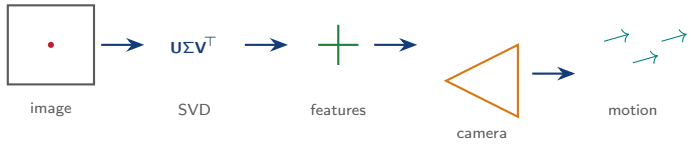


ES666 Computer Vision

Shanmuganathan Raman

Professor, EE and CSE, IIT Gandhinagar

Imaging · Linear algebra & SVD · Convolution & Fourier · Filtering · Edges & corners
Scale space & SIFT · Projective geometry · Homography, DLT & RANSAC · Cameras
Epipolar geometry (F & E) · Multi-view stereo & SfM · Optical flow · Photometric stereo and inverse rendering
MLP & CNNs · VAEs, GANs, diffusion & flows · Transformers · Representation & Self-Supervised Learning ·
Ten vision applications

Twenty-eight hours of teaching a computer to do what your retina does for free — with proofs where we can prove, and gradients where we can't.

Course outline — twenty-eight hours, six parts

Part I — Foundations (Hr 1–3): image formation; the SVD; least squares.

Part II — Filtering (Hr 4–6): convolution; Fourier; smoothing.

Part III — Features (Hr 7–10): edges; Canny/Harris; scale space; SIFT/SURF/ORB.

Part IV — Geometry, motion & light (Hr 11–18): projective geometry; homography; DLT/RANSAC; calibration; epipolar geometry; multi-view stereo & SfM; optical flow; photometric stereo. *Supplement (Lecture 18a)*: depth from focus/defocus; shape from interreflections & polarization.

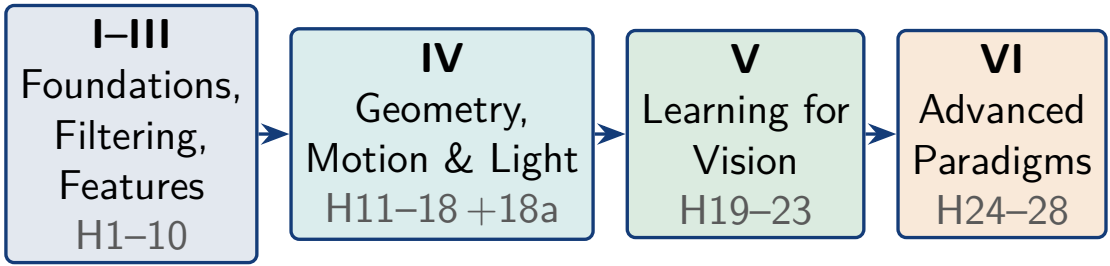
Part V — Learning for vision (Hr 19–23): the MLP; generalization & optimizers; CNNs; the VAE; vision applications.

Part VI — Advanced paradigms (Hr 24–28): GANs; transformers; diffusion; flows; representation & self-supervised learning.

Parts I–IV *design* the function (classical, proof-complete); Parts V–VI *learn* it.

The whole course on one slide

Twenty-eight lecture-hours in six parts — classical first, learned second:

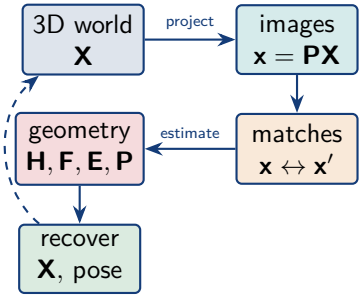


Real talk

Math basics (linear algebra, calculus, optimization) thread through Parts I–III; ten vision *applications* thread through Parts V–VI. Read in order — each part assumes the last.

Sources: Course structure.

The big picture: geometric vision



Forward: the world projects to images (Hr 14). **Inverse**: from images and matches we estimate geometry (Hr 11–15) and recover structure & motion (Hr 15–16). Part IV is this inverse problem.

Sources: Hartley & Zisserman (2004); Szeliski (2022).

Part I Roadmap

1 — Introduction & image formation

CV vs. image processing / graphics / generative AI; image as a function; sampling, quantization & Nyquist; why CV is hard; **pinhole model & perspective projection**; lens & depth of field; the sensing pipeline; demosaicing, gamma, CRF.

