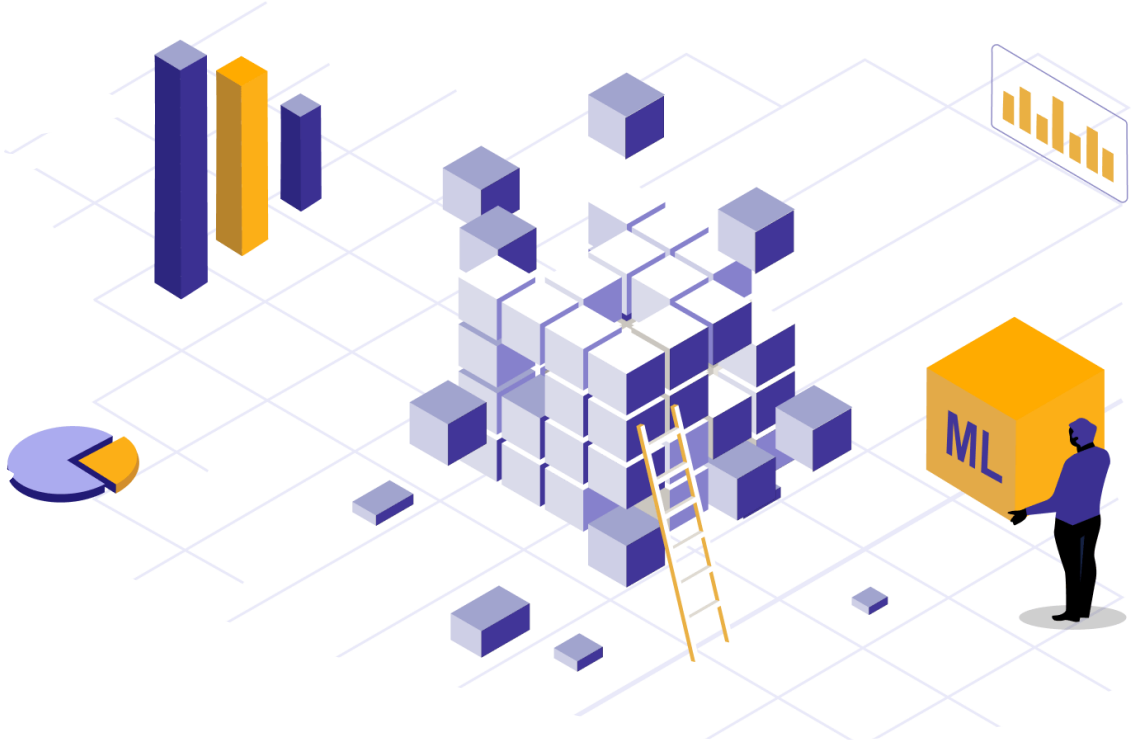


Feature Engineering

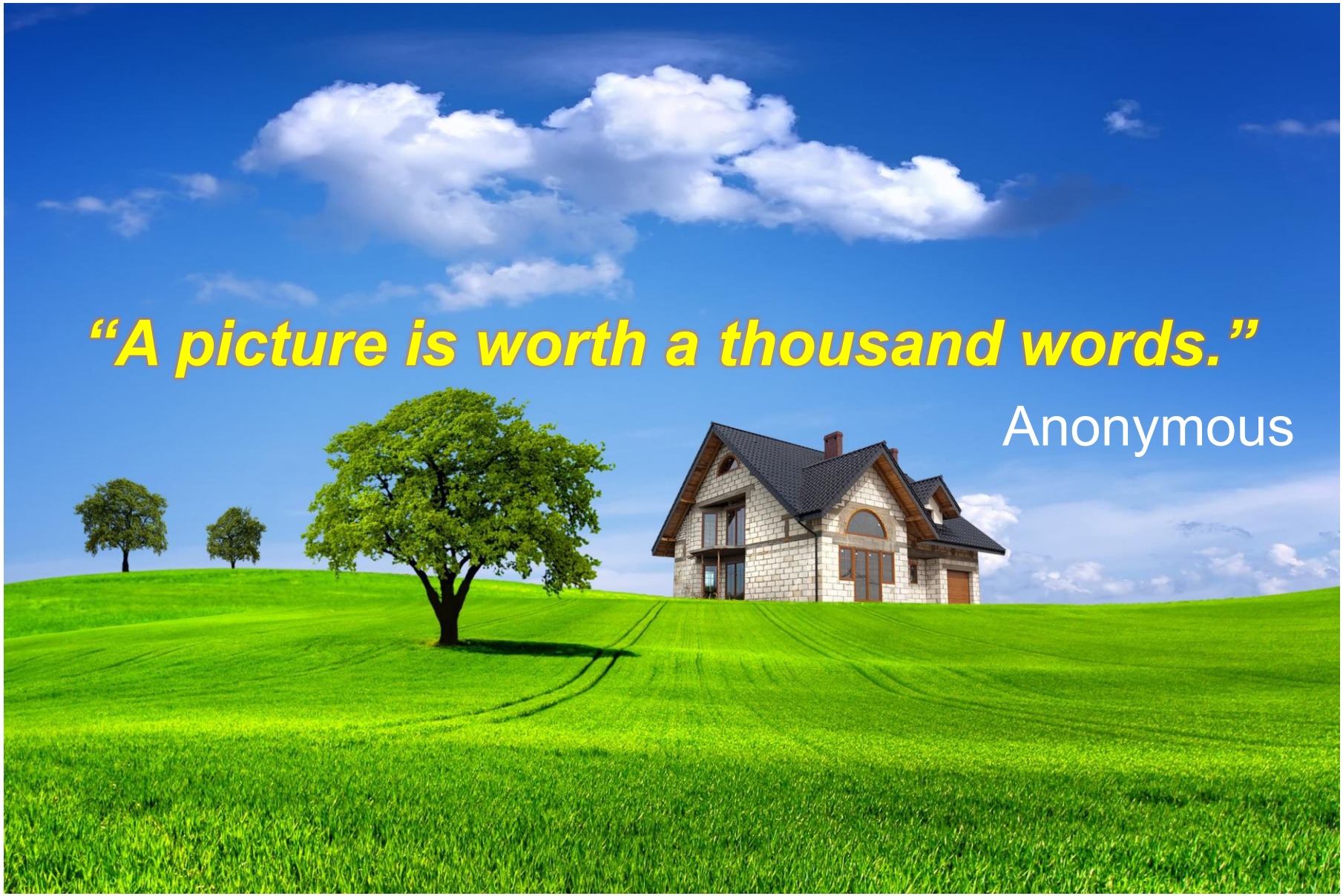
Part II

Nguyen Ngoc Thao

nnthao@fit.hcmus.edu.vn

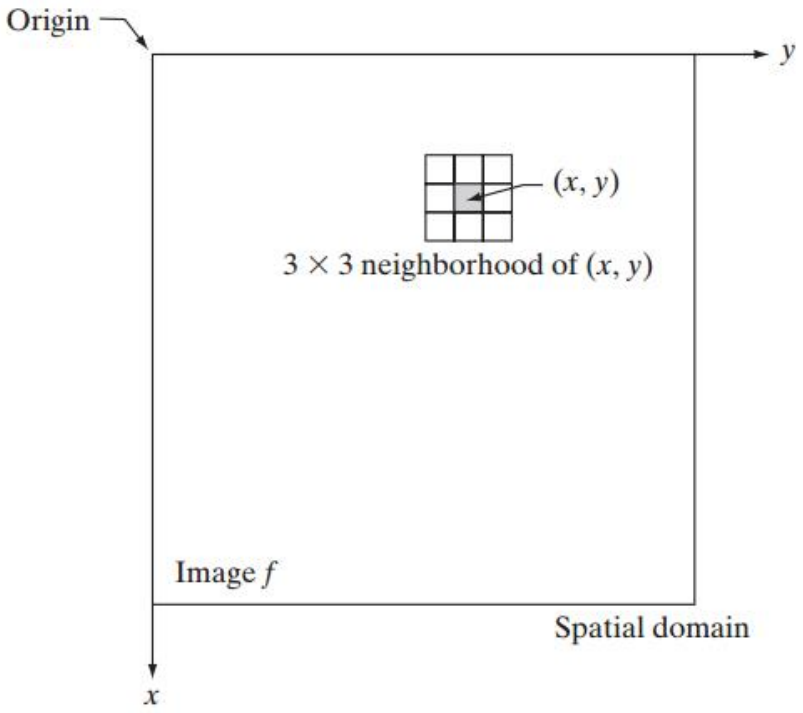
Content outline

- Conventional image features
- Modern visual embeddings
- Similarity metrics for vision


“A picture is worth a thousand words.”
Anonymous

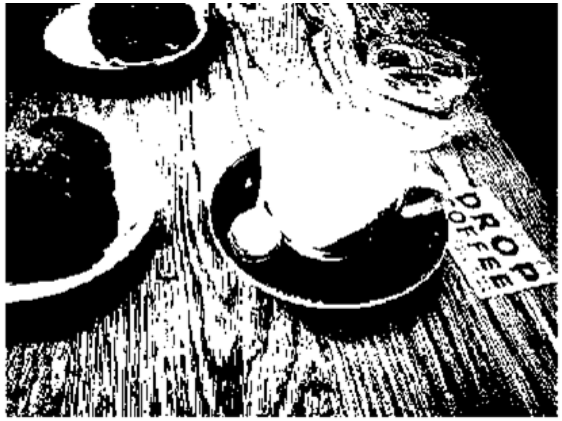
Digital images: A definition

- An image is a 2D function, $f(x, y)$, whose amplitude at each coordinate (x, y) represents the image intensity at that point.



Digital images: Channels

Black and White
1 channel



RGB
3 channels

Grayscale
1 channel

Red channel

Green channel

Blue channel

The electromagnetic spectrum

- Regular images capture only the visible spectrum, yet cover many applications.

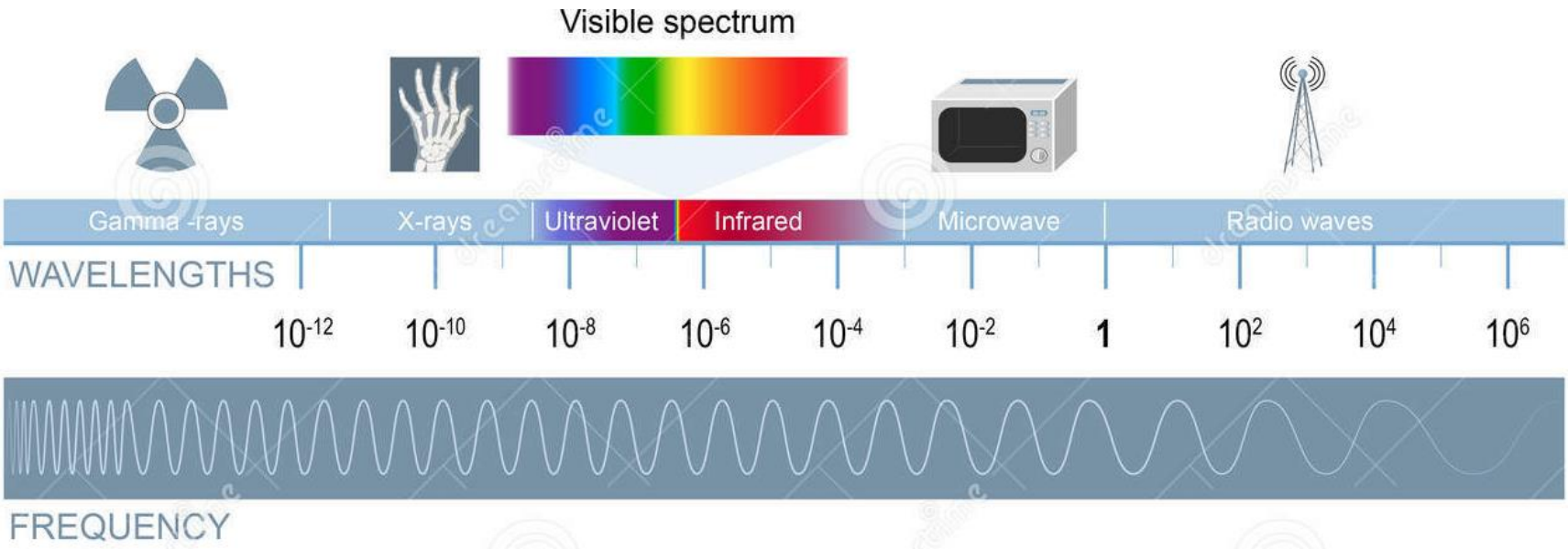


Image features

Histogram

- The histogram of a digital image with intensity levels in the range $[0, L-1]$ is a discrete function of the form

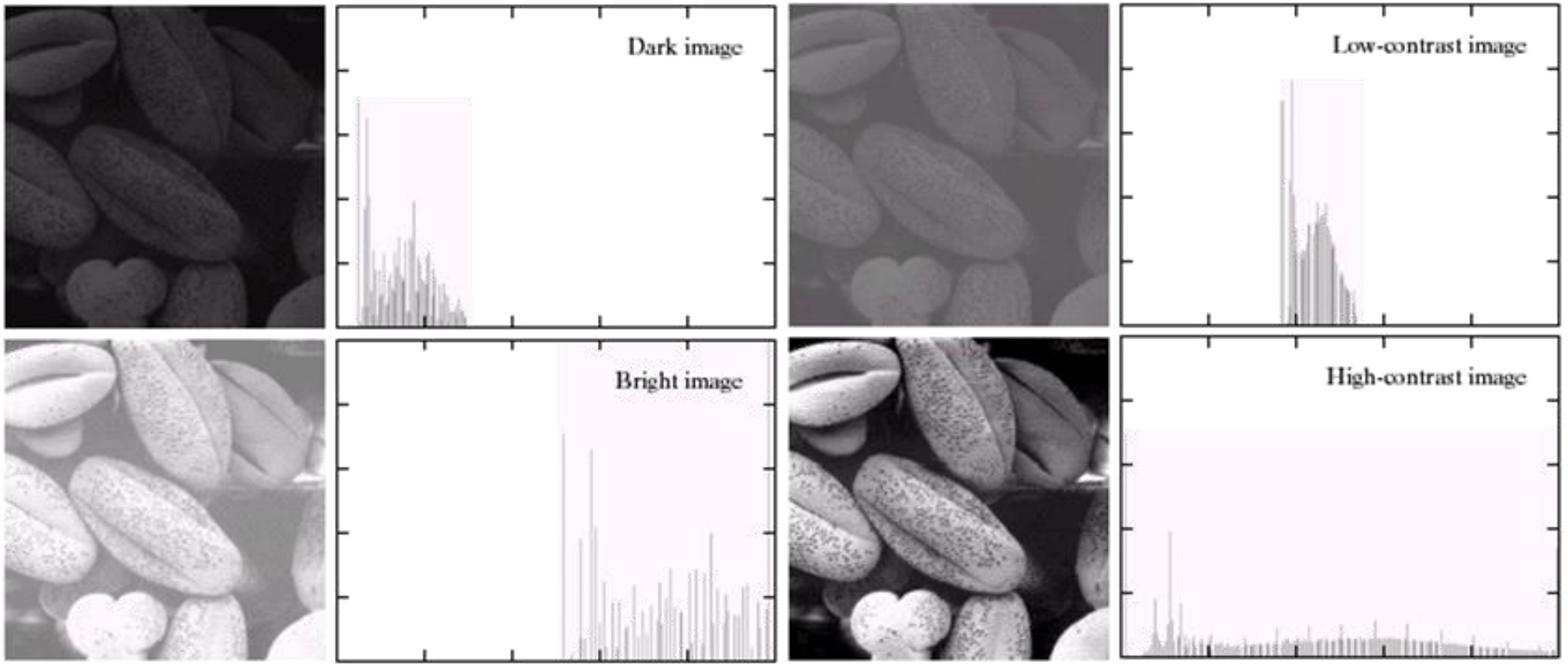
$$\boldsymbol{h(r_k) = n_k}$$

  - n_k is the number of pixels in the image with intensity value r_k.

