Error-Decoupled Learning for Knowledge Distillation

Dongqin Liu[1,2], Hongchang Yang[1,2], Zhaoxing Li[1,2*], Jiao Dai[1,2], Jizhong Han[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{liudongqin, yanghongchang, lizhaoxing, daijiao, hanjizhong}@iie.ac.cn

***Abstract*—Knowledge distillation aims to transfer the embedded knowledge of complex models (teacher models) to simpler models (student models). It is known that teacher models can produce both correct and incorrect predictions, making it essential to extract valuable knowledge from accurate predictions while minimizing the impact of errors. However, current research either ignores this problem or utilizes methods like temperature scaling to smooth the teacher's output, which may somewhat reduce the negative effects of errors but also results in a loss of detailed correct information from the teacher. In this work, we propose Error-Decoupled Learning for Knowledge Distillation (EDLKD), a novel framework that explicitly disentangles correct and incorrect teacher predictions during the distillation process. By reinforcing reliable signals while suppressing erroneous ones, EDLKD ensures reliable knowledge transfer while minimizing noise. Although incorrect predictions may contain limited information, our ablation studies confirm that their suppression consistently enhances student performance. Extensive experiments on image classification and object detection demonstrate that EDLKD achieves state-of-the-art results in logit-based distillation.**

***Index Terms*—Knowledge distillation, Image classification, Object detection.**

I. Introduction

Deep neural networks have shown remarkable success in tasks such as image classification , object detection , and segmentation . Increasing the depth and width of these networks generally enhances model performance but also brings substantial computational and memory costs. Such demands pose significant challenges for resource-limited scenarios, including edge computing and smart devices. Model compression techniques have become a focal point of research to address this issue. Recently, in addition to traditional methods like network pruning [1] and quantization [2], knowledge distillation has also achieved notable progress.

Knowledge distillation is an effective model compression technique that transfers knowledge from a complex, high-capacity model (teacher model) to a simpler, compact model (student model), facilitating application across various fields. Early research [3] formally introduced knowledge distillation using a scaled softmax function at temperature scaled to soften teacher networks logs, which then serve as guidance for the student network. Over time, knowledge distillation approaches have evolved beyond sole reliance on logits, incorporating intermediate feature representations from the teacher model. Logit-based distillation remains a focal area due to its advantage of not requiring insight into the teacher's internal structure, with recent advancements [4]–[7] significantly boosting distillation effectiveness and driving renewed interest.

Logit-based distillation primarily relies on a distillation loss calculation based on Kullback-Leibler (KL) divergence. Critical directions for optimizing this approach include: (1) Enhancing distillation performance by adjusting the temperature parameter T in the softmax function, such as through dynamic temperature adjustment techniques [8], [9]. (2) Refining the model's logit outputs, e.g., via logit standardization [5], entropy-guided adaptive loss reweighting teacher and student output distributions [10], or decoupling target and non-target logits [7]. (3) Extracting more knowledge from the logits, such as y leveraging inter-sample, relationships [11]–[13].

Existing methods typically use the logits of the teacher model for distillation without handling the knowledge of incorrect predictions, which reduces distillation efficiency. Although temperature scaling smooths the probability distribution, reducing the influence of incorrect predictions, it does not eliminate them. When erroneous knowledge is transferred to the student model, it can degrade predictive performance.

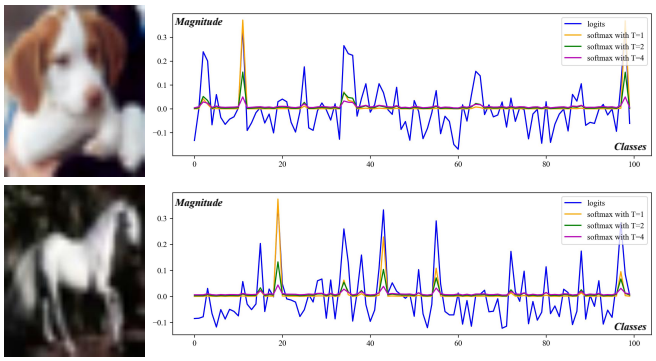

Fig. 1: The visualization shows the normalized logits from the teacher model ResNet32x4 on CIFAR-100, along with softmax outputs at varying temperatures. As temperature increases, the distribution becomes smoother but loses fine-grained detail.

