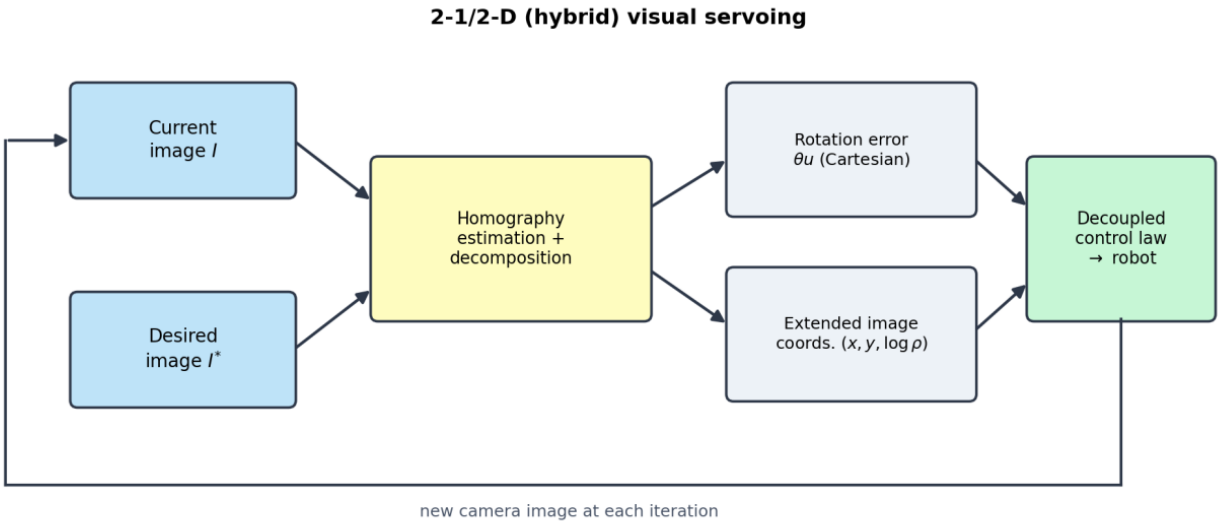


Figure 3 Architecture of 2-1/2-D Hybrid Visual Servoing (Malis et al., 1999)

Table 1 Comparison of Three Traditional Visual Servoing Approaches

Comparison item	IBVS	PBVS	Hybrid (2–1/2-D)
Error space	Image plane	Cartesian pose	Rotation in Cartesian space, translation in extended image coordinates
Model required	Intrinsics; depth of each feature	Intrinsics, hand-eye and 3-D object model	Intrinsics only; no object model (Malis et al., 1999)
Convergence	Local; image singularities and local minima (Chaumette & Hutchinson, 2006)	Essentially global in pose space, but the target may leave the field of view	No singularity in the task space; convergence from any initial pose under exact intrinsics (Malis et al., 1999)
Calibration error	Perturbs the path, not the equilibrium	Biases the estimated pose, hence the task error	Necessary and sufficient local stability conditions derived (Malis et al., 1999)
Main weakness	Unpredictable Cartesian trajectory; camera retreat (Corke, 2001)	Dependence on an accurate object model and calibration	Visibility needs an added adaptive law; estimation is noise-sensitive (Malis et al., 1999)

Taken together, the two contributions bracket the classical pipeline from both ends. Hosoda and Asada replace the *model* inside an otherwise classical loop, keeping the control law, the error definition, and the stability analysis (Hosoda & Asada, 1994); Lampe and Riedmiller replace the *law*, keeping only the loop (Lampe & Riedmiller, 2013). Figure 4 shows the two routes, which the deep-learning literature follows almost without exception, with far greater representational capacity, and correspondingly greater data requirements.

5. DEEP LEARNING APPROACHES

5.1 Deep Relative-Pose Regression: Learning Inside the Classical Loop

The most conservative use of deep learning keeps Eq. (2) and replaces only what precedes it. Bateux et al. train a convolutional network to estimate the relative pose between the current and the desired image and feed that estimate to a position-based control law, so the network occupies exactly the position of the feature-extraction and pose-estimation blocks of Figure 2 (b) (Bateux et al., 2018). The substantive contribution is the training data. Rather than collecting a labelled dataset, the authors generate one from a *single* real image of the scene, synthesising large numbers of views with known relative pose and applying simulated perturbations, occlusions and lighting changes, so that the network learns invariance to precisely the disturbances that defeat classical feature tracking. The reported result is that the method converges robustly under strong lighting variation and occlusion, and achieves a positioning error below one millimetre on a six-degree-of-freedom robot (Bateux et al., 2018). Figure 5 shows the architecture.

The design has an engineering property worth stating explicitly: the interface between perception and control survives. The network emits a pose, a quantity with physical units that can be bounded, logged, and sanity-checked, and the controller downstream of it is the same analysable proportional law as in Section 3.3. The classical failure mode, loss of tracking, is replaced by a distributional one in which the network returns a confident but wrong pose for a scene unlike anything it was trained on. Both are failures, but only the first announces itself.

A second variant learns the *features* rather than the pose. Lee, Levine, and Abbeel servo in the space of deep features taken from a network trained for object classification, pairing them with a bilinear predictive model of how those features respond to camera motion, the learned counterpart of the interaction matrix, and using fitted Q-iteration to determine which feature channels are informative for the task (Lee & Levine, 2017). Weighting the channels matters: servoing on raw pixels or on unweighted deep features performs substantially worse. The sample efficiency is the headline result. An effective servo is learned on a synthetic target-following benchmark from twenty training trajectories, an improvement of more than two orders of magnitude over standard model-free deep reinforcement learning, and the resulting controller is robust to changes in viewing angle, appearance, and partial occlusion (Lee & Levine, 2017). The work is a bridge in both directions: it is structurally IBVS, with a learned feature vector and a learned predictive model, and it is reinforcement learning, but of the batch, sample-efficient kind rather than the online kind.

5.2 End-to-End and Simulation-to-Reality Policies

The alternative is to discard the decomposition. Levine et al. pose the question directly: whether training perception and control jointly outperforms training them separately, and answer it with policies that map raw images to motor torques through convolutional networks of about 92 000 parameters (Levine et al., 2016). Training uses guided policy search, which converts policy search into supervised learning: a trajectory-centric reinforcement-learning procedure produces successful trajectories and the network is trained to reproduce them. The tasks are real manipulation tasks requiring tight coordination between vision and control, such as screwing a cap onto a bottle (Levine et al., 2016). The contrast with the classical pipeline is the point: no features, no interaction matrix, no pose, no separately specified low-level controller.

