# DTO-KD: Dynamic Trade-off Optimization for Effective Knowledge Distillation



Zeeshan Hayder[1,2]


Ali Cheraghian[1,2]


Lars Petersson[1]

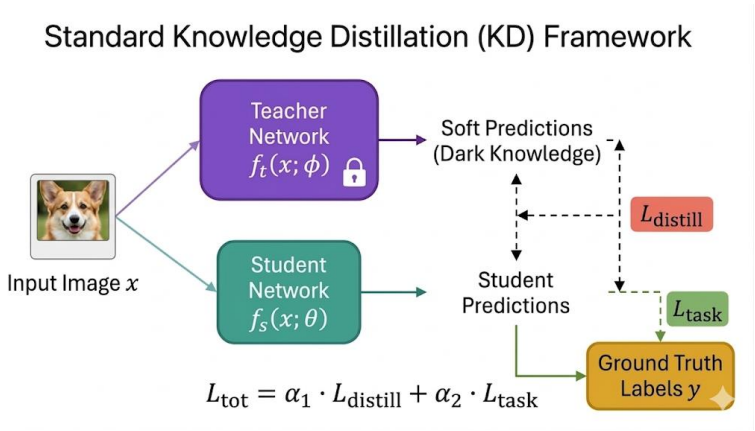

Mehrtash Harandi[3]


Richard Hartley[2]

[1]Data61/CSIRO, [2]Australian National University, [3]Monash University,

# Knowledge Distillation — Background

- **A model compression paradigm:** transfer knowledge from a large teacher to a compact student
- **Why it matters:** Enables deployment on resource-constrained devices (mobile, edge, embedded)
- **Standard approach:** Student is trained with a weighted combination of two losses:
  - Task loss: fit ground-truth lables
  - Distillation loss: mimic teacher's soft predictions
- **Key tension:** These two objectives can conflict.



# Motivation: Why Current KD Fails

## Gradient conflict

$\nabla L_{distill}$ and $\nabla L_{task}$ often point in opposing directions

- Task loss pulls student toward ground truth
- Distillation loss pulls student toward the teacher

## Gradient dominance

One gradient's magnitude overwhelms the other

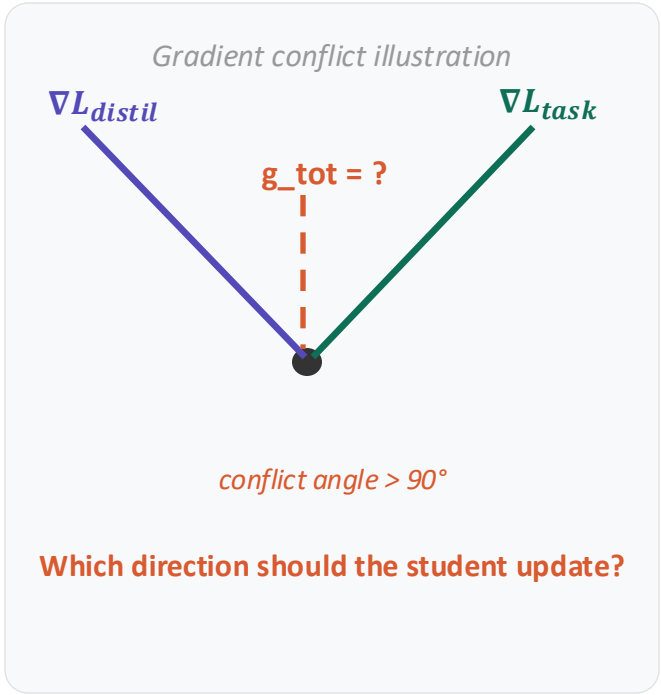
- The weaker objective gets suppressed during training

## Dynamic training landscape

Gradient magnitudes shift across epochs

- A fixed $\alpha_1$, $\alpha_2$ cannot adapt to these evolving dynamics

**Bottom line:** Manual loss weighting is fundamentally brittle



# Limitations of Existing KD Approaches

ICLR International Conference On Learning Representations

## Logit-level (predictive distributions)

- Transfers only final predictions
- Cannot encode teacher's inductive biases or intermediate reasoning