In this work, we re-examine the current knowledge distillation framework and introduce a novel method, Error-Decoupled Learning for Knowledge Distillation (EDLKD). EDLKD decouples the correct and incorrect predictions of the teacher model during distillation, enhancing the reliability of knowledge transfer. By reinforcing the loss term L_c, correct

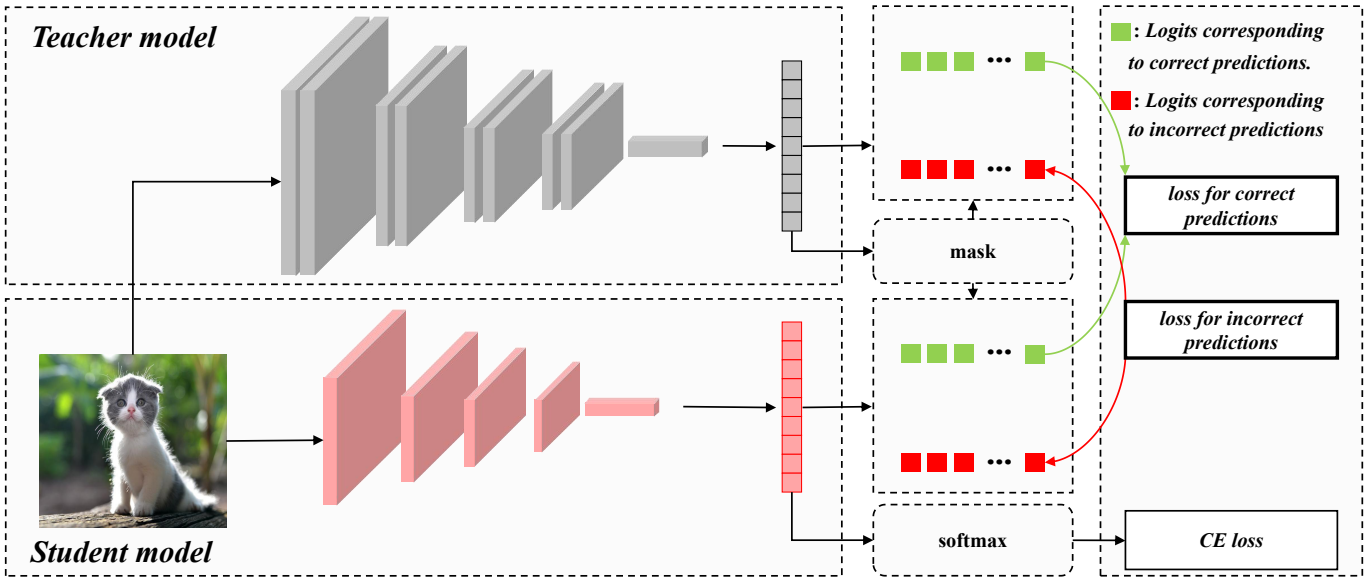

Fig. 2: Overview of the Error-Decoupled Learning for Knowledge Distillation (EDLKD) Architecture.

knowledge dominates the optimization, improving convergence stability and final accuracy, while erroneous knowledge is suppressed via L_{inc} to reduce low signal-to-noise transfer. Although incorrect predictions may contain a small number of valuable patterns (e.g., hard negatives), experiments demonstrate that decoupling erroneous knowledge and applying differentiated distillation weights with moderate suppression of incorrect predictions can significantly improve the accuracy of the student model. Furthermore, to preserve informative logit relationships that temperature scaling may over-smooth, we employ cosine distance in the logit space, maintaining angular relationships and capturing subtle differences more effectively than KL divergence. Our contributions are threefold:

1. Error Decoupling: We propose an innovative strategy that explicitly decouples correct and incorrect predictions during distillation, ensuring that correct knowledge dominates transfer while erroneous knowledge is actively suppressed.

2. Distillation: The student is encouraged to align closely with correct predictions via L_c while diverging from incorrect ones via L_{inc}, minimizing interference from low-quality signals while preserving useful high-difficulty patterns.

