Distilling Knowledge from Heterogeneous Architectures for Semantic Segmentation

Yanglin Huang[1], Kai Hu[1,*], Yuan Zhang[1], Zhineng Chen[2], Xieping Gao[3]

[1]Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Xiangtan University
[2]School of Computer Science, Fudan University
[3]Key Laboratory for Artificial Intelligence and International Communication, Hunan Normal University
ylhuang@smail.xtu.edu.cn, {kaihu, yuanz}@xtu.edu.cn, zhinchen@fudan.edu.cn, xpgao@hunnu.edu.cn

Abstract

Current knowledge distillation (KD) methods for semantic segmentation focus on guiding the student to imitate the teacher's knowledge within homogeneous architectures. However, these methods overlook the diverse knowledge contained in architectures with different inductive biases, which is crucial for enabling the student to acquire a more precise and comprehensive understanding of the data during distillation. To this end, we propose for the first time a generic knowledge distillation method for semantic segmentation from a heterogeneous perspective, named *HeteroAKD*. Due to the substantial disparities between heterogeneous architectures, such as CNN and Transformer, directly transferring cross-architecture knowledge presents significant challenges. To eliminate the influence of architecture-specific information, the intermediate features of both the teacher and student are skillfully projected into an aligned logits space. Furthermore, to utilize diverse knowledge from heterogeneous architectures and deliver customized knowledge required by the student, a teacher-student knowledge mixing mechanism (KMM) and a teacher-student knowledge evaluation mechanism (KEM) are introduced. These mechanisms are performed by assessing the reliability and its discrepancy between heterogeneous teacher-student knowledge. Extensive experiments conducted on three main-stream benchmarks using various teacher-student pairs demonstrate that our *HeteroAKD* outperforms state-of-the-art KD methods in facilitating distillation between heterogeneous architectures.

1 Introduction

Knowledge Distillation (KD), as a model compression technique, has been extensively researched in the field of semantic segmentation and has achieved remarkable progress (Liu et al. 2019; He et al. 2019; Wang et al. 2020; Shu et al. 2021; Yang et al. 2022; Fan et al. 2023). According to the distillation position of the segmenters, existing KD methods can be roughly classified into two categories: logits-based and feature-based. Logits-based methods (See Figure 1a) follow the idea proposed in (Hinton, Vinyals, and Dean 2015), which forces the student to mimic the prediction distribution of the teacher to acquire more accurate knowledge. Differently, feature-based methods (See Figure 1b), inspired by (Romero et al. 2015), extend the form of taught knowledge from the prediction distribution to the feature representation of the model. It aims to enforce the feature consistency between the teacher-student pair.

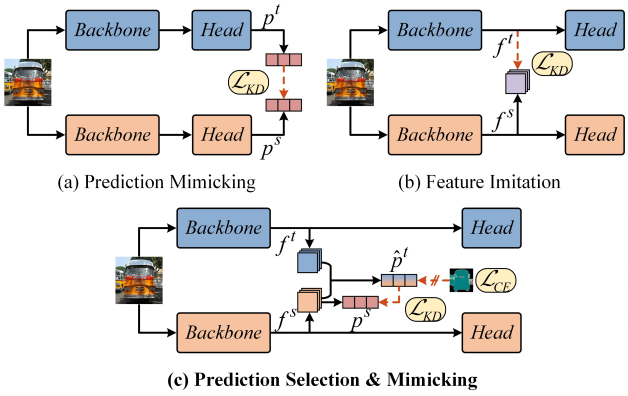


Figure 1: Comparison of the vanilla KD methods ((a) and (b)) with our *HeteroAKD* (c).

Currently, both logits-based (Shu et al. 2021; Baek et al. 2022) and feature-based (He et al. 2019; Liu, Zhang, and Wang 2022) KD approaches for semantic segmentation focus on knowledge transfer between teacher-student pairs in homogeneous architectures, while the distillation of heterogeneous architectures has not been explored. However, it is crucial to distill knowledge from heterogeneous architectures in practical scenarios. In general, architectures with different inductive biases tend to focus on distinct patterns, enabling them to understand the data from various perspectives to attain diverse knowledge (Ren et al. 2022). Therefore, gaining diverse knowledge from heterogeneous architectures enables students to achieve a more precise and comprehensive understanding of the data during distillation. For example, when distilling the student model on the ADE20K (Zhou et al. 2019) dataset, as our experiments will demonstrate, transferring knowledge from DeepLabV3-ResNet-101 to SegFormer-Mix Transformer-B1 (our *HeteroAKD* ΔmIoU: +3.66%) can easily surpass the performance increment achieved by transferring knowledge to DeepLabV3-ResNet-18 (Af-DCD (Fan et al. 2023)

*Corresponding author.

recorded ΔmIoU: +2.30%). Distilling knowledge from heterogeneous architectures thus provides another viable solution. Moreover, the continuous emergence of new architectures (Chen et al. 2022; Gu and Dao 2023) brings deeper understanding of the data, allowing researchers to enhance their own models using pre-trained teachers of different architectures.

Due to the substantial disparities between heterogeneous architectures, directly transferring knowledge from the teacher to the student presents significant challenges. This prompts us to consider: *how can a student effectively extract knowledge while retaining its own expertise when faced with a heterogeneous teacher?* Through an in-depth investigation of two types of main-stream architectures in KD, *i.e.*, CNN and Transformer, we argue that existing KD approaches face two key challenges: **(i)** the substantial disparities in the features learned by teachers and students with different inductive biases, as illustrated in Figure 2; **(ii)** the uncritical imitation may lead students to acquire erroneous knowledge which is caused by the fact that prediction made by teachers are not invariably superior to those made by students (See Figure 3).

