# Heterogeneous Complementary Distillation

**Liuchi Xu[1,4], Hao Zheng[2], Lu Wang[*1], Lisheng Xu[3], Jun Cheng[*4,5]**

[1]School of Computer Science and Engineering, Northeastern University, China;
[2]School of Computer Science, South China Normal University, China;
[3]College of Information Science and Engineering, Northeastern University, China;
[4]Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China;
[5]The Chinese University of Hong Kong, Hong Kong, China
{xuliuchi@stumail,wanglu@mail}.neu.edu.cn, zzeo.zheng@gmail.com,
xuls@mail.neu.edu.cn, Jun.cheng@siat.ac.cn

## Abstract

Knowledge distillation (KD) transfers the "dark knowledge" from a complex teacher model to a compact student model. However, heterogeneous architecture distillation, such as Vision Transformer (ViT) to ResNet18, faces challenges due to differences in spatial feature representations. Traditional KD methods are mostly designed for homogeneous architectures and hence struggle to effectively address the disparity. Although heterogeneous KD approaches have been developed recently to solve these issues, they often incur high computational costs and complex designs, or overly rely on logit alignment, which limits their ability to leverage the complementary features. To overcome these limitations, we propose Heterogeneous Complementary Distillation (HCD), a simple yet effective framework that integrates complementary teacher and student features to align representations in shared logits. These logits are decomposed and constrained to facilitate diverse knowledge transfer to the student. Specifically, HCD processes the student's intermediate features through convolutional projector and adaptive pooling, concatenates them with teacher's feature from the penultimate layer and then maps them via the Complementary Feature Mapper (CFM) module, comprising fully connected layer, to produce shared logits. We further introduce Sub-logit Decoupled Distillation (SDD) that partitions the shared logits into $n$ sub-logits, which are fused with teacher's logits to rectify classification. To ensure sub-logit diversity and reduce redundant knowledge transfer, we propose an Orthogonality Loss (OL). By preserving student-specific strengths and leveraging teacher knowledge, HCD enhances robustness and generalization in students. Extensive experiments on the CIFAR-100, Fine-grained (e.g., CUB200, Aircraft) and ImageNet-1K datasets demonstrate that HCD outperforms state-of-the-art KD methods, establishing it as an effective solution for heterogeneous KD.

**Code** — https://github.com/yema-web/HCD

## Introduction

Knowledge distillation (KD), proposed by Hinton et al. (Hinton et al. 2014), aligns the soft label distributions of a complex teacher model (abbreviated teacher) and a compact student model (abbreviated student) using Kullback-Leibler (KL) divergence (Kullback and Leibler 1951). This technique enables the student's performance to approach that of the teacher's. Due to its simplicity and its ability to preserve the original architecture, KD is widely applied to resource-constrained computer vision tasks, such as image classification (He et al. 2016; Yang et al. 2022a; Zheng et al. 2025), object detection (Li et al. 2024b; Zhao et al. 2024; Li et al. 2023a) and semantic segmentation (Zou et al. 2024; Liang et al. 2023). However, traditional variants of KD, such as FitNets (Romero et al. 2015), DKD (Zhao et al. 2022), and RKD (Park et al. 2019), mainly focus on distillation between homogeneous architectures (e.g., ResNet34 to ResNet18). This limits the improvement of the student's performance due to the lack of knowledge diversity in the KD strategy. In contrast, Vision Transformers (ViTs) (Dosovitskiy et al. 2021; Tolstikhin et al. 2021), with stronger representational capacity, provide new avenues for heterogeneous architecture (or cross-architecture) distillation (Liu et al. 2022) and allow the student to acquire diverse feature representations from the teacher. Intuitively, heterogeneous architecture distillation, such as ViT-to-ResNet18, leverages the complementary information of inductive biases from the teacher and the student. As a result, it enhances the student's robustness and generalization beyond the limitations of homogeneous distillation methods. However, when traditional feature-based distillation methods (Romero et al. 2015; Park et al. 2019) in homogeneous settings are applied to heterogeneous settings, distillation result often fails due to mismatched spatial feature representations (e.g., differences in receptive field size), which results in the degradation of the student's performance.

To overcome the heterogeneous feature representation gap, existing methods, such as OFA-KD (Hao et al. 2023), adopt a one-for-all strategy at each student stage to map intermediate representations to the logits space and align them with the teacher's logits. However, relying solely on logits alignment between teacher and student fails to leverage their complementary strengths, thus limiting potential performance gains. Similarly, PAT proposes a region-aware attention (RAA) module (Lin et al. 2025) that restructures stu-

[*]Corresponding authors

dent intermediate features to align with teacher representations, improving student's performance but incurring high computational costs and complex designs. These limitations highlight the challenge of transferring heterogeneous knowledge effectively, primarily due to the complex spatial feature patterns that students struggle to fully capture in heterogeneous distillation. In particular, over-reliance on the teacher's logits, as seen in OFA-KD, may cause the student to overlook its own feature extraction strengths. Therefore, our research focuses on tailoring logit knowledge to integrate the feature representations from both teacher's and the student's architectural strengths, facilitating more effective knowledge transfer in heterogeneous distillation.

To this end, we propose **H**eterogeneous **C**omplementary **D**istillation (HCD), a novel approach that leverages complementary teacher-student feature representations to map them into a shared logits space, where logits are decomposed and constrained for diverse knowledge transfer to the student. Specifically, outputs from the student's intermediate $i$-th stages are processed through a convolutional module and adaptive pooling, and then concatenated with the teacher's penultimate layer features. The concatenated features are then fed into the **C**omplementary **F**eature **M**apper (CFM) module, comprising fully connected layer, to produce shared logits output that integrates teacher and student spatial representations while preserving student-specific features. However, the direct learning of shared logits knowledge presents challenges, as they contain global contextual information that is difficult for the student to grasp. To solve this problem, we propose **S**ub-logit **D**ecoupled **D**istillation (SDD). It decomposes shared logits into $n$ sub-logits, which are then fused with the teacher's logits to ensure classification consistency. Furthermore, to promote diversity and non-redundant sub-logit representations, we introduce an **O**rthogonality **L**oss (OL) to ensure the $n$ sub-logits capture distinct characteristics. The HCD approach enables the student to effectively assimilate the teacher's feature knowledge while preserving its own strengths, thereby improving robustness and generalization. Extensive experiments on datasets, such as CIFAR-100, Fine-grained (e.g., CUB-200-2011, FGVC-Aircraft), and ImageNet-1K, have confirmed the effectiveness of our HCD method. We summarize the contributions of this paper as follows:

- We design a Complementary Feature Mapper (CFM) to produce shared logits space, which addresses the disparate features between the teacher and the student.
- We propose Sub-logit Decoupled Distillation (SDD) to decompose shared logits space and obtain various sub-logit, facilitating more effective and specialized knowledge transfer to the student.
- We introduce an orthogonal loss to emphasize sub-logit diversity, thereby improving the efficiency of knowledge transfer by promoting non-redundant feature learning.
- Extensive experiments on CIFAR-100, Fine-grained, and ImageNet-1K datasets consistently demonstrate that our proposed method outperforms current state-of-the-art KD methods, achieving superior performance.

## Related Work

In this section, we review existing homogeneous and heterogeneous knowledge distillation approaches, highlighting their differences below.