2 — Linear algebra I (with proofs)

Four subspaces; spectral theorem; **SVD existence proof**; eigendecomposition link; rank/range/null space, norms, det, κ (all proved); worked 2×2 SVD; **Eckart–Young proof**.

3 — Linear algebra II (with proofs)

Normal-equations derivation; residual orthogonality; projection matrix; pseudo-inverse via SVD; **homogeneous $\mathbf{Ax} = \mathbf{0}$ proof**; minimum-norm (Lagrange); regression; total least squares; conditioning & normalization.

Why this first?

Every later topic reduces to *solving a linear system* or *minimizing a residual*. Master this and the rest is geometry.

Sources: Szeliski (2022); Hartley & Zisserman (2004); Trefethen & Bau (1997).

Part II roadmap (Hours 4–6 of 18)

Hour 4 — Convolution & LTI

Sifting; **LTI** $\Rightarrow$ **convolution** (proof); commutativity / associativity / length (proofs); FIR/IIR; **BIBO stability** proof; correlation & NCC; 2D **separability** proof; Toeplitz/circulant view.

Hour 5 — Fourier analysis

Complex exponentials as **eigenfunctions** (proof); CTFT/DTFT/DFT/FFT; **convolution theorem** (proof); shift/scaling/derivative (proofs); circular convolution; 2D separability; **Gaussian self-duality** (proof); sampling & aliasing.

Hour 6 — Image filtering

Box filter (sinc response); **separability & integral images** (proofs); Gaussian **separability + semigroup** (proofs); binomial/CLT; derivative-of-Gaussian; **noise-variance proof**; median / bilateral / NL-means; anti-aliased pyramids; sharpening.

Thread

One idea runs through all three hours: *convolution in space = multiplication in frequency*. Hour 4 builds it, Hour 5 explains it, Hour 6 exploits it.

Sources: Oppenheim & Schafer; Bracewell; Gonzalez & Woods (2018); Szeliski (2022).

Part III roadmap (Hours 7–10 of 18)

Hour 7 — Gradients, edges, LoG, DoG

Finite differences (accuracy proof); Sobel separability; noise/derivative trade-off; **LoG closed form**; heat equation; **DoG** $\approx$ **LoG** proof; structure tensor; Hough.

Hour 8 — Canny & Harris

Canny criteria & optimal detector; NMS; hysteresis; **auto-correlation** $\rightarrow$ **structure tensor** derivation; eigenvalue classification; $R = \det \mathbf{M} - k\,\mathrm{tr}^2\,\mathbf{M}$; **rotation-invariance proof**.

Hour 9 — Scale space & blobs

Scale-space axioms; **diffusion-equation proof**; pyramids & reconstruction; **scale normalization** & characteristic-scale proof; DoG scale space; 26-neighbour extrema.

Hour 10 — SIFT / SURF / ORB

DoG keypoints; **sub-pixel & edge-rejection** proofs; orientation; 128-D descriptor; normalization; **ratio test**; SURF; FAST/BRIEF/**ORB**; **RANSAC** count proof.

Sources: Marr & Hildreth (1980); Canny (1986); Harris & Stephens (1988); Lindeberg (1994); Lowe (2004).

Part IV roadmap (Hours 11–14 of 28)

Hour 11 — Projective geometry of the plane

Homogeneous coordinates; points/lines **duality**; **join & meet** as cross products; ideal points & the line at infinity; the transformation hierarchy; **cross-ratio** invariance; vanishing points/lines; conics & duals; transforms-as-convolutions; **image warping** & implicit reps.

Hour 12 — Transformations & the homography

The 3×3 homography; **planar** & **rotational** homographies (proofs); forward/backward warping & bilinear resampling; when $\mathbf{H}$ applies vs. the **epipolar** constraint; cross-product DLT equations; $\mathbf{H}$ without correspondences.

Hour 13 — DLT & robust estimation (RANSAC)

DLT & the **null-space** (SVD) solution; Hartley **normalization**; algebraic vs. geometric error & **Sampson**; nonlinear refinement; **RANSAC** & the adaptive iteration count; MSAC/PROSAC; **linear triangulation**; bundle adjustment; learned matching.

Hour 14 — Camera models & calibration

Pinhole & the 3×4 $\mathbf{P}$; intrinsics $\mathbf{K}$ ($a = fN/w$); **centre** = **null space of P**; lens distortion; **DLT calibration** + RQ; **Zhang** plane-based; vanishing-point calibration; horizon & tilt; back-projection to rays.