- It is common practice to normalize a histogram: $\boldsymbol{p(r_k) = \frac{n_k}{MN}}$
  - M and N are the row and column dimensions of the image
  - $k = 0, 1, 2, \ldots, L-1.$
- $p(r_k)$ estimates the probability of occurrence of the intensity level r_k in an image.

a b
c d

Four image types: dark, light, low contrast, high contrast, and their corresponding histograms



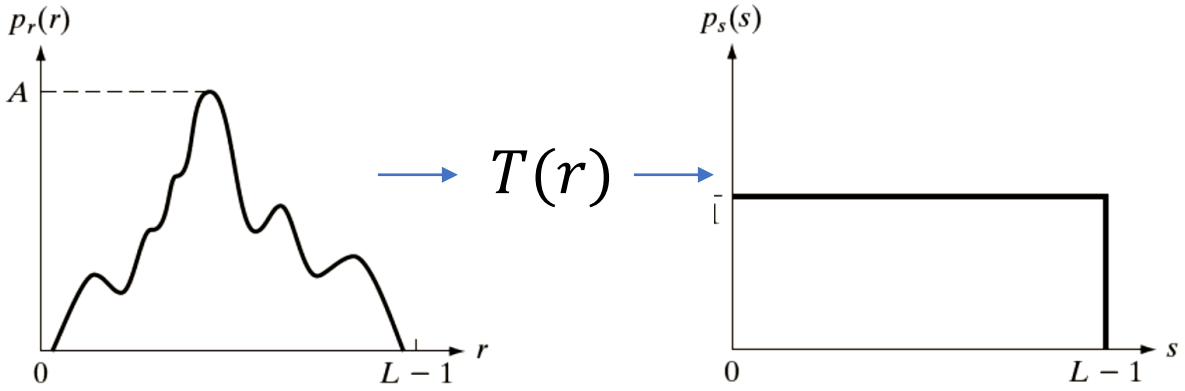
Histogram equalization

- Each pixel in the input image with intensity r_k is mapped into a corresponding pixel of level s_k in the output image.

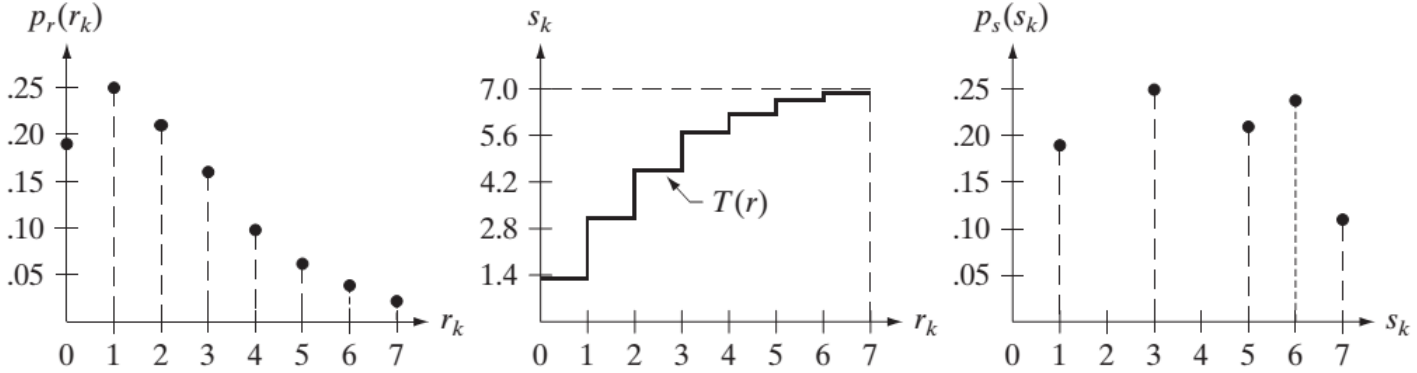
$$\boldsymbol{s_k = T(r_k) = (L-1)\sum_{j=0}^{k} p_r(r_j)}$$

  - where $k = 0, 1, 2, \ldots, L-1$

- The transformation function T is determined **automatically** to produce an output image that has a uniform histogram.

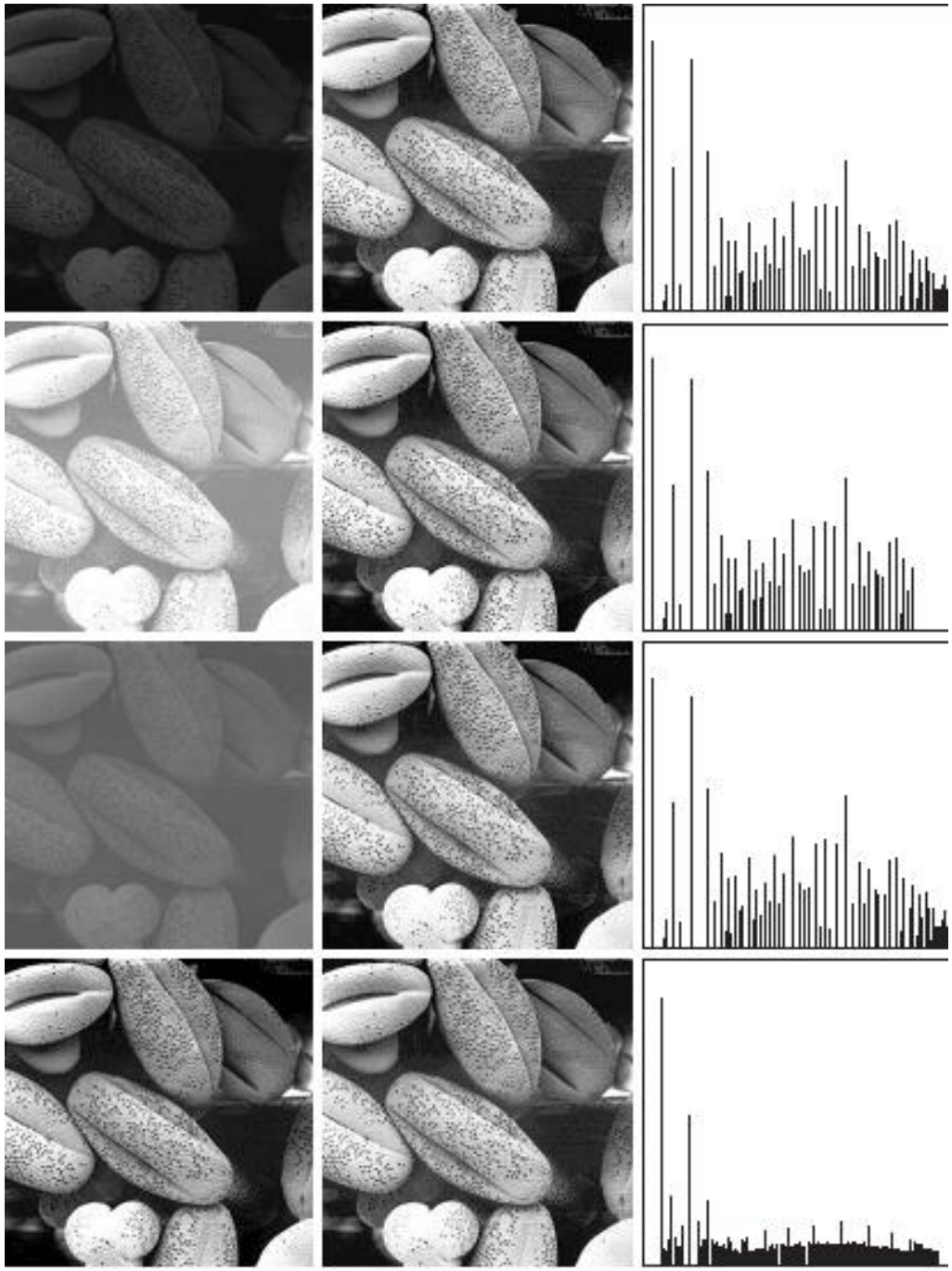


a b (a) An arbitrary PDF. (b) Result of applying histogram equalization to all intensity levels, r. The resulting intensities, s, have a uniform PDF, independently of the form of the PDF of the r's.



c d e Illustration of histogram equalization of a 3-bit (8 intensity levels) image. (c) Original histogram. (d) Transformation function. (e) Equalized histogram.

Note: *Discrete histogram equalization cannot guarantee a uniform histogram, but it generally spreads the input histogram.*

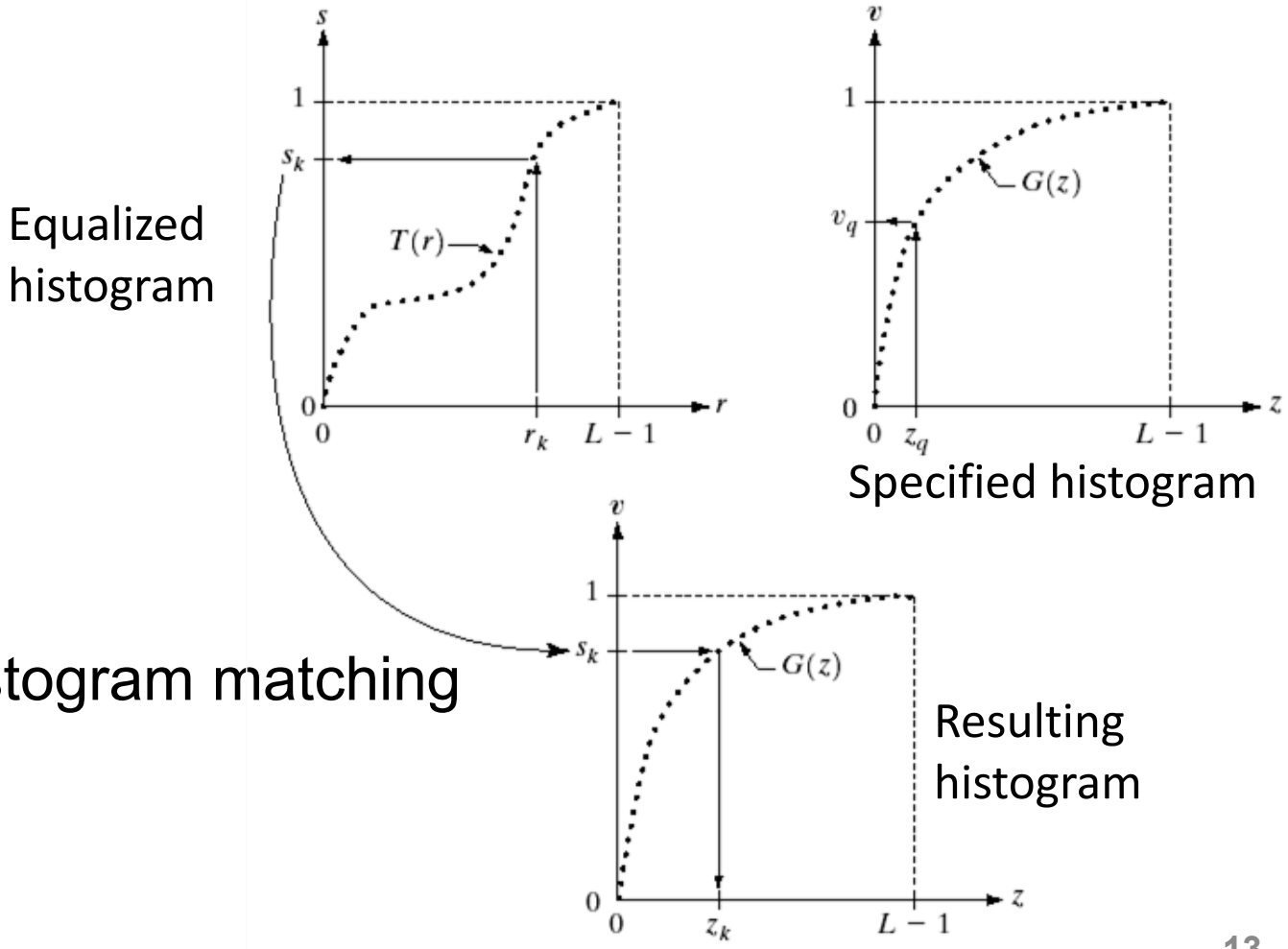


a b c

Left column: original images.
Center column: corresponding histogram-equalized images.
Right column: histograms of the images in the center column.