To this end, we propose for the first time a generic **Hetero**geneous **A**rchitecture **K**nowledge **D**istillation framework for semantic segmentation, named *HeteroAKD* (See Figure 1c). To tackle the first challenge, instead of using any fancy tricks to bridge the intermediate feature gap between heterogeneous teacher-student pairs, we transfer the feature representations into the aligned logits space, which contains less architecture-specific information. By matching the output of the student's intermediate features with that of the teacher's in logits space, the student is constrained to approximate the teacher's performance. This manner to knowledge transfer in logits space circumvents directly imposing constraints on the students' intermediate features, thereby allowing the student more flexibility in learning intermediate feature representations that are conducive to downstream tasks (Zheng et al. 2023b).

To address the second challenge, we utilize human knowledge (*i.e.*, labels) to serve as the "textbook", guiding the process of knowledge transfer for students. Inspired by human educational practices (Midgley 2014), we propose a teacher-student knowledge mixing mechanism (KMM) and a teacher-student knowledge evaluation mechanism (KEM), to utilize diverse knowledge from heterogeneous architectures and deliver customized knowledge desired by the student. Specifically, prior to targeted instruction, the KMM assesses the reliability of knowledge by calculating the loss between intermediate feature outputs of both teacher and student against labels. This assessment guides the dynamic generation of more precise teacher-student hybrid knowledge, which incorporates contributions from both the teacher and student. Due to varying levels of student mastery of different knowledge at different times (Yang et al. 2024), directly imitating teacher-student hybrid knowledge may not be an optimal choice. The KEM further utilizes the knowledge reliability discrepancy between teacher and student to evaluate the relative importance of knowledge, which can deliver the customized knowledge according to the student's ability. As the learning progresses, the KEM progressively guides the student to master more difficult knowledge to increase the upper performance limit.

In summary, our main contributions are listed as follows:

- We propose a novel *HeteroAKD* framework, which transfers heterogeneous architecture knowledge in the logits space, to eliminate the influence of architecture-specific information. To the best of our knowledge, this is the first generic knowledge distillation method for semantic segmentation explored from a heterogeneous perspective.
- We propose a teacher-student knowledge mixing mechanism and a teacher-student knowledge evaluation mechanism based on human knowledge guidance to utilize diverse knowledge from heterogeneous architectures and deliver customized knowledge desired by the student.
- Extensive experiments on three main-stream benchmarks demonstrate the superiority of our *HeteroAKD* in facilitating distillation between heterogeneous architectures.

2 Related Work

Knowledge Distillation. KD is an effective method for transferring valuable knowledge from a complex teacher model to a simpler student model. Currently, KD methods can be broadly categorized into logits-based and feature-based approaches according to the distillation position. Logits-based KD methods (Hinton, Vinyals, and Dean 2015; Zhou et al. 2021) required the student model to replicate the class probability distribution of the teacher model. Feature-based KD methods (Romero et al. 2015; Chen et al. 2021a,b; Hao et al. 2022) transferred detailed feature activation from the teacher model to supervise the learning process of the student model. As models with diverse inductive biases offer a more comprehensive depiction of the data, recent methods have begun to incorporate distillation techniques based on heterogeneous architectures, leading to promising performance across various tasks, such as classification (Ren et al. 2022; Hao et al. 2023), face recognition (Zhao et al. 2023) and monocular depth estimation (Zheng et al. 2024).

Knowledge Distillation in Semantic Segmentation. Since semantic segmentation is an intensive predictive task, direct application of KD methods designed for other tasks may not yield satisfactory results. Thus, specific KD methods have been proposed for semantic segmentation. For example, SKD (Liu et al. 2019) directly aligned the similarity between teacher-student pairs at the pixel-wise level, while CWD (Shu et al. 2021) transferred meaningful knowledge by simply minimizing the channel-wise pixel distribution between teacher-student pairs. In addition, Af-DCD (Fan et al. 2023) proposed a contrastive distillation learning paradigm to utilize feature partitions across both channel and spatial dimensions for knowledge transfer. Furthermore, some methods explore the inherent knowledge among different samples. Among them, IFVD (Wang et al. 2020) forced the student to mimic teacher intra-class relations by assessing distances with prototypes from different classes. Similarly, CIRKD (Yang et al. 2022) built pixel dependencies

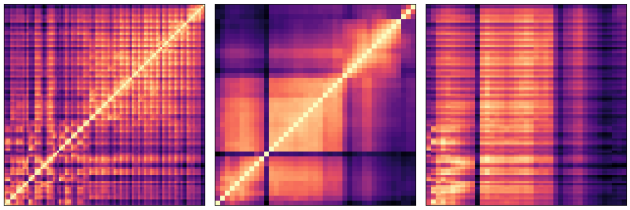


Figure 2: Similarity heatmap of intermediate features measured by centered kernel alignment (CKA). We compare features from ResNet-101 (CNN) and Mix Transformer-B4 (Transformer). Best viewed with zoom in.

across global samples to transfer structured relations knowledge. Despite achieving remarkable performance in existing distillation methods for semantic segmentation, they assume that the student and teacher architectures are homogeneous. However, when the architectures are heterogeneous, these methods may fail due to significant variability between the student and teacher. C2VKD (Zheng et al. 2023a) attempts to learn a compact Transformer-based model from a cumbersome yet high-performance CNN-based model, but this approach is still limited to transforming knowledge in a single mode. Therefore, how to distill knowledge from any heterogeneous architectures for semantic segmentation remains an open problem.

3 Methodology

3.1 Preliminary

Notations of Knowledge Distillation. Logits and features are the most common used types of knowledge in KD. A naive logits-based method is to train the student to mimic the class probability distribution of each pixel of the teacher, which can be defined as:

$$\mathcal{L}_{kd} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} KL(\sigma(\frac{\mathbf{Z}^s_{h,w}}{\tau}) \| \sigma(\frac{\mathbf{Z}^t_{h,w}}{\tau})), \quad (1)$$

where $\sigma(\mathbf{Z}^s_{h,w}/\tau)$ and $\sigma(\mathbf{Z}^t_{h,w}/\tau)$ denote the soft class probabilities of the student and teacher models on the (h, w)-th pixel, respectively. $KL(\cdot)$ represents the Kullback-Leibler divergence function. τ is a temperature parameter.