The obstacle to end-to-end policies is data, and the standard response is to move training into simulation. Sadeghi et al. address the harder version of the problem, in which the camera viewpoint is not fixed and the policy must

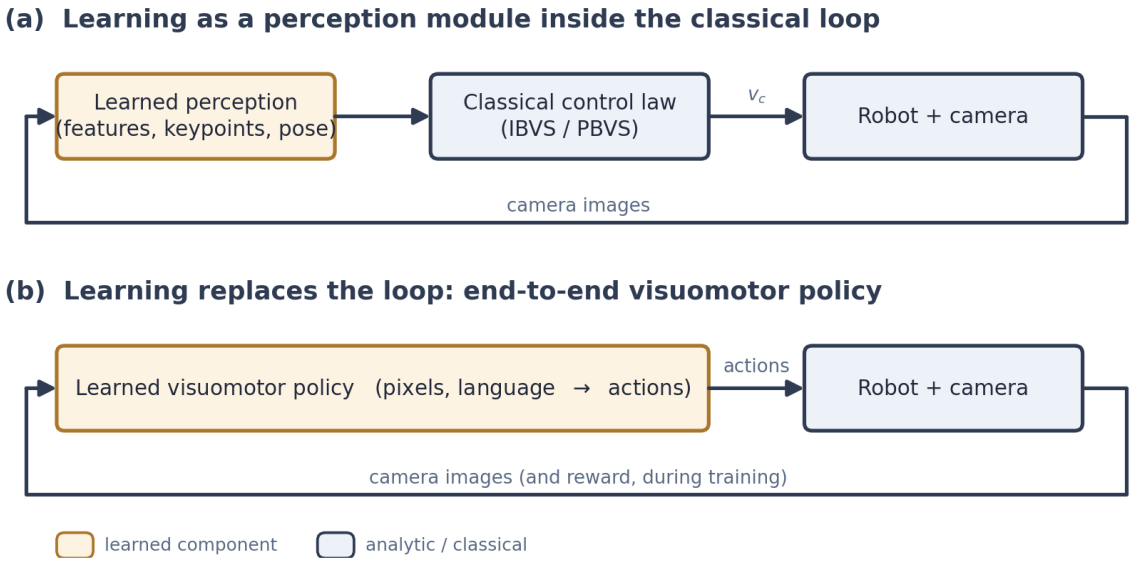


Figure 4 Two Technical Paths for Integrating Learning into Visual Servoing (Amber: Learned Components; Blue-Grey: Analytic Components, This Color Convention Applies to Subsequent Figures) (Hosoda & Asada, 1994; Bateux et al., 2018; Lampe & Riedmiller, 2013; Levine et al., 2016)

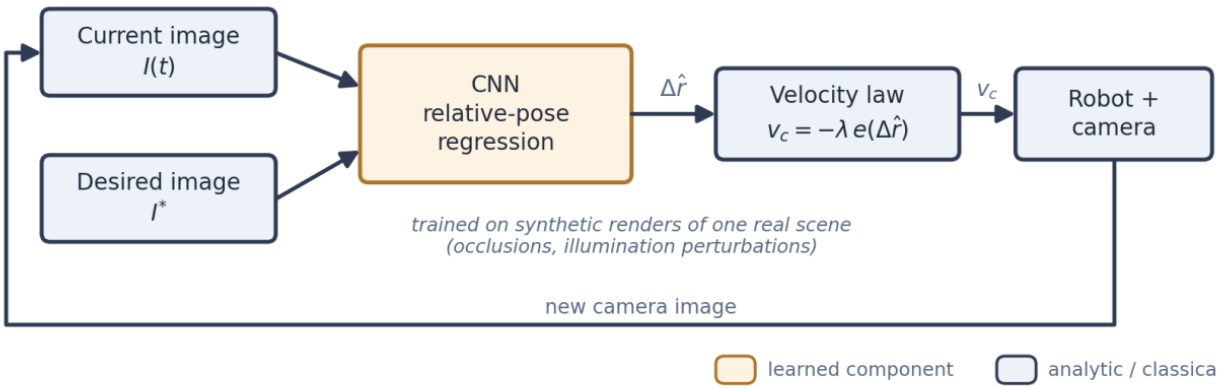


Figure 5 Architecture of Deep Relative-Pose Regression Within Servo Loop (Bateux et al., 2018)

therefore infer the effect of its own actions from experience gathered during the current trial; a purely reactive controller cannot resolve this ambiguity, so the controller is recurrent and uses its memory of past motions to interpret the current view (Sadeghi et al., 2018). The policy is trained in simulation from synthetic demonstrations combined with reinforcement learning, and transferred by disentangling perception from control and adapting only the visual layers on real images while the control layers are kept, after which the system servos to previously unseen objects from novel viewpoints on a real Kuka IIWA arm (Sadeghi et al., 2018). The modular transfer strategy is notable in the context of this review: even in an end-to-end policy, reintroducing a boundary between perception and control is what makes the system transferable. Figure 6 summarises the loop and the transfer strategy.

5.3 Generative Policies and Vision-Language-Action Models

The most recent phase changes what a policy *is*. Diffusion Policy represents the visuomotor policy as a conditional denoising diffusion process over action sequences: instead of regressing one action, the network learns the score function of the action distribution and produces a sequence by iterative refinement, executed in receding-horizon fashion (Chi et al., 2023). The motivation is that demonstration data are multimodal, since there are several acceptable ways to perform a task, and a regression policy averages them into something that may satisfy no mode at all, whereas a generative policy represents the modes explicitly. Across twelve tasks drawn from four robot manipulation benchmarks, the method improves on prior state-of-the-art robot learning methods by an average of 46.9%, and the authors identify receding-horizon control, visual conditioning and a time-series diffusion transformer as the components required to make the formulation work on physical robots (Chi et al., 2023).

The final step generalises the task specification itself. RT-2 co-finetunes a pretrained vision-language model on robot trajectories together with Internet-scale vision-language data, expressing robot actions as text tokens so that actions and language occupy a single output vocabulary and the model requires no architectural change (Zitkovich et al., 2023). The reported evidence rests on more than six thousand robotic trials. Performance on tasks seen in the robot data is retained, while performance on previously unseen objects, backgrounds and environments rises from 32% for the RT-1 baseline to 62%; on categories designed to require web knowledge, namely symbol understanding, reasoning about object relations and human recognition, generalisation improves by more than a factor of three over baselines pretrained on robotic or visual data alone; and chain-of-thought variants can plan a short sequence of sub-goals before emitting actions (Zitkovich et al., 2023). Figure 7 shows both architectures.

In the terms of Section 2.1, these systems are the limiting case of the template. There is no feature vector, no interaction matrix, and no error signal; the desired configuration is not shown to the robot but described to it, and the map from images and language to actions is learned in its entirety. What is gained is semantic generality that no analytic method could express. What is given up is every quantity the classical analysis operated on, and with it the ability to state, before deployment, what the system will do.