**Homogeneous Architecture Knowledge Distillation.** Knowledge distillation (KD) transfers the teacher's "dark knowledge" to a compact student, enabling the student to achieve performance comparable to the teacher's while reducing the model size. KD methods are broadly categorized into feature-based KD (Li et al. 2023c; Yang et al. 2022b; Zhao et al. 2024; Romero et al. 2015; Liu et al. 2023b; Tian, Krishnan, and Isola 2020; Li et al. 2023b; Chen et al. 2021; Liu et al. 2023a; Dong, Li, and Wei 2023; Guo et al. 2023; Huang et al. 2023; Kim et al. 2024; Yang et al. 2022c, 2021; Wang et al. 2019; Li and Jin 2022) and logit-based KD (Zhao et al. 2022; Yang et al. 2023; Li et al. 2022; Huang et al. 2022; Xu et al. 2024; Wei, Luo, and Luo 2024; Peng et al. 2024; Zheng and Yang 2024; Sun et al. 2024; Jin, Wang, and Lin 2023). Feature-based KD methods use the teacher's intermediate features to improve the student's feature representations. For example, FitNets (Romero et al. 2015) aligns intermediate features across stages using mean squared error. NORM (Liu et al. 2023b) enables efficient many-to-one feature matching via expanded student representations. CRD (Tian, Krishnan, and Isola 2020) incorporates contrastive learning to improve knowledge transfer efficiency. ReviewKD (Chen et al. 2021) introduces a review learning mechanism that improves the efficiency of knowledge transfer from the teacher. Despite their effectiveness, these methods often incur high computational and memory cost, prompting a shift toward logit-based distillation approaches. For instance, DKD (Zhao et al. 2022) decouples logit knowledge into target and non-target class KD. LDRLD (Xu et al. 2025) mines fine-grained logit relationships to improve the student's performance. IKD (Wang et al. 2025) uses intra-class distillation to mitigate prediction biases within the same class. WTTM (Zheng and Yang 2024) refines the temperature scaling in teacher matching. RLD (Sun et al. 2025) uses labeling information to dynamically refine teacher knowledge. TeKAP (Hossain et al. 2025) employs multi-perspective teacher knowledge via random sampling. Following the success of KD methods in homogeneous architecture, researchers have shifted to address architectural disparities between the teacher and the student.

**Heterogeneous Architecture Knowledge Distillation.** Architectural disparities between the teacher and student often limit the effectiveness of homogeneous distillation methods in heterogeneous settings. To tackle this challenge, researchers have proposed various methods (Hao et al. 2023; Li et al. 2024a; Lin et al. 2025; Zhao, Song, and Liang 2023; Zhou, Zhu, and Wu 2025; Huang et al. 2025; Lee et al. 2025; Zhang et al. 2025). For example, MLDR-KD (Yang et al. 2025) proposes a novel heterogeneous relational KD framework that retains dark knowledge while boosting confidence in the correct target. RSD (Zhang et al. 2025) extracts architecture-agnostic knowledge by reducing redundant architecture-specific information. PAT (Lin et al. 2025) introduces a region-aware attention (RAA) module that aligns teacher representations by restructuring multilayer student features. LFCC (Wu et al. 2024) uses a multi-scale low-pass filter and its learnable derivative to compress low-frequency components of teacher and student features, reducing representational disparity and mitigating the impact of spatial noise. TCS (Zhou, Zhu, and Wu 2025) tailors a coordinate system to transfer "dark knowledge" from a task-agnostic teacher to task-specific student networks. HeteroAKD (Huang et al. 2025) transforms heterogeneous knowledge into the logit space, mitigating architecture-specific influences. CustomKD (Lee et al. 2025) customizes the generalized features of large vision foundation models (LVFMs) for the student, minimizing spatial representation differences. However, methods like HeteroAKD, LFCC, and PAT, which rely on all intermediate stages, incur high computational costs and decrease the efficiency of knowledge transfer. Consequently, heterogeneous KD approaches require further optimization to overcome these limitations. To address these challenges, our proposed method integrates complementary feature mapping from the teacher and student, sub-logit decoupled distillation, and orthogonal constraints to enhance efficiency and effectiveness.

# Method

In this section, we first review knowledge distillation (KD) methods and our motivation. We then propose Heterogeneous Complementary Distillation (HCD) for heterogeneous KD. Finally, we optimize the total loss function.

## Preliminary

KD was originally proposed by Hinton et al. (Hinton et al. 2014). By minimizing the output probability distributions of the teacher and the student via KL divergence, KD transfers the complex teacher's "dark knowledge" to the compact student. Specifically, the KD loss, $\mathcal{L}_{KL}$, uses KL divergence to measure the difference between the soft label distributions of the student and the teacher, while the cross-entropy loss, $\mathcal{L}_{CE}$, supervises the student's classification performance. The total distillation loss combines two terms as follows:

$$\begin{aligned} \mathcal{L}_{total} &= \alpha \mathcal{L}_{CE}(\sigma(\mathbf{z}^s), \mathbf{y}) + (1-\alpha)\mathcal{L}_{KL}(\boldsymbol{p}^t || \boldsymbol{p}^s) \\ &= -\alpha \times \mathbf{y} \log \sigma(\mathbf{z}^s) + (1-\alpha)\boldsymbol{p}^t \log(\boldsymbol{p}^t/\boldsymbol{p}^s), \end{aligned} \tag{1}$$

where $\boldsymbol{p}^s = \sigma(\mathbf{z}^s/\tau)$ and $\boldsymbol{p}^t = \sigma(\mathbf{z}^t/\tau)$ denote the student's and the teacher's soft label distributions derived from their logits $\mathbf{z}^s$ and $\mathbf{z}^t$ scaled by temperature $\tau$, $\sigma(\cdot)$ denotes the softmax function. The vector $\mathbf{y}$ represents the one-hot labels, and the parameter $\alpha$ balances the two loss terms.

## Motivation

Heterogeneous architectures, such as convolutional neural networks (CNNs), vision transformers (ViTs), and MLP-Mixers, generate distinct spatial feature representations due to their inherent inductive biases. For instance, CNNs emphasize locality and translation invariance, whereas ViTs excel at modeling global contextual relationships. Direct alignment of the intermediate feature representations using FitNets (Romero et al. 2015) between a student (e.g., CNN) and a teacher (e.g., ViT) often results in failure due to distinct representations. This consequently hinders effective knowledge transfer and degrades student performance. Intuitively, heterogeneous distillation can enrich knowledge diversity by leveraging differences in feature representation between architectures, thereby improving the student's performance. However, traditional homogeneous distillation methods, such as feature alignment in FitNets or logit distillation in KD, fail to address these representation disparities, leading to suboptimal student performance. As a result, they cannot fully exploit the inductive bias information of each architecture to enhance the student's performance. While existing heterogeneous distillation methods attempt to address this issue, they still have limitations. For example, OFA's single-stage logit mapping oversimplifies the process, limiting the student's feature representation capacity. Conversely, PAT's RAA mitigates feature representation misalignment effectively but incurs significant computational overhead. To address these issues, The proposed **H**eterogeneous **C**omplementary **D**istillation (HCD) method. The HCD integrates the teacher's deep feature representations with the student's intermediate feature representations and enhances the student's performance.

## Heterogeneous Complementary Distillation (HCD)

An overview of the proposed HCD framework is illustrated in Fig. 1, including CFM, SDD, and orthogonal loss.