3. Experimental Validation: We comprehensively evaluate various student and teacher models on the CIFAR-100, ImageNet, and COCO2017 datasets to demonstrate the effectiveness and superiority of our method, which achieves state-of-the-art performance in logit distillation.

II. Methodology

In this chapter, we present our proposed Error-Decoupled Learning for Knowledge Distillation (EDLKD) method (Fig. 2), designed to enhance the precision and effectiveness of knowledge transfer from teacher models to student models. The method addresses critical limitations in existing logit-based distillation techniques, particularly the issue of incorrect predictions in the teacher model that can negatively impact the student model's learning process.

A. Preliminaries

Before delving into the specifics of our approach, we first define the key concepts and notations used throughout this methodology, particularly related to mainstream knowledge distillation and its limitations.

Knowledge Distillation. Knowledge distillation is a popular model compression technique, typically in which smaller student models learn from larger, pre-trained teacher models. This process typically involves the transfer of knowledge in the form of soft logits, representing the teacher model's output probability distribution over the class labels. In traditional distillation methods, the student model attempts to match the teacher's soft predictions by minimizing the KL divergence loss after applying the temperature-scaled softmax function, with such methods combining true labels and softened logits.

Logits and Softmax Function. Let $T(x)$ and $S(x)$ represent the teacher and student models respectively, with a single input x. The logits $z_T(x) \in \mathbb{R}^C$ from the teacher model are passed through the softmax function to generate the softened probabilities for C classes:

$$p_T(x) = \text{softmax}(z_T(x)/\tau) \tag{1}$$

where τ is the temperature parameter that controls the softness of the output distribution. The student model's logits $z_S(x)$ are similarly transformed into probabilities $p_S(x)$. Standard distillation minimizes the divergence between $p_T(x)$ and $p_S(x)$ using KL divergence, which penalizes differences in the two probability distributions.

$$L_{KD} = KL(p_T(x)||p_S(x)) \quad = \sum_{j=1}^{C} p_{T,j}(x) \log \left(\frac{p_{T,j}(x)}{p_{S,j}(x)} \right) \tag{2}$$

$p_{T,j}(x)$ and $p_{S,j}(x)$ denote the probabilities of the j-th class predicted by the teacher and student models, respectively. **Limitations of Traditional Distillation.** In mainstream knowledge distillation, the goal is to minimize the KL-divergence between the softened teacher logits $p_T(x)$ and the student logits $p_S(x)$. This objective function assumes all predictions from the teacher model are correct and valuable for the student's learning. However, treating both correct and incorrect predictions indiscriminately can be problematic. As noted in [9] erroneous teacher predictions can introduce "genetic errors" that misguide the student. It is crucial to systematically investigate how teacher mispredictions affect the quality and robustness of the distilled student model.

Let $y \in \{1, 2, \ldots, C\}$ denote the true label for input x. If the teacher model makes an incorrect prediction by assigning a higher probability to some wrong class j^* instead of the correct class y, then although the student is guided to mimic the teacher's probability distribution $p_T(x)$, this does not constitute entirely reliable knowledge—at least along the dimension corresponding to class j^*, the teacher's prediction is erroneous.For a correct prediction, the teacher assigns high confidence to the true label y:

$$p_{T,y}(x) > p_{T,j}(x) \quad \text{for} \quad j \neq y \tag{3}$$

For an incorrect prediction, the teacher assigns higher confidence to an incorrect class j^*:

$$p_{T,j^*}(x) > p_{T,y}(x) \quad \text{for} \quad j^* \neq y \tag{4}$$

In traditional distillation, the student learns by minimizing the divergence between the teacher and its own output distribution. However, when the teacher assigns a higher probability to an incorrect class j^*, the student is encouraged to replicate this error, which results in suboptimal performance. Specifically, the KL-divergence term penalizes the student for deviating from the teacher's incorrect prediction, as shown in the following case:

$$\begin{aligned} L_{KD}(x) &= \sum_{j=1}^{C} p_{T,j}(x)\, \log \frac{p_{T,j}(x)}{p_{S,j}(x)} \\ &= \underbrace{p_{T,j^*}(x)\, \log \frac{p_{T,j^*}(x)}{p_{S,j^*}(x)}}_{\text{error term from wrong class } j^*} + \sum_{\substack{j=1 \\ j \neq j^*}}^{C} p_{T,j}(x)\, \log \frac{p_{T,j}(x)}{p_{S,j}(x)}. \end{aligned} \tag{5}$$

When the teacher assigns a high confidence $p_{T,j^*}(x)$ to the incorrect class j^*, the first term becomes dominant, causing the student's loss to be largely determined by this erroneous prediction. In other words, using the KL divergence alone as the distillation term may enforce the transfer of the teacher's "incorrect knowledge" to the student, which can degrade the distillation performance when the teacher makes significant prediction errors.