Histogram specification

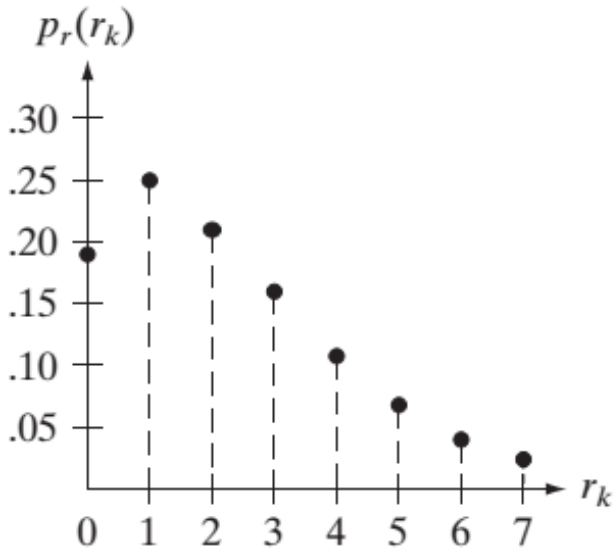
- Create an image whose histogram shape is specified

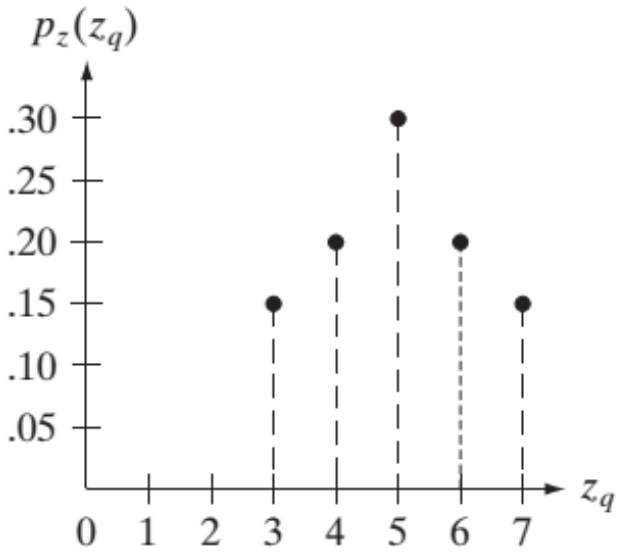


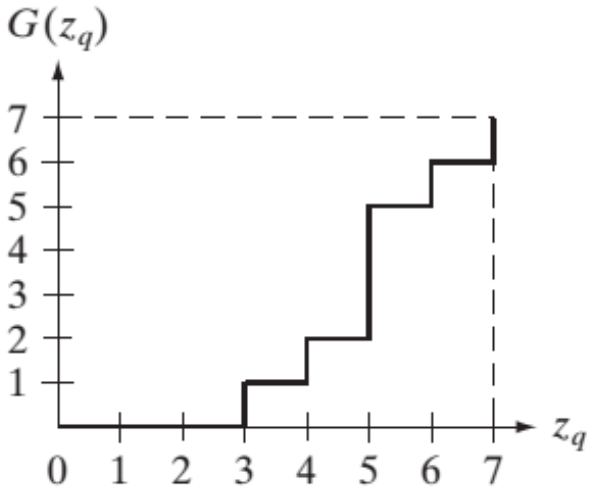
- Also called histogram matching

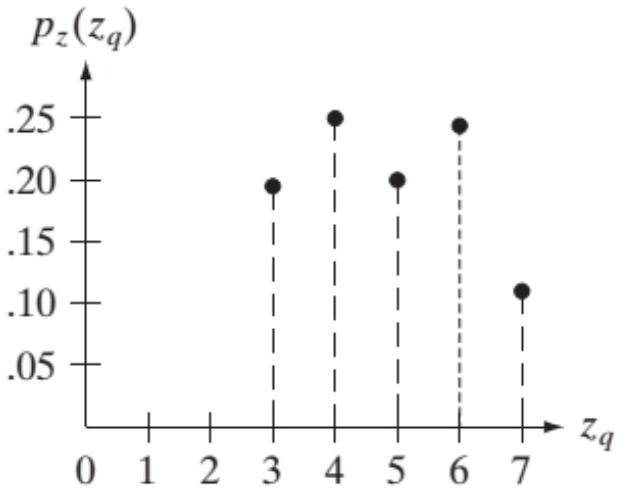
a b
c d

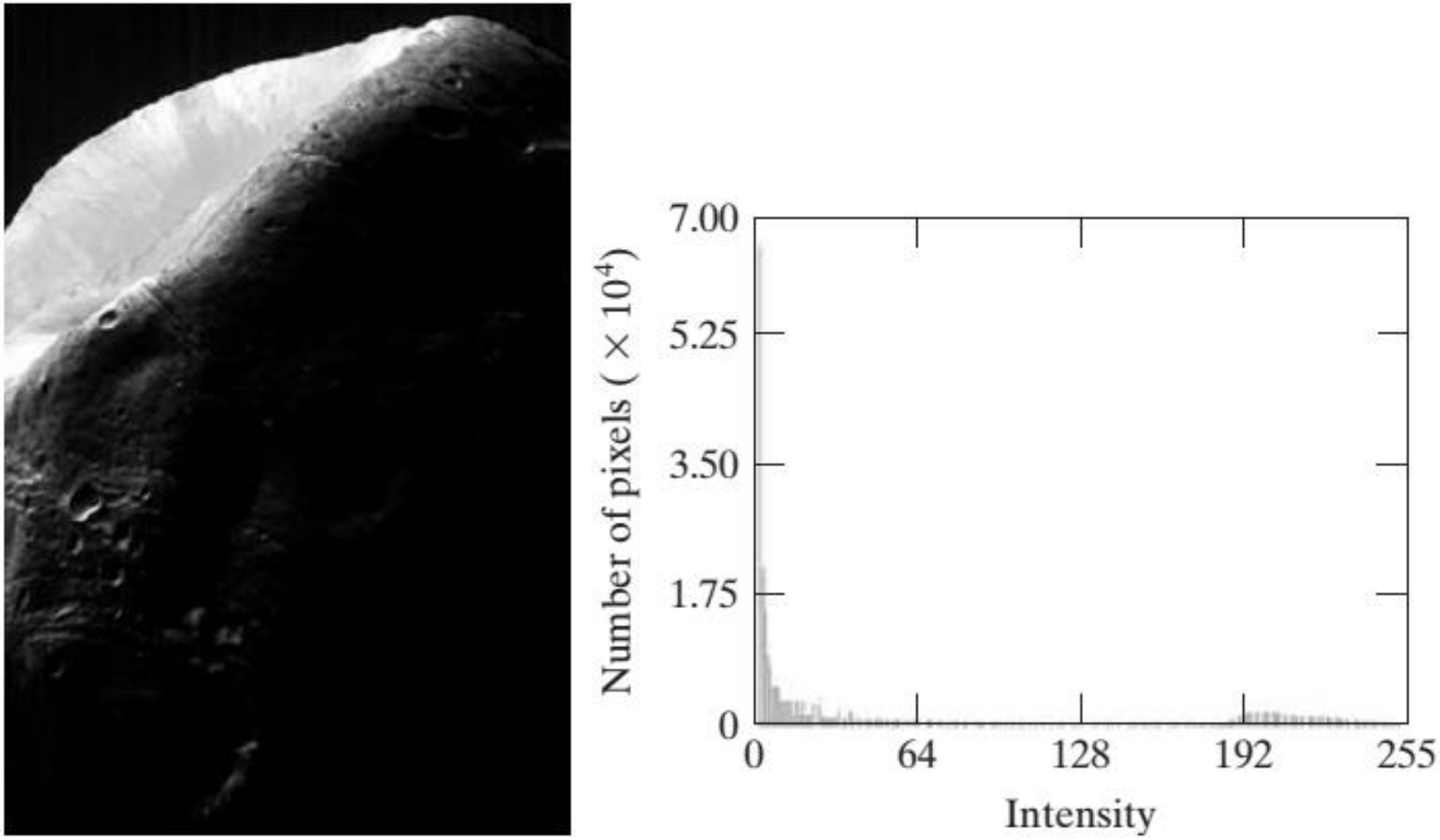
(a) Histogram of a 3-bit image.
(b) Specified histogram.
(c) Transformation function obtained from the specified histogram.
(d) Result of performing histogram specification.
Compare (b) and (d).





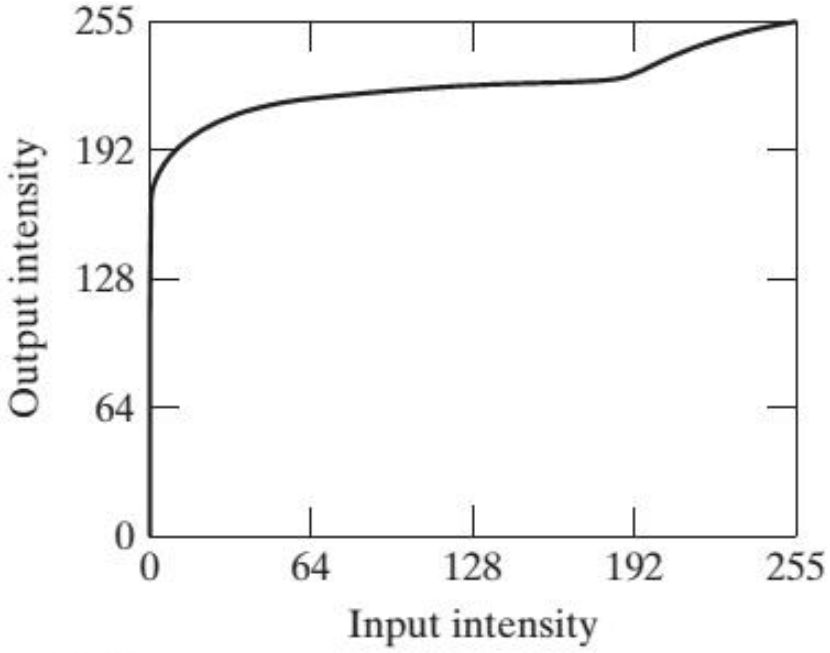


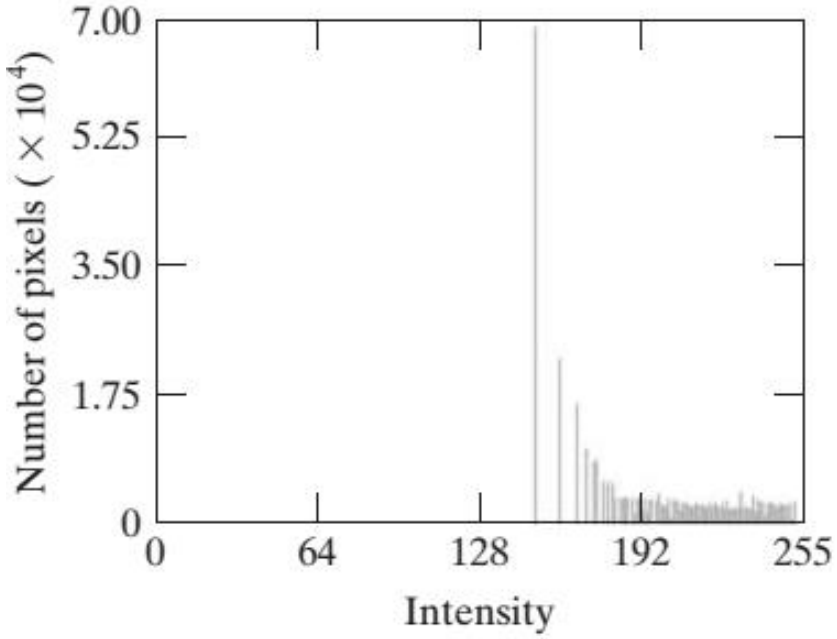




a b

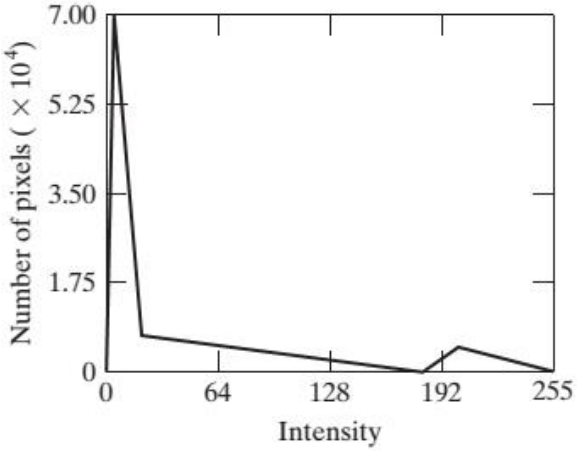
(a) Image of the Mars moon Phobos taken by NASA's Mars Global Surveyor.
(b) Histogram. (Original image courtesy of NASA.)

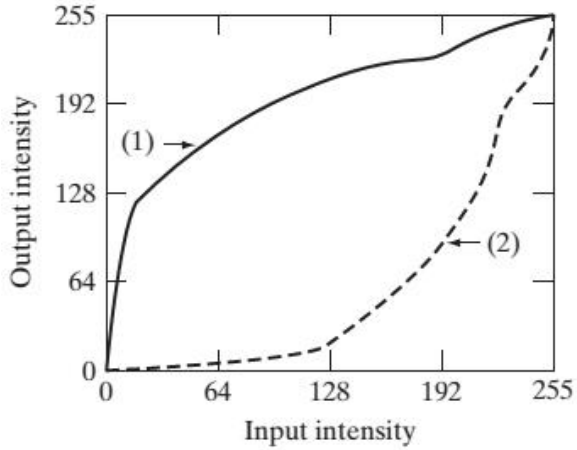




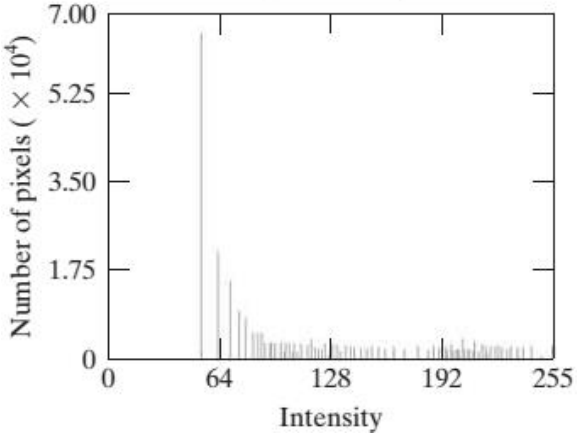
a b
c

(a) Transformation function for histogram equalization.
(b) Histogram-equalized image (note the washed-out appearance).
(c) Histogram of (b).