## Complementary Feature Mapper

To address the challenge of aligning intermediate layer feature in the heterogeneous architecture KD, caused by differences in inductive biases, we propose a simple yet effective module designated as the **C**omplementary **F**eature **M**apper (CFM). The CFM concatenates the features from different intermediate stages of the student (e.g., $i$-th stage output) with the penultimate layer features of the teacher and maps them into shared logits space. Specifically, given a batch $x$ with $B$ samples for a ViT teacher, we extract the feature from the penultimate layer of the teacher's encoder $\mathcal{F}^t$, denoted as $\boldsymbol{f}^t = \mathcal{F}^t(x) \in \mathbb{R}^{B\times d}$, where $d$ represents the feature dimension (e.g., $d = 1024$ for ViT-L). For a CNN student, we extract the intermediate feature from the student's encoder $\mathcal{F}^s$, denoted as $\mathcal{G}_i^s = \mathcal{F}_i^s(x) \in \mathbb{R}^{B\times C_i \times H_i \times W_i}$ at the $i$-th stage output of the student, where $C_i$, $H_i$, and $W_i$ represent the number of channel, height, and width, respectively. The intermediate feature $\mathcal{G}_i^s$ is fed into a pair of convolutional blocks, each consisting of a $3 \times 3$ convolution (Conv), batch normalization (BN), and ReLU function, followed by adaptive average pooling (Pool) to yield $\boldsymbol{f}_i^s \in \mathbb{R}^{B\times m}$, where $m$ is the feature dimension (e.g., $m = 256$), is formulated as follows:

$$\boldsymbol{f}_i^s = \mathrm{Pool}\left(2 \times \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}\left(\mathcal{G}_i^s\right)\right)\right)\right) \in \mathbb{R}^{B\times m}. \tag{2}$$

To effectively exploit the complementary inductive biases of heterogeneous architectures, we propose the CFM module. In particular, we concatenate the mapped features $\boldsymbol{f}_i^s$ from the $i$-th stage of the student with the penultimate layer features of the teacher $\boldsymbol{f}^{\mathrm{t}}$, as formulated below:

$$\boldsymbol{f}_i^{\mathrm{cat}} = \mathrm{Concat}\left(\boldsymbol{f}_i^s, \boldsymbol{f}^t\right) \in \mathbb{R}^{B\times(d+m)}, \tag{3}$$

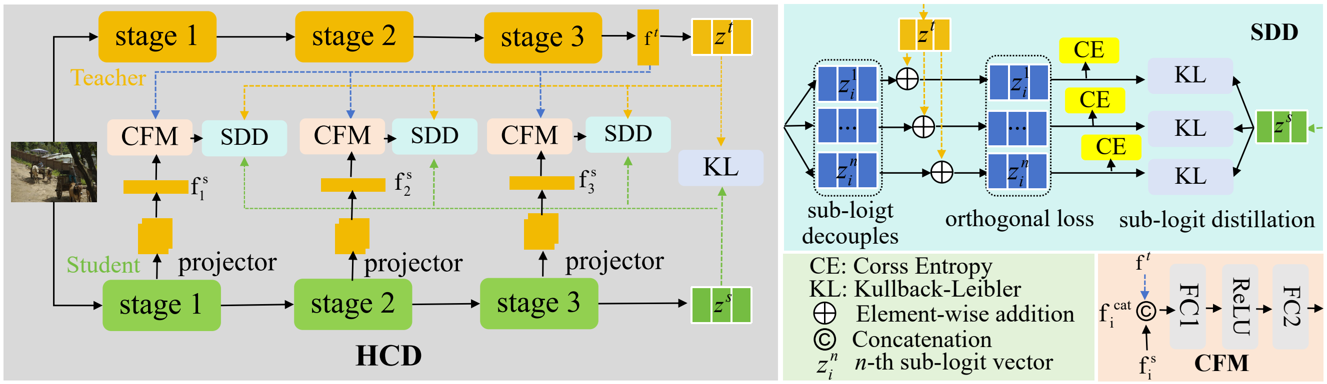


Figure 1: Overview of the proposed HCD framework, which includes the three components.

where $\mathrm{Concat}(\cdot)$ denotes the concatenation operation along the feature dimension. We then project the concatenated features, $\boldsymbol{f}_i^{\text{cat}}$, through fully connected layer (FC), such as two layers, to map them into the shared logits space, $\mathbf{z}_i$, as expressed below:

$$\mathbf{z}_i = \mathrm{FC}_2\left(\mathrm{ReLU}\left(\mathrm{FC}_1\left(\boldsymbol{f}_i^{\text{cat}}\right)\right)\right) \in \mathbb{R}^{B\times K}, \tag{4}$$

where $\mathrm{FC}_1 : \mathbb{R}^{d+m} \to \mathbb{R}^d$ and $\mathrm{FC}_2 : \mathbb{R}^d \to \mathbb{R}^K$, and $K$ represents the number of classes. As a result, this process can reduce the information gap between the teacher and the student. The proposed module effectively mitigates misalignment between spatial feature representations and leverages the complementary inductive biases of the teacher and the student to improve performance.

In this way, CFM effectively leverages the strengths of both the teacher and the student in two key aspects. First, CFM integrates low-level feature maps from the student (e.g., edges and textures in the CNN) with high-level semantic features from the teacher (e.g., class-discriminative features in the ViT), thus enhancing the diversity of feature representations. Second, CFM combines the complementary inductive biases of heterogeneous architectures, such as the locality and translation invariance in CNN student and the global contextual relationships with attention mechanism of the ViT teacher, to generate more comprehensive feature representations. Consequently, CFM improves the effectiveness of heterogeneous knowledge distillation and enhances the robustness of the student.

## Sub-logit Decoupled Distillation

**Sub-logit Decomposition:** Directly aligning the student's logits with the CFM-mapped shared logits presents a challenge for optimization, as these logits incorporate global contextual representations that are difficult for the student to learn. To address this, we extend the CFM's $\mathrm{FC}_2$ output dimension from $\mathbb{R}^{B\times K}$ to $\mathbb{R}^{B\times(n\times K)}$ in Eq. (4), and refer to it as $\overline{\mathrm{FC}_2}$. Then, we decompos the shared logits output into $n$ sub-logits tailored to the student's representation capacity. Such decomposition enables the student to effectively capture the complementary knowledge. The decomposition process is defined as follows:

$$\mathbf{z}_i = \overline{\mathrm{FC}_2}\left(\mathrm{ReLU}\left(\mathrm{FC}_1\left(\boldsymbol{f}_i^{\text{cat}}\right)\right)\right) \in \mathbb{R}^{B\times(n\times K)}, \tag{5}$$

where $\mathbf{z}_i = \left[\mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^n\right]$, and $\mathbf{z}_i^j \in \mathbb{R}^{B\times K}$, $1 \leq j \leq n$.

**Sub-logit Augmentation:** To ensure each sub-logit $\mathbf{z}_i^j$ remains consistent with the teacher's predictions and to prevent deviation from the task objective, we fuse each sub-logit with the teacher's logits $\mathbf{z}^t$, expressed as follows:

$$\mathbf{z}_i^j \leftarrow \mathbf{z}_i^j \oplus \mathbf{z}^t, \quad \text{for each} \quad j = 1, 2, \ldots, n, \tag{6}$$

where the symbol $\oplus$ denotes element-wise addition. The fusion update of each sub-logit $\mathbf{z}_i^j$ is concatenated column-wise to obtain the final decomposed logits $\mathbf{z}_i'$ as follows:

$$\mathbf{z}_i' = \left[\mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^n\right] \in \mathbb{R}^{B\times(n\times K)}. \tag{7}$$