## Feature-level (intermediate repr.)

- Requires heuristic layer matching and normalization
- Does not resolve gradient conflict — two losses still compete

## Token-based (transformer semantics)

- Naïve token matching ignores contextual and relational structure
- Misses higher-order dependencies between tokens

★

## Multi-objective optimization

- Previously not applied to knowledge distillation

**This is the gap we fill → DTO-KD**

**Common thread:** None of these approaches address the fundamental gradient conflict between competing KD objectives

# Our Contributions

**Key idea:** Formulate KD as a dynamic multi-objective optimization (MOO) problem and solve it analytically at each iteration

1

### KD as dynamic MOO problem

Jointly optimize task loss and distillation loss as competing objectives with per-iteration trade-off

2

### Closed-form optimal trade-off

Analytically compute $\pi_1$, $\pi_2$ at each step — no grid search, no extra hyperparameters

3

### Gradient conflict & dominance resolved

Update direction always aligned with both objectives; equal contribution ensures neither is suppressed

**Result:** New state-of-the-art on ImageNet-1K, CIFAR-100, and COCO across homogeneous & heterogeneous teacher-student pairs

# Problem Formulation

## Goal of knowledge distillation

Train student to approximate teacher on all samples:

$$f_s(x_i;\theta) \approx f_t(x_i;\phi), \qquad \forall i = 1, \ldots, N.$$

## Conventional KD: weighted sum of two losses

$$\mathrm{L}_{\mathrm{tot}}(\boldsymbol{\theta}) \triangleq \alpha_1 \mathrm{L}_{\mathrm{distill}}(\boldsymbol{\theta}) + \alpha_2 \mathrm{L}_{\mathrm{task}}(\boldsymbol{\theta})$$

$\alpha_1$, $\alpha_2$ are **fixed scalar weights** — manually tuned, cannot adapt during training

### Distillation loss

$$\mathrm{L}_{\mathrm{distill}}(\boldsymbol{\theta}) \triangleq \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{S}}\, \ell_{\mathrm{distill}}\big(f_s(\boldsymbol{x};\boldsymbol{\theta}), f_t(\boldsymbol{x};\boldsymbol{\phi})\big)$$

### Task loss

$$\mathrm{L}_{\mathrm{task}}(\boldsymbol{\theta}) \triangleq \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{S}}\, \ell_{\mathrm{task}}\big(f_s(\boldsymbol{x};\boldsymbol{\theta}), f_t(\boldsymbol{x};\boldsymbol{\phi})\big)$$

# Gradient Conflict and Dominance

**Gradient of total loss:**

$$g_{tot} = \nabla L_{tot}(\theta) = \alpha_1 g_{dist} + \alpha_2 g_{task}$$

## Gradient Conflict (GrC)

Occurs when the gradients of the two losses point in opposing directions:

$$\langle g_{dist}, g_{task} \rangle < 0$$

→ Negative inner product means the losses pull the student in opposing directions