Error Propagation in Distillation. The error propagation problem arises when the teacher's incorrect predictions are transferred to the student through the distillation process, thereby degrading the student's generalization capability. To quantify this effect, we analyze the expected performance degradation caused by distilling erroneous predictions. Assume that the student learns to mimic both correct and incorrect outputs from the teacher.

Let the teacher's overall error rate on the training distribution be

$$q_e = \Pr_{x \sim D}[\arg\max_j p_{T,j}(x) \neq y], \tag{6}$$

which measures how often the teacher produces an incorrect top-1 prediction. The expected contribution of erroneous predictions to the total loss is then

$$\mathbb{E}[L_{KD}^{\text{error}}] = q_e \cdot p_{T,j^*}(x), \log! \left(\frac{p_{T,j^*}(x)}{p_{S,j^*}(x)} \right), \tag{7}$$

where j^* denotes the wrongly predicted class. When either q_e or $p_{T,j^*}(x)$ is large, the student's optimization is dominated by the teacher's erroneous knowledge, leading to amplified error propagation.

Gradient Analysis and Buffering Effect. To better understand how this occurs during optimization, consider the gradient of the KD loss with respect to the student logits z_S (ignoring constants):

$$\frac{\partial L_{KD}}{\partial z_S} = \lambda\big(p_S(T) - q_T(T)\big). \tag{8}$$

Even when the teacher's top-1 prediction is incorrect, $q_T(T)$ often assigns proportionally higher probabilities to several semantically related "secondary" classes. This guides the student's non-target channels toward a more reasonable relative ranking (i.e., a pairwise or differential logits structure), thereby providing a smoother and more generalizable gradient field near the decision boundary. Meanwhile, the term $(1-\lambda)CE(y, p_S)$ still pulls the student toward the true label, mitigating the negative effect of the teacher's overconfident but incorrect predictions.

B. Error Decoupling

We first categorize the teacher model's predictions into two groups: correct and incorrect, determined by comparing the predicted class index with the ground-truth label. Predictions matching the ground truth are deemed correct, while mismatches are deemed incorrect.

Formally, consider a dataset of N examples $\{(x_i,\, y_{\text{true},i})\}_{i=1}^{N}$. For each input x_i with label $y_{\text{true},i} \in \{1, \ldots, C\}$, let the teacher logits be $y_{t,i} \in \mathbb{R}^C$ and the student logits be $y_{s,i} \in \mathbb{R}^C$. We define a binary mask:

$$M_i = \begin{cases} 1 & \text{if } \arg\max(y_{t,i}) = y_{\text{true},i}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

where M_i indicates whether the teacher's prediction on sample i is correct. We then split the teacher logits $y_{t,i}$ into $y_{t,i}^{c} = M_i\, y_{t,i}$ for correctly classified examples and $y_{t,i}^{\text{inc}} = (1 - M_i)\, y_{t,i}$ for misclassified ones, and apply the same mask to the student logits $y_{s,i}$ to obtain $y_{s,i}^{c} = M_i\, y_{s,i}$ and $y_{s,i}^{\text{inc}} = (1 - M_i)\, y_{s,i}$, thus partitioning both models' outputs into the "correct" pair $\big(y_t^{c},\, y_s^{c}\big)$ and the incorrect pair $\big(y_t^{\text{inc}},\, y_s^{\text{inc}}\big)$.