**Sub-logit Knowledge Transfer:** The goal of sub-logit is to enable the student to fully acquire the complementary knowledge from the teacher. Thus, we use KL divergence ($\mathcal{L}_{\text{KL}}$) to transfer decomposed logits knowledge from $\mathbf{z}_i'$ to the student's $\mathbf{z}^s$, and ensure accurate sub-logit classification using cross-entropy loss ($\mathcal{L}_{\text{CE}}$). Note that instead of aligning the teacher's raw logits $\mathbf{z}^t$, we aim to preserve the student's knowledge while incorporating the teacher's heterogeneous information to enhance comprehension. The sub-logit losses, $\mathcal{L}_{KL}^{sub}$ and $\mathcal{L}_{CE}^{sub}$, are expressed as follows:

$$\mathcal{L}_{KD}^{sub} = \frac{1}{l\times n}\sum_{i=1}^{l}\sum_{j=1}^{n}\mathcal{L}_{KL}(\boldsymbol{p}_i^j||\boldsymbol{p}^s), \tag{8}$$

$$\mathcal{L}_{CE}^{sub} = \frac{1}{l\times n}\sum_{i=1}^{l}\sum_{j=1}^{n}\mathcal{L}_{\text{CE}}(\sigma(\mathbf{z}_i^j), \mathbf{y}), \tag{9}$$

where $\boldsymbol{p}^s = \sigma(\mathbf{z}^s/\tau)$ and $\boldsymbol{p}_i^j = \sigma(\mathbf{z}_i^j/\tau)$ denote the student's and sub-logit's output probability distributions, respectively, which are obtained by normalizing their logits $\mathbf{z}^s$ and $\mathbf{z}_i'$, with the batch dimension $B$ omitted for simplicity. $l$ is the total number of stages of the student.

## Orthogonal Loss for Sub-logit Diversity

**Masking Ground-truth Label Position:** According to Eq. (9), cross-entropy loss enforces uniform probability distributions of sub-logits relative to the ground-truth label, leading to similar predictions for the same label across all sub-logits. To alleviate this problem and promote diversity among the sub-logits $\mathbf{z}_i^j$, while excluding the influence of ground-truth label position prediction, we modify the sub-logit by suppressing the value at the ground-truth label index $Y \in \{0, 1, \ldots, K-1\}$ for each sub-logit. We adjust the sub-logit vector, $\mathbf{z}_i^j \in \mathbb{R}^K$, using a mask-based approach as follows:

$$\mathbf{z}_i^j \leftarrow \mathbf{z}_i^j \odot (1 - \mathbf{m}_i) - \epsilon \cdot \mathbf{m}_i, \tag{10}$$

where the symbol $\odot$ denotes element-wise multiplication, $\mathbf{m}_i \in \{0,1\}^K$ is a one-hot vector, where $\mathbf{m}_{i,k} = \delta_{k,Y}$, and $\delta_{k,Y} = 1$ if $k = Y$, otherwise 0, $k$ represents the index of the maximum value of $\mathbf{z}_i^j$, and $\epsilon = 10^{-6}$ prevents zero values in the ground-truth labels to ensure numerical stability. This modification ensures that the ground-truth label index does not contribute to the diversity loss, enabling effective diversification of non-target class predictions across sub-logits.

**Orthogonal Loss:** To fully utilize the discriminative information among sub-logits while reducing their similarity, we impose orthogonal constraints to promote the diversity of the sub-logits, thereby enabling the student to effectively acquire the teacher's global spatial feature representation. Following Eq. (7) and Eq. (10) that $\mathbf{z}'_i = \left[\mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^n\right] \in \mathbb{R}^{n \times K}$, with the batch dimension $B$ omitted for simplicity, each sub-logit $\mathbf{z}_i^j$ is normalized to obtain its normalized vector $\bar{\mathbf{z}}_i^j$, which is defined as:

$$\bar{\mathbf{z}}_i^j = \frac{\mathbf{z}_i^j}{||\mathbf{z}_i^j||_2}, \quad \left\|\mathbf{z}_i^j\right\|_2 = \sqrt{\sum_{k=1}^{K} \left(\mathbf{z}_{i,k}^j\right)^2}, \tag{11}$$

where the vector $\bar{\mathbf{z}}_i^j$ satisfies $||\bar{\mathbf{z}}_i^j||_2 = 1$. Next, we compute the dot product of any two sub-logits $\bar{\mathbf{z}}_i^p$ and $\bar{\mathbf{z}}_i^q$, denoted by the matrix $\mathcal{A}_{pq}$ as follows:

$$\mathcal{A}_{pq} = \bar{\mathbf{z}}_i^p \cdot \bar{\mathbf{z}}_i^q = \frac{\mathbf{z}_i^p}{||\mathbf{z}_i^p||_2} \cdot \frac{\mathbf{z}_i^q}{||\mathbf{z}_i^q||_2}, \tag{12}$$

where $1 \leq q \leq n$ and $1 \leq p \leq n$. For $p = q$, the diagonal elements represent the self-correlation, where $\mathcal{A}_{pq} = 1$. To calculate the dot product between different vectors, we exclude the diagonal elements of the matrix $\mathcal{A}_{pq}$, obtaining a new matrix $\mathcal{A}'$, which contains all the off-diagonal elements:

$$\mathcal{A}' = \{\mathcal{A}_{pq} \mid p \neq q\}. \tag{13}$$

The orthogonal loss function computes the mean squared value of the off-diagonal elements of the matrix $\mathcal{A}'$. We enforce these off-diagonal elements to be small, ensuring that distinct sub-logits remain orthogonal. The orthogonal loss, $\mathcal{L}_{orth}$, is formulated as follows:

$$\mathcal{L}_{orth} = \frac{1}{i \times n(n-1)} \sum_i \sum_{p \neq q} \left[\max(0, \mathcal{A}' - \theta)\right]^2, \tag{14}$$

where $\theta$ serves as a threshold to prevent excessive orthogonality between sub-logits, with $\theta$= 0.5. The $\max(\cdot)$ term uses ReLU activation function as the loss.

## Overall Objective Function

Our proposed HCD method combines cross-entropy loss, KL divergence, and a diversity constraint on sub-logit, enabling the student model to capture the teacher's heterogeneous knowledge effectively and improve classification accuracy. The total HCD loss is defined as follows:

$$\mathcal{L}_{HCD} = \mathcal{L}_{CE} + \mathcal{L}_{CE}^{sub} + \lambda \mathcal{L}_{KL} + \beta \mathcal{L}_{KL}^{sub} + \omega \mathcal{L}_{orth}, \tag{15}$$

where $\mathcal{L}_{KL}$ denotes the standard KL divergence. $\lambda$, $\beta$, and $\omega$ are the weight coefficients used to balance their contributions, and detailed algorithm in the *Supplementary Material*.

# Experiment

## Experiment settings

**Datasets.** We evaluate the performance of the HCD method using the CIFAR-100 (Krizhevsky, Hinton et al. 2009) and ImageNet-1K (Deng et al. 2009) datasets. Specifically, CIFAR-100 consists of 100 classes, with a total of 50,000 samples, including 500 training images and 100 test images per class, each with a resolution of 32×32 pixels. To ensure consistency with our model's input requirements in heterogeneous distillation, we resize all CIFAR-100 images to 224×224 pixels before training. In contrast, ImageNet-1K contains 1,000 classes, with approximately 1.28 million training images and 50,000 validation images, each with a resolution of 224×224 pixels. CUB-200-2021 (Wah et al. 2011) consists of 11,788 images from 200 bird species, widely used for fine-grained classification with annotations like bounding boxes and keypoints. FGVC-Aircraft (Maji et al. 2013) consists of 10,000 images across 100 aircraft models, used for fine-grained classification tasks.

**Implementation details**, including comparative methods, model architectures, and training details, can be found in the supplementary material due to page constraints.