Arc

Represent geometry (11) → *estimate* it (12–13) → *calibrate* the camera (14).

Sources: Hartley & Zisserman (2004); Szeliski (2022); Zhang (2000); Torralba et al. (2024), Ch. 38–41.

Part IV roadmap (Hours 15–18 + 18a of 28)

Hour 15 — Epipolar geometry: F & E

The **epipolar constraint**; **essential** $\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$ & **fundamental F**; **8-point** (normalized) & the **rank-2** projection; 7-/5-point; $\mathbf{R}, \mathbf{t}$ from $\mathbf{E}$ & cheirality; rectification & disparity; cost volumes & learned stereo.

Hour 16 — Multi-view stereo & SfM

Reprojection error & **bundle adjustment** (GN/LM, Schur); incremental vs. global SfM; **gauge freedom**; **Tomasi–Kanade factorization**; projective→metric; MVS photo-consistency & plane-sweep; **scale drift** & loop closure.

Hour 17 — Optical flow

Motion field vs. flow; **brightness constancy** & the aperture problem; **Lucas–Kanade** & the structure tensor; **Horn–Schunck**; pyramidal coarse-to-fine; **time-to-contact**; self-supervised flow; event cameras.

Hour 18 — Photometric stereo & shape

Radiometry & the **BRDF**; Lambertian shading; **shape-from-shading**; calibrated/uncalibrated **photometric stereo**; **GBR ambiguity**; integrability & Poisson; **inverse rendering**; differentiable & neural rendering (**NeRF**, Gaussian splatting).

Lecture 18a — shape-from-X (supplement)

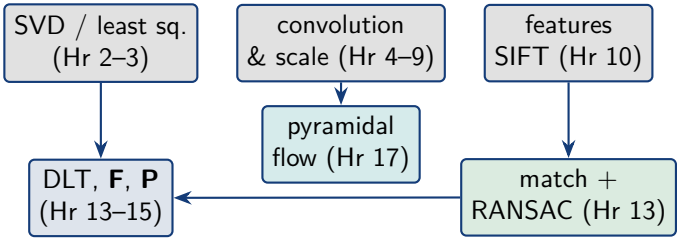
Depth from **focus**/**defocus** (coded apertures, dual pixels); **shape from interreflections** (direct–global separation); **shape from polarization** (angle/degree → normal).

Arc

Relate two views (15) → *reconstruct* scenes (16) → *track* motion (17) → *recover shape from light* (18, 18a).

Sources: Hartley & Zisserman (2004); Szeliski (2022); Torralba et al. (2024), Ch. 40, 44–45, 47–48.

How Part IV builds on Parts I–III



Every tool from the first three parts is load-bearing here: SVD solves the homogeneous systems, scale-space feeds coarse-to-fine flow, and SIFT supplies the correspondences that geometry consumes.

Sources: Parts I–III of this course; Szeliski (2022).

Lecture 18a — more routes to shape (supplement)

A companion to Hour 18: four **single-view** cues that read shape from optics, radiometry, and wave optics rather than triangulation.

Optical-blur depth	Radiometric & wave-optics shape
Depth from Focus — pick the sharpest frame in a focal stack.	**Shape from Interreflections** — bounced light as a cue; direct–global separation.
Depth from Defocus — invert the blur $\sigma(u)$; coded apertures, dual pixels, deep optics.	**Shape from Polarization** — normals from the angle & degree of polarization.

All four are monocular and *complement* stereo/SfM (Hr 15–16) and photometric stereo (Hr 18): coarse depth + polarization/focus detail $\rightarrow$ crisp surfaces.

Sources: Nayar–Ikeuchi–Kanade (1991); Levin et al. (2007); Atkinson & Hancock (2006); Kadambi et al. (2015); Torralba et al. (2024), Ch. 5–6.

Prerequisites in one slide

Four things from earlier parts to keep within arm's reach:

- **SVD** (Hr 2–3): $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$; smallest singular vector solves $\min_{\|\mathbf{h}\|=1} \|\mathbf{A}\mathbf{h}\|$; Eckart–Young gives best low-rank.
- **Eigenvalues** & the **structure tensor** (Hr 8): corner strength from λ_1, λ_2 of $\mathbf{M}$.
- **Scale space & pyramids** (Hr 9): Gaussian/Laplacian pyramids, coarse-to-fine.
- **Features** (Hr 10): SIFT detection & matching, the ratio test.