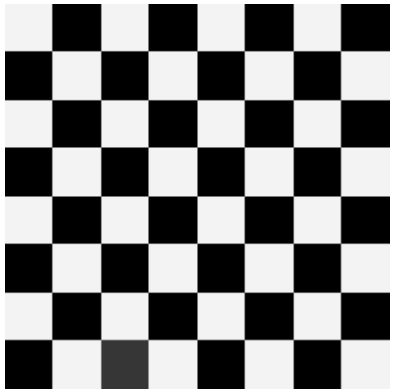
a c
b
d

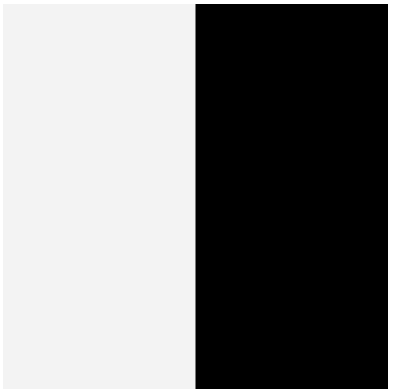


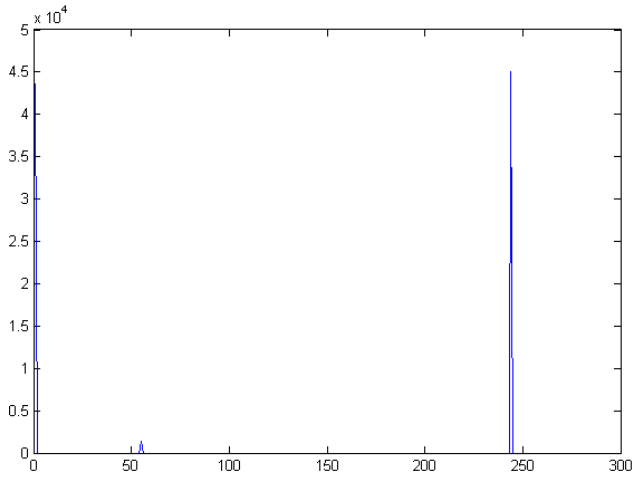
(a) Specified histogram.
(b) Transformations.
(c) Enhanced image using mappings from curve (2).
(d) Histogram of (c).

Can histogram be a good feature?

- The histogram, along with related statistical measures, is considered a weak feature due to its overly general nature.

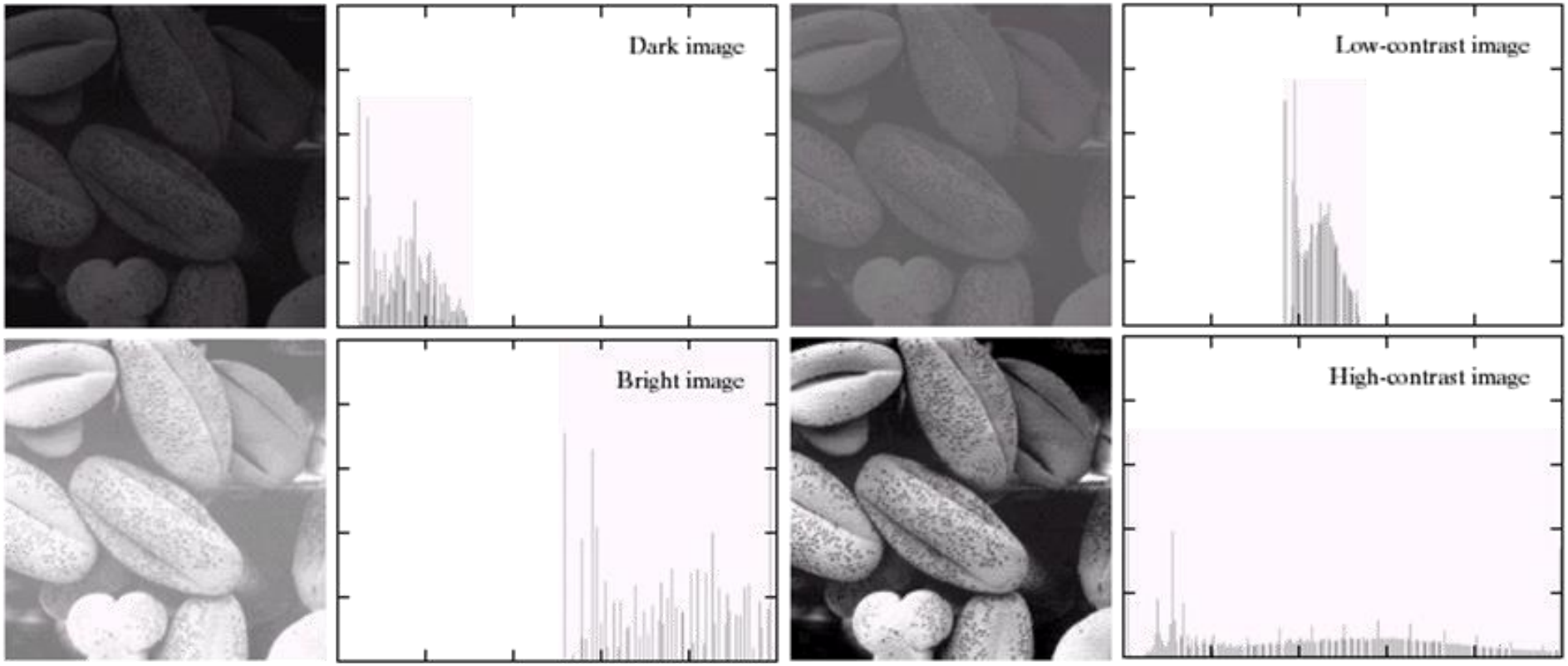






Can intensity be a good feature?

- Intensity is sensitive to noise and illumination changes.
- The relation between intensities may reduce the effect of illumination changes yet the effect of noise remains.



Can color be a good feature?

- No. Color is sensitive to noise and illumination changes

- Some color spaces (e.g., CIE-Lab, CIE-Luv) are empirically shown to be more discriminative than the others (e.g., RGB).

Edge features

Edge detection

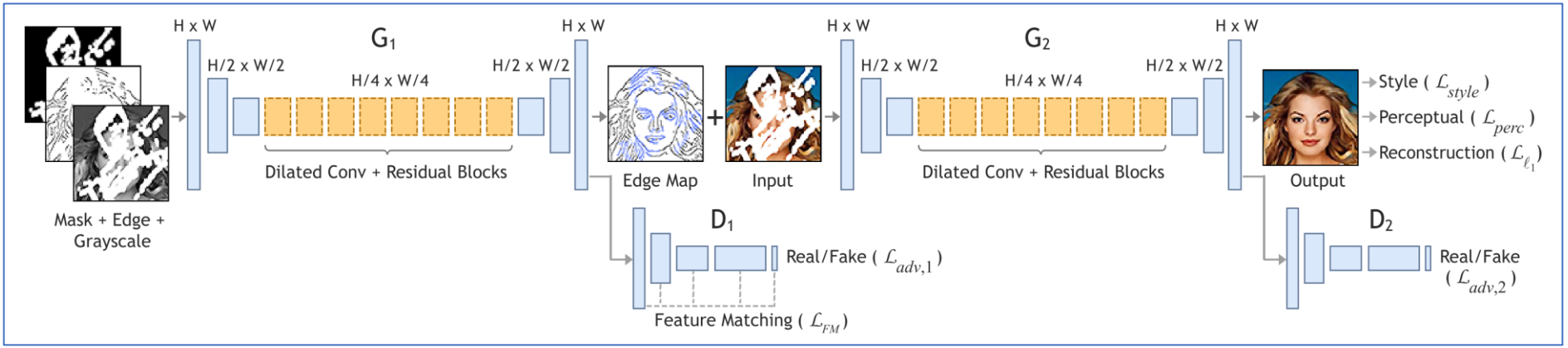
- Edges are defined as curves at which the pixel brightness changes sharply or, more formally, has discontinuities.
- Edge detection includes a variety of mathematical methods that aim at identifying edges.

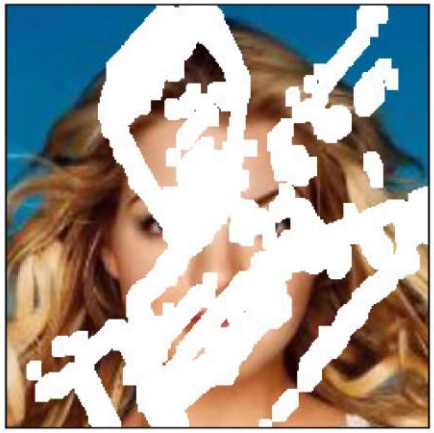
Input image

Edges detected by Prewritt detector

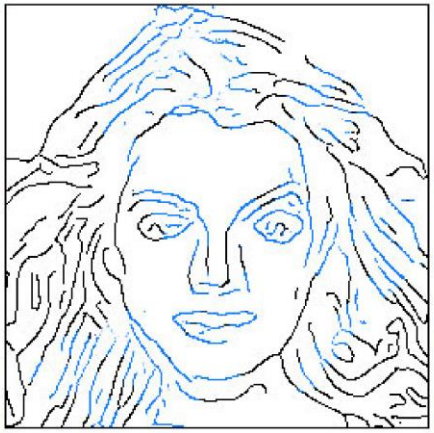
Image credit: RoboFlow

Edges as augmented features





Input image



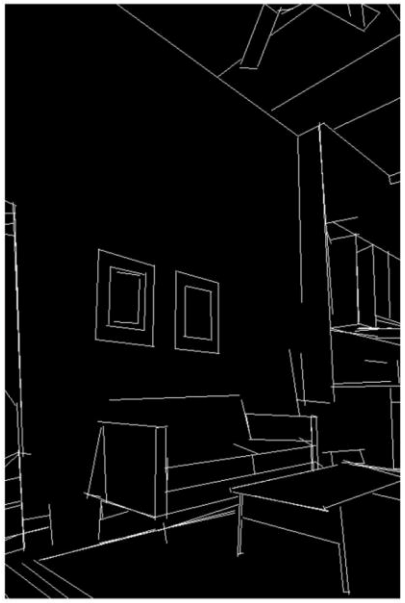
Edge map

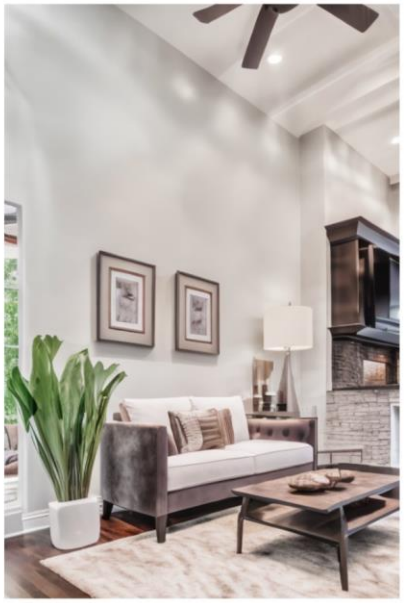
Resulting image

Edges drawn in black are computed (for the available regions) using Canny edge detector; whereas edges shown in blue are hallucinated (for the missing regions) by the edge generator network.

Nazeri, K. "EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning." arXiv preprint arXiv:1901.00212 (2019).

Edges as augmented features





Input image

Edge map

Resulting image

Ye, Hu, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models." arXiv preprint arXiv:2308.06721 (2023).

Feature descriptors

Gradient-based features

- Consider an image $I(x, y)$.
- The gradient $\nabla I = \left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}\right)$ of image I is a vector that show how quickly the intensity changes at each pixel.

Two types of gradients, with blue arrows to indicate the direction of the gradient. Light areas indicate higher pixel values

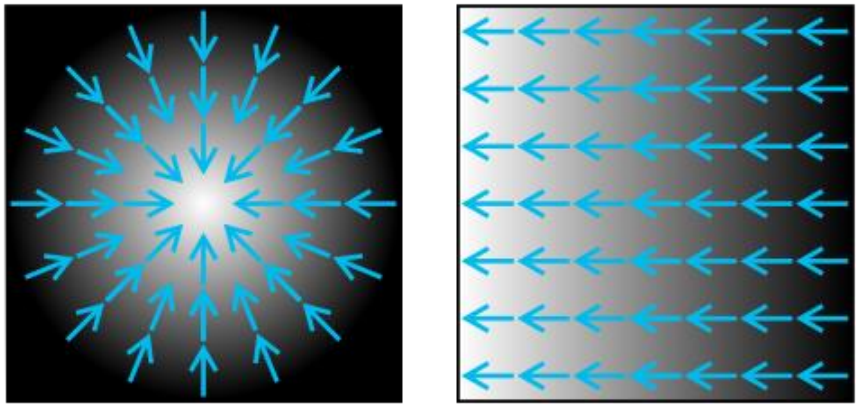
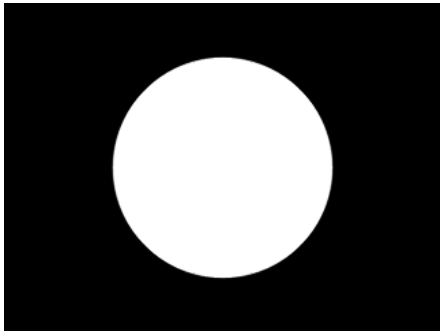


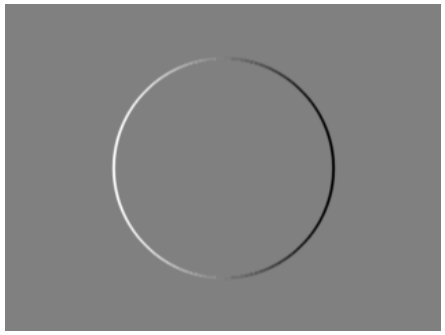
Image credit: Wikipedia

Gradient-based features

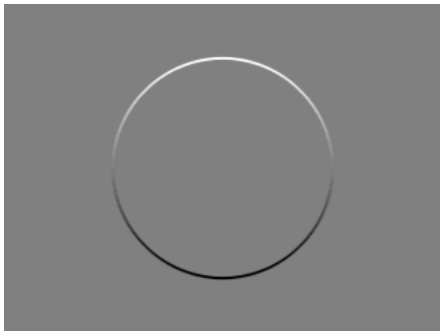
- Gradient's magnitude represents the strength of the change, while its orientation shows the direction of change.



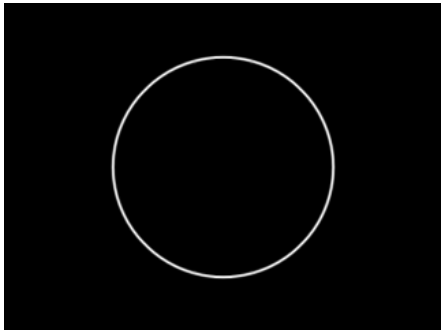
Original image



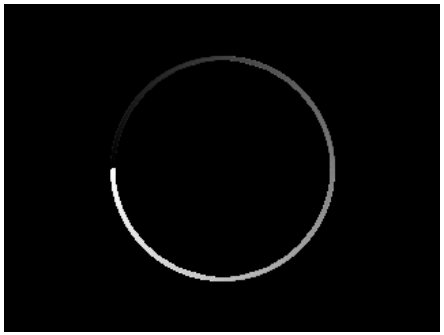
x-derivative



y-derivative



magnitude



orientation

Image credit: stack overflow

Gradient-based features

- Feature descriptors based on gradients encode the edges and textures in local image regions.

✓ Shape information: Edge orientation encodes geometry.