## Main Results and Analysis

**Image Classification Results on CIFAR-100.** To evaluate the effectiveness of our proposed HCD method, we first conducted heterogeneous KD experiments on the CIFAR-100 dataset. Our approach, based on logit distillation, employs various characteristics of both the teacher and student. Teacher models include Vision Transformer-based architectures (e.g., Swin-T, ViT-S, Mixer-B/16) and student models consist of ResNet18, MobileNetV2. As shown in Table 1, HCD outperforms existing logit-based KD methods, including OFA, DKD, TeKAP, and LDRLD, achieving a significant accuracy improvement of 4-8% over traditional knowledge distillation (Vanilla KD), demonstrating its superiority in heterogeneous distillation tasks. This effectiveness stems from the ability of HCD to leverage complementary information from heterogeneous architectures by mapping intermediate student features and deep teacher features to shared logits space, followed by sub-logit decomposition and orthogonal constraints to reduce semantic discrepancies, thereby effectively enhancing knowledge transfer.

**Image Classification Results on ImageNet-1K.** To evaluate the effectiveness of our proposed HCD method, we conducted heterogeneous and homogeneous KD experiments on

| Teacher | Student | From Scratch | | Feature-based | | | | | Logit-based | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | T:Acc | S:Acc | FitNet | RKD | CRD | PAT | RSD | KD | DKD | DIST | OFA | TeKAP | LDRLD | HCD | Δ |
| Swin-T | ResNet18 | 89.26 | 74.01 | 78.87 | 74.11 | 77.63 | 81.22 | **83.92** | 78.74 | 80.26 | 77.75 | 80.54 | 81.38 | 82.17 | **82.78** | **+4.04** |
| ViT-S | ResNet18 | 92.04 | 74.01 | 77.71 | 73.72 | 76.60 | 80.11 | **81.50** | 77.26 | 78.10 | 76.49 | 80.15 | 79.06 | 80.36 | **81.33** | **+4.07** |
| Mixer-B/16 | ResNet18 | 87.29 | 74.01 | 77.15 | 73.75 | 76.42 | 80.07 | **81.85** | 77.79 | 78.67 | 76.36 | 79.39 | 80.05 | 80.69 | **81.53** | **+4.24** |
| Swin-T | MobileNetV2 | 89.26 | 73.68 | 74.28 | 69.00 | 79.80 | 78.78 | **83.68** | 74.68 | 71.07 | 72.89 | 80.98 | 80.23 | 81.64 | **82.19** | **+7.51** |
| ViT-S | MobileNetV2 | 92.04 | 73.68 | 73.54 | 68.46 | 78.14 | 78.87 | **81.68** | 72.77 | 69.80 | 72.54 | 78.45 | 78.41 | 79.21 | **80.81** | **+8.04** |
| Mixer-B/16 | MobileNetV2 | 87.29 | 73.68 | 73.78 | 68.95 | 78.15 | 78.62 | **81.74** | 73.33 | 70.20 | 73.26 | 78.78 | 79.89 | 80.64 | **81.09** | **+7.76** |

Table 1: Evaluation of the top-1 accuracy (%) of student using ViT-based heterogeneous models on the CIFAR100.

| Teacher | Student | From Scratch | | Feature-based | | | | | Logit-based | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | T:Acc | S:Acc | FitNet | CC | RKD | CRD | RSD | KD | DKD | DIST | OFA | TeKAP | HCD | Δ |
| Swin-T | ResNet18 | 81.38 | 69.75 | 71.18 | 70.07 | 68.89 | 69.09 | **72.13** | 71.14 | 71.10 | 70.91 | 71.85 | 71.25 | **71.91** | **+0.77** |
| Mixer-B/16 | ResNet18 | 76.62 | 69.75 | 70.78 | 70.05 | 69.46 | 68.40 | **71.41** | 70.89 | 69.89 | 70.66 | 71.38 | 70.95 | **71.66** | **+0.77** |
| DeiT-T | MobileNetV2 | 72.17 | 68.87 | 70.95 | 70.69 | 69.72 | 69.60 | **72.18** | 70.87 | 70.14 | 71.08 | **71.39** | 70.23 | 71.38 | +0.51 |
| Swin-T | MobileNetV2 | 81.38 | 68.87 | 71.75 | 70.69 | 67.52 | 69.58 | **72.36** | 72.05 | 71.71 | 71.76 | 72.32 | 72.50 | **72.72** | **+0.67** |
| Mixer-B/16 | MobileNetV2 | 76.62 | 68.87 | 71.59 | 70.79 | 69.86 | 68.89 | **71.90** | 71.92 | 70.93 | 71.74 | 72.12 | 72.01 | **72.32** | **+0.40** |

Table 2: Evaluation of the top-1 accuracy (%) of student using ViT-based heterogeneous models on the ImageNet-1K.

| ResNet34 (teacher): 73.31% Top-1, 91.42% Top-5 accuracy. ResNet18 (student): 69.75% Top-1, 89.07% Top-5 accuracy. | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Features** | AT | OFD | CRD | ReviewKD | FCFD | CAT-KD | RSD | **Logits** | KD | CTKD | DKD | LSKD | SDD | WTTM | TeKAP | RLD | LDRLD | HCD | Δ |
| Top-1 | 70.69 | 70.81 | 71.17 | 71.61 | **72.24** | 71.61 | 72.18 | Top-1 | 70.66 | 71.32 | 71.70 | 71.42 | 71.44 | **72.19** | 71.35 | 71.91 | 71.88 | 72.18 | **+1.52** |
| Top-5 | 90.01 | 90.34 | 90.13 | 90.51 | **90.74** | 90.45 | - | Top-5 | 89.88 | 90.27 | 90.31 | 90.29 | 90.05 | - | 90.54 | 90.54 | 90.59 | **90.64** | **+0.76** |
| ResNet50 (teacher): 76.16% Top-1, 92.87% Top-5 accuracy. MobileNetV1 (student): 68.87% Top-1, 88.76% Top-5 accuracy. | | | | | | | | | | | | | | | | | | | |
| **Features** | AT | OFD | CRD | ReviewKD | FCFD | CAT-KD | RSD | **Logits** | KD | IPWD | DKD | LSKD | SDD | WTTM | TeKAP | RLD | LDRLD | HCD | Δ |
| Top-1 | 70.18 | 71.25 | 71.32 | 72.56 | **73.37** | 72.24 | 73.08 | Top-1 | 70.49 | 72.65 | 72.05 | 72.18 | 72.24 | 73.09 | 72.87 | 72.75 | 73.12 | **73.23** | **+2.74** |
| Top-5 | 89.68 | 90.34 | 90.41 | 91.00 | **91.35** | 91.13 | - | Top-5 | 89.92 | 91.08 | 91.05 | 90.80 | 90.71 | - | 91.05 | 91.18 | 91.43 | **91.46** | **+1.54** |

Table 3: Evaluation of the top-1 and top-5 accuracy (%) of student using same-architecture on the ImageNet-1K.