Different from the logits-based method, the feature-based method encourages the student to mimic the more fine-grained teacher feature activation. The formulation can be expressed as:

$$\mathcal{L}_{fd} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} (\mathbf{F}^t_{h,w} - \psi(\mathbf{F}^s_{h,w}))^2, \quad (2)$$

where $\mathbf{F}^t_{h,w}$ and $\mathbf{F}^s_{h,w}$ denote the (h, w)-th pixel in features produced from the teacher and student models, respectively. $\psi(\cdot)$ is a feature projector that maps student model features to match the dimension of teacher model features.

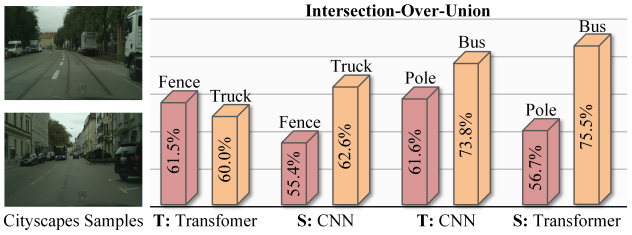


Figure 3: Analysis of IoU metrics for class probabilities predicted by CNN-based and Transformer-based architectures. We choose the first pair of teacher-student models for each mode in Table 1b for our analysis.

3.2 Analysis of Knowledge Distillation for Heterogeneous Architectures

To explore the impacts of the intrinsic differences of heterogeneous architectures (*i.e.*, CNN and Transformer) on knowledge distillation for semantic segmentation, we provide an analysis of logits-based and feature-based methods of knowledge distillation.

Centered Kernel Alignment Analysis. Inspired by (Hao et al. 2023), we employ minibatch centered kernel alignment (CKA) (Kornblith et al. 2019; Nguyen, Raghu, and Kornblith 2021) to compare the feature representations extracted by heterogeneous architectures in semantic segmentation. Suppose $\mathbf{X}_i \in \mathbb{R}^{n \times d_1}$ and $\mathbf{Y}_i \in \mathbb{R}^{n \times d_2}$ are features of the i-th minibatch of n samples extracted by CNN-based and Transformer-based models, with d_1 and d_2 neurons respectively. Let $\mathbf{K}_i = \mathbf{X}_i \mathbf{X}_i^\top$ and $\mathbf{L}_i = \mathbf{Y}_i \mathbf{Y}_i^\top$ denote the Gram matrices for the two feature representations (which reflects the similarities between a pair of samples according to feature representations), CKA can be computed as:

$$\text{CKA} = \frac{\frac{1}{k}\sum_{i=1}^{k} \text{HSIC}(\mathbf{K}_i, \mathbf{L}_i)}{\sqrt{\frac{1}{k}\sum_{i=1}^{k} \text{HSIC}(\mathbf{K}_i, \mathbf{K}_i)} \sqrt{\frac{1}{k}\sum_{i=1}^{k} \text{HSIC}(\mathbf{L}_i, \mathbf{L}_i)}}, \quad (3)$$

where k denotes the number of minibatch. HSIC is the Hilbert-Schmidt independence criterion (Gretton et al. 2007). In our implementation, we use an unbiased estimator of HSIC as proposed in (Song et al. 2012).

From Figure 2, we can observe that homogeneous architectures prefer to learn similar feature representations at layers of similar positions, whereas heterogeneous architectures only achieve similar feature representations at shallow layers. Existing feature-based distillation methods directly project teacher and student features to the same dimension, which is not a universal solution for aligning feature representations of heterogeneous architectures. *How to project features into a space that is unaffected by architecture-specific information is a key aspect in designing heterogeneous distillation methods.*

Class Probabilities Analysis. Heterogeneous architectures exhibit significant differences in their inner features

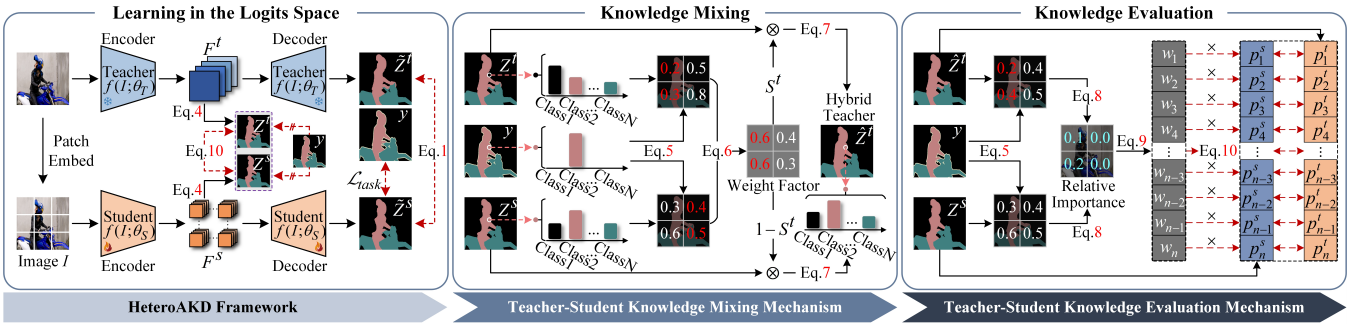


Figure 4: An overview of the *HeteroAKD* framework. Here, we take the "CNN→Transformer" mode as an example.