✓ Robust to illumination changes: Gradients focus on differences, not absolute intensity.

× Computationally heavy

Gradient-based features

- **Convolution kernels** (e.g., Sobel, Scharr, etc.) are used to estimate horizontal and vertical gradients.

Image

Sobel filter

-1	0	1
-2	0	2
-1	0	1

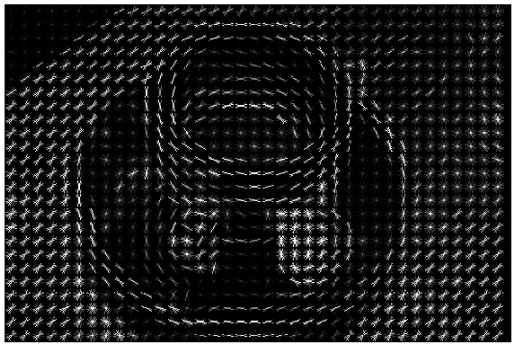
Gx

1	2	1
0	0	0
-1	-2	-1

Gy

- **Histogram of Oriented Gradients (HOG)** counts gradient orientations in local cells.

Image



HOG

Image credit: [1](), [2]()

Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, pp. 886-893. Ieee, 2005.

Gradient-based features

- **SIFT** (Scale-Invariant Feature Transform): Compute gradient histograms around key points to produce invariant descriptors.
- **SURF** uses Haar-wavelet responses (fast approximation of gradients).
- **CNN first-layer filters** (implicit gradient features): Early convolutional filters often learn edge detectors, which are effectively gradient-based.

Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60.2 (2004): 91-110.

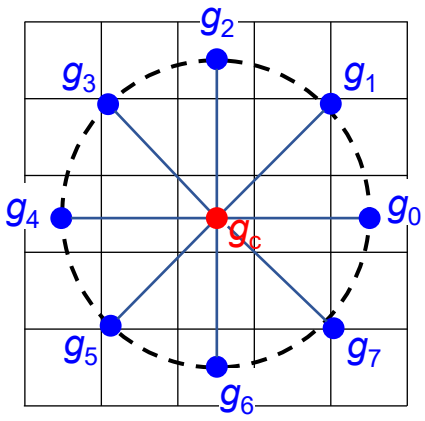
Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. "Surf: Speeded up robust features." In European conference on computer vision, pp. 404-417. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

Local Binary Patterns (LBP)

- Local Binary Patterns (LBP) is a texture operator that encodes the local information around each pixel.

$$\mathrm{LBP}_{P,R}(x,y) = \sum_{p=0}^{P-1} s(g_p - g_c)\, 2^p,$$

$$s(z) = \begin{cases} 1 & z \geq 0 \\ 0 & otherwise \end{cases}$$



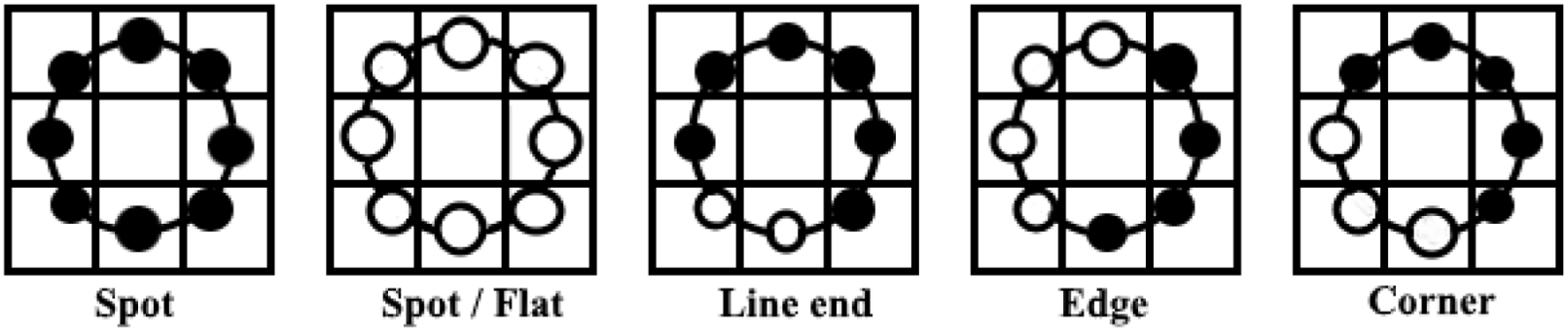
R: radius of the neighborhood, P: number of neighbors

g_c, g_p : the gray value of the center pixel and of p^{th} neighboring pixels

Ojala, Timo, Matti Pietikainen, and Topi Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." IEEE Transactions on pattern analysis and machine intelligence 24.7 (2002): 971-987.

Local Binary Patterns (LBP)

- LBP can detect several simple yet essential patterns.



Input image

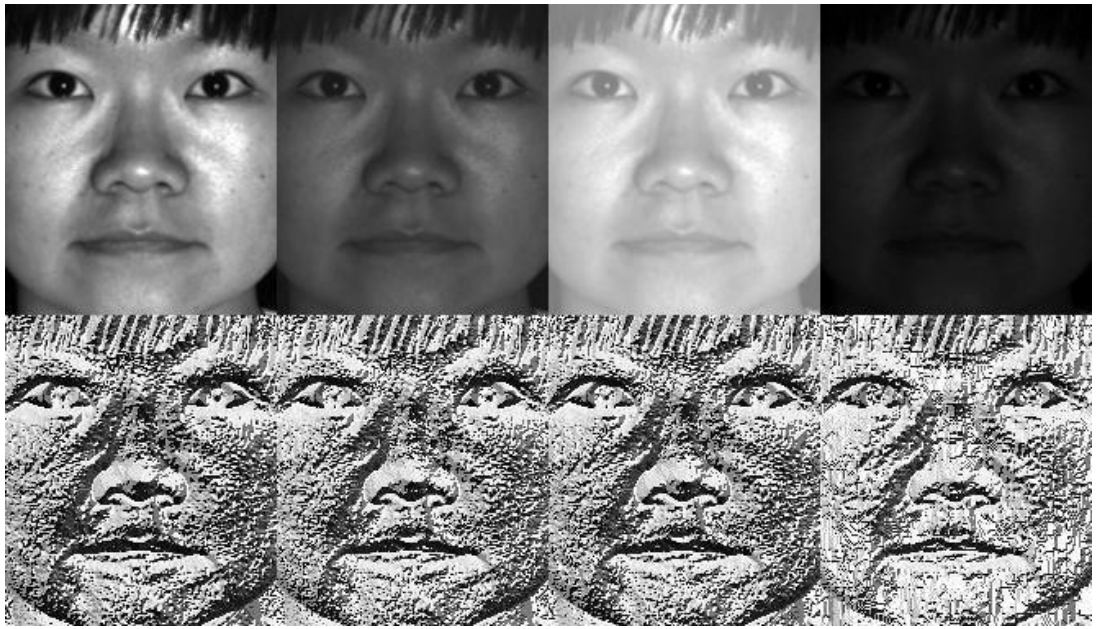
gray values in the $n \times n$ neighborhood

137	140	143
144	140	139
132	135	136

Output image

LBP: Advantages and Drawbacks

1. Invariant to any monotonic gray-level transformation
2. Nonparametric method
3. Highly discriminative against illumination changes
4. The operator is intuitive and computationally simple
5. The LBP code is quantized by its nature



LBP: Advantages and Drawbacks

- Thresholding function $s(g_p - g_c)$
  - Unstable on noisy or near-uniform regions
  - Fail to deal with image details whose $g_p - g_c$ are of the same sign yet different magnitudes.

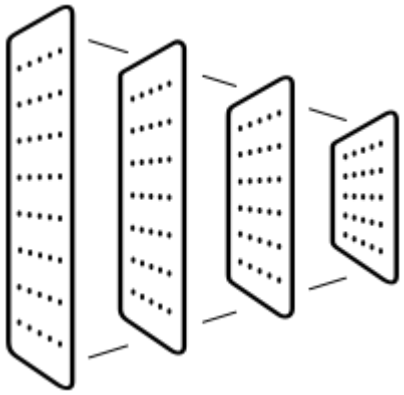
$$s(g_p - g_c) = \begin{cases} 1 & g_p - g_c \geq 0 \\ 0 & otherwise \end{cases}$$

$g_c = 29, g_p = 30$ $s(g_c - g_p) = 0$

$g_c = 30, g_p = 30$ $s(g_c - g_p) = 1$

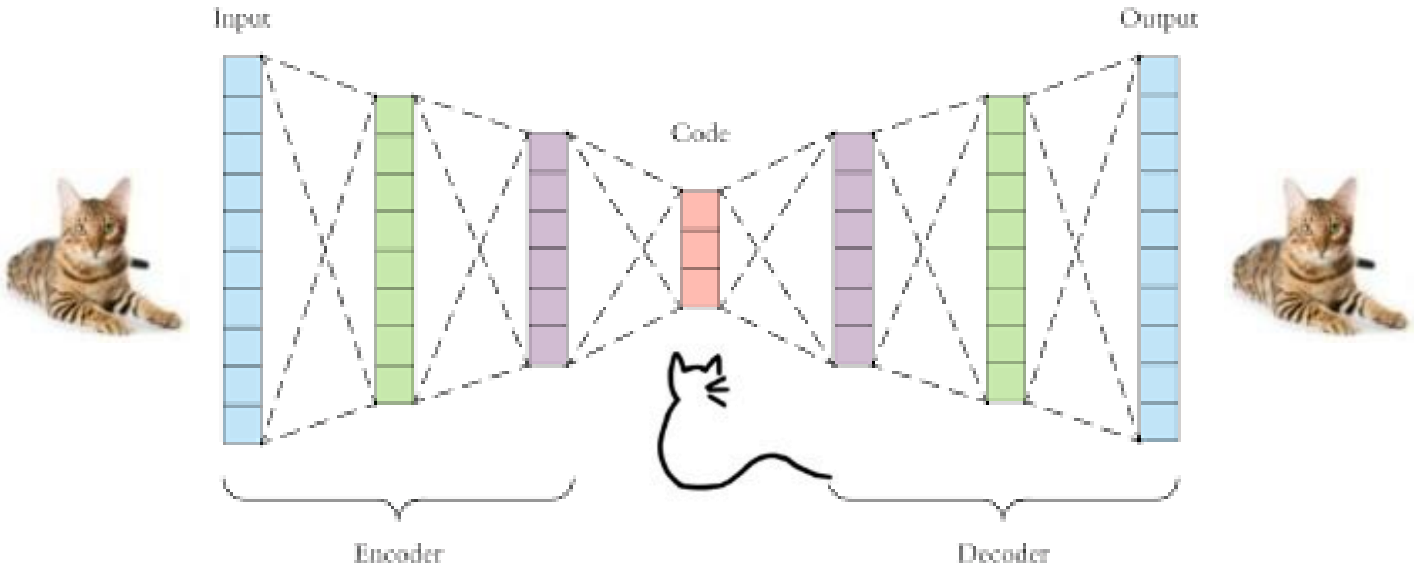
- The feature vectors are usually high dimensional.
  - $LBP_{8,R}$ has 2^8 (256) dimensions.

Visual Embeddings



Features from an Autoencoder

- An autoencoder is a neural network that is trained to attempt to copy its input to its output.

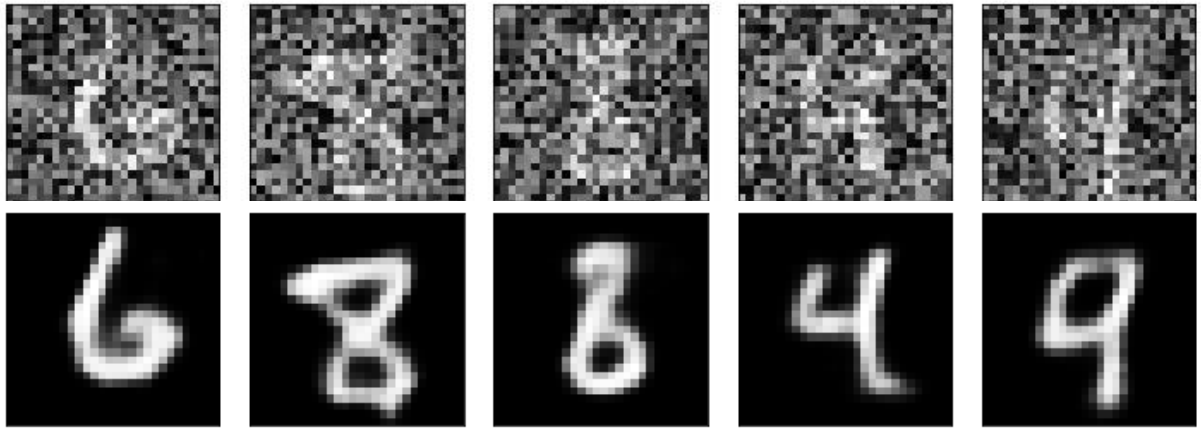


- Once trained, the decoder is discarded, and the encoder is used as needed to create compact representations of input.