If a step ever feels like it fell from the sky, it almost certainly traces back to one of these four. Look up, not out the window.

Sources: Parts I–III; Trefethen & Bau (1997); Lowe (2004).

Part V roadmap (Hours 19–23 of 28)

Hour 19 — Neural nets & the MLP

Discriminative vs. generative; perceptron & the XOR problem; activations; **universal approximation**; backpropagation; SGD, initialization & vanishing gradients; nets as distribution transformers.

Hour 20 — Generalization, regularization & optimizers

Bias–variance & the over/under-fitting U; capacity & double descent; train/val/test; L2/L1, dropout, early stopping, augmentation; batch & layer norm; momentum, RMSProp, **Adam**; LR schedules.

Hour 21 — Convolutional neural networks

Locality, weight sharing & translation equivariance; receptive fields; pooling; LeNet→VGG→ResNet; transfer learning; CNNs as image-to-image; the ViT bridge.

Hour 22 — Generative models & the VAE

Generator $g(\mathbf{z}, \mathbf{y})$ & latent variables; autoencoders; variational inference & the **ELBO**; reparameterization trick; the VAE & β-VAE; density vs. energy models; a map of GANs, flows & diffusion.

Hour 23 — Computer-vision applications

One backbone, many heads: classification, semantic segmentation (FCN/U-Net), detection (Faster R-CNN/YOLO); homography, depth & optical-flow estimation; restoration & super-resolution; relighting; NeRF; HDR.

Bridge

Part V keeps the geometry of Parts I–IV as *inductive bias* and lets an optimizer fit the rest.

Sources: Torralba, Isola & Freeman (2024), Ch. 9–14, 23–24, 30–34; Goodfellow et al. (2016).

Part VI roadmap (Hours 24–28 of 28)

Hour 24 — Generative Adversarial Networks

The adversarial game; minimax & the optimal discriminator $\rightarrow$ JS divergence; non-saturating loss; mode collapse; **WGAN**; DCGAN, cGAN, pix2pix, CycleGAN, SRGAN, StyleGAN; FID.

Hour 25 — Transformers

Tokens & attention; softmax$(\mathbf{Q}\mathbf{K}^\top/\sqrt{d})\mathbf{V}$; multi-head; the transformer block; **ViT**; "CNNs in disguise"; positional codes; DETR, Swin, SegFormer, DPT.

Hour 26 — Diffusion models

Forward noising; predict the noise ε_θ; reverse sampling (DDPM/DDIM); U-Net & time embedding; classifier-free guidance; **latent diffusion**; cross-attention text-to-image; ControlNet.

Hour 27 — Flow models

Exact likelihood; change of variables & the log-det Jacobian; coupling layers (RealNVP); Glow 1×1 conv; continuous flows (Neural ODE); **flow matching**; SRFlow, compression, anomalies.

Hour 28 — Representation & self-supervised learning

Encoders & embeddings; autoencoders; pretext tasks; contrastive **InfoNCE**; alignment & uniformity; SimCLR, MoCo, BYOL; MAE; **CLIP**, zero-shot & open-vocabulary vision.

Close

Four ideas — a learned loss, attention, score/transport, and representation — power today's foundation models.

Sources: Torralba, Isola & Freeman (2024), Ch. 26, 30, 32–34, 51; Goodfellow et al. (2014); Ho et al. (2020).

Ten applications, threaded through the course

Real tasks reappear as each new tool arrives — solved classically in Parts I–IV, then learned in Parts V–VI:

Recognition & dense prediction

Image classification · semantic segmentation · object detection · depth estimation · optical-flow estimation

Geometry, low-level & synthesis

Homography estimation · image restoration · image relighting · neural radiance fields (NeRF) · HDR imaging

Real talk

Each application is solved at least twice — once with a hand-built pipeline, once with a learned model. Watching the two meet is the point of the course.

Sources: Course applications thread; Torralba et al. (2024).

Four throughlines

Beneath the twenty-eight hours run four recurring ideas:

- **Geometry** — the camera is a projection; vision inverts it. (Parts I, IV)
- **Learning** — replace a designed function with one fit to data by minimizing a loss. (Part V)
- **Generation** — model the data distribution itself; synthesize, don't just classify. (Hours 22, 24, 26, 27)
- **Representation** — the right embedding makes every task easier; learn it without labels. (Hours 21, 25, 28)

Spot these threads and the field stops looking like a pile of tricks and starts looking like one idea seen from many angles.