and output paradigms, which often leads to different class distributions (Huang et al. 2024). An intuitive idea is that a complex teacher is not always superior to a simple student. Instead, we believe that architectures with different inductive biases tend to learn more precise knowledge on particular patterns. To this end, we analyze the IoU metrics for class probabilities predicted by heterogeneous architectures, as shown in Figure 3. We can observe that teachers are inferior to the corresponding heterogeneous students in specific classes (*e.g.*, truck and bus). This indicates that heterogeneous architectures produces inconsistent understanding of knowledge from different perspectives, even if they are learning from the same dataset. Existing logits-based methods naively mimic the teacher's class probability distribution, which may lead to students acquiring erroneous knowledge. *How to utilize the diverse knowledge from heterogeneous architectures and deliver the knowledge required by the student is another key aspect in designing heterogeneous distillation methods.*

3.3 Proposed Heterogeneous Architecture Knowledge Distillation

An overview of the proposed *HeteroAKD* framework is illustrated in Figure 4. Our aim is to train a compact student model $f(I;\theta_S)$ by transferring diverse knowledge from a heterogeneous teacher model $f(I;\theta_T)$. This student model $f(I;\theta_S)$ possesses a more precise and comprehensive understanding of the data, enabling accurate assignment of a pixel-wise label $y_{h,w} \in 1,...,C$ to each pixel $p_{h,w}$ in image $i \in I$. Next, we will elaborate in detail on the key components that drive our framework.

Learning in the Logits Space. Given the input images I, we can obtain the intermediate feature representations ($\mathbf{F}^t \in \mathbb{R}^{H_1\times W_1\times d_1}$ and $\mathbf{F}^s \in \mathbb{R}^{H_2\times W_2\times d_2}$) from the heterogeneous teacher model $f(I;\theta_T)$ and student model $f(I;\theta_S)$. As analyzed in Section 3.2, directly aligning feature representations $\mathbf{F}^t$ and $\mathbf{F}^s$ is extremely challenging. To this end, we propose to project the intermediate features of the teacher $\mathbf{F}^t$ and student $\mathbf{F}^s$ into the logits space, thereby obtaining their respective categorical logit maps, designated as $\mathbf{Z}^t \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{Z}^s \in \mathbb{R}^{H\times W\times C}$, respectively. Here, H and W are the height and width of image $i \in I$, and C is the number of classes. $\mathbf{Z}^t$ and $\mathbf{Z}^s$ eliminate redundant architecture-specific information, and thus provide an ideal form of transferring knowledge from heterogeneous architectures (Hao et al. 2023). Moreover, performing knowledge distillation in the logits space circumvents directly imposing constraints on student's intermediate features $\mathbf{F}^s$, thereby allowing the student model $f(I;\theta_S)$ more flexibility in learning feature representations that are conducive for downstream tasks (Zheng et al. 2023b). This process can be formulated as:

$$\mathbf{Z}^t = \mathcal{G}_{proj}(\mathbf{F}^t), \quad \mathbf{Z}^s = \mathcal{G}_{proj}(\mathbf{F}^s), \tag{4}$$

where $\mathcal{G}_{proj}(\cdot)$ denotes a feature projector that is composed of 1×1 convolutional layer with BN and ReLU.

Teacher-Student Knowledge Mixing Mechanism. As analyzed in Section 3.2, the teacher may not outperform the student on a particular pattern. Our objective is to generate a teacher-student hybrid knowledge, which has a more precise and comprehensive understanding of the data. To this end, we treat labels as "textbook" that contain reliable knowledge generated by human intelligence. Thereby, the knowledge reliability of each pixel can be obtained by calculating the cross-entropy between the pixel-wise label $\mathbf{y}_{h,w}$ and class probability distribution $\sigma(\mathbf{Z}_{h,w})$ as:

$$\begin{aligned}\mathcal{H}(\mathbf{Z}_{h,w|c}) = -\,(&\mathbf{y}_{h,w}\log(\sigma(\mathbf{Z}_{h,w|c})) \\ &+ (1-\mathbf{y}_{h,w})\log(1-\sigma(\mathbf{Z}_{h,w|c}))),\end{aligned} \tag{5}$$

where $\mathbf{Z}_{h,w|c}$ denotes the categorical logit map at the position of (h,w) of the c-th channel. $\sigma(\cdot)$ is the sigmoid function. A lower cross-entropy value indicates a greater degree of similarity between the probability distribution of $\sigma(\mathbf{Z}_{h,w|c})$ and $\mathbf{y}_{h,w}$. Accordingly, we argue that a lower cross-entropy value reflects a higher degree of knowledge reliability. Following the established criteria for knowledge reliability, we perform a preference selection of teacher's and student's knowledge, which can be formulated as follows:

$$\mathbf{S}^t_{h,w|c} = 1 - \frac{\mathcal{H}(\mathbf{Z}^t_{h,w|c})}{\mathcal{H}(\mathbf{Z}^t_{h,w|c}) + \mathcal{H}(\mathbf{Z}^s_{h,w|c})}, \tag{6}$$

where $\mathbf{S}^t_{h,w|c}$ records the weight factor of pixels from the teacher, while $(1-\mathbf{S}^t_{h,w|c})$ denotes the weight factor of pixels from the student at the corresponding position. Accord-

Method	Params	FLOPs	Val mIoU
Mode: Transformer→CNN			
T: DeepLabV3-MiT-B4	63.5M	980.1G	75.89
S: DeepLabV3-Res18	13.6M	572.0G	74.53
+SKD			73.55
+IFVD			74.54
+CWD			73.39
+CIRKD			73.88
+Af-DCD			<u>75.64</u>
+HeteroAKD (Ours)			**76.35**
S: DeepLabV3-MBV2	4.1M	164.9G	73.92
+SKD			71.50
+IFVD			73.25
+CWD			71.59
+CIRKD			73.20
+Af-DCD			<u>73.69</u>
+HeteroAKD (Ours)			**74.91**
Mode: CNN→Transformer			
T: DeepLabV3-Res101	61.1M	2371.7G	78.34
S: DeepLabV3-MiT-B1	15.8M	275.9G	70.91
+SKD			72.04
+IFVD			72.43
+CWD			73.31
+CIRKD			72.87
+Af-DCD			<u>73.81</u>
+HeteroAKD (Ours)			**74.28**
S: DeepLabV3-PVT-B1	16.1M	293.9G	71.90
+SKD			72.74
+IFVD			73.32
+CWD			<u>73.94</u>
+CIRKD			73.69
+Af-DCD			73.81
+HeteroAKD (Ours)			**74.65**