## Gradient Dominance (GrD)

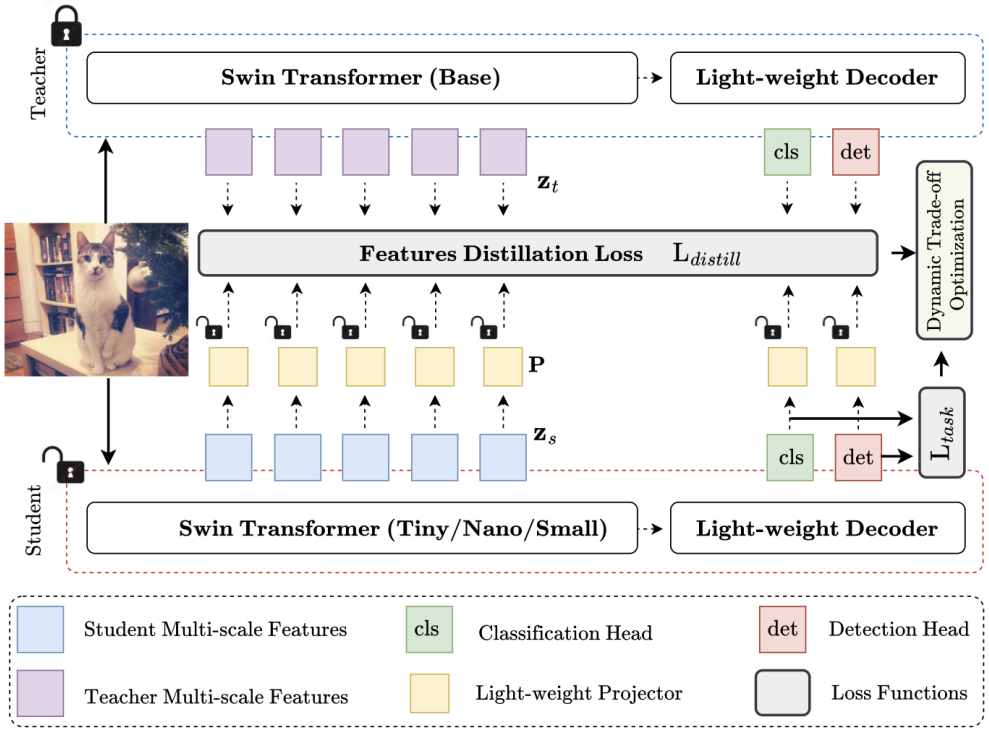
Arises when one gradient's magnitude overwhelms the other:

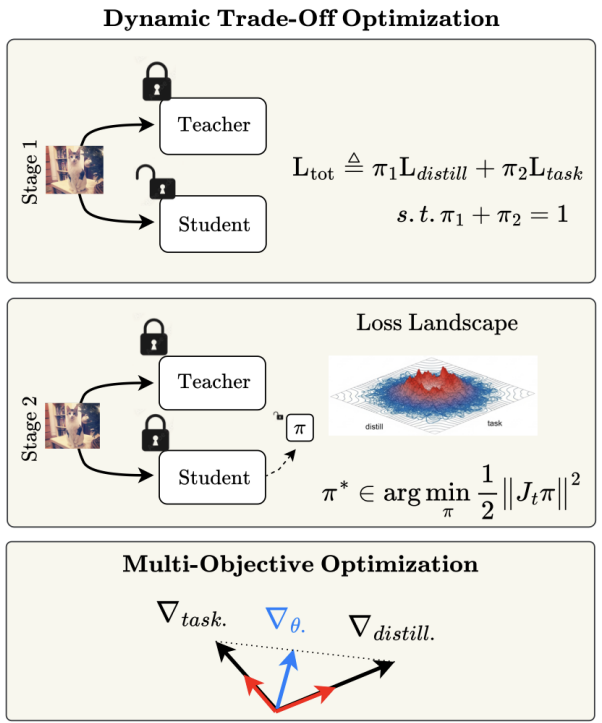
$$\| g_{dist} \| \gg \| g_{task} \|$$

→ The stronger gradient dominates the update, suppressing the weaker objective entirely

**DTO-KD resolves both:** dynamically compute $\pi = (\pi_1, \pi_2)$ so the update is aligned with both objectives

# DTO-KD Framework



# KD as Dynamic Trade-off Optimization

- Formulate KD as multi-objective optimization: $\mathbf{L}_{\text{tot}}(\boldsymbol{\theta}) = \left(\mathrm{L}_{\text{distill}}(\boldsymbol{\theta}), \mathrm{L}_{\text{task}}(\boldsymbol{\theta})\right)^{\top}$

## Stage 1: Update student parameters

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\, \boldsymbol{g}_t$$

Rate of improvement:

$$r_{\text{task}}(\boldsymbol{g}_t) = \frac{\mathrm{L}_{\text{task}}(\boldsymbol{\theta}_t) - \mathrm{L}_{\text{task}}(\boldsymbol{\theta}_{t+1})}{\mathrm{L}_{\text{task}}(\boldsymbol{\theta}_t)} \qquad r_{\text{dist}}(\boldsymbol{g}_t) = \frac{\mathrm{L}_{\text{distill}}(\boldsymbol{\theta}_t) - \mathrm{L}_{\text{distill}}(\boldsymbol{\theta}_{t+1})}{\mathrm{L}_{\text{distill}}(\boldsymbol{\theta}_t)}$$