| | Dataset | CUB-200-2011 | | | | | | | | FGVC-Aircraft | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Teacher | Teacher's Acc | Top-1,Top-5 for ViT-B: 82.76, 97.50; ViT-L: 85.28, 98.03 | | | | | | | | Top-1,Top-5 for ViT-B: 74.56, 95.20; ViT-L: 88.48, 99.25 | | | | | | | |
| | Method | VGG8 | | ResNet20 | | MobileNetV2 | | ShuffleNetV2 | | VGG8 | | ResNet20 | | MobileNetV2 | | ShuffleNetV2 | |
| | | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 |
| SigLIP2 (ViT-B) | From Scratch | 43.91 | 70.97 | 44.14 | 72.68 | 38.23 | 65.53 | 53.49 | 78.34 | 66.06 | 88.39 | 63.81 | 88.21 | 70.24 | 90.46 | 62.26 | 90.21 |
| | KD | 56.08 | 83.81 | 42.04 | 73.70 | 44.13 | 57.66 | 42.77 | 72.94 | 70.75 | 93.31 | 56.04 | 87.40 | 66.64 | 91.39 | 61.65 | 89.80 |
| | DKD | 61.06 | 86.52 | 38.21 | 70.19 | 53.33 | 80.95 | 49.78 | 77.87 | 71.41 | 93.42 | 52.18 | 86.77 | 70.99 | 93.31 | 65.74 | 90.82 |
| | WTTM | 59.35 | 84.76 | 42.84 | 73.27 | 56.23 | 82.45 | 55.63 | 81.53 | 52.60 | 87.22 | 60.81 | 88.98 | 76.12 | **94.84** | 8.00 | 20.98 |
| | RKKD | 59.06 | 83.85 | 38.70 | 68.64 | 50.76 | 78.98 | 46.32 | 74.09 | 73.30 | 93.24 | 62.35 | 88.96 | 72.88 | 92.77 | 70.18 | 91.78 |
| | LDRLD | **63.34** | **87.00** | 50.60 | 77.60 | **59.15** | **84.07** | **55.80** | **81.79** | 70.30 | 91.33 | 59.08 | 86.44 | 75.28 | 94.18 | 71.18 | 92.17 |
| | HCD (Ours) | 62.16 | 86.23 | **50.98** | **78.89** | 55.32 | 81.25 | 54.70 | 80.74 | **75.04** | **92.53** | **66.17** | **90.11** | **76.23** | 93.97 | **72.76** | **92.33** |
| SigLIP2 (ViT-L) | From Scratch | 43.91 | 70.97 | 44.14 | 72.68 | 38.23 | 65.53 | 53.49 | 78.34 | 66.06 | 88.39 | 63.81 | 88.21 | 70.24 | 90.46 | 62.26 | 90.21 |
| | KD | 56.32 | 84.02 | 43.18 | 74.58 | 44.24 | 74.09 | 44.25 | 73.51 | 75.66 | 94.54 | 62.50 | 90.19 | 73.66 | 93.76 | 72.00 | 92.47 |
| | DKD | 58.31 | 84.82 | 35.23 | 67.42 | 52.19 | 80.86 | 49.55 | 77.30 | 75.55 | **94.59** | 53.53 | 87.46 | 75.28 | 94.18 | 72.04 | 93.49 |
| | WTTM | 60.17 | 85.38 | 40.94 | 70.61 | 55.63 | 81.48 | 54.09 | 79.89 | 62.14 | 90.18 | 65.38 | 89.86 | 78.37 | 94.95 | 70.42 | 90.22 |
| | RKKD | 56.53 | 82.72 | 36.95 | 67.74 | 49.83 | 77.70 | 47.17 | 74.28 | 76.57 | 94.54 | 66.16 | 90.82 | 76.12 | 94.30 | 75.64 | 93.43 |
| | LDRLD | 62.55 | 86.28 | 49.23 | 76.32 | **58.75** | **85.89** | **56.20** | **81.91** | 74.89 | 93.73 | 58.85 | 86.01 | 78.76 | 94.69 | 77.05 | 94.00 |
| | HCD (Ours) | **63.21** | **86.90** | **50.04** | **77.34** | 56.14 | 82.38 | 54.36 | 80.50 | **77.08** | 93.34 | **68.47** | **90.82** | **80.02** | **95.53** | **77.65** | **94.23** |

Table 4: Image classification accuracy (%) on fine-grained datasets using SigLIP2 as a teacher and CNN as a student.

the ImageNet-1K dataset using Top-1 and Top-1&5 accuracy metrics, respectively. The homogeneous architectures used in our experiments include the ResNet and MobileNet series. As shown in Table 2, our method outperforms existing logit-based KD approaches, such as OFA, with a Top-1 accuracy improvement of 0.40% to 0.77% over traditional KD. We further expand experiments on homogeneous distillation, Table 3 demonstrates that our method achieves a Top-1 accuracy improvement of 1.52% and 2.74%, and a Top-5 accuracy improvement of 0.76% and 1.54% for ResNet18 and MobileNetV1, further confirming its superiority. This performance stems from leveraging complementary information (e.g., low-level texture information and high-level discriminative features) through feature mapping and sub-logit decomposition, exhibiting strong scalability across heterogeneous and homogeneous architectures.

**Image Classification Results on Fine-grained datasets.** To further evaluate the performance of HCD and assess its generalizability, we also extended the experiments to fine-grained datasets. Leveraging recent advancements in the powerful feature representation capabilities of vision foundation models (VFMs), we employed the visual encoders of DINOv2 (Oquab et al. 2024) and SigLIP2 (Tschannen et al. 2025) as teacher models, with various CNNs acting as student models. As shown in Tables 4 and 5, compared to exsiting KD methods (e.g., DKD,LDRLD), HCD achieves supe-

| Teacher | Dataset | CUB-200-2011 | | | | | | | | FGVC-Aircraft | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Teacher's Acc | Top-1,Top-5 for ViT-S: 84.35, 96.29; ViT-L: 88.99, 98.12 | | | | | | | | Top-1,Top-5 for ViT-S: 71.68, 94.96; ViT-L: 83.41, 97.06 | | | | | | | |
| | Method | VGG8 | | ResNet20 | | MobileNetV2 | | ShuffleNetV2 | | VGG8 | | ResNet20 | | MobileNetV2 | | ShuffleNetV2 | |
| | | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 |
| DINOv2 (ViT-S) | From Scratch | 46.50 | 72.76 | 50.50 | 76.77 | 50.38 | 76.30 | 53.49 | 78.34 | 68.20 | 88.66 | 62.19 | 87.57 | 69.19 | 90.25 | 72.61 | 91.75 |
| | KD | 60.48 | 84.98 | 50.78 | 77.55 | 60.75 | 84.17 | 60.22 | 83.22 | 67.75 | 90.82 | 59.74 | 86.80 | 69.88 | 92.14 | 70.72 | 90.34 |
| | DKD | 64.82 | 87.02 | 46.48 | 77.34 | 64.03 | 86.78 | 62.63 | 85.61 | 55.03 | 87.91 | 24.21 | 65.96 | 52.18 | 87.43 | 47.46 | 84.91 |
| | WTTM | **70.71** | **89.21** | 52.66 | 78.37 | 61.06 | 84.16 | 56.32 | 80.62 | 70.78 | 91.78 | 63.19 | 87.46 | 72.52 | 93.03 | 69.34 | 90.49 |
| | RKKD | 61.41 | 85.17 | 51.14 | 77.22 | 59.77 | 83.34 | 60.77 | 83.78 | 67.33 | 90.52 | 58.54 | 86.26 | 70.51 | 92.02 | 68.71 | 90.34 |
| | LDRLD | 68.43 | 87.97 | 56.66 | 82.38 | **71.30** | **90.70** | **68.81** | **88.40** | 68.89 | 91.60 | 50.77 | 83.98 | 70.12 | 93.01 | 66.28 | 90.49 |
| | HCD (Ours) | 69.16 | 88.23 | **57.58** | **82.78** | 67.12 | 87.25 | 64.48 | 86.64 | **74.20** | **92.17** | **65.17** | **88.57** | **78.49** | **94.66** | **75.76** | **92.56** |
| DINOv2 (ViT-L) | From Scratch | 46.50 | 72.76 | 50.50 | 76.77 | 50.38 | 76.30 | 53.49 | 78.34 | 68.20 | 88.66 | 62.19 | 87.57 | 69.19 | 90.25 | 72.61 | 91.75 |
| | KD | 60.94 | 83.95 | 50.69 | 76.65 | 57.75 | 81.83 | 62.16 | 84.47 | 72.94 | 92.53 | 59.50 | 85.51 | 76.36 | 93.67 | 73.71 | 92.05 |
| | DKD | 66.31 | 86.90 | 39.70 | 71.21 | 62.80 | 85.28 | 63.65 | 85.12 | 61.33 | 90.19 | 27.90 | 71.86 | 53.41 | 88.48 | 49.95 | 85.45 |
| | WTTM | 69.24 | **88.92** | 52.73 | 79.15 | 60.75 | 82.71 | 56.20 | 81.36 | 73.78 | 92.56 | 63.25 | 87.43 | 76.18 | 93.10 | 70.42 | 90.22 |
| | RKKD | 60.99 | 84.88 | 51.61 | 78.31 | 60.39 | 83.79 | 60.65 | 83.24 | 72.22 | 92.13 | 62.77 | 87.61 | 76.45 | 93.46 | 73.21 | 91.60 |
| | LDRLD | **69.38** | 85.89 | 52.78 | 78.15 | 66.21 | 85.89 | 64.62 | 84.85 | 72.52 | 92.44 | 56.47 | 85.30 | 75.31 | 93.37 | 70.57 | 91.10 |
| | HCD (Ours) | 68.48 | 86.80 | **57.04** | **81.34** | **67.14** | **88.38** | **65.65** | **87.16** | **74.04** | **92.86** | **63.41** | **88.51** | **78.73** | **94.09** | **74.50** | **92.83** |