(a) The same segmentation head with different backbone architectures

Method	Params	FLOPs	Val mIoU
Mode: Transformer→CNN			
T: SegFormer-MiT-B4	64.1M	1230.1G	78.80
S: DeepLabV3-Res18	13.6M	572.0G	74.53
+SKD			74.28
+IFVD			75.16
+CWD			73.53
+CIRKD			74.68
+Af-DCD			<u>75.46</u>
+HeteroAKD (Ours)			**76.42**
S: PSPNet-Res18	12.9M	507.4G	73.19
+SKD			71.19
+IFVD			72.94
+CWD			<u>73.74</u>
+CIRKD			72.60
+Af-DCD			71.97
+HeteroAKD (Ours)			**74.26**
Mode: CNN→Transformer			
T: DeepLabV3-Res101	61.1M	2371.7G	78.34
S: SegFormer-MiT-B1	13.7M	240.3G	74.91
+SKD			70.68
+IFVD			73.73
+CWD			74.80
+CIRKD			74.25
+Af-DCD			<u>75.20</u>
+HeteroAKD (Ours)			**76.34**
S: PSPNet-MiT-B1	15.1M	247.0G	71.29
+SKD			67.06
+IFVD			73.29
+CWD			<u>73.41</u>
+CIRKD			72.68
+Af-DCD			72.99
+HeteroAKD (Ours)			**74.25**

(b) Different segmentation heads with the same backbone architecture

Table 1: Comparison with state-of-the-art distillation methods on Cityscapes validation set. 'T' and 'S' denote the teacher and student, respectively. Params and FLOPs are measured according to CIRKD (Yang et al. 2022). The **best**/<u>second best</u> results are marked in bold/underline.

ing to the obtained weight factor, we can select a more accurate hybrid knowledge $\hat{\mathbf{Z}}^t_{h,w|c}$ from both the teacher and student. It can be formulated as follows:

$$\hat{\mathbf{Z}}^t_{h,w|c} = \mathbf{S}^t_{h,w|c} \odot \mathbf{Z}^t_{h,w|c} + (1 - \mathbf{S}^t_{h,w|c}) \odot \mathbf{Z}^s_{h,w|c}, \quad (7)$$

where $\odot$ denotes Hadamard product. Notably, it is difficult to obtain valuable information by directly utilizing $\mathbf{Z}^s_{h,w|c}$ generated by a naive student trained from scratch. In fact, such an approach may even compromise the accuracy of the hybrid knowledge $\hat{\mathbf{Z}}^t_{h,w|c}$. Therefore, we warm up the student model $f(I;\theta_S)$ under the full supervision of labels $\mathbf{y}$ before distillation.

Teacher-Student Knowledge Evaluation Mechanism. During distillation, an important pixel is one that the student has not yet fully grasped, but which can be acquired through learning from the teacher. To this end, we propose to evaluate the relative importance of pixels through the discrepancy between the hybrid teacher and student knowledge reliability, which can be utilized as a guidance to provide customized knowledge to the student. It is formulated as:

$$\Delta\mathcal{H}(\mathbf{Z}_{h,w|c}) = \mathbb{1}_+ \times (\mathcal{H}(\mathbf{Z}^s_{h,w|c}) - \mathcal{H}(\hat{\mathbf{Z}}^t_{h,w|c})), \quad (8)$$

where $\mathbb{1}_+$ is an indicator function which returns 1 if $\mathcal{H}(\mathbf{Z}^s_{h,w|c}) > \mathcal{H}(\hat{\mathbf{Z}}^t_{h,w|c})$ else 0. $\Delta\mathcal{H}(\mathbf{Z}_{h,w|c})$ denotes the relative importance of the pixel at the position (h,w) of the c-th channel. We further transform the relative importance of pixels into weight values by:

$$\mathbf{W}_{:,:|c} = \begin{cases} \frac{\exp(\mathcal{H}(\mathbf{Z}^s_{:,:|c}) + \Delta\mathcal{H}(\mathbf{Z}_{:,:|c}))}{\sum_{i=1}^{C} \exp(\mathcal{H}(\mathbf{Z}^s_{:,:|i}) + \Delta\mathcal{H}(\mathbf{Z}_{:,:|i}))}, & \Delta\mathcal{H}(\mathbf{Z}_{:,:|c}) > 0 \\ \frac{\exp(\mathcal{H}(\mathbf{Z}^s_{:,:|c}))}{\sum_{i=1}^{C} \exp(\mathcal{H}(\mathbf{Z}^s_{:,:|i}) + \Delta\mathcal{H}(\mathbf{Z}_{:,:|i}))}, & \Delta\mathcal{H}(\mathbf{Z}_{:,:|c}) \leq 0 \end{cases} \quad (9)$$

where $\mathbf{W}_{:,:|c}$ denotes the weight matrix of c-th category. According to $\mathbf{W}_{:,:|c}$, we reweight the original distillation loss (Eq. 1) to enhance the important information desired by the