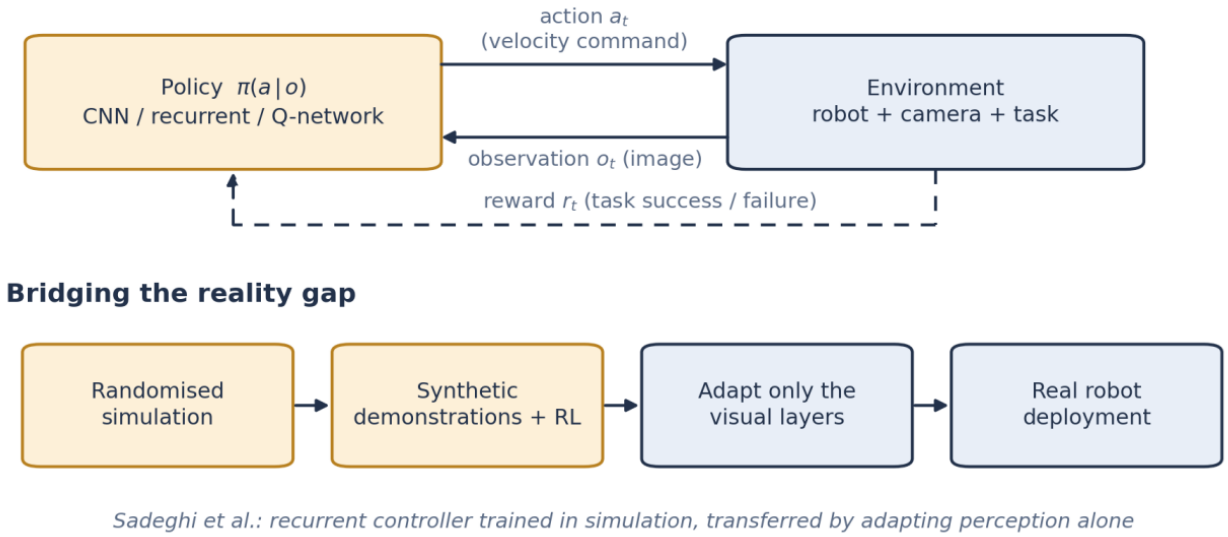


Figure 6 Reinforcement-Learning Framework for Visual Servoing and Sim-To-Real Transfer Strategy (Lampe & Riedmiller, 2013; Lee & Levine, 2017; Levine et al., 2016; Sadeghi et al., 2018)

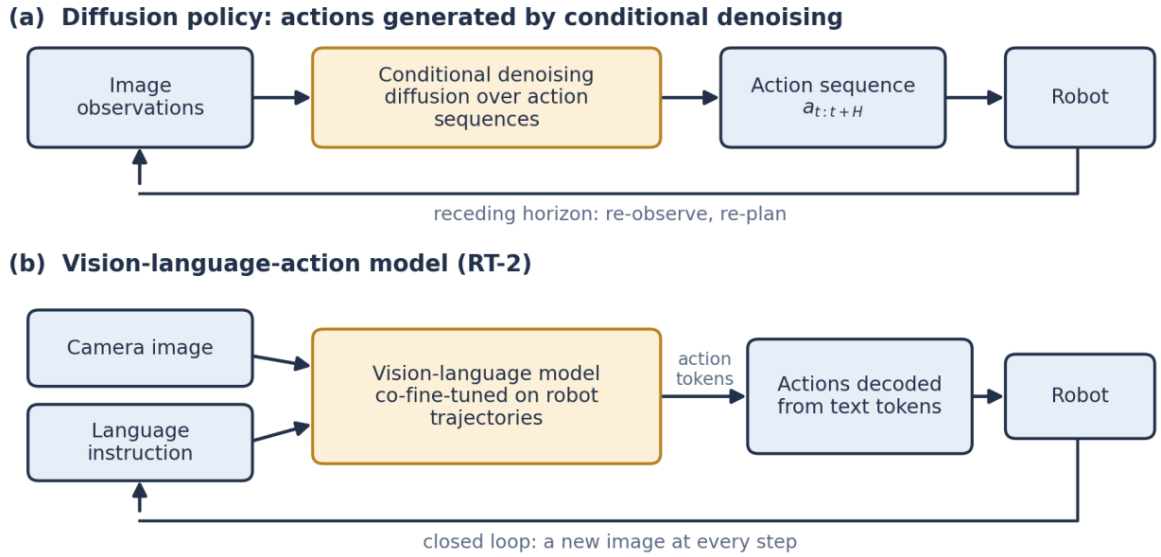


Figure 7 Generative Visuomotor Policies: Diffusion Policy and Vision-Language-Action Model RT-2 (Chi et al., 2023; Zitkovich et al., 2023)

6. RECENT DEVELOPMENTS (2024–2026)

The five works discussed in this section postdate the material of Sections 3 to 5. They do not constitute a fourth methodological phase. They are better read as the clearest attempt so far to answer the question the earlier phases left open, namely how the guarantees that analytic control provides can be combined with the generality of learned perception. Two of them retain the servo loop and replace only the representation on which it operates (Gupta et al., 2026; Scherl et al., 2025); one inserts learned components into an analytically defined task-priority structure (Auddy et al., 2025); two continue the generalist-policy line and change what a learned policy is made of (Black et al., 2025; Kim et al., 2024).

6.1 Foundation-Model Features Inside the Classical Loop

The most direct development is the use of features taken from large pretrained vision models as the feature vector in Eq. (1), leaving the control law of Eq. (2) untouched. Scherl et al. extract semantic features with a pretrained vision transformer and use them in an otherwise classical image-based scheme (Scherl et al., 2025). What makes the result significant is the combination of properties it achieves. The method requires no task-specific or object-specific training, which is a property classical schemes have and learned schemes lack, and it nevertheless reports full convergence in unperturbed scenarios and up to a 31.2% relative improvement over classical IBVS in perturbed ones, matching the convergence rates of learning-based methods (Scherl et al., 2025). Real-world evaluation covers end-effector positioning, industrial box manipulation, and the grasping of unseen objects using only a reference from the same category (Scherl et al., 2025).

Gupta et al. apply the same idea on the perception side of the loop, in their case for a mobile manipulator (Gupta et al., 2026). Rather than learning the features, they use vision foundation models to generate three-dimensional targets that a closed servo loop then pursues, so that open-vocabulary object detectors supply semantic targets such as knobs and point-tracking methods supply interaction sites indicated by a user click. The obstacle they identify is specific to a camera that approaches what it is looking at: the end-effector occludes the target. Their solution is to use a vision model to remove the end-effector from the image and fill in what it was hiding, which they report as making the target much easier to locate (Gupta et al., 2026). Evaluated across 10 novel environments in 6 buildings and 72 object instances, the system reaches a 71% zero-shot success rate on unseen objects in the real world, exceeding an open-loop control method by 42 percentage points and an imitation-learning baseline trained on more than 1000 demonstrations by 50 (Gupta et al., 2026).

Both results relate to the difference between analytic and learned robustness discussed later. The robustness being exploited is distributional, since it derives from the pretraining distribution of the vision model, but it is inserted into a loop whose convergence behaviour remains that of the

analytic scheme. Neither work offers a stability argument for the combined system, which is precisely the gap this review identifies.

6.2 Learned Components Inside an Analytic Priority Structure

Auddy et al. (2025) proceed from the opposite direction. Their starting point is dynamical-system-based imitation learning, which allows a manipulator to perform complex tasks stably and without explicit programming, and their concern is that such systems must be closed around visual feedback if they are to cope with environmental change. They close the loop directly on the image, using off-the-shelf deep-learning-based perception modules to extract robust features from the raw input and an imitation-learning strategy to execute the motion, and they integrate the two learned blocks using the large projection task-priority formulation (Auddy et al., 2025). The method is validated experimentally on a robotic manipulator.

What matters here is the structure, not the numbers. The large projection formulation is an analytic device that lets a secondary objective run without disturbing a primary one, so using it to join the learned parts holds them inside a control structure whose properties are known independently of what was learned. That is a more disciplined combination than swapping a network into a pipeline stage, and the closest any work reviewed here comes to the synthesis argued for in the conclusion.

