# Knowledge Distillation

- transfer knowledge from a teacher model to a student model
  - student learns to match teacher outputs or intermediate features
- why used
  - compression: smaller model with similar accuracy
  - speed: faster inference (mobile / real-time)
  - regularization: teacher soft targets can improve generalization
- common objectives
  - logit distillation (soft targets)
    - $L = \mathrm{CE}(y, p_s) + \lambda T^2, \mathrm{KL}(p_t^{(T)} | p_s^{(T)})$
    - $p^{(T)} = \mathrm{softmax}(z/T)$
  - feature distillation
    - match hidden representations (e.g., L2 / cosine loss)
- key knobs
  - temperature $T$ (softness of targets)
  - weight $\lambda$ (balance with supervised loss)

초반에는 낮게 → 후반 갈수록 높게
softening sharpening

## DINO

label 필요 X teacher → student

- self-supervised distillation for vision transformers (teacher-student)
  - student learns to predict teacher's output distribution
  - no labels required
- core idea
  - two augmented views of the same image
  - teacher sees global crops, student sees global + local crops
  - teacher targets are sharpened (low entropy) with temperature scheduling

multi-crops
local → receptive field ↓
일부만 보고 global 정보 유추
model이 local-to-global correspondence 학습

- objective
  - minimize cross-entropy between teacher and student outputs
  - encourages consistent representations across views
- why it works
  - teacher is an EMA of the student (stable target)
  - centering + temperature prevent collapse (trivial constant outputs)
- outcome
  - learns strong visual features that transfer well to downstream tasks

---

## ControlNet

- add conditioning control to a pretrained diffusion model without destroying its generative ability
  - conditions: edge, depth, pose, segmentation, scribble, etc.
- core idea
  - copy the U-Net and attach a trainable control branch
  - keep the original pretrained U-Net frozen (or mostly frozen)
  - inject control features into the main U-Net via residual additions
- how it works (high-level)
  - control input $c$ is encoded into multi-scale features
  - at each U-Net block, add control residuals to the main activations
  - "zero-conv" (initialized near zero) makes training stable
    - starts from “no control effect” and gradually learns control strength
- benefits
  - strong controllability with minimal degradation of text alignment and realism
  - works as a plug-in for many conditions
- limitations

    - extra compute and memory (another branch)
    - need paired data (image + condition) or condition extraction pipeline

## T2I Adapter

- lightweight conditioning adapter for pretrained text-to-image diffusion
    - similar use cases: edges, depth, pose, segmentation, style hints
- core idea
    - add a small adapter network that encodes the condition into feature maps
    - inject those features into the pretrained U-Net (often via residual connections)
    - typically fewer parameters than ControlNet
    - pretrained diffusion backbone stays frozen
- how it works (high-level)
    - condition encoder produces multi-scale representations
    - features are fused into U-Net blocks (shallow additions)
    - train only the adapter (and sometimes small projection layers)
- benefits
    - parameter-efficient, faster to train, lower memory overhead
    - easier to deploy when compute is tight
- limitations
    - usually weaker controllability than full ControlNet
    - capacity may be insufficient for complex spatial constraints

---