Method	Params	FLOPs	Val mIoU
Mode: Transformer→CNN			
T: SegFormer-MiT-B4	64.1M	485.8G	80.27
S: DeepLabV3-Res18	13.6M	305.0G	74.53
+SKD			74.08
+IFVD			73.75
+CWD			71.43
+CIRKD			<u>74.87</u>
+Af-DCD			74.18
+HeteroAKD (Ours)			**75.44**
Mode: CNN→Transformer			
T: DeepLabV3-Res101	61.1M	1294.6G	78.82
S: SegFormer-MiT-B1	13.7M	89.0G	75.66
+SKD			72.70
+IFVD			74.70
+CWD			74.79
+CIRKD			75.23
+Af-DCD			<u>75.73</u>
+HeteroAKD (Ours)			**76.11**

(a) Pascal VOC

Method	Params	FLOPs	Val mIoU
Mode: Transformer→CNN			
T: SegFormer-MiT-B4	64.1M	485.8G	46.20
S: DeepLabV3-Res18	13.6M	305.0G	33.70
+SKD			34.38
+IFVD			34.54
+CWD			33.09
+CIRKD			<u>35.05</u>
+Af-DCD			34.68
+HeteroAKD (Ours)			**35.73**
Mode: CNN→Transformer			
T: DeepLabV3-Res101	61.1M	1294.6G	42.47
S: SegFormer-MiT-B1	13.7M	89.0G	35.18
+SKD			33.57
+IFVD			34.95
+CWD			33.74
+CIRKD			34.71
+Af-DCD			<u>36.74</u>
+HeteroAKD (Ours)			**38.84**

(b) ADE20K

Table 2: Comparison with state-of-the-art distillation methods on Pascal VOC and ADE20K validation sets. Params and FLOPs are measured according to CIRKD (Yang et al. 2022). The **best**/<u>second best</u> results are marked in bold/underline.

student as:

$$\mathcal{L}_{hakd} = -\frac{1}{C}\sum_{c=1}^{C}\sigma(\frac{\hat{\mathbf{Z}}^t_{:,:|c}}{\tau})\log(\sigma(\frac{\mathbf{Z}^s_{:,:|c}}{\tau})) \times \mathbf{W}_{:,:|c}, \quad (10)$$

where $\sigma(\hat{\mathbf{Z}}^t_{:,:|c}/\tau)$ and $\sigma(\mathbf{Z}^s_{:,:|c}/\tau)$ denote the soft class probabilities of the hybrid teacher and student on the c-th category. τ is a temperature parameter.

3.4 Optimization Objective

The overall loss for optimization can be formulated as the weighted sum of the task loss $\mathcal{L}_{task}$, class probability KD loss $\mathcal{L}_{kd}$ (Eq. 1), and heterogeneous architecture KD loss $\mathcal{L}_{hakd}$ (Eq. 10), written as:

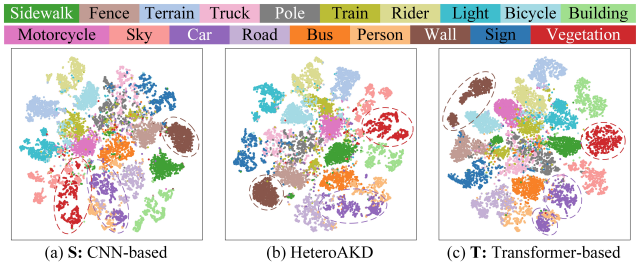
$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda_1\mathcal{L}_{kd} + \lambda_2\mathcal{L}_{hakd}, \quad (11)$$

where $\mathcal{L}_{task}$ is the cross-entropy loss for semantic segmentation task. λ_1 and λ_2 are weight factors used to balance the relationship between losses. In distillation losses, the projection heads used for teacher-student pairwise dimension matching are composed of 1×1 convolutional layer with BN and ReLU. They are discarded at the inference phase without introducing extra costs. Notably, the features from the last layer of the backbone architecture are used for distillation in our $\mathcal{L}_{hakd}$.

4 Experiments

4.1 Experimental Setups

Datasets. Our experiments are conducted on three popular semantic segmentation datasets, including Cityscapes (Cordts et al. 2016), Pascal VOC (Everingham et al. 2010) and ADE20K (Zhou et al. 2019).



(a) **S:** CNN-based (b) HeteroAKD (c) **T:** Transformer-based

Figure 5: T-SNE visualization of learned feature embeddings (*i.e.*, SegFormer-MiT-B4→DeepLabV3-Res18) on the Cityscapes dataset. We outline some classes with dash circles in their colors for a clearer view.

Implementation Details. Following the previous methods (Liu et al. 2019; Yang et al. 2022; Fan et al. 2023), we adopt DeepLabV3 (Chen et al. 2018), PSPNet (Zhao et al. 2017) and SegFormer (Xie et al. 2021) for segmentation heads, ResNet-101 (Res101) (He et al. 2016) and Mix Transformer-B4 (MiT-B4) (Xie et al. 2021) for teacher backbone architectures, ResNet-18 (Res18), MobileNetV2 (MBV2) (Sandler et al. 2018), Mix Transformer-B1 (MiT-B1) and Pyramid Vision Transformer v2-B1 (PVT-B1) (Wang et al. 2022) for student backbone architectures and group various teacher-student pairs. We compare the proposed *HeteroAKD* with state-of-the-art (SOTA) knowledge distillation methods for semantic segmentation: SKD (Liu et al. 2019), IFVD (Wang et al. 2020), CWD (Shu et al. 2021), CIRKD (Yang et al. 2022) and Af-DCD (Fan et al. 2023). We re-implemented all methods on both CIRKD codebase (Yang et al. 2022) and Af-DCD codebase (Fan et al. 2023). For crop size during the training phase, we use