6.3 Generalist Policies After RT-2

The generalist-policy line of Section 5.3 continued along two axes: accessibility and action representation. Kim et al. address the first with OpenVLA, a 7-billion-parameter open-source vision-language-action model trained on 970,000 real-world robot demonstrations and built on a Llama 2 language model with a visual encoder that fuses pretrained DINOv2 and SigLIP features (Kim et al., 2024). Its reported results complicate the assumption that scale alone determines capability: OpenVLA exceeds the closed 55-billion-parameter RT-2-X by 16.5 percentage points in absolute task success rate across 29 tasks and several robot embodiments while using seven times fewer parameters, and outperforms from-scratch imitation-learning methods such as Diffusion Policy (Chi et al., 2023) by 20.4 points (Kim et al., 2024). It can be fine-tuned on consumer GPUs through low-rank adaptation and served under quantisation without a loss in success rate, and its checkpoints, fine-tuning notebooks, and training code were released (Kim et al., 2024).

Black et al. address the second axis (Black et al., 2025). Their model places a flow matching architecture on top of a pretrained vision-language model, so that the policy inherits Internet-scale semantic knowledge while producing actions through a continuous generative process. It is trained on a large and diverse dataset drawn from several kinds of robot platform, including single-arm robots, dual-arm robots and mobile manipulators, and is evaluated on zero-shot performance after pretraining, on following language instructions issued both by people and by a high-level vision-language policy, and on acquiring new skills by fine-tuning, across tasks such as laundry folding, table cleaning and box assembly (Black et al., 2025). In the terms of this review, it continues the generative line opened by Diffusion Policy (Chi et al., 2023) and the semantic line opened by RT-2 (Zitkovich et al., 2023), now inside a single model.

Neither generalist model is evaluated in the vocabulary used for the analytic schemes. The quantities reported are task success rates over instruction sets rather than terminal pose error, convergence domain or settling behaviour, and neither is accompanied by a stability argument. The trend identified in Section 5 therefore continues uncorrected: as policies become more general, the guarantees available for them weaken, and the measurement vocabulary itself diverges, so that the two families grow harder to compare rather than easier.

6.4 What the Recent Period Changes

Taken together, the five works split along the line this review has drawn: three keep an explicit control structure and use learning inside it (Auddy et al., 2025; Gupta et al., 2026; Scherl et al., 2025), while two dissolve that structure into a single trained policy (Black et al., 2025; Kim et al., 2024). The first group is where this review's central claim gains direct evidence. Reference (Scherl et al., 2025) shows that a classical scheme with pretrained features can beat its classical ancestor and match learned competitors' convergence rates without task-specific training; (Auddy et al., 2025) shows that learned blocks can sit inside an analytic priority formulation rather than merely replace one of its stages. The second group shows that generality keeps improving and, with (Kim et al., 2024), that it is now reproducible outside the few laboratories able to train such models.

What has not appeared in this period is a stability or convergence result for any of the combined systems. The open problem set out in the following discussion is therefore unchanged by the recent literature. What has changed is that the components required to attack it, a servo loop with known properties and a perception front end that generalises without retraining, now exist in the same systems and, in the case of (Scherl et al., 2025) and (Kim et al., 2024), are publicly available.

7. COMPARATIVE DISCUSSION

7.1 Comparison on Common Axes

Table 2 places the three phases side by side. Read across the rows, it shows that no phase dominates. A fourth column summarises the recent period of Section 6, which is shown separately because it works to recombine the three phases rather than forming a fourth.

Accuracy. The most precise results come from opposite ends of the spectrum. Analytic IBVS converges to the configuration at which the image error vanishes, so it is accurate in the image whatever the calibration (Chaumette & Hutchinson, 2006), while deep pose regression reaches sub-millimetre positioning on a 6-DOF arm (Bateux et al., 2018). Generalist policies report success rates instead: 62% on unseen scenarios for RT-2 (Zitkovich et al., 2023), and a 46.9% average improvement over prior methods for Diffusion Policy (Chi et al., 2023). The two literatures therefore cannot be compared directly, and no system reports both breadth and millimetre precision. The recent period narrows the gap from the analytic side: pretrained vision-transformer features give full convergence when undisturbed and a 31.2% relative gain over classical IBVS when disturbed, with no task-specific training (Scherl et al., 2025), while (Gupta et al., 2026; Kim et al., 2024) and (Black et al., 2025) are still reported as success rates.

Convergence domain. Analytic results are sharp but narrow: IBVS is locally stable with characterised singularities (Chaumette & Hutchinson, 2006), PBVS is essentially global in pose space but unbounded in the image, and 2–1/2-D servoing is provably larger than IBVS (Malis et al., 1999). Learned results are broad but empirical, reported as a success rate over a test distribution (Sadeghi et al., 2018; Zitkovich et al., 2023); this covers object and viewpoint variation that no analytic method addresses, but says nothing about where the guarantee ends. Reference (Scherl et al., 2025) places both kinds of claim in one system, while (Gupta et al., 2026; Kim et al., 2024) and (Black et al., 2025) report measured results only. No convergence-domain result is derived for any of them.

Model and data requirements. The trend across the three phases runs in one direction only, and is one of the clearest findings (Figure 8). Analytic PBVS needs intrinsics, hand-eye calibration and an object model (Wilson & Hulls, 1996); IBVS needs intrinsics and depth (Chaumette & Hutchinson, 2006); 2–1/2-D servoing removes the object model (Malis et al., 1999); online estimation removes the remaining parameters but requires persistent excitation (Hosoda & Asada, 1994); deep pose regression removes calibration but needs training data, though as little as one real image plus synthesis (Bateux et al., 2018); sim-to-real policies need a simulator and a randomisation scheme (Sadeghi et al., 2018); and vision-language-action models need web-scale pretraining plus large demonstration sets (Zitkovich et al., 2023). Calibration has not been eliminated so much as moved into the training distribution, where it can no longer be inspected. The recent period is the first to interrupt the trend: pretrained representations are used with no task-specific training (Gupta et al., 2026; Scherl et al., 2025), while the generalist policies go the other way, with 970,000 demonstrations on top of vision-language pretraining (Kim et al., 2024) and multi-platform data collection (Black et al., 2025).

Computation and loop rate. Perception latency is phase lag in a feedback loop, so cost is not a secondary concern. Analytic laws are a pseudo-inverse of a small matrix and are negligible (Espiau et al., 1992); regression networks and small learned policies run at camera rate (Bateux et al., 2018; Sadeghi et al., 2018); and diffusion policies need many denoising iterations