Autoencoder: Applications

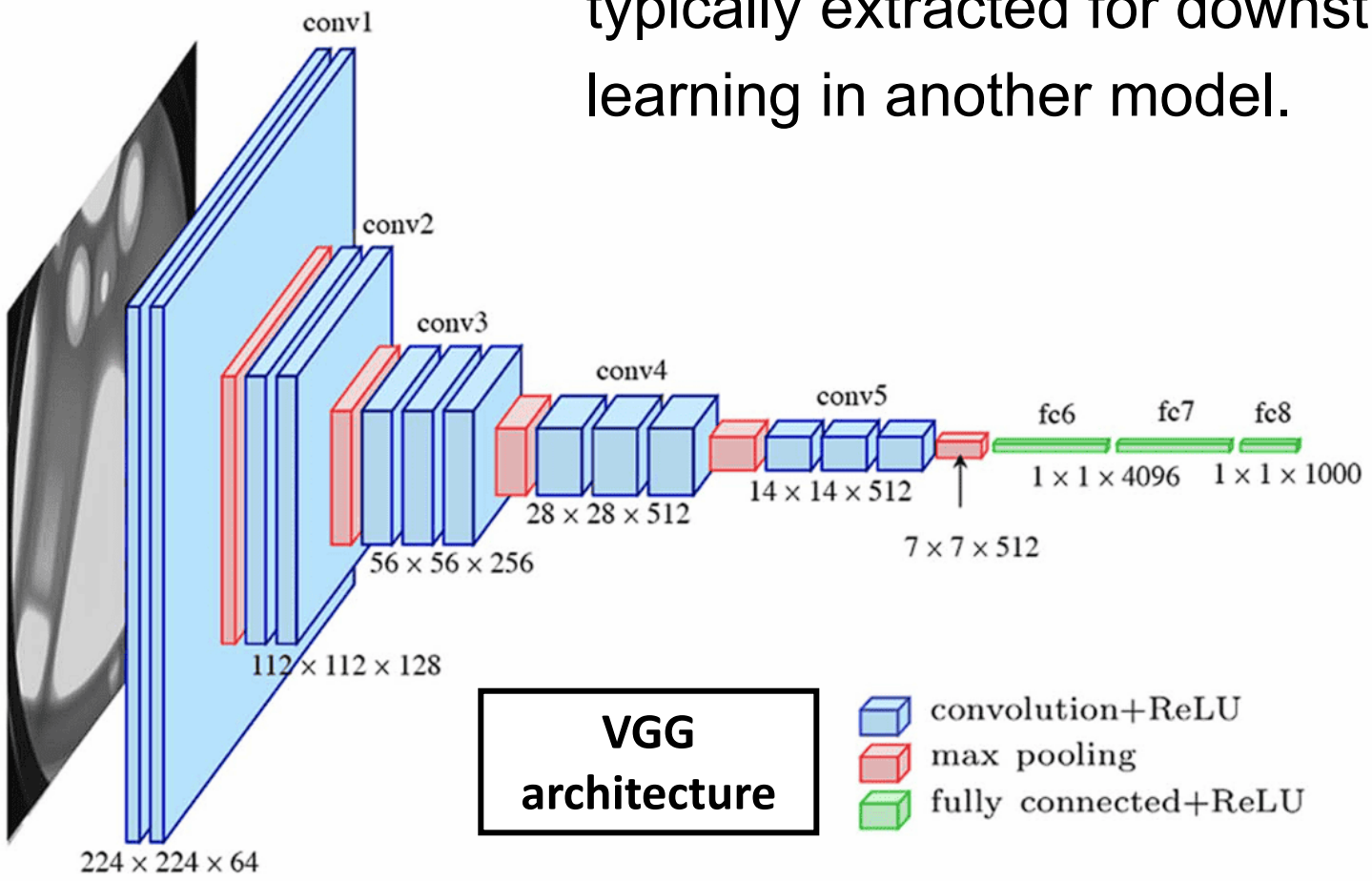
- Image colorization

- Image denoising



Features from a CNN

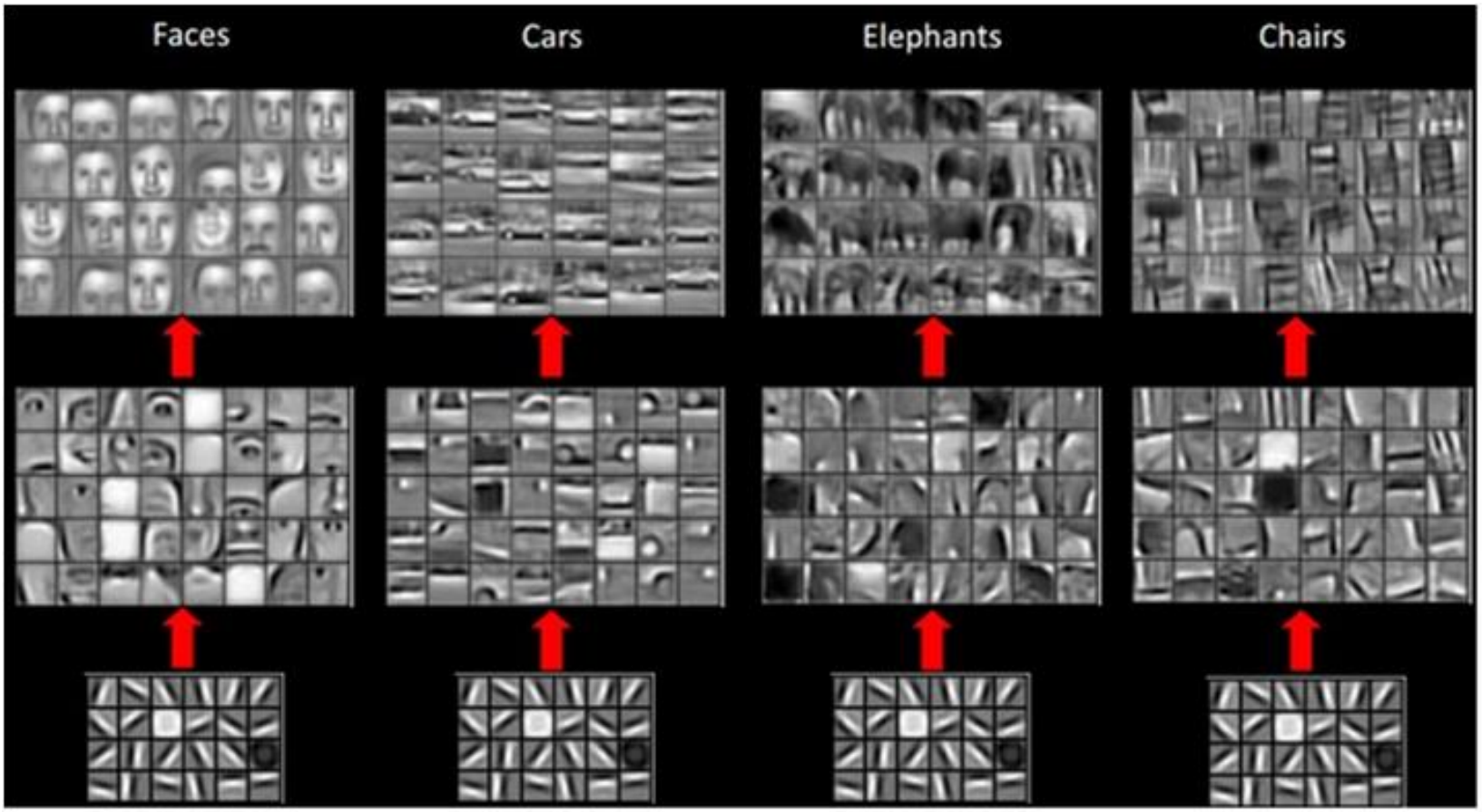
- Features from the final CNN layer are typically extracted for downstream learning in another model.



VGG architecture

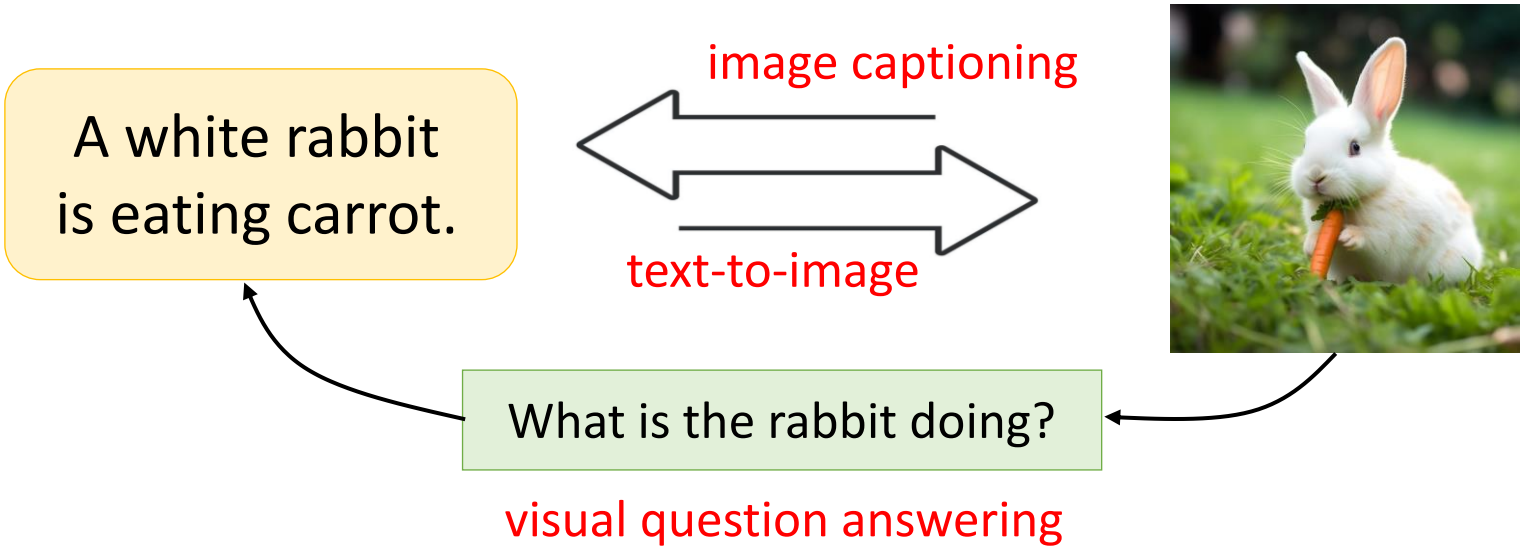
Features from a CNN

- CNN extracts features from the input image in a hierarchical way, from low-level cues to more abstract shapes.



Vision-Language Model (VLM)

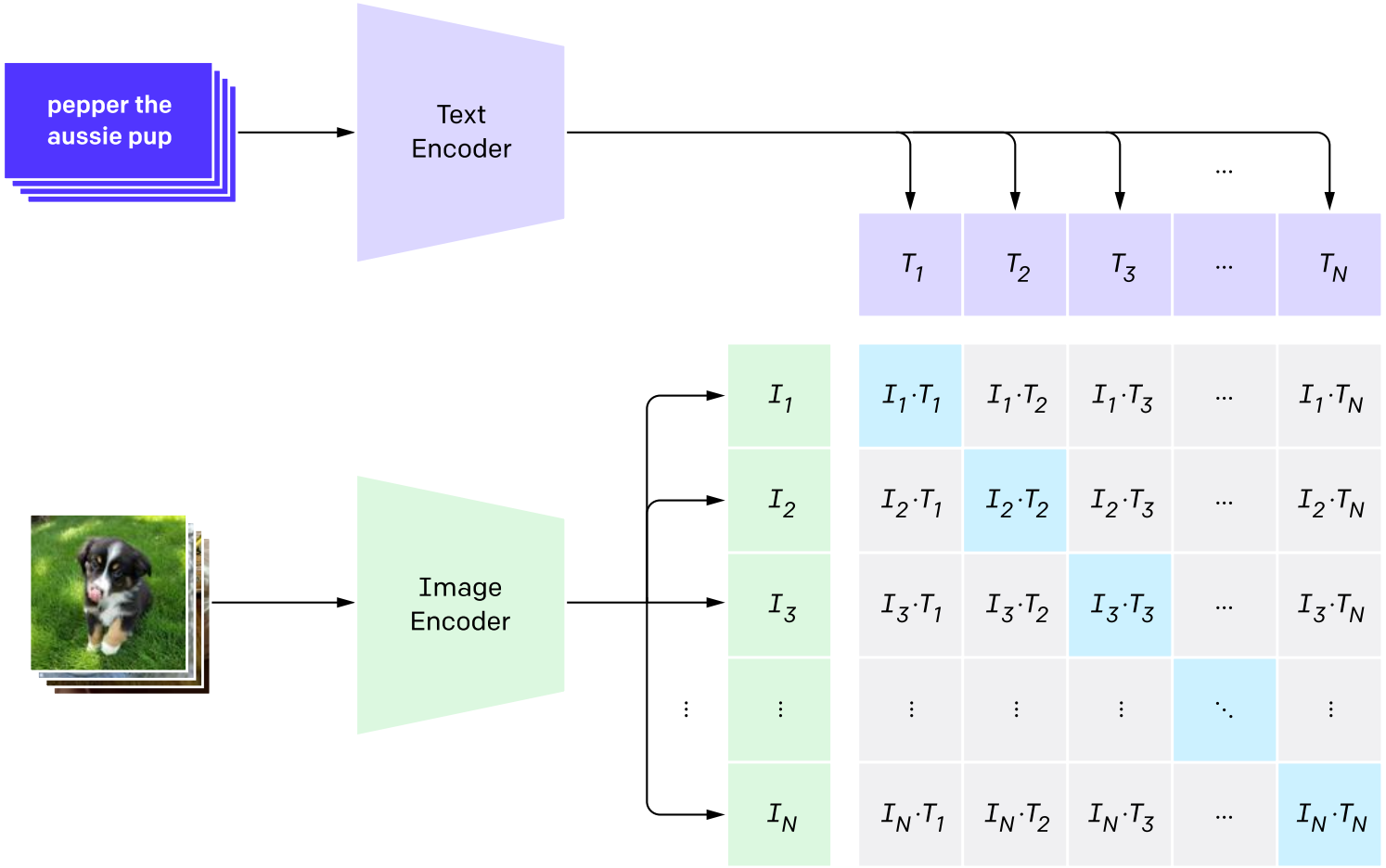
- VLM jointly trains models on both visual and textual data, enabling them to understand and generate contextually relevant outputs across both modalities.
- It serves as a base for tasks requiring a deep understanding of the relationship between images and text.



OpenAI CLIP model

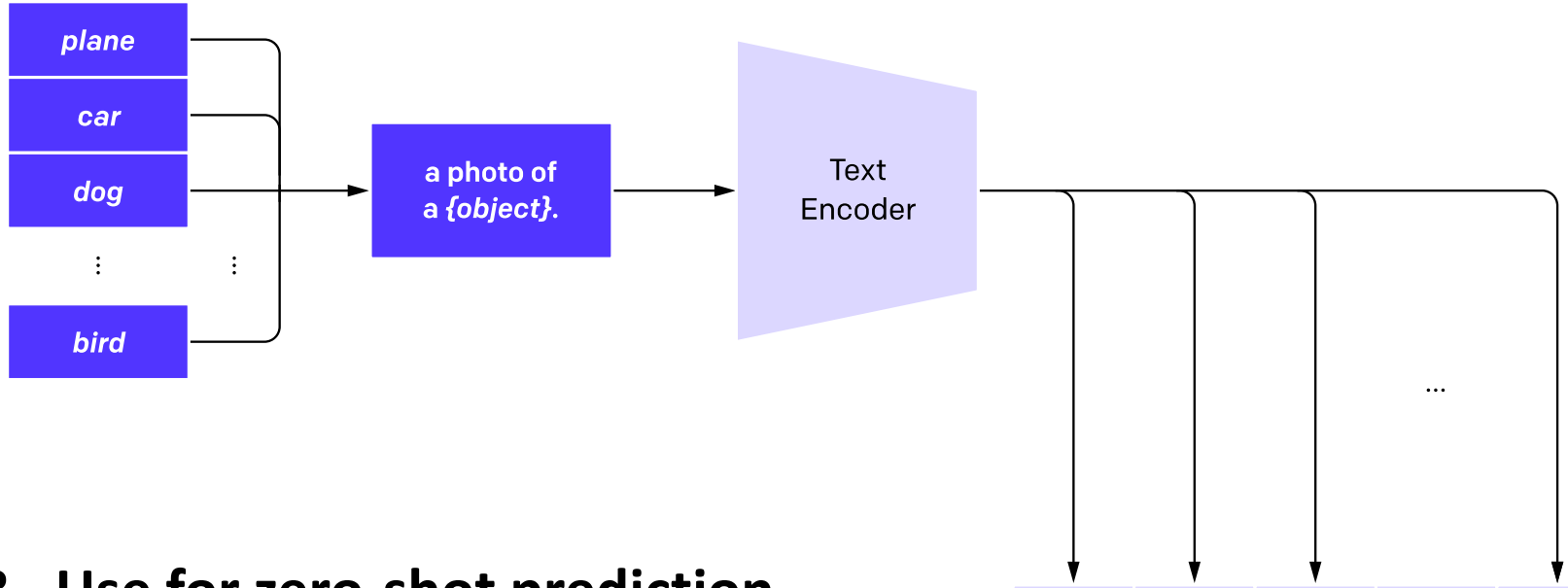
- CLIP (Contrastive Language-Image Pre-Training) is a neural network trained on diverse (image, text) pairs.
- It demonstrates strong zero-shot ability, closely matching the original ResNet50's on ImageNet without additional training
  - CLIP does not rely on the 1.28 million labeled examples.