Method	mIoU (%)	ΔmIoU (%)
Mode: Transformer→CNN		
Baseline	74.53	n/a
+$\mathcal{L}_{kd}$	75.67	+1.14
+$\mathcal{L}_{hakd}$	76.03	+1.50
+$\mathcal{L}_{kd}$ + $\mathcal{L}_{hakd}$	76.42	+1.89
+$\mathcal{L}_{kd}$ + $\mathcal{L}_{hakd}$ *w/o KMM*	76.19	+1.66
+$\mathcal{L}_{kd}$ + $\mathcal{L}_{hakd}$ *w/o KEM*	75.82	+1.29
Mode: CNN→Transformer		
Baseline	74.91	n/a
+$\mathcal{L}_{kd}$	75.64	+0.73
+$\mathcal{L}_{hakd}$	75.56	+0.65
+$\mathcal{L}_{kd}$ + $\mathcal{L}_{hakd}$	76.34	+1.43
+$\mathcal{L}_{kd}$ + $\mathcal{L}_{hakd}$ *w/o KMM*	75.87	+0.96
+$\mathcal{L}_{kd}$ + $\mathcal{L}_{hakd}$ *w/o KEM*	75.96	+1.05

Table 3: Ablation studies of loss terms and key components on Cityscapes validation set. The results are obtained using the first teacher-student pair for each mode in Table 1b.

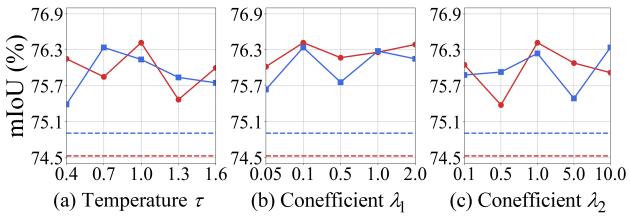
512×1024, 512×512 and 512×512 for Cityscapes, Pascal VOC and ADE20K, respectively.

4.2 Comparison with State-of-the-Art Methods

Results on Cityscapes. Table 1 presents the quantitative results of four backbone architectures and three segmentation heads on Cityscapes dataset. Our *HeteroAKD* consistently outperforms the baseline across all backbone architectures, with the maximum mIoU and average mIoU margin by 3.37% and 2.04%, respectively. Notably, in some cases (*e.g.*, DeepLabV3-MiT-B4→DeepLabV3-Res18), the student's performance after distillation is superior to that of the teacher by 0.46%. This indicates that students are not simply imitating their teachers to learn knowledge. In contrast, existing SOTA KD methods are heavily influenced by the challenges analyzed in Section 3.2, making it difficult to benefit from heterogeneous teachers.

As shown in Figure 5, we further analyze the feature embeddings learned by our *HeteroAKD* using T-SNE visualization. The visual results indicate that our *HeteroAKD* maintains its own advantageous feature embeddings while pushing students to imitate the teacher's feature embeddings. This facilitates students to achieve better intra-class compactness and inter-class separability, thus improving segmentation performance.

Results on Two Other Datasets. In Table 2, we compare the proposed *HeteroAKD* with the existing SOTA KD methods on PASCAL VOC and ADE20K datasets to validate the generalization of our method in solving different semantic segmentation tasks. According to the results shown in Table 2a-2b, our *HeteroAKD* consistently achieves the best performance in different heterogeneous distillation modes for different datasets. Compared to the SOTA KD methods, our model gains the maximal mIoU and average mIoU margin by 2.10% and 0.93%, respectively. The results demonstrate that our *HeteroAKD* is effective in facilitating knowledge distillation between heterogeneous teacher-student pairs in different semantic segmentation tasks.



(a) Temperature τ (b) Conefficient λ_1 (c) Conefficient λ_2

Figure 6: Ablation studies of (a) temperature τ, (b) $\mathcal{L}_{kd}$ coefficient λ_1 and (c) $\mathcal{L}_{hakd}$ coefficient λ_2 on Cityscapes validation set. Red and blue lines indicate "Transformer→CNN" and "CNN→Transformer" modes, respectively.

4.3 Ablation Studies

Ablation Study on Different Loss Terms. We analyze the contribution of each distillation loss. From the results shown in Table 3, we can get following observations: (i) Compared to *Baseline*, the introduction of either $\mathcal{L}_{kd}$ (0.94% average mIoU gain) or $\mathcal{L}_{hakd}$ (1.08% average mIoU gain) individually improves the performance of both knowledge transfer modes. (ii) The baseline continues to demonstrate improvement (1.66% average mIoU gain), with the combined contribution of both losses $\mathcal{L}_{kd} + \mathcal{L}_{hakd}$. This indicates that simultaneous learning from teacher intermediate features and output logits is beneficial for improving the student performance.

Ablation Study on Key Components. We verify the validity of the proposed KMM and KEM in $\mathcal{L}_{hakd}$. As shown in Table 3, we can see that removing either KMM (0.35% average mIoU reduction) or KEM (0.49% average mIoU reduction) brings a significant negative impact on performance. This illustrates that both components are instrumental in distilling knowledge from heterogeneous architectures.

Ablation Studies on Hyper-parameters. We investigate the impact of different hyper-parameter settings. As illustrated in Figure 6, our method consistently enables students to benefit from heterogeneous teachers, with the minimum mIoU gain of 0.48%. Different hyper-parameter settings have different impacts on distillation efficiency, this difference in optimal hyper-parameters can be attributed to the varying strengths of the teacher and student.

5 Conclusion

In this paper, we propose a generic knowledge distillation framework for semantic segmentation from a heterogeneous perspective, named *HeteroAKD*. Compared to previous methods, our *HeteroAKD* can help students learn more diverse knowledge from the heterogeneous teacher. Extensive experiments on three main-stream benchmarks demonstrate the superiority of our *HeteroAKD* framework in facilitating distillation between heterogeneous architectures. While our method makes significant progress in facilitating distillation between heterogeneous architectures, it is worth noting that in certain cases, the efficiency of knowledge distillation from a heterogeneous teacher may be lower than that achieved by a homogeneous teacher.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62272404 and 62372170, in part by the Natural Science Foundation of Hunan Province of China under Grant 2023JJ40638, and in part by the Research Foundation of Education Department of Hunan Province of China under Grant 23A0146.