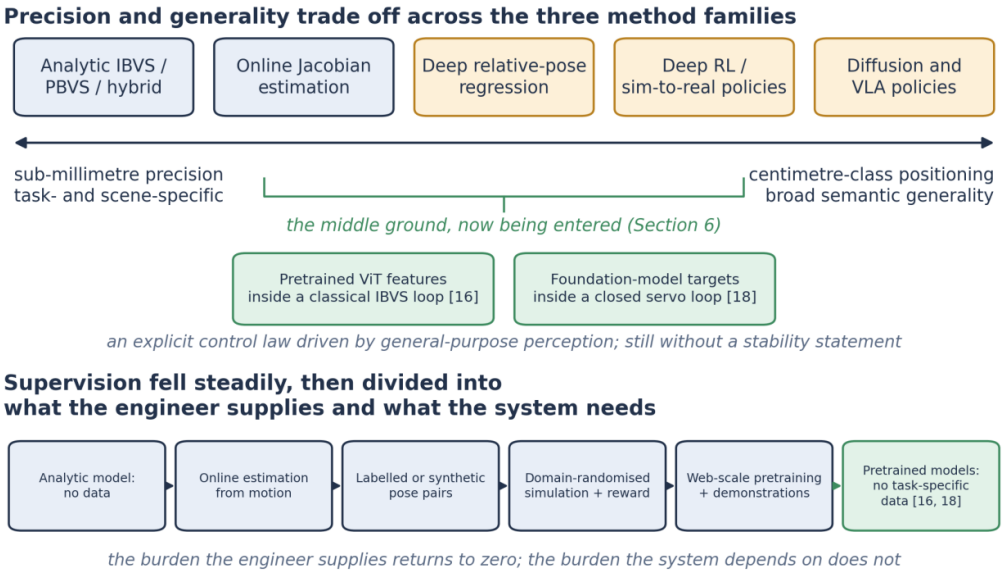


Figure 8 Precision-Generality Trade-Off and Evolution of Supervision Requirements for Different Approaches (Scherl et al., 2025; Gupta et al., 2026)

per action, made viable by generating a whole sequence and executing it in receding-horizon form (Chi et al., 2023). Vision-language-action models are built on billion-parameter backbones (Black et al., 2025; Kim et al., 2024; Zitkovich et al., 2023) that cannot run at classical loop rates on embedded hardware, which is why such models command setpoints rather than close a tight loop themselves. The recent works inherit this cost rather than resolve it, and neither (Scherl et al., 2025) nor (Gupta et al., 2026) reports a loop rate.

Verifiability. This axis separates the phases most sharply. The analytic schemes come with Lyapunov-based arguments (Chaumette & Hutchinson, 2006; Espiau et al., 1992; Malis et al., 1999), and so does online Jacobian estimation (Hosoda & Asada, 1994), which shows that learning a model need not mean giving up a stability proof. From (Lampe & Riedmiller, 2013) onward, the guarantee is replaced by measured success rates, and no deep-learning system reviewed here comes with a convergence proof or a bounded-error statement. That is decisive in safety-rated settings, and it is the main reason the deployed pattern remains learned perception feeding a monitored classical controller. The recent period does not close the gap but moves it: (Auddy et al., 2025; Scherl et al., 2025) and (Gupta et al., 2026) retain a control structure for which such statements exist, so what is missing is no longer a proof about an arbitrary network but a proof about a known controller with a learned input, the same problem (Hosoda & Asada, 1994) solved for an estimated Jacobian.

7.2 Robustness: Two Different Concepts

The two traditions use the same word for different things, and conflating them causes most of the confusion in cross-era comparison. Classical robustness is *structural*: the closed loop tolerates a bounded discrepancy between $\hat{\mathbf{L}}_\mathbf{s}$ and $\mathbf{L}_\mathbf{s}$, and this is established by analysis (Chaumette & Hutchinson, 2006; Espiau et al., 1992). It is narrow, saying nothing about a new object or a changed illuminant, but it is checkable, and the system knows when it is violated, because a tracker that loses its target reports failure.

Learned robustness is *distributional*: the network was trained on enough variation that a new observation resembles something it has seen. Bateux et al. obtain it by synthesising occlusions and lighting changes from one image (Bateux et al., 2018), Sadeghi et al. by randomising viewpoints in simulation (Sadeghi et al., 2018), RT-2 by inheriting it from web-scale pretraining (Zitkovich et al., 2023). This covers perturbations no classical tracker survives. Its defect is silence: outside the training distribution, a regression network still returns a well-formed pose, and nothing in the signal says it is wrong. The two notions are complements, and the architectures that combine them- a learned estimator whose output is bounded and monitored, feeding an analytic controller, have the best ratio of capability to integration risk. The recent works are the clearest instance of this combination so far. In (Scherl et al., 2025) and (Gupta et al., 2026), the distributional robustness of a pretrained vision model supplies the input to a loop whose structural robustness is that of the classical scheme, which is precisely the architecture just described. What they still lack is the bound and the monitor, since neither reports a confidence signal or a failure flag on the learned front end; the silence identified above is inherited along with the robustness.

7.3 Deployment and Open Problems

Sorted by setting, the literature is unambiguous. Structured industrial cells, where the objects are known and the tolerance is tight, are served by analytic schemes and by narrowly trained precision regressors (Bateux et al., 2018; Malis et al., 1999; Wilson & Hulls, 1996). Semi-structured settings with viewpoint and object variation but moderate precision are served by learned perception in a classical loop and by sim-to-real policies (Lee & Levine, 2017; Sadeghi et al., 2018). The semi-structured class is the one the recent works have most enlarged, since (Scherl et al., 2025) and (Gupta et al., 2026) operate on unseen objects and in novel environments without task-specific training, which was not previously available at this level of generality. Open-world tasks specified in language are served only by generalist policies (Black et al., 2025; Chi et al., 2023; Kim et al., 2024; Zitkovich et al., 2023), at centimetre-class precision and without any safety argument.

Four problems follow directly. *Verifiable learned servoing*: the precedent of (Hosoda & Asada, 1994), an estimated model with a Lyapunov proof, shows that data-driven components and stability arguments are not inherently incompatible, and extending that reasoning to learned feature spaces is the most valuable open direction. *The precision-generality handoff*: no reviewed system is both semantically general and millimetre-accurate, and the obvious architecture, in which a foundation-model policy supplies semantics and coarse alignment and hands off to a high-rate analytic servo for the terminal phase, remains largely unexplored. *Failure awareness*: learned perception needs the equivalent of the tracker's failure flag or the Kalman filter's covariance (Wilson & Hulls, 1996), so that distributional failure becomes detectable rather than silent. *Reporting practice*: progress at the seam requires papers that report latency, worst-case behaviour and the boundary of the operating envelope alongside task success. Section 6 shows the first two of these problems being approached rather than solved. References (Scherl et al., 2025) and (Gupta et al., 2026) place a general-purpose representation inside a loop with an explicit control law, and (Auddy et al., 2025) composes learned blocks inside an analytic priority structure instead of replacing it; but neither line yet yields the stability statement that would make the combination verifiable, and neither addresses failure awareness or reporting practice at all.

8. SUMMARY OF THE REVIEWED METHODS

Table 1 compares the three traditional families, and Table 2 compares the methodological phases on common axes. Table 3 reorganises the same

Table 2 Multi-Dimensional Comparison Among Three Methodological Phases and Recent Advances