Table 5: Image classification accuracy (%) on fine-grained datasets using DINOv2 as a teacher and CNN as a student.

rior performance on the FGVC-Aircraft. This results demonstrate that HCD effectively uses heterogeneous complementary information to enhance feature representation in transfer learning for VFMs. However, the FGVC-Aircraft images exhibit distinct category differences, while CUB-200 presents greater learning difficulty due to its fine-grained features, resulting in better performance in heterogeneous distillation for the former in Tables 4 and 5. However, HCD outperforms LDRLD slightly on the CUB200 dataset, likely due to the fine-grained nature of the task, where LDRLD excels in capturing inter-class relationships.

## Ablation Studies

**Impact of the number of sub-logits ($n$).** We conducted an ablation study to investigate the impact of the number of decoupled sub-logits on the student's performance, using Swin-Tiny as the teacher and ResNet18 as the student, with $n$ ranging from 1 to 8. Table 6 indicates that $n$=4 yields the optimal performance, as four sub-logits strike a balance between student's performance and computational cost, optimizes diversity and capacity while enhancing heterogeneous knowledge transfer. Experimental analysis further revealed that at $n$=1, sub-logits compressed teacher's global features and the student's local patterns, leading to insufficient knowledge transfer, increased learning difficulty, and reducing generalization. However, at $n$=8, excessive sub-logits introduced redundant noise and computational overhead, resulting in performance degradation.

| T:Swin-Tiny | 89.26 | $n = 1$ | $n = 2$ | $n = 4$ | $n = 6$ | $n = 8$ |
|---|---|---|---|---|---|---|
| S:ResNet18 | 74.01 | 81.53 | 81.84 | **82.78** | 82.66 | 82.60 |
| Time (s/epoch) | - | **13.66** | 13.78 | 14.32 | 14.56 | 14.82 |

Table 6: Ablation study of $n$ on the CIFAR-100 dateset with 8 NVIDIA RTX 3090 GPUs.

**Impact of different loss functions.** We kept the weight coefficients of the cross-entropy loss ($\mathcal{L}_{CE}^{sub}$ and $\mathcal{L}_{CE}$) constant and conducted ablation experiments to evaluate the impact of different loss functions on the CIFAR-100 dataset, as shown in Table 7. Ablating the loss components reveals that each individual loss function yields a significant performance improvement over the baseline (e.g., over 3.0% increase in Top-1 accuracy). Combining the three losses ($\mathcal{L}_{KD}$, $\mathcal{L}_{KL}^{sub}$, and $\mathcal{L}_{orth}$) yields the best performance, highlighting the effectiveness of our multi-loss optimization.

| $\mathcal{L}_{KD}$ | $\mathcal{L}_{KL}^{sub}$ | $\mathcal{L}_{orth}$ | Swin-T ResNet18 | Mixer-B/16 MobileNetV2 | Swin-T MobileNetV2 | ViT-S MobileNetV2 |
|---|---|---|---|---|---|---|
| - | - | - | 74.01 | 73.68 | 73.68 | 73.68 |
| ✓ | | | 78.74 | 73.33 | 74.68 | 72.77 |
| ✓ | ✓ | | 82.32 (+3.58) | 80.52 (+7.19) | 81.42 (+6.74) | 80.39 (+7.62) |
| ✓ | ✓ | ✓ | **82.78** (+4.04) | **81.09** (+7.76) | **82.19** (+7.51) | **80.81** (+8.04) |

Table 7: Ablation study of different losses on student's performance on the CIFAR100 dataset.

## Conclusion

In this work, we propose Heterogeneous Complementary Distillation (HCD), a novel framework that leverages complementary teacher-student feature representations through the Complementary Feature Mapper module to align them in shared logits space. We also propose Sub-logit Decoupled Distillation that decomposes these logits into $n$ sub-logits, fused with teacher logits for specialized knowledge transfer, while an orthogonality loss ensures sub-logit diversity and non-redundancy knowledge. Extensive experiments on CIFAR-100, ImageNet-1K, and Fine-grained datasets demonstrate that HCD outperforms state-of-the-art KD methods, enhancing student robustness and generalization by effectively integrating heterogeneous knowledge while preserving student-specific strengths.


## Acknowledgments

The project was supported in part by Guangdong Major Project of Basic and Applied Basic Research (2023B0303000016), in part by National Natural Science Foundation of China (U21A20487, 62273082), in part by Shenzhen Technology Project (GJHZ20240218112504008), in part by Guangdong Technology Project (2023TX07Z126), in part by Shenzhen High-tech Zone Development Special Plan Innovation Platform Construction Project, the proof of concept center for high precision and high resolution 4D imaging.

## References


Chen, P.; Liu, S.; Zhao, H.; and Jia, J. 2021. Distilling knowledge via knowledge review. In *CVPR*, 5008–5017.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Dong, P.; Li, L.; and Wei, Z. 2023. Diswot: Student architecture search for distillation without training. In *CVPR*, 11898–11908.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Guo, Z.; Yan, H.; Li, H.; and Lin, X. 2023. Class Attention Transfer Based Knowledge Distillation. In *CVPR*, 11868–11877.

Hao, Z.; Guo, J.; Han, K.; Tang, Y.; Hu, H.; Wang, Y.; and Xu, C. 2023. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. *NeurIPS*, 36: 79570–79582.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2014. Distilling the knowledge in a neural network. In *NeurIPS Workshop*.

Hossain, M. I.; Akhter, S.; Hong, C. S.; and Huh, E.-N. 2025. Single Teacher, Multiple Perspectives: Teacher Knowledge Augmentation for Enhanced Knowledge Distillation. In *ICLR*.

Huang, T.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2022. Knowledge distillation from a stronger teacher. *NeurIPS*, 35: 33716–33727.

Huang, T.; Zhang, Y.; Zheng, M.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2023. Knowledge Diffusion for Distillation. In *NeurIPS*, volume 36, 65299–65316.