1. Contrastive pre-training

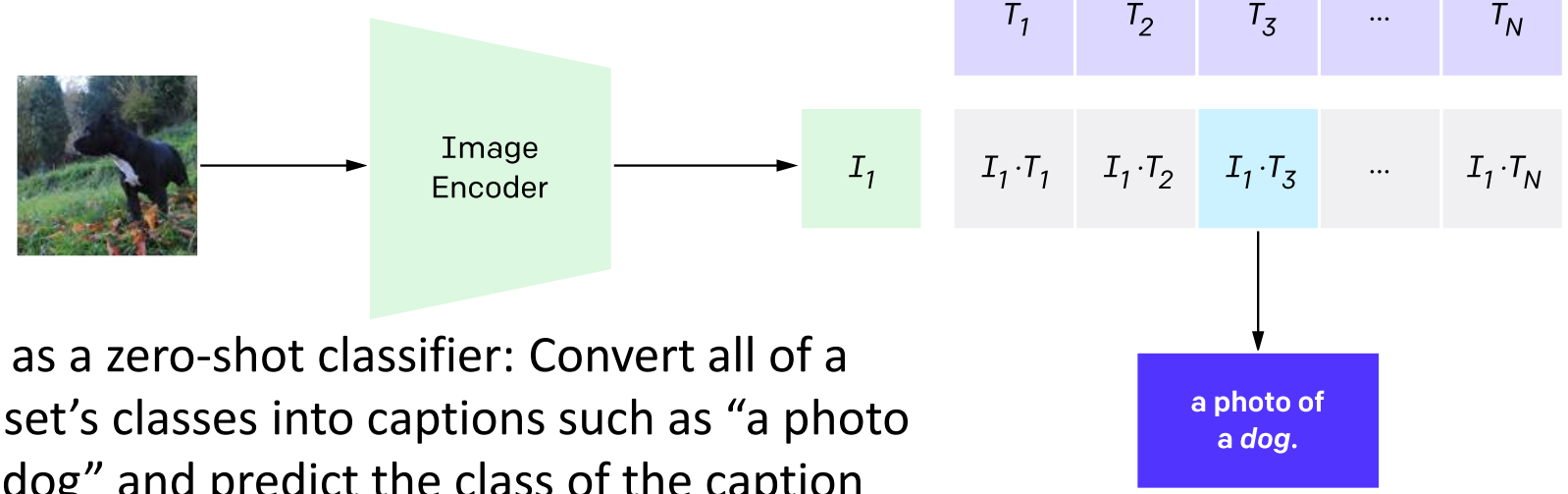


CLIP pre-trains an image encoder and a text encoder to predict which images were paired with which texts in the dataset.

2. Create dataset classifier from label text

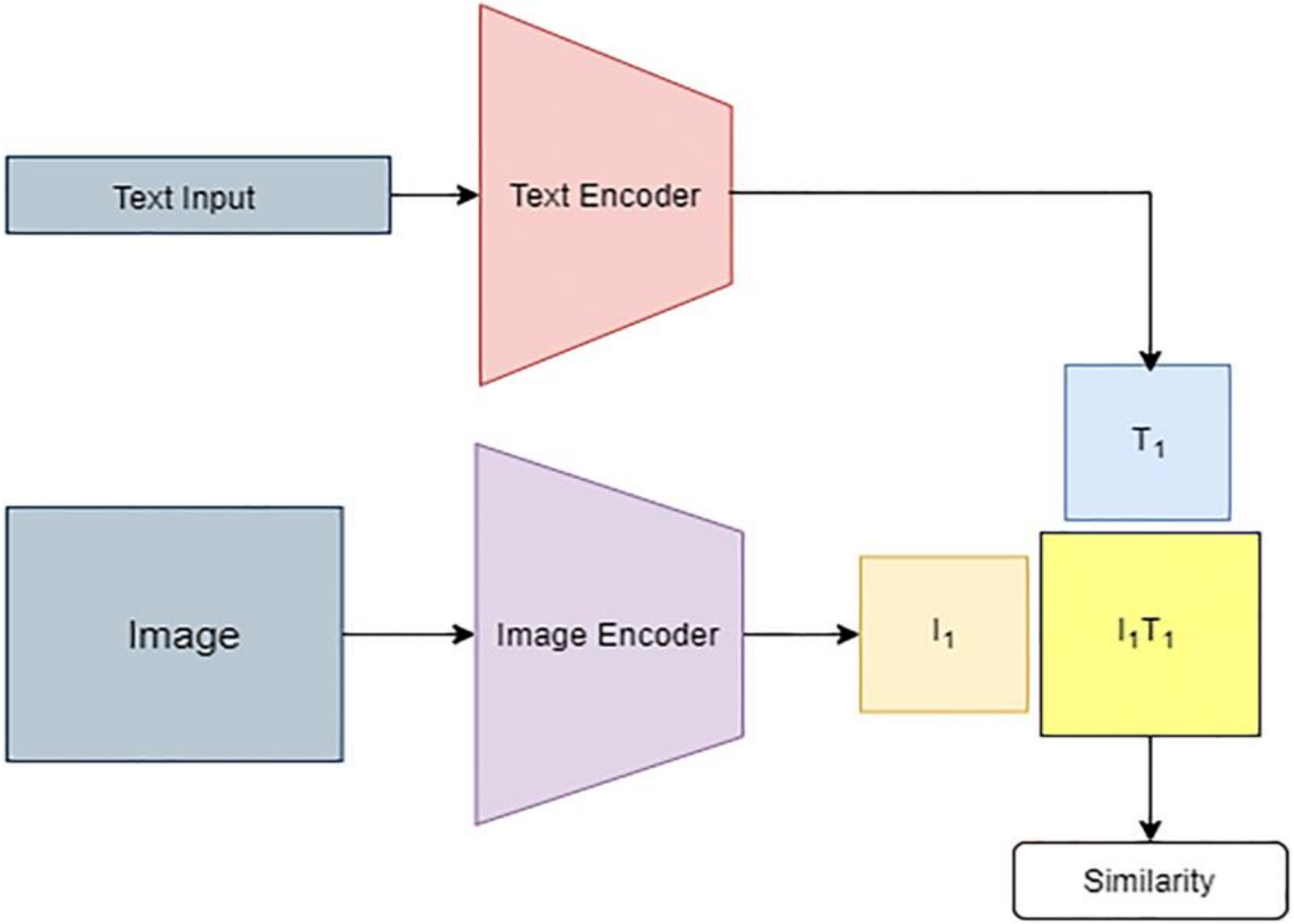


3. Use for zero-shot prediction



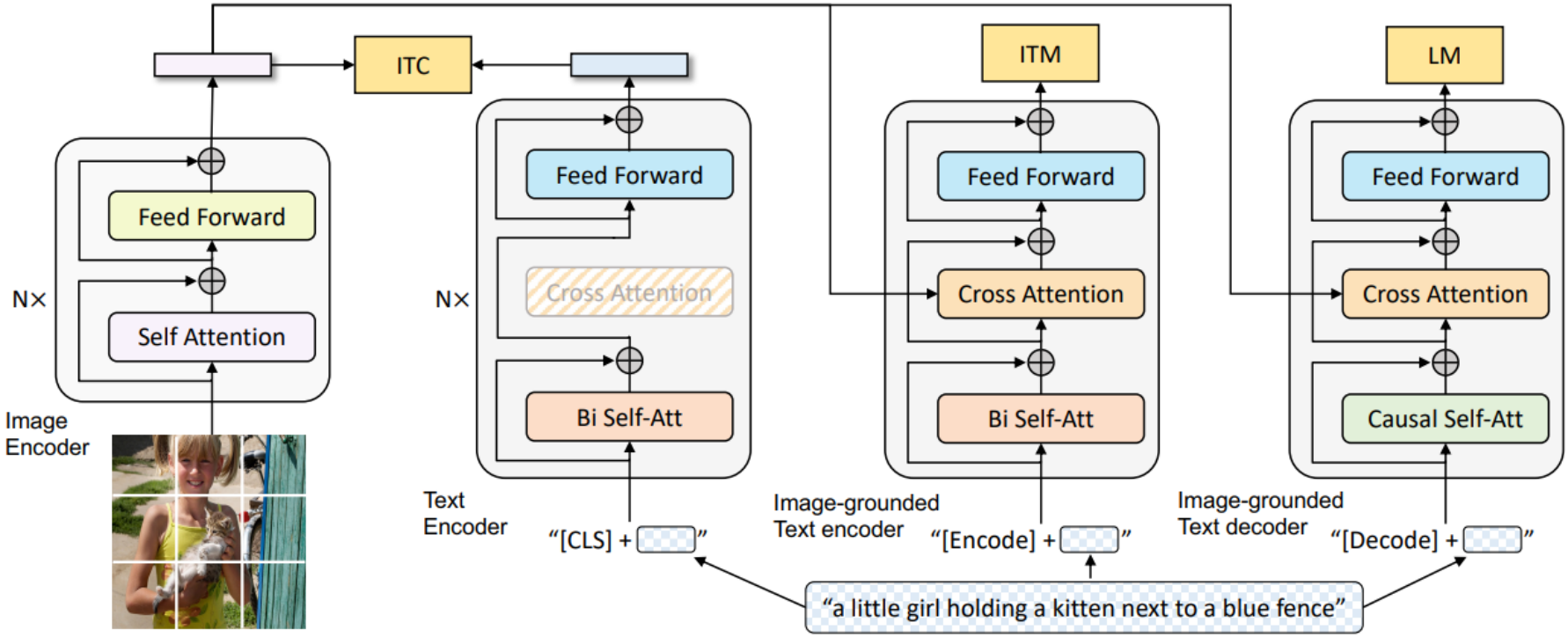
CLIP as a zero-shot classifier: Convert all of a dataset's classes into captions such as "a photo of a dog" and predict the class of the caption CLIP estimates best pairs with a given image.

OpenAI CLIP: Compute similarity



BLIP model (ICML, 2022)

- BLIP (Bootstrapping Language-Image Pre-training) utilizes the noisy web data by bootstrapping the captions.



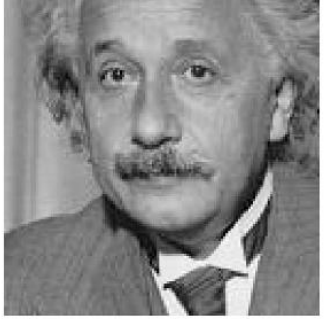
Li, Junnan, Dongxu Li, Caiming Xiong, and Steven Hoi. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." In *ICLR*, pp. 12888-12900. PMLR, 2022.

Similarity metric for vision

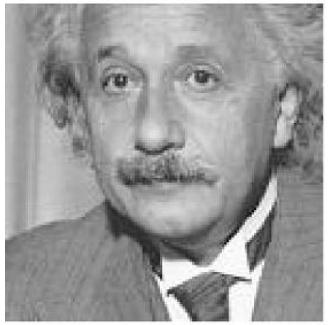
Pixel-based metrics: MAE and MSE

- Let $\mathbf{x}$ be the original image with n pixels x_i. Similarly, let $\mathbf{y}$ be the image after inpainting (with the same size) with pixels y_i.
- Mean absolute error (MAE): $$\boldsymbol{MAE(x,y) = \frac{1}{n}\sum_{i=1}^{n} \|x_i - y_i\|_1}$$
  - The changes in MAE are linear and therefore intuitive.
- Mean squared error (MSE): $$\boldsymbol{MSE(x,y) = \frac{1}{n}\sum_{i=1}^{n} (x_i - y_i)^2}$$
  - MSE punishes larger errors more than smaller errors.

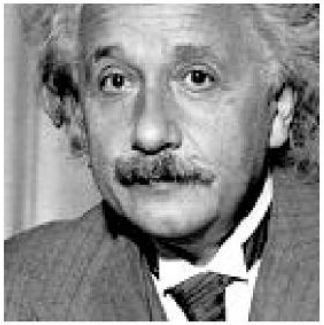
Einstein image altered with different types of distortions



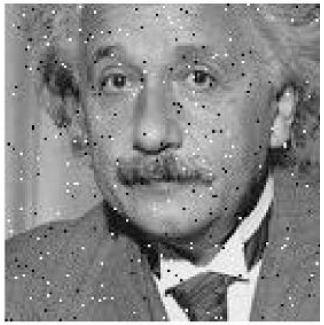
(a)



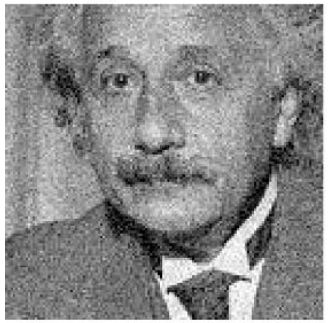
(b) MSE = 309



(c) MSE = 306

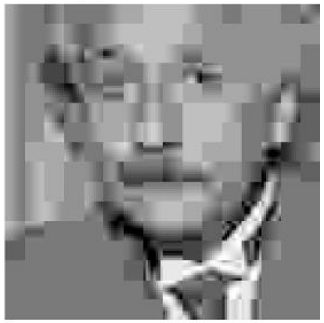


(d) MSE = 313

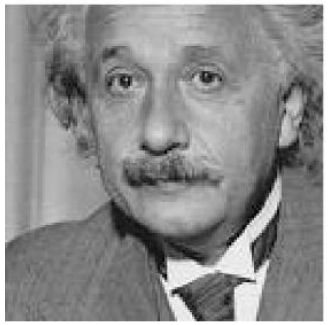


(e) MSE = 309

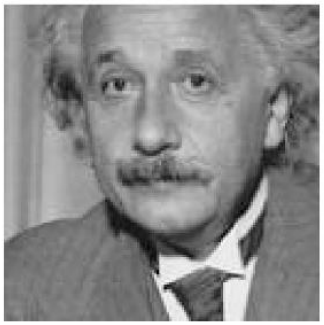
(f) MSE = 308



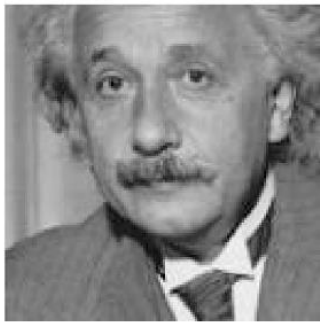
(g) MSE = 309



(h) MSE = 871



(i) MSE = 694



(j) MSE = 590

(a) "original image"
(b) mean luminance shift
(c) a contrast stretch
(d) impulsive noise contamination
(e) white Gaussian noise contamination
(f) Blurring
(g) JPEG compression
(h) a spatial shift (to the left)
(i) spatial scaling (zooming out);
(j) a rotation (counterclockwise).

