Problem 1 (Easy)

What is an image encoding?
An image encoding is a numerical representation of an image that captures important visual information such as shapes, textures, and patterns in a form that can be processed by machine learning models.

Why are raw pixel values generally a poor encoding for visual recognition tasks?
Raw pixel values are highly sensitive to small changes such as lighting variations, noise, rotation, or translation. They do not explicitly capture meaningful structures like edges or shapes and result in very high-dimensional data, which makes learning inefficient and prone to overfitting.

Problem 2 (Easy)

Why is edge information useful for recognizing objects in images?
Edges define object boundaries and shapes, which are critical for identifying objects. Many objects can be recognized primarily by their outlines, and edge information is relatively invariant to illumination changes. Edges also serve as basic building blocks from which more complex visual patterns are formed.

Problem 3 (Medium)

Why do early convolutional layers in CNNs learn edge-like filters, and how does this relate to HOG?
Early convolutional layers learn edge-like filters because edges are fundamental visual primitives present in almost all images. Capturing edges allows the network to represent important spatial intensity changes efficiently. This behavior is closely related to Histogram of Oriented Gradients (HOG), which explicitly encodes gradient directions. The key difference is that HOG uses fixed, hand-crafted features, while CNNs learn similar features automatically and optimize them end-to-end for the task.

Problem 4 (Medium)

Why do CNN embeddings separate image classes better than PCA applied to raw pixels? Give two reasons.

First, CNN embeddings are learned using labeled data and are optimized to maximize class separability, whereas PCA is an unsupervised method that only preserves variance and does not consider class labels.

Second, CNNs learn nonlinear and hierarchical representations, progressing from edges to shapes to object-level features, while PCA is a linear method and cannot capture such complex structures.

Problem 5 (Hard)

Why is learning spatially-aware embeddings essential for object detection, and why would global image embeddings fail?

Object detection requires identifying both what the object is and where it is located. Spatially-aware embeddings preserve local spatial relationships, enabling the model to predict object positions, sizes, and multiple objects within the same image. YOLO uses convolutional feature maps that retain spatial layout for this purpose. A pipeline based only on global image embeddings, such as PCA followed by a classifier, discards spatial information and therefore cannot localize objects or detect multiple instances, making it unsuitable for object detection tasks.