Huang, Y.; Hu, K.; Zhang, Y.; Chen, Z.; and Gao, X. 2025. Distilling Knowledge from Heterogeneous Architectures for Semantic Segmentation. In *AAAI*, 3824–3832.

Jin, Y.; Wang, J.; and Lin, D. 2023. Multi-level logit distillation. In *CVPR*, 24276–24285.

Kim, J.; You, J.; Lee, D.; Kim, H. Y.; and Jung, J.-H. 2024. Do Topological Characteristics Help in Knowledge Distillation? In *ICML*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Toronto, ON, Canada*.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1): 79–86.

Lee, J.; Das, D.; Hayat, M.; Choi, S.; Hwang, K.; and Porikli, F. 2025. CustomKD: Customizing Large Vision Foundation for Edge Model Improvement via Knowledge Distillation. *CVPR*.

Li, G.; Wang, Q.; Yan, K.; Ding, S.; Gao, Y.; and Xia, G.-S. 2024a. TAS: Distilling Arbitrary Teacher and Student via a Hybrid Assistant. *arXiv preprint arXiv:2410.12342*.

Li, L.; Bao, Y.; Dong, P.; Yang, C.; Li, A.; Luo, W.; Liu, Q.; Xue, W.; and Guo, Y. 2024b. Detkds: Knowledge distillation search for object detectors. In *ICML*.

Li, L.; Dong, P.; Li, A.; Wei, Z.; and Yang, Y. 2023a. Kd-zero: Evolving knowledge distiller for any teacher-student pairs. *NeuIPS*, 36: 69490–69504.

Li, L.; Dong, P.; Wei, Z.; and Yang, Y. 2023b. Automated knowledge distillation via monte carlo tree search. In *ICCV*, 17413–17424.

Li, L.; and Jin, Z. 2022. Shadow knowledge distillation: Bridging offline and online knowledge transfer. *NeuIPS*, 35: 635–649.

Li, X.-C.; Fan, W.-S.; Song, S.; Li, Y.; Yunfeng, S.; Zhan, D.-C.; et al. 2022. Asymmetric temperature scaling makes larger networks teach well again. *NeurIPS*, 35: 3830–3842.

Li, Z.; Li, X.; Yang, L.; Zhao, B.; Song, R.; Luo, L.; Li, J.; and Yang, J. 2023c. Curriculum temperature for knowledge distillation. In *AAAI*, 1504–1512.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, 7061–7070.

Lin, J.-H.; Yao, Y.; Hsu, C.-F.; Xie, H.-X.; Shuai, H.-H.; and Cheng, W.-H. 2025. Perspective-Aware Teaching: Adapting Knowledge for Heterogeneous Distillation. *ICCV*, 4178–4187.

Liu, D.; Kan, M.; Shan, S.; and CHEN, X. 2023a. Function-Consistent Feature Distillation. In *ICLR*.

Liu, X.; LI, L.; Li, C.; and Yao, A. 2023b. NORM: Knowledge Distillation via N-to-One Representation Matching. In *ICLR*.

Liu, Y.; Cao, J.; Li, B.; Hu, W.; Ding, J.; and Li, L. 2022. Cross-architecture knowledge distillation. In *ACCV*, 3396–3411.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; HAZIZA, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *CVPR*, 3967–3976.

Peng, B.; Fang, Z.; Zhang, G.; and Lu, J. 2024. Knowledge Distillation with Auxiliary Variable. In *ICML*, volume 235 of *Proceedings of Machine Learning Research*, 40185–40199. PMLR.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. Fitnets: Hints for thin deep nets. In *ICLR*.

Sun, S.; Ren, W.; Li, J.; Wang, R.; and Cao, X. 2024. Logit Standardization in Knowledge Distillation. In *CVPR*, 15731–15740.

Sun, W.; Chen, D.; Lyu, S.; Chen, G.; Chen, C.; and Wang, C. 2025. Knowledge distillation with refined logits. *ICCV*.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive representation distillation. In *ICLR*.

Tolstikhin, I. O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *NeurIPS*, 34: 24261–24272.

Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M. F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset. *CNS-TR-2011-001*.

Wang, L.; Xu, L.; Yang, X.; Huang, Z.; and Cheng, J. 2025. Debiased Distillation for Consistency Regularization. In *AAAI*, 7799–7807.

Wang, T.; Yuan, L.; Zhang, X.; and Feng, J. 2019. Distilling object detectors with fine-grained feature imitation. In *CVPR*, 4933–4942.

Wei, S.; Luo, C.; and Luo, Y. 2024. Scaled Decoupled Distillation. In *CVPR*, 15975–15983.

Wu, H.; Xiao, L.; Zhang, X.; and Miao, Y. 2024. Aligning in a Compact Space: Contrastive Knowledge Distillation between Heterogeneous Architectures. *arXiv preprint arXiv:2405.2405*.

Xu, L.; Liu, K.; Liu, J.; Wang, L.; Xu, L.; and Cheng, J. 2025. Local Dense Logit Relations for Enhanced Knowledge Distillation. *ICCV*.

Xu, L.; Ren, J.; Huang, Z.; Zheng, W.; and Chen, Y. 2024. Improving Knowledge Distillation via Head and Tail Categories. *TCSVT*, 34(5): 3465–3480.

Yang, C.; Zhou, H.; An, Z.; Jiang, X.; Xu, Y.; and Zhang, Q. 2022a. Cross-image relational knowledge distillation for semantic segmentation. In *CVPR*, 12319–12328.

Yang, J.; Martinez, B.; Bulat, A.; Tzimiropoulos, G.; et al. 2021. Knowledge distillation via softmax regression representation learning. In *ICLR*.

Yang, Y.; Ye, P.; Lin, W.; Li, K.; Wen, Y.; Hao, J.; and Chen, T. 2025. Multi-Level Decoupled Relational Distillation for Heterogeneous Architectures. *arXiv preprint arXiv:2502.06189*.

Yang, Z.; Li, Z.; Jiang, X.; Gong, Y.; Yuan, Z.; Zhao, D.; and Yuan, C. 2022b. Focal and global knowledge distillation for detectors. In *CVPR*, 4643–4652.

Yang, Z.; Li, Z.; Shao, M.; Shi, D.; Yuan, Z.; and Yuan, C. 2022c. Masked generative distillation. In *ECCV*, 53–69.

Yang, Z.; Zeng, A.; Li, Z.; Zhang, T.; Yuan, C.; and Li, Y. 2023. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *ICCV*, 17185–17194.

Zhang, W.; Liu, Y.; Ran, W.; and Ma, C. 2025. Cross-Architecture Distillation Made Simple with Redundancy Suppression. *ICCV*.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled Knowledge Distillation. In *CVPR*, 11953–11962.

Zhao, B.; Song, R.; and Liang, J. 2023. Cumulative spatial knowledge distillation for vision transformers. In *ICCV*, 6146–6155.

Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; and Chen, J. 2024. Detrs beat yolos on real-time object detection. In *CVPR*, 16965–16974.

Zheng, H.; Yang, S.; He, Z.; Yang, J.; and Huang, Z. 2025. Hierarchical cross-modal prompt learning for vision-language models. In *ICCV*, 1891–1901.

Zheng, K.; and Yang, E.-H. 2024. Knowledge Distillation Based on Transformed Teacher Matching. In *ICLR*.

Zhou, J.; Zhu, K.; and Wu, J. 2025. All You Need in Knowledge Distillation Is a Tailored Coordinate System. In *AAAI*, 22946–22954.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2024. Segment everything everywhere all at once. *NeurIPS*, 36: 19769–19782.