Comparison dimension	Traditional	Machine learning	Deep learning	Recent (2024–2026)
What is learned	Nothing; derived from geometry (Espiau et al., 1992)	The Jacobian (Hosoda & Asada, 1994) or the policy (Lampe & Riedmiller, 2013)	Features, pose, or the entire policy (Bateux et al., 2018; Lee & Levine, 2017; Levine et al., 2016)	Pretrained visual features (Scherl et al., 2025); target proposals (Gupta et al., 2026); the motion (Auddy et al., 2025); the entire policy (Black et al., 2025; Kim et al., 2024)
Task specified by	Desired image or desired pose (Hutchinson & Hager, 1996)	Desired features; reward (Lampe & Riedmiller, 2013)	Desired image, reward, demonstrations, or language (Zitkovich et al., 2023)	A same-category reference image (Scherl et al., 2025); a click or an open-vocabulary name (Gupta et al., 2026); language (Black et al., 2025; Kim et al., 2024)
Reported accuracy	Convergence to the taught configuration (Chaumette & Hutchinson, 2006)	Convergence of features to desired trajectories (Hosoda & Asada, 1994)	Sub-mm positioning (Bateux et al., 2018); success rates otherwise (Chi et al., 2023; Zitkovich et al., 2023)	Full convergence unperturbed, +31.2% relative over classical IBVS when perturbed (Scherl et al., 2025); success rates otherwise (Black et al., 2025; Gupta et al., 2026; Kim et al., 2024)
Generalisation	None beyond the taught task	None beyond the trained task	Viewpoint (Sadeghi et al., 2018); objects and instructions (Zitkovich et al., 2023)	Unseen objects with no task-specific training (Scherl et al., 2025); novel environments and instances (Gupta et al., 2026); embodiments and instructions (Black et al., 2025; Kim et al., 2024)
Data needed	None	Online motion data; reward (Hosoda & Asada, 1994; Lampe & Riedmiller, 2013)	One image plus synthesis (Bateux et al., 2018); 20 trajectories (Lee & Levine, 2017); simulation (Sadeghi et al., 2018); web scale (Zitkovich et al., 2023)	None task-specific; pretrained models only (Gupta et al., 2026; Scherl et al., 2025); demonstrations (Auddy et al., 2025); 970k demonstrations plus pretraining (Kim et al., 2024)
Computation	Negligible	Low	Camera-rate to far below it, by model size (Chi et al., 2023; Zitkovich et al., 2023)	Foundation-model inference per iteration (Gupta et al., 2026; Scherl et al., 2025); billion-parameter backbones (Black et al., 2025; Kim et al., 2024)
Verifiability	Lyapunov stability (Espiau et al., 1992; Malis et al., 1999)	Lyapunov stability retained (Hosoda & Asada, 1994)	Empirical success rates only	No proofs reported; an analytic control structure is retained (Auddy et al., 2025; Gupta et al., 2026; Scherl et al., 2025)

material into the form most useful when a method has to be chosen for a particular task, listing each family of reviewed work against its advantages, its limitations and the applications for which the primary sources actually demonstrate it. Every entry is taken from a result reported in the cited work rather than from general expectation, and the citation in each cell identifies where the claim comes from. Read as a whole, the table restates the argument of Section 7 in compact form: the advantages column moves from analytic guarantees towards semantic breadth, the limitations column moves in the opposite direction, and no single row offers both.

9. BENCHMARKS AND DATASETS

The comparison in Section 7 repeatedly runs into the fact that the analytic and learned literatures report different quantities. Part of the explanation is that they evaluate on different things. Analytic servoing is assessed on a positioning task defined by the experimenter, so there is nothing to standardise; learned policies are assessed on shared benchmarks, and it is those benchmarks that make the success rates quoted throughout this review comparable to one another. Table 4 lists the three that the works reviewed here rely on. Each is peer-reviewed, each is openly released, and each is used by a work already discussed, so no new method is introduced by naming them.

The protocol behind a reported number matters as much as the number itself. Diffusion Policy, for example, is evaluated on 12 tasks drawn from four benchmarks, of which robomimic (Mandlekar et al., 2021) is one; success rate is the metric for every task except Push-T, which is scored by target area coverage, and each result is averaged over the last ten checkpoints, three training seeds, and fifty environment initialisations (Chi et al., 2023). The 46.9% average improvement quoted in Section 7 is an average over that protocol, not a single trial. OpenVLA reports separately on the four LIBERO suites (Liu et al., 2023), which vary the spatial arrangement, the objects, the goal and the horizon independently, so that a drop in success rate can be attributed to a particular kind of change (Kim et al., 2024).

Table 5 collects the results the reviewed methods report on these three benchmarks. Only methods that were actually benchmarked on them are listed, which restricts the table to the generative and vision-language-action family: the benchmarks postdate the rest of the reviewed work, and the analytic schemes are not posed as task suites that a benchmark could score. Where two reviewed methods do meet on the same suites, the comparison is informative rather than decisive. On LIBERO, they divide the four suites between them, Diffusion Policy taking Object at 92.5% against 88.4%, OpenVLA taking Spatial and Long, and the higher average at 76.5% against 72.4%, which is a margin of about four points spread unevenly across conditions rather than a uniform advantage.

Two limitations follow, and they bear directly on the argument of this review. None of these benchmarks measures terminal pose error, so a method that positions to a fraction of a millimetre and one that succeeds at a task are scored on incomparable scales; and none of them evaluates the analytic schemes of Sections 3 and 4 at all, because those schemes are not posed as task suites. A benchmark reporting both terminal accuracy and task success on the same manipulator would make the comparison in Section 7 quantitative rather than argumentative, and its absence is one reason the precision and generality literatures remain difficult to compare.

10. CONCLUSION

This review has traced visual servoing for robotic manipulators through three methodological phases. The traditional phase produced a complete analytic theory: the interaction matrix and the task-function formulation (Espiau et al., 1992), the taxonomy and tutorial synthesis that made the field learnable (Chaumette & Hutchinson, 2006; Chaumette & Hutchinson, 2007; Hutchinson & Hager, 1996), recursive pose estimation that made PBVS practical (Wilson & Hulls, 1996), partitioned control that removed the camera retreat (Corke, 2001), and hybrid servoing with a model-free formulation and an analytically enlarged convergence domain (Malis et al., 1999). Its limitation was never the control law but everything around it, namely brittle feature tracking, hand-supplied models, and task specification by demonstration, none of which generalises.

The machine-learning phase attacked the models. Estimating the visual-motor Jacobian online removed the dependence on camera and link parameters while keeping the Lyapunov argument intact (Hosoda & Asada, 1994), and neural reinforcement learning showed that the policy itself could be acquired from success and failure alone (Lampe & Riedmiller, 2013). These two works define the two learning targets, model and policy, that the deep-learning phase inherited and scaled.

The deep-learning phase then attacked perception and, progressively, the control law. Networks regressing relative pose reached sub-millimetre positioning while leaving the classical controller in place (Bateux et al., 2018); servoing on weighted deep features with a learned bilinear model