Note that images (b)–(g) have almost the same MSE values but drastically different visual quality. Also, note that the MSE is highly sensitive to spatial translation, scaling, and rotation [Images (h)–(j)].

Image credit: Modern Image Quality Assessment

Pixel-based metrics: PSNR

- Peak Signal to Noise Ratio (PSNR) is the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation.
- It is usually derived from MSE, using the decibel (dB) unit.

$$PSNR(\mathbf{x}, \mathbf{y}) = 10 \log_{10} \frac{x_{max}^2}{MSE(\mathbf{x}, \mathbf{y})}$$

  - x_{max} is the maximum signal in the image (e.g., 255 for 8-bit images)
  - Identical images $\rightarrow$ MSE = 0 and PSNR undefined.

- A high-quality image has a high PSNR, but not reverse.

Pixel-based metrics: SSIM

- Structural Similarity Index Measure (SSIM) assesses the structural similarity between two images by comparing their local patches after normalizing luminance and contrast.
- Let $\mathbf{x}$ and $\mathbf{y}$ be the local patches extracted from the same position in the two images, $\mathbf{X}$ and $\mathbf{Y}$, respectively.
- $SSIM(\mathbf{x}, \mathbf{y}) = l(\mathbf{x}, \mathbf{y}) \cdot c(\mathbf{x}, \mathbf{y}) \cdot s(\mathbf{x}, \mathbf{y})$ (luminance, contrast, structure)

$$= \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right) \cdot \left(\frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\right) \cdot \left(\frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}\right)$$

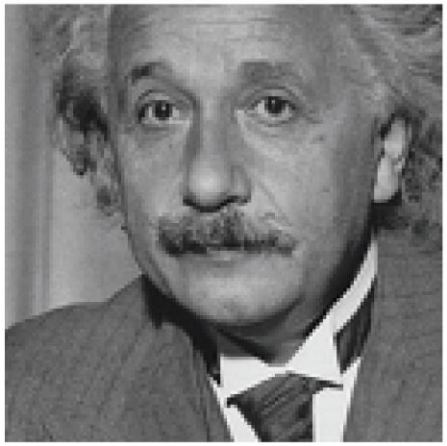
  - μ_x, μ_y, σ_x, and σ_y are means and standard deviations of the local patches $\mathbf{x}$ and $\mathbf{y}$, respectively. C_1, C_2, and C_3 are constants.

Pixel-based metrics: SSIM

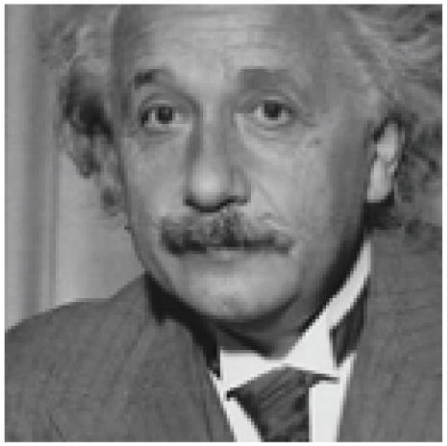
- Patches are extracted by applying a sliding window onto the image, with a stride of one pixel.
- The mean SSIM for the images $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$SSIM(\mathbf{X}, \mathbf{Y}) = \frac{1}{M} \sum_{j=1}^{M} SSIM(\mathbf{x}_j, \mathbf{y}_j)$$

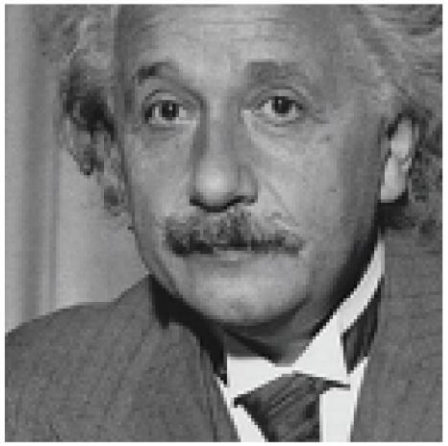
  - M is the total number of patches in the images.



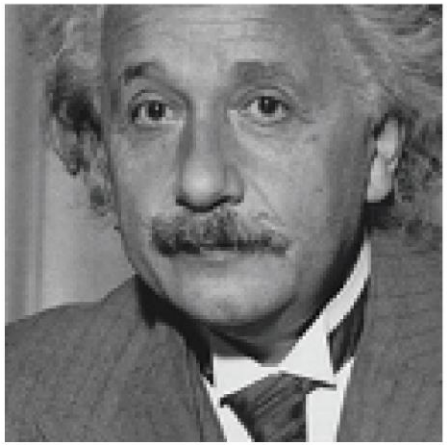
(a) SSIM = 1



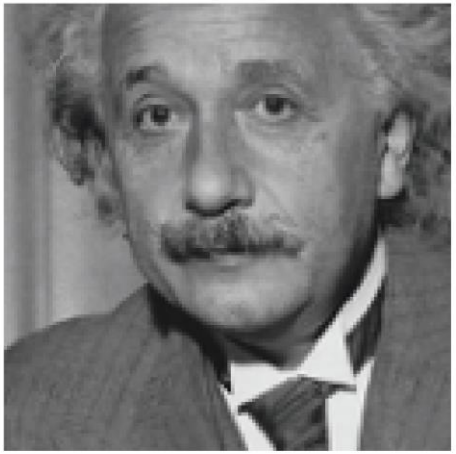
(b) SSIM = 0.505



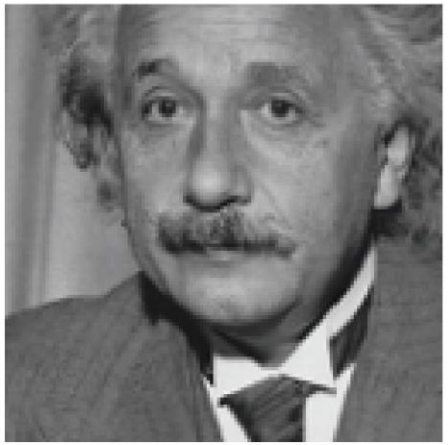
(c) SSIM = 0.404



(d) SSIM = 0.399



(e) SSIM = 0.549



(f) SSIM = 0.551

Some examples to show that SSIM may not estimate the geometric distortions correctly, leading to low SSIM scores. (a) Original image. The other images are processed via: (b) Scaling (shrink), (c) Anti-clockwise rotation, and (d) Clockwise rotation.

Image credit: Modern Image Quality Assessment

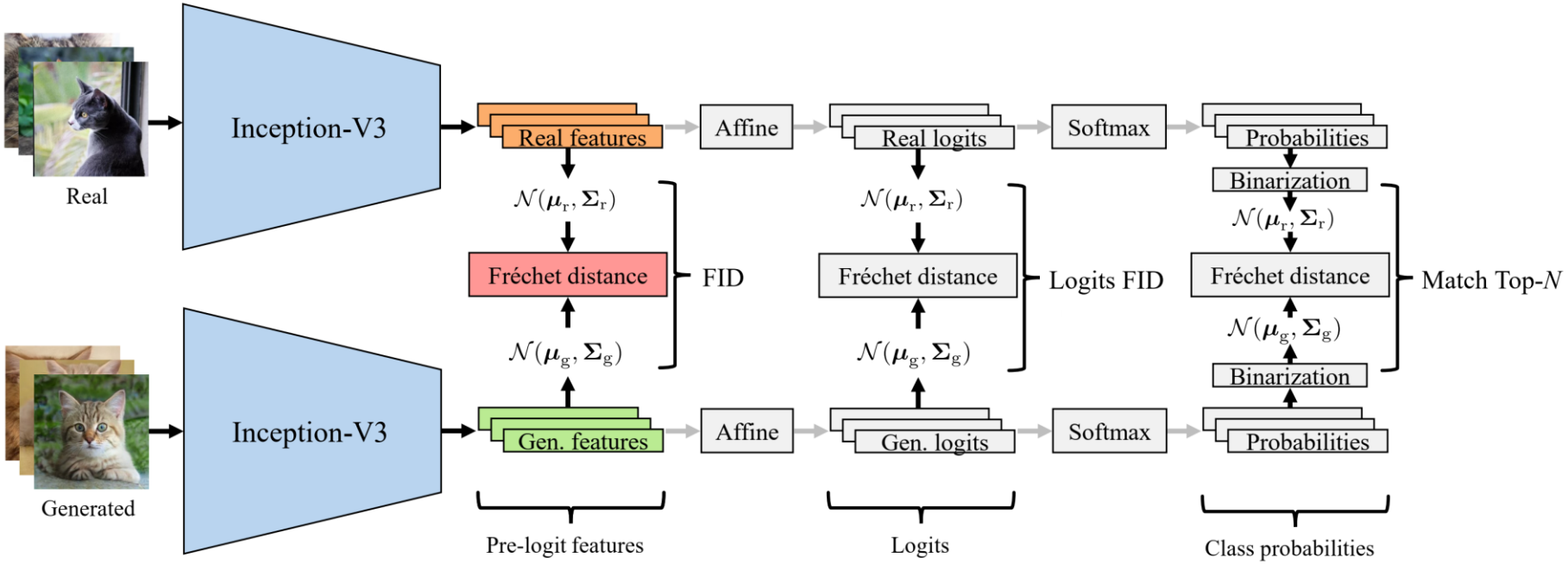
Perceptual metrics: FID

- Fréchet inception distance (FID) assesses the quality of images created by a generative model.
- For two multidimensional Gaussian distributions, $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\mu', \Sigma')$, FID is explicitly solvable as

$$\boldsymbol{d_F\left(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma')\right)^2 = \|\mu - \mu'\|_2^2 + \mathrm{tr}\left(\Sigma + \Sigma' - 2(\Sigma\Sigma')^{1/2}\right)}$$

- Consider a function $\boldsymbol{f}$: $\Omega_X \to \mathbb{R}^n$ and two datasets $S, S' \subset \mathbb{R}^n$.
- Compute $\boldsymbol{f}(S), \boldsymbol{f}(S') \subset \mathbb{R}^n$ and fit two Gaussian distributions, $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\mu', \Sigma')$, respectively.
- Return $d_F\left(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma')\right)^2$

Perceptual metrics: FID

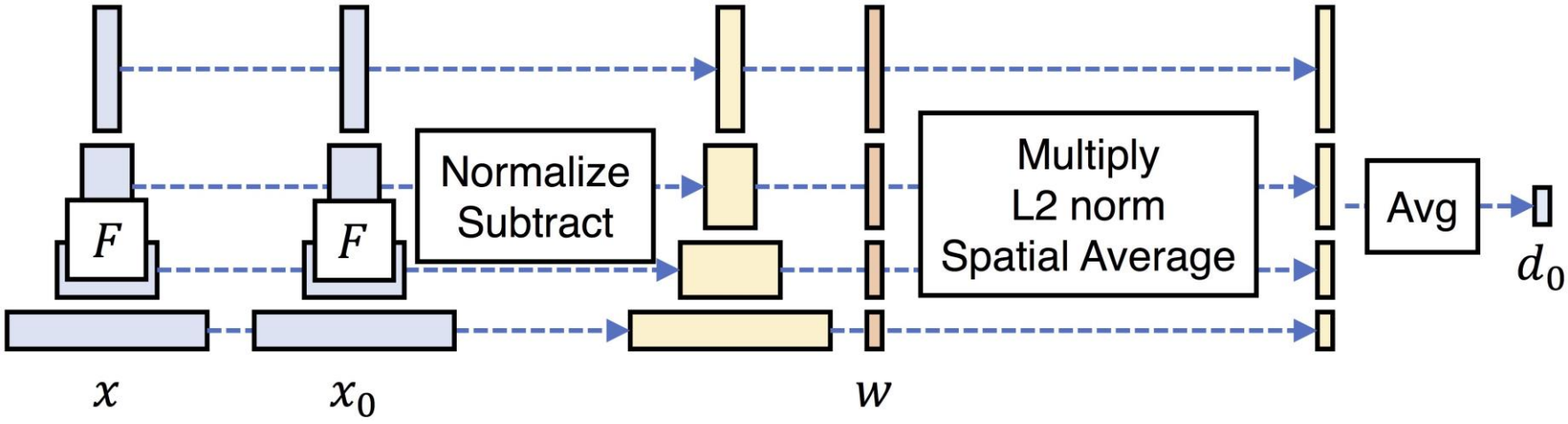


First, the real and generated images are separately passed through a pre-trained network, typically the InceptionV3, to produce two sets of feature vectors. Then, both distributions of features are estimated with multivariate Gaussians, and FID is defined as the Fréchet distance between the two Gaussians.

Kynkäänniemi, Tuomas, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. "The Role of ImageNet Classes in Fréchet Inception Distance". ICLR 2023.

Perceptual metrics: LPIPS

Learned Perceptual Image Patch Similarity



$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})||_2^2$$

Zhang, Richard, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. "The unreasonable effectiveness of deep features as a perceptual metric." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 586-595. 2018.

References

- Rafael C. Gonzalez, Richard E. Woods, “Digital Image Processing”, 3rd edition, 2008. Chapter 3
- Images are obtained from the above materials and Google


...the end.