## Stage 2: Optimize trade-off weights

Define: $J_t = [\nabla \log \mathrm{L}_{\text{distill}}(\boldsymbol{\theta}_t) \mid \nabla \log \mathrm{L}_{\text{task}}(\boldsymbol{\theta}_t)]$

Solve:

Update: $\boldsymbol{g}^* = \pi_1 \nabla \log \mathrm{L}_{\text{distill}}(\boldsymbol{\theta}_t) + \pi_2 \nabla \log \mathrm{L}_{\text{task}}(\boldsymbol{\theta}_t)$

## Theorem 3.1 — Closed-form solution:

$$\pi_1^* = \frac{g_{22} - g_{12}}{g_{11} + g_{22} - 2g_{12}} \qquad \pi_2^* = \frac{g_{11} - g_{12}}{g_{11} + g_{22} - 2g_{12}}$$

Practical: Amortized training — single backprop per iteration via gradient-based π update (Eq. 18)

$$\boldsymbol{\pi}(t+1) = \boldsymbol{\pi}(t) - \eta_\pi \nabla_{\boldsymbol{\pi}} \frac{1}{2} \left\| \pi_{\text{dist}}(t) \nabla \log \mathrm{L}_{\text{distill}} + \pi_{\text{task}}(t) \nabla \log \mathrm{L}_{\text{task}} \right\|^2$$

# Quantitative Results

ICLR International Conference On Learning Representations

## ImageNet-1K (Classification)

**DeiT-Tiny:**

- +1.4% over SOTA
- +5.2% over baseline

**DeiT-Small:**

- +0.8% over SOTA

| Method | Venue | Top@1 | Teacher | #Param. |
|---|---|---|---|---|
| RegNetY-160 (Radosavovic et al., 2020) | *CVPR20* | 82.6 | None | 84M |
| CaiT-S24 (Touvron et al., 2021b) | *ICCV21* | 83.4 | None | 47M |
| DeiT3-B (Touvron et al., 2022) | *ECCV22* | 83.8 | None | 87M |
| DeiT-Ti (Touvron et al., 2021a) | *ICML21* | 72.2 | None | 5M |
| DeiT-Ti (KD) (Touvron et al., 2021a) | *ICML21* | 74.5 | Regnety-160 | 6M |
| ↳ 1000 epochs | *ICML21* | 76.6 | Regnety-160 | 6M |
| CivT-Ti (Ren et al., 2022) | *CVPR22* | 74.9 | Regnety-600m | 6M |
| Manifold (Hao et al., 2022) | *NeurIPS22* | 76.5 | CaiT-S24 | 6M |
| DearKD (Chen et al., 2022) | *CVPR22* | 74.8 | Regnety-160 | 6M |
| ↳ 1000 epochs | *CVPR22* | 77.0 | Regnety-160 | 6M |
| USKD (Yang et al., 2023) | *ICCV23* | 75.0 | Regnety-160 | 6M |
| MaskedKD (Son et al., 2024) | *ECCV24* | 75.4 | CaiT-S24 | 6M |
| SRD (Miles & Mikolajczyk, 2024) | *AAAI24* | 77.2 | Regnety-160 | 6M |
| $V_k$D-Ti (Roy Miles & Deng, 2024) | *CVPR24* | 78.3 | Regnety-160 | 6M |
| **DTO-KD (Ti)** | | **79.7** | Regnety-160 | 6M |
| DeiT-S (Touvron et al., 2021a) | *ICML21* | 79.8 | None | 22M |
| DeiT-S (KD) (Touvron et al., 2021a) | *ICML21* | 81.2 | Regnety-160 | 22M |
| ↳ 1000 epochs | *ICML21* | 82.6 | Regnety-160 | 22M |
| CivT-S (Ren et al., 2022) | *CVPR22* | 82.0 | Regnety-4gf | 22M |
| DearKD (Chen et al., 2022) | *CVPR22* | 81.5 | Regnety-160 | 22M |
| ↳ 1000 epochs | *CVPR22* | 82.8 | Regnety-160 | 22M |
| USKD (Yang et al., 2023) | *ICCV23* | 80.8 | Regnety-160 | 22M |
| MaskedKD (Son et al., 2024) | *ECCV24* | 81.4 | Deit3-B | 22M |
| SRD (Miles & Mikolajczyk, 2024) | *AAAI24* | 82.1 | Regnety-160 | 22M |
| $V_k$D-S (Roy Miles & Deng, 2024) | *CVPR24* | 82.3 | Regnety-160 | 22M |
| **DTO-KD (S)** | | **83.1** | Regnety-160 | 22M |