Table 3 Advantages, Limitations and Typical Applications of Reviewed Methods

Method family	Advantages	Limitations	Typical applications
IBVS (Chaumette & Hutchinson, 2006; Corke, 2001; Espiau et al., 1992; Hutchinson & Hager, 1996)	Error formed in the image, so calibration and depth error perturb the path but not the equilibrium (Chaumette & Hutchinson, 2006). No object model and no pose reconstruction.	Locally stable only, with image singularities and local minima (Chaumette & Hutchinson, 2006). Unpredictable Cartesian path, including camera retreat, removed by partitioned control (Corke, 2001). Needs reliable feature tracking (Chaumette & Hutchinson, 2007; Hutchinson & Hager, 1996).	Positioning against a taught image in structured cells with known objects and imperfect calibration.
PBVS (Chaumette & Hutchinson, 2006; Wilson & Hulls, 1996)	Cartesian error decouples translation and rotation: straight-line path, essentially global convergence in pose space (Chaumette & Hutchinson, 2006). Estimator separate from the controller, with a covariance that can be monitored (Wilson & Hulls, 1996).	Calibration and model error bias the pose and appear as steady-state task error (Chaumette & Hutchinson, 2006). No field-of-view guarantee. Requires a 3-D model of each object (Wilson & Hulls, 1996).	Relative end-effector positioning where a predictable Cartesian path matters and object geometry is known in advance (Wilson & Hulls, 1996).
Hybrid, 2–1/2-D visual servoing (Chaumette & Hutchinson, 2007; Malis et al., 1999)	Upper-triangular interaction matrix, no singularity in the task space, convergence from any initial position under exact intrinsics. No 3-D object model; necessary and sufficient local stability conditions derived (Malis et al., 1999).	Visibility needs an added adaptive law; the displacement estimation is sensitive to image noise (Malis et al., 1999).	Positioning over large displacements without a full object model.
Model-free estimation of the visual-motor map (Hosoda & Asada, 1994)	Visual-motor Jacobian identified online by recursive least squares from the motion the loop already produces; no kinematic, camera, or link parameters needed. Convergence proved by Lyapunov analysis, so the estimated model keeps the stability argument (Hosoda & Asada, 1994).	The estimate is local and needs continuous refresh; it degrades under poor excitation, and one mis-tracked feature can corrupt it (Hosoda & Asada, 1994).	Uncalibrated or reconfigurable set-ups where hand-eye parameters are unknown or change during operation.
Reinforcement learning for servoing (Lampe & Riedmiller, 2013; Lee & Levine, 2017)	Policy learned from success and failure signals alone (Lampe & Riedmiller, 2013). Servoing on weighted deep features with a learned bilinear model needs only twenty trajectories, over two orders of magnitude better than model-free deep RL, and tolerates viewpoint and appearance change (Lee & Levine, 2017).	Reward is weak supervision, so interaction cost is high. The controller is a network rather than an equation, so the proof available to (Hosoda & Asada, 1994) is lost, and behaviour is specific to the task trained on (Lampe & Riedmiller, 2013).	Reaching and grasping learned from scratch (Lampe & Riedmiller, 2013); target following with a learned feature space (Lee & Levine, 2017).
Learned components inside an explicit control law (Auddy et al., 2025; Bateux et al., 2018; Gupta et al., 2026; Scherl et al., 2025)	The analytic controller survives, so the network output can be bounded and logged (Bateux et al., 2018). Sub-millimetre positioning on a 6-DOF arm from one real image plus synthesis (Bateux et al., 2018). Pretrained ViT features give full convergence undisturbed and up to 31.2% gain over classical IBVS when disturbed, without task-specific training (Scherl et al., 2025). Learned blocks sit inside the large projection law (Auddy et al., 2025).	Failure becomes distributional and silent: a confident but wrong output off-distribution (Bateux et al., 2018). No stability proof for any combined system. Foundation-model inference on every iteration, with loop rates unreported (Gupta et al., 2026; Scherl et al., 2025).	Precise positioning under lighting change and occlusion (Bateux et al., 2018); end-effector positioning, industrial box manipulation and grasping of unseen objects (Scherl et al., 2025); mobile manipulation of knobs and light switches, 71% zero-shot (Gupta et al., 2026).
End-to-end and simulation-to-reality policies (Levine et al., 2016; Sadeghi et al., 2018)	Perception and control trained jointly, pixels to torques, with no features, interaction matrix or pose (Levine et al., 2016). Randomised simulation, transferred by adapting only the visual layers, servos to unseen objects from novel viewpoints on real hardware (Sadeghi et al., 2018).	The data requirement is met only by simulation and a randomisation scheme (Sadeghi et al., 2018). No stability argument; results given as success rates rather than positioning error.	Tasks needing tight vision-control coordination, such as screwing a cap onto a bottle (Levine et al., 2016); servoing when the viewpoint is not fixed (Sadeghi et al., 2018).
Generative and vision-language-action policies (Black et al., 2025; Chi et al., 2023; Kim et al., 2024; Zitkovich et al., 2023)	Conditional denoising generates whole action sequences, capturing multimodal distributions, with a 46.9% average gain on twelve benchmark tasks (Chi et al., 2023). The task is given as a language instruction (Zitkovich et al., 2023). An open 7B model beats the closed 55B RT-2-X by 16.5 points on 29 tasks (Kim et al., 2024); actions are continuous under flow matching (Black et al., 2025).	Billion-parameter backbones cannot be evaluated at classical loop rates on embedded hardware (Zitkovich et al., 2023). Precision is centimetre-class, and results are success rates with no convergence proof or error bound.	Open-world tasks specified in language (Kim et al., 2024; Zitkovich et al., 2023); benchmark manipulation with multimodal actions (Chi et al., 2023); multi-stage household tasks such as laundry folding, table cleaning and box assembly (Black et al., 2025).

Table 4 Benchmarks and Datasets Adopted in the Reviewed Works

Benchmark	What it provides	Used by
Open X-Embodiment / RT-X (Collaboration et al., 2024)	Real robot demonstrations pooled from 22 robot embodiments and 21 institutions, covering 527 skills across 160,266 tasks, released in a standardised format.	Pretraining corpus for OpenVLA (Kim et al., 2024); the RT-X models trained on it are the baseline OpenVLA is compared against.
Robomimic (Mandlekar et al., 2021)	Human teleoperated demonstration datasets of varying quality for five simulated and three real multi-stage manipulation tasks, with a study of six offline learning algorithms over them.	One of the four benchmarks on which Diffusion Policy (Chi et al., 2023) is evaluated.
LIBERO (Liu et al., 2023)	A lifelong manipulation benchmark of four task suites that vary spatial arrangement, objects, goal, and horizon independently, designed to separate declarative from procedural knowledge transfer.	Evaluation of OpenVLA (Kim et al., 2024) under parameter-efficient fine-tuning.

achieved order-of-magnitude gains in sample efficiency (Lee & Levine, 2017); end-to-end training mapped pixels directly to torques (Levine et al., 2016); recurrent policies trained in simulation transferred to real hardware and generalised across viewpoints (Sadeghi et al., 2018); generative policies modelled multimodal action distributions (Chi et al., 2023); and vision-language-action models replaced the taught image with a spoken instruction (Zitkovich et al., 2023).

The works of the most recent period do not add a fourth phase so much as begin to fold the first three together. A classical image-based law supplied with pretrained vision-transformer features converges without any task-specific training and improves on its classical ancestor under perturbation (Scherl et al., 2025); a servo loop driven by targets from vision foundation models manipulates unseen objects in unseen environments (Gupta et al., 2026); learned perception and learned motion are composed inside the large projection task-priority formulation rather than replacing it (Auddy et al., 2025); and the generalist policies became open and, in their action representation, continuous (Black et al., 2025; Kim et al., 2024). What none of them supplies is a stability statement for the combination.