References

Baek, D.; Oh, Y.; Lee, S.; Lee, J.; and Ham, B. 2022. Decomposed Knowledge Distillation for Class-Incremental Semantic Segmentation. In *Advances in Neural Information Processing Systems*, 10380–10392.

Chen, D.; Mei, J.-P.; Zhang, Y.; Wang, C.; Wang, Z.; Feng, Y.; and Chen, C. 2021a. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7028–7036.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European Conference on Computer Vision*, 801–818.

Chen, P.; Liu, S.; Zhao, H.; and Jia, J. 2021b. Distilling Knowledge via Knowledge Review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5008–5017.

Chen, S.; Xie, E.; Ge, C.; Chen, R.; Liang, D.; and Luo, P. 2022. CycleMLP: A MLP-like Architecture for Dense Prediction. In *Proceedings of the International Conference on Learning Representations*.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3213–3223.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2): 303–338.

Fan, J.; Li, C.; Liu, X.; Song, M.; and Yao, A. 2023. Augmentation-Free Dense Contrastive Knowledge Distillation for Efficient Semantic Segmentation. In *Advances in Neural Information Processing Systems*, 51359–51370.

Gretton, A.; Fukumizu, K.; Teo, C.; Song, L.; Schölkopf, B.; and Smola, A. 2007. A Kernel Statistical Test of Independence. In *Advances in Neural Information Processing Systems*, 585–592.

Gu, A.; and Dao, T. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752.

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. In *Advances in Neural Information Processing Systems*, 79570–79582.

Hao, Z.; Guo, J.; Jia, D.; Han, K.; Tang, Y.; Zhang, C.; Hu, H.; and Wang, Y. 2022. Learning Efficient Vision Transformers via Fine-Grained Manifold Distillation. In *Advances in Neural Information Processing Systems*, 9164–9175.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

He, T.; Shen, C.; Tian, Z.; Gong, D.; Sun, C.; and Yan, Y. 2019. Knowledge Adaptation for Efficient Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 578–587.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531.

Huang, H.; Huang, Y.; Xie, S.; Lin, L.; Tong, R.; Chen, Y.-W.; Li, Y.; and Zheng, Y. 2024. Combinatorial CNN-Transformer Learning with Manifold Constraints for Semi-supervised Medical Image Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2330–2338.

Kornblith, S.; Norouzi, M.; Lee, H.; and Hinton, G. 2019. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, 3519–3529.

Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Luo, Z.; and Wang, J. 2019. Structured Knowledge Distillation for Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2604–2613.

Liu, Y.; Zhang, W.; and Wang, J. 2022. Multi-Knowledge Aggregation and Transfer for Semantic Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1837–1845.

Midgley, C., ed. 2014. *Goals, Goal Structures, and Patterns of Adaptive Learning*. Routledge.

Nguyen, T.; Raghu, M.; and Kornblith, S. 2021. Do Wide and Deep Networks Learn the Same Things? Uncovering How Neural Network Representations Vary with Width and Depth. In *Proceedings of the International Conference on Learning Representations*.

Ren, S.; Gao, Z.; Hua, T.; Xue, Z.; Tian, Y.; He, S.; and Zhao, H. 2022. Co-Advise: Cross Inductive Bias Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16773–16782.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. FitNets: Hints for Thin Deep Nets. In *Proceedings of the International Conference on Learning Representations*.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Shu, C.; Liu, Y.; Gao, J.; Yan, Z.; and Shen, C. 2021. Channel-Wise Knowledge Distillation for Dense Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5311–5320.

Song, L.; Smola, A.; Gretton, A.; Bedo, J.; and Borgwardt, K. 2012. Feature Selection via Dependence Maximization. *Journal of Machine Learning Research*, 13(1): 1393–1434.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022. PVT v2: Improved baselines with Pyramid Vision Transformer. *Computational Visual Media*, 8(3): 415–424.

Wang, Y.; Zhou, W.; Jiang, T.; Bai, X.; and Xu, Y. 2020. Intra-Class Feature Variation Distillation for Semantic Segmentation. In *Proceedings of the European Conference on Computer Vision*, 346–362.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Advances in Neural Information Processing Systems*, 12077–12090.

Yang, C.; Zhou, H.; An, Z.; Jiang, X.; Xu, Y.; and Zhang, Q. 2022. Cross-Image Relational Knowledge Distillation for Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12319–12328.

Yang, S.; Yang, J.; Zhou, M.; Huang, Z.; Zheng, W.-S.; Yang, X.; and Ren, J. 2024. Learning From Human Educational Wisdom: A Student-Centered Knowledge Distillation Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6): 4188–4205.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid Scene Parsing Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2881–2890.

Zhao, W.; Zhu, X.; He, Z.; Zhang, X.-Y.; and Lei, Z. 2023. Cross-Architecture Distillation for Face Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8076–8085.

Zheng, X.; Luo, Y.; Zhou, P.; and Wang, L. 2023a. Distilling Efficient Vision Transformers from CNNs for Semantic Segmentation. arXiv:2310.07265.

Zheng, Z.; Huang, T.; Li, G.; and Wang, Z. 2024. Promoting CNNs with Cross-Architecture Knowledge Distillation for Efficient Monocular Depth Estimation. arXiv:2404.16386.

Zheng, Z.; Ye, R.; Hou, Q.; Ren, D.; Wang, P.; Zuo, W.; and Cheng, M.-M. 2023b. Localization Distillation for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 10070–10083.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic Understanding of Scenes Through the ADE20K Dataset. *International Journal of Computer Vision*, 127(3): 302–321.

Zhou, H.; Song, L.; Chen, J.; Zhou, Y.; Wang, G.; Yuan, J.; and Zhang, Q. 2021. Rethinking Soft Labels for Knowledge Distillation: A Bias–Variance Tradeoff Perspective. In *Proceedings of the International Conference on Learning Representations*.