# Quantitative Results

- New SOTA across homogeneous & heterogeneous CNNs (CIFAR100)

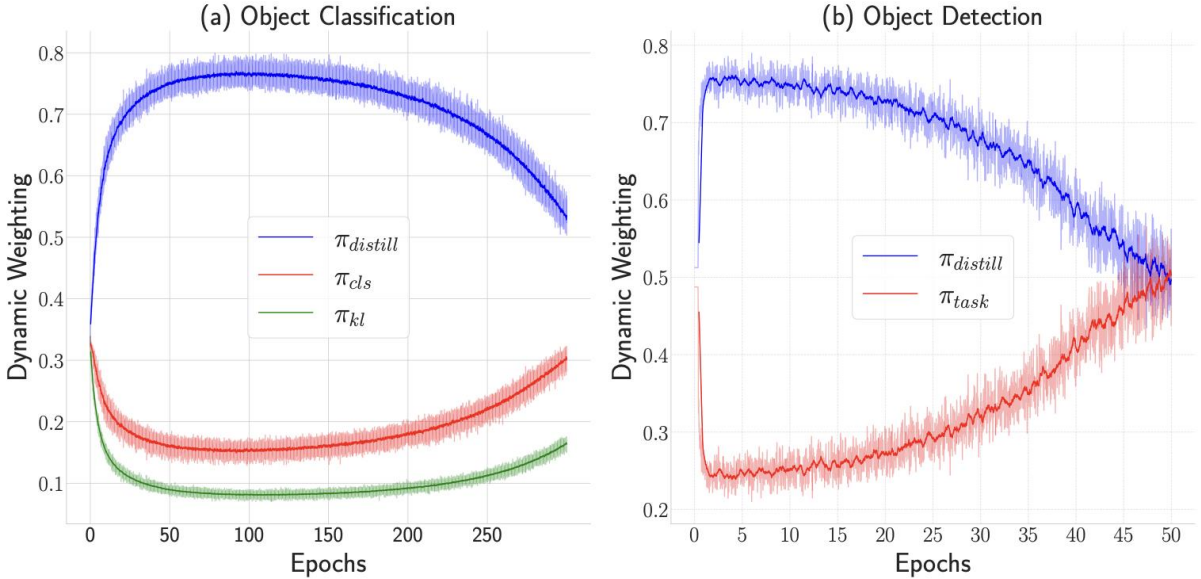
| Methods | Homogeneous | | | Heterogeneous | | |
|---|---|---|---|---|---|---|
| | ResNet-56<br>ResNet-20 | WRN-40-2<br>WRN-40-1 | ResNet-32×4<br>ResNet-8×4 | ResNet-50<br>MobileNet-V2 | ResNet-32×4<br>ShuffleNet-V1 | ResNet-32×4<br>ShuffleNet-V2 |
| Teacher | 72.34 | 75.61 | 79.42 | 79.34 | 79.42 | 79.42 |
| Student | 69.06 | 71.98 | 72.50 | 64.60 | 70.50 | 71.82 |
| FitNet (Romero et al., 2015) | 69.21 | 72.24 | 73.50 | 63.16 | 73.59 | 73.54 |
| RKD (Park et al., 2019) | 69.61 | 72.22 | 71.90 | 64.43 | 72.28 | 73.21 |
| PKT (Passalis et al., 2020) | 70.34 | 73.45 | 73.64 | 66.52 | 74.10 | 74.69 |
| KD (Hinton et al., 2015) | 70.66 | 73.54 | 73.33 | 67.65 | 74.07 | 74.45 |
| OFD (Heo et al., 2019b) | 70.98 | 74.33 | 74.95 | 69.04 | 75.98 | 76.82 |
| CRD (Tian et al., 2019) | 71.16 | 74.14 | 75.51 | 69.11 | 75.11 | 75.65 |
| DIST (Huang et al., 2022) | 71.78 | 74.42 | 75.79 | 69.17 | 75.23 | 76.08 |
| ReviewKD (Chen et al, 2021) | 71.89 | 75.09 | 75.63 | 69.89 | 77.45 | 77.78 |
| DKD (Zhao et al., 2022b) | 71.97 | 74.81 | 75.44 | 70.35 | 76.45 | 77.07 |
| ReviewKD++ (Wang et al, 2024) | 72.05 | 75.66 | 76.07 | 70.45 | 77.68 | 77.93 |
| **DTO-KD** (ours) | **72.35** | **75.68** | **76.40** | **70.90** | **77.95** | **78.22** |