The synthesis is not that one phase superseded another. Each solved the problem the previous one could not, and each gave up something the previous one had. Analytic control still offers the only stability argument and, with a well-conditioned feature set, the best terminal precision; learning offers perception that survives occlusion, lighting change, and object variation, and task specifications no interaction matrix can express. The unsolved problem lies exactly at the seam: a system that is told what to do in language, that locates the object with a learned representation robust to the real world, and that closes the final millimetres with a controller whose behaviour can be stated in advance. Every ingredient exists in the literature reviewed here, and the works of the most recent period have begun to assemble them (Auddy et al., 2025; Gupta et al., 2026; Scherl et

Table 5 Experimental Performance of Reviewed Methods on Benchmark Datasets

Benchmark	Reviewed method	Reported result	What the number is
LIBERO (Liu et al., 2023)	OpenVLA (Kim et al., 2024), fine-tuned	84.7, 88.4, 79.2 and 53.7 on the Spatial, Object, Goal and Long suites; average 76.5.	Success rate in per cent, reported with its uncertainty, over the four task suites.
LIBERO (Liu et al., 2023)	Diffusion Policy (Chi et al., 2023) trained from scratch, as reported in (Kim et al., 2024)	78.3, 92.5, 68.3 and 50.5 on the same four suites; average 72.4.	Same metric and protocol; run as a baseline by the authors of (Kim et al., 2024) rather than reported in (Chi et al., 2023).
Robomimic (Mandlekar et al., 2021)	Diffusion Policy (Chi et al., 2023)	46.9% mean improvement over prior methods.	Averaged over 12 tasks drawn from four benchmarks, of which Robomimic is one; success rate for every task except Push-T, which is scored by target area coverage.
Open X-Embodiment (Collaboration et al., 2024)	OpenVLA (Kim et al., 2024)	16.5 points over RT-2-X (55B) and 20.4 points over from-scratch imitation learning.	Absolute task success rate across 29 tasks and several robot embodiments, after pretraining on the dataset.

al., 2025). What does not yet exist is the argument that holds when they are put together.

Four directions follow. The most consequential is a stability result for the combination: (Hosoda & Asada, 1994) established convergence for an estimated Jacobian, and the natural generalisation is to state the conditions a learned feature space must satisfy for that argument to survive, which would give the systems of Section 6 a guarantee rather than only a measured success rate. The second is a confidence signal for learned perception. The extended Kalman filter of (Wilson & Hulls, 1996) carries a covariance that can be monitored; no learned front end reviewed here reports an equivalent, and without one, distributional failure stays silent. The third is an explicit handoff between a generalist policy supplying semantics and coarse alignment and an analytic servo closing the final millimetres, together with a criterion for when to switch; the components exist (Gupta et al., 2026; Scherl et al., 2025), the switching argument does not. The fourth is reporting practice: loop rates, worst-case behaviour and the boundary of the operating envelope published alongside success rates, and shared benchmarks measuring terminal pose error and semantic breadth on the same task, so that the analytic and learned literatures can at last be compared directly.

REFERENCES

Alhaddad, M., Shaukifeh, B., & Joukhadar, A. (2019). Adaptive LQ based computed torque controller for robotic manipulator. In *Proceedings of the IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)* (pp. 2169–2173).

Auddy, S., Paolillo, A., Piater, J. H., & Saveriano, M. (2025). Imitation learning-based Direct Visual Servoing using the large projection formulation. *Robotics and Autonomous Systems, 190*, 104971.

Bateux, Q., Marchand, E., Leitner, J., Chaumette, F., & Corke, P. (2018). Training deep neural networks for visual servoing. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (pp. 3307–3314).

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M. R., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L. X., Smith, L., Tanner, J., Vuong, Q., Wang, H., & Zhilinsky, U. (2025). π_0: A vision-language-action flow model for general robot control. In *Proceedings of Robotics: Science and Systems XXI*. https://roboticsproceedings.org/rss21/p010.html

Chaumette, F., & Hutchinson, S. (2006). Visual servo control, Part I: Basic approaches. *IEEE Robotics & Automation Magazine, 13*(4), 82–90.

Chaumette, F., & Hutchinson, S. (2007). Visual servo control, Part II: Advanced approaches. *IEEE Robotics & Automation Magazine, 14*(1), 109–118.

Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., & Song, S. (2023). Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems XIX*. https://doi.org/10.15607/RSS.2023.XIX.026

Open X-Embodiment Collaboration, O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., Tung, A., Bewley, A., Herzog, A., Irpan, A., Khazatsky, A., Rai, A., Gupta, A., Wang, A., Singh, A., Lin, Z. (2024). Open X-Embodiment: Robotic learning datasets and RT-X models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 6892–6903). IEEE. https://doi.org/10.1109/ICRA57147.2024.10611477

Corke, P. I., & Hutchinson, S. A. (2001). A new partitioned approach to image-based visual servo control. *IEEE Transactions on Robotics and Automation, 17*(4), 507–515.

Espiau, B., Chaumette, F., & Rives, P. (1992). A new approach to visual servoing in robotics. *IEEE Transactions on Robotics and Automation, 8*(3), 313–326.

Gupta, A., Sathua, R., & Gupta, S. (2026). Precise mobile manipulation of small everyday objects. *IEEE Robotics and Automation Letters*.

Hosoda, K., & Asada, M. (1994). Versatile visual servoing without knowledge of true Jacobian. In *Proceedings of the IEEE/RSJ/GI International Conference on Intelligent Robots and Systems (IROS)* (pp. 186–193).

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.

Joukhadar, A., Hanna, K., & Abo AlIzam, E. (2020). UKF-based image filtering and 3D reconstruction. In O. Sergiyenko, W. Flores-Fuentes, & P. Mercorelli (Eds.), *Machine vision and navigation* (pp. 267–289). Springer International Publishing.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E. P., Sanketi, P. R., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., & Finn, C. (2025). OpenVLA: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning* (Vol. 270, pp. 2679–2713). Proceedings of Machine Learning Research.

Laalou, A., Joukhadar, A., Alnaddaf, H., & Chamilothoris, G. (2024). Toward a computationally efficient solution of the inverse kinematics problem using machine learning. In *Frontiers of artificial intelligence, ethics, and multidisciplinary applications (FAIEMA)* (pp. 471–485). Springer.

Lampe, T., & Riedmiller, M. (2013). Acquiring visual servoing reaching and grasping skills using neural reinforcement learning. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8).

Lee, A. X., Levine, S., & Abbeel, P. (2017). Learning visual servoing with deep features and fitted Q-iteration. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research, 17*(39), 1–40.

Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., & Stone, P. (2023). LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in neural information processing systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*.

Malis, E., Chaumette, F., & Boudet, S. (1999). 2–1/2-D visual servoing. *IEEE Transactions on Robotics and Automation, 15*(2), 238–250.

Mandlekar, A., Xu, D., Wong, J., Nasiriany, S., Wang, C., Kulkarni, R., Fei-Fei, L., Savarese, S., Zhu, Y., & Martín-Martín, R. (2021). What matters in learning from offline human demonstrations for robot manipulation. In *Proceedings of the 5th Conference on Robot Learning*.

Sadeghi, F., Toshev, A., Jang, E., & Levine, S. (2018). Sim2Real viewpoint invariant visual servoing by recurrent control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4691–4699).

Scherl, A., Thalhammer, S., Neuberger, B., Wöber, W., & García-Rodríguez, J. (2025). ViT-VS: On the applicability of pretrained vision transformer

features for generalizable visual servoing. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Wilson, W. J., Williams Hulls, C. C., & Bell, G. S. (1996). Relative end-effector control using Cartesian position-based visual servoing. *IEEE Transactions on Robotics and Automation, 12*(5), 684–696.

Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning* (Vol. 229, pp. 2165–2183). Proceedings of Machine Learning Research.