Sources: Deck throughlines.

Textbooks & references

Primary

- Torralba, Isola & Freeman, *Foundations of Computer Vision*, MIT Press 2024 (free: `visionbook.mit.edu`)
- Szeliski, *Computer Vision: Algorithms & Applications*, 2nd ed., Springer 2022

Geometry & classical

- R. Hartley & A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed., Cambridge University Press 2004; Horn, *Robot Vision*, MIT Press 1986

Shape-from-X (Lecture 18a): Nayar–Ikeuchi–Kanade (1991); Levin et al. (2007); Atkinson & Hancock (2006); Kadambi et al. (2015).

Deep learning

- Bishop & Bishop, *Deep Learning: Foundations & Concepts*, Springer 2024; S. J. D. Prince, *Understanding Deep Learning*, MIT Press 2023 (free: `udlbook.com`)
- Zhang et al., *Dive into Deep Learning* 2023; Goodfellow, Bengio & Courville, *Deep Learning*, MIT Press 2016

Sources: Course reference list.

Notation & conventions

- scalars a; vectors $\mathbf{x}$ (bold, columns); matrices $\mathbf{A}$
- $\mathbf{A}^\top$ transpose; $\mathbf{A}^{-1}$ inverse; $\mathbf{A}^+$ pseudo-inverse
- $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^\top \mathbf{x}}$; $\|\mathbf{A}\|_2, \|\mathbf{A}\|_F$ matrix norms
- $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$; $\mathbf{x} \perp \mathbf{y} \iff \mathbf{x}^\top \mathbf{y} = 0$
- $\mathbf{I}$ identity; $\mathbf{0}$ zero vector; $\mathbf{e}_i$ standard basis
- Col, Null, rank, span, tr, diag as named
- σ_i singular values; λ_i eigenvalues
- "$\simeq$" equality *up to non-zero scale* (homogeneous)
- $I(i, j)$ image intensity; $\mathbf{X} \in \mathbb{R}^3$ scene point; $\mathbf{x} \in \mathbb{R}^2$ image point

Real talk

Grey footnotes list each slide's **primary sources**. Boxed results are the ones to memorize; *Proof* boxes are self-contained.

Notation & conventions (signals edition)

- $x[n], h[n]$: discrete signals; $x(t)$: continuous
- $\delta[n]$ Kronecker, $\delta(t)$ Dirac (impulse)
- $*$ convolution; $\circledast$ *circular* convolution; $\star$ correlation
- h: impulse response; $H(\omega)$: frequency response
- $\mathcal{F}\{\cdot\}$ Fourier transform; $X(\omega)$ spectrum
- $\mathrm{DFT}, \mathrm{FFT}$: discrete / fast Fourier transform
- $\widehat{f}$: transform of f; $|X|, \angle X$: magnitude, phase
- G_σ: Gaussian of std. dev. σ; $\mathrm{sinc}(x) = \sin x / x$
- $I[m, n]$: image; k: kernel size; N: pixel count

Real talk

Grey footnotes give each slide's **primary sources**; boxed results are the keepers; *Proof* boxes are self-contained and end with ■.

Notation & conventions (features edition)

- $I(x, y)$: image; I_x, I_y: partial derivatives
- $\nabla I = (I_x, I_y)$: gradient; $\nabla^2 I$: Laplacian
- G_σ: Gaussian, scale σ; G'_σ: derivative of Gaussian
- $L(x, y, \sigma) = G_\sigma * I$: scale space
- $\mathrm{LoG} = \nabla^2 G_\sigma$; $\mathrm{DoG} = G_{k\sigma} - G_\sigma$
- $\mathbf{M}$: structure / second-moment tensor; λ_1, λ_2 its eigenvalues
- $\mathcal{H}$: Hessian; R: Harris response
- $\mathbf{d}$: descriptor vector; θ: orientation
- NMS: non-maximum suppression

Real talk

Grey footnotes give each slide's **primary sources**; boxed results are the keepers; *Proof* boxes are self-contained and end with ■.