# Quantitative Results

- COCO (Object Detection)

| | ViDT Model | Epochs | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ | #Params | FPS |
|---|---|---|---|---|---|---|---|---|---|---|
| **Teacher** | Swin-base (Song et al., 2021) | 50 | 49.4 | 69.6 | 53.4 | 31.6 | 52.4 | 66.8 | 0.1B | 9.0 |
| **Student** | Swin-nano (Song et al., 2021) | 50 | 40.4 | 59.6 | 43.3 | 23.2 | 42.5 | 55.8 | 16M | 20.0 |
| | Token-Matching (Song et al., 2022) | 50 | 41.9 | 61.2 | 44.7 | 23.6 | 44.1 | 58.7 | | |
| | $V_k$D-nano (Roy Miles & Deng, 2024) | 50 | 43.0 | 62.3 | 46.2 | 24.8 | 45.3 | 60.1 | | |
| | **DTO-KD (nano)** | 50 | **43.7** | **63.1** | **46.8** | **25.1** | **46.2** | **61.9** | | |
| | Swin-tiny (Song et al., 2021) | 50 | 44.8 | 64.5 | 48.7 | 25.9 | 47.6 | 62.1 | 38M | 17.2 |
| | Token-Matching (Song et al., 2022) | 50 | 46.6 | 66.3 | 50.4 | 28.0 | 49.5 | 64.3 | | |
| | $V_k$D-tiny (Roy Miles & Deng, 2024) | 50 | 46.9 | 66.6 | 50.9 | 27.8 | 49.8 | 64.6 | | |
| | **DTO-KD (tiny)** | 50 | **47.4** | **67.2** | **51.3** | **28.0** | **50.7** | **65.8** | | |
| | Swin-small (Song et al., 2021) | 50 | 47.5 | 67.7 | 51.4 | 29.2 | 50.7 | 64.8 | 61M | 12.1 |
| | Token-Matching (Song et al., 2022) | 50 | 49.2 | 69.2 | 53.6 | 30.7 | 52.3 | 66.8 | | |
| | $V_k$D-small (Roy Miles & Deng, 2024) | 50 | 48.5 | 68.4 | 52.4 | 30.8 | 52.2 | 66.0 | | |
| | **DTO-KD (small)** | 50 | **49.6** | **69.4** | **53.9** | **31.6** | **53.1** | **67.1** | | |

# Effectiveness of Dynamic Balancing



# Summary

- DTO-KD brings a principled, gradient-aware, multi-objective optimization framework into KD

- It resolves fundamental issues that have limited KD for years
  - no fixed weights
  - no gradient conflict
  - no gradient dominance
  - stable, fast convergence

- State-of-the-art across multiple vision benchmarks

# Thank you