Notation I — projective geometry

$\mathbf{x} = (x, y, 1)^\top$	homogeneous image point
$\boldsymbol{\ell} = (a, b, c)^\top$	line; incidence $\boldsymbol{\ell}^\top \mathbf{x} = 0$
$\mathbf{x}_1 \times \mathbf{x}_2$	join (line through two points)
$\boldsymbol{\ell}_1 \times \boldsymbol{\ell}_2$	meet (intersection of two lines)
$\mathbb{P}^2,\ \boldsymbol{\ell}_\infty$	projective plane, line at infinity
$\mathbf{H}$ (3×3)	homography; $\mathbf{x}' = \mathbf{H}\mathbf{x}$, 8 DOF
$\mathbf{C},\ \mathbf{C}' = \mathbf{H}^{-\top}\mathbf{C}\mathbf{H}^{-1}$	conic and its transform
$\mathrm{Cross}(\cdot)$	cross-ratio (projective invariant)

Sources: Hartley & Zisserman (2004), Ch. 2.

Notation II — cameras & two views

$\mathbf{P} = \mathbf{K}[\mathbf{R} \mid \mathbf{t}]$	3×4 camera, 11 DOF
$\mathbf{K}$	intrinsics (f_x, f_y, c_x, c_y, s)
$\mathbf{R}, \mathbf{t}, \ \tilde{\mathbf{c}} = -\mathbf{R}^\top \mathbf{t}$	extrinsics, camera centre
$\boldsymbol{\omega} = \mathbf{K}^{-\top}\mathbf{K}^{-1}$	image of the absolute conic
$\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$	essential matrix (calibrated)
$\mathbf{F} = \mathbf{K}'^{-\top}\mathbf{E}\mathbf{K}^{-1}$	fundamental matrix
$\mathbf{e}, \mathbf{e}'$	epipoles $=$ Null $\mathbf{F}$, Null $\mathbf{F}^\top$
$\boldsymbol{\ell}' = \mathbf{F}\mathbf{x}$	epipolar line of $\mathbf{x}$

Sources: Hartley & Zisserman (2004), Ch. 6, 9.

Notation III — estimation & motion

$\mathbf{A}\mathbf{h} = \mathbf{0}$	homogeneous DLT system
$\mathbf{h} =$ smallest sing. vec.	DLT solution (SVD)
$\mathbf{T}, \mathbf{T}'$	Hartley normalizing transforms
$N \geq \frac{\log(1-p)}{\log(1-w^s)}$	RANSAC iterations
$\mathcal{E}_{\text{Samp}}$	Sampson (geometric) error
(u, v)	optical flow at a pixel
$I_x u + I_y v + I_t = 0$	flow constraint (OFCE)
$\mathbf{M} = \sum \left[\begin{smallmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{smallmatrix} \right]$	structure tensor (Harris = LK)

Sources: Hartley & Zisserman (2004), §4; Lucas & Kanade (1981).

Notation IV — learning & generative models

$\mathbf{x}, \mathbf{y}, \mathbf{z}$	input, label/target, latent code
$f_{\boldsymbol{\theta}},\ \boldsymbol{\theta}$	network with learnable parameters
$\mathcal{L},\ J = \frac{1}{N}\sum_i \mathcal{L}_i$	per-example loss, total cost
$g(\mathbf{z}, \mathbf{y}),\ \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$	generator & its latent prior
$\mathbb{E}_q[\cdot],\ \mathrm{KL}(q \,\|\, p)$	expectation, KL divergence (ELBO)
$\mathbf{T} \in \mathbb{R}^{N \times d}$	a set of N tokens of width d
$\mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}$	scaled dot-product self-attention
$\varepsilon_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$	diffusion noise prediction

Sources: Torralba et al. (2024), notation; Ch. 9–14, 26, 32.

How to survive these slides

Proof

Self-contained derivations, ending in ■. If you black out mid-proof, that little square is where you come to.

Boxed result

Memorize these. Pull one out and the entire field quietly collapses back into a pile of pixels.

Real talk

The honest asides: intuition, caveats, and whatever the formula politely declined to mention.

Slides marked "Worked" contain real numbers — we did the arithmetic so you can nod as if you would have. The grey text at the bottom credits the people who worked this out decades before any of us showed up.

Sources: Deck conventions.

Twenty-eight hours. One pipeline.

From a pinhole in a dark room to a model that paints from noise —
and the single thread of ideas connecting them.

ES666 Computer Vision · Shanmuganathan Raman · IIT Gandhinagar