By explicitly decoupling the teacher's logits into correct and incorrect subsets, EDLKD establishes differentiated optimization objectives for each part. The *correct* predictions are emphasized to ensure that high-fidelity knowledge dominates the transfer process, leading to more stable convergence and improved final accuracy. In contrast, the *incorrect* predictions are regarded as low signal-to-noise components and are down-weighted to mitigate the propagation of teacher-specific biases. While these erroneous predictions may occasionally include informative patterns (e.g., hard negatives), empirical evidence shows that their contribution is insufficient to outweigh the accompanying noise. Ablation studies further confirm that suppressing this component consistently improves performance, validating both the limited utility and the suppression necessity of erroneous knowledge. Consequently, EDLKD fosters a distillation paradigm in which reliable knowledge exerts primary influence, whereas misleading signals are systematically reduced.

C. Distillation

The decoupled predictions are more reliable, so we discard the softmax processing to retain more accurate knowledge, which prevents the application of KL divergence to represent the similarity between the teacher and student outputs, as they are not probability distributions. Instead, we consider the logits from both models as features and distill the logits using a cosine distance loss function. The use of cosine distance is intended to preserve the angular relationships between the logits, ensuring that the student model captures the intricate patterns and relationships in the teacher model's outputs.

We independently represent the cosine distance loss of correct and incorrect predictions between the logits of the teacher and student models:

$$L_{\mathrm{c}} = 1 - \frac{y_t^{\mathrm{c}} \cdot y_s^{\mathrm{c}}}{\|y_t^{\mathrm{c}}\| \, \|y_s^{\mathrm{c}}\|}, \quad L_{\mathrm{inc}} = 1 - \frac{y_t^{\mathrm{inc}} \cdot y_s^{\mathrm{inc}}}{\|y_t^{\mathrm{inc}}\| \, \|y_s^{\mathrm{inc}}\|} \tag{10}$$

D. Final Loss Function

The final loss function for the student model is a combination of the cosine distance loss for correct predictions and the cosine distance loss for incorrect predictions. The overall loss L_{EDLKD} is defined as:

$$_{\mathrm{EDLKD}} = L_{\mathrm{CE}} + \lambda \cdot L_{\mathrm{c}} + \beta \cdot L_{\mathrm{inc}} \tag{11}$$

where L_{CE} denotes the standard cross-entropy loss. λ and β are hyperparameters controlling the contributions of the correct and incorrect prediction terms, respectively.

The positive weight λ encourages the student to align its correct predictions with the teacher's, thereby promoting the transfer of meaningful feature representations. The smaller weight β, when combined with the cross-entropy supervision that pulls the student toward the ground-truth labels, ensures that the student still receives limited guidance from the teacher's incorrect predictions but at a substantially reduced influence. In practice, the L_{CE} term realigns the student's logits toward the true label while a down-weighted β prevents teacher-specific erroneous signals from being reinforced; this combination softly suppresses noisy or misleading information yet preserves occasional useful cues (e.g., hard negatives). By explicitly weighting correct and incorrect knowledge in this way, the scheme implements a controlled knowledge-transfer regime: the student absorbs reliable patterns from the teacher while appropriately moderating contributions from less reliable signals, yielding improved generalization and robustness without overfitting the teacher's errors.

III. Experimental Evaluation

In this section, we comprehensively evaluate the proposed Error-Decoupled Learning for Knowledge Distillation (EDLKD) method through extensive experiments on diverse datasets and model architectures. Our results show that effectively handling correct and incorrect knowledge during distillation further enhances overall performance. The evaluation spans classification and object detection tasks, comparing EDLKD with state-of-the-art knowledge distillation techniques.

Datasets: We evaluate our method on three widely-used datasets: (1) CIFAR-100 [24], a dataset consisting of 60,000 images across 100 classes, with 50,000 images used for training and 10,000 for testing. (2) ImageNet [25], a large-scale dataset with over 1.2 million training images and 50,000 validation images across 1,000 classes. (3) COCO2017 [26], a challenging object detection dataset, contains over 200,000 labeled images with 80 object categories.

Settings: We focus on knowledge distillation with three different settings in our experiment section. (1) Homogenous architecture, where the teacher and student model are in the same type of architecture, and (2) Heterogeneous architecture, where the two models are different in architecture.We evaluate a diverse set of architectures, including VGG, ResNet, WideResNet, MobileNet, ShuffleNet [4], and modern models such as ConvNeXt, ViT, DeiT, MLP-Mixer, and Swin Transformer [23].

- *Feature distillation*: FitNet [14], CC [21], AT [15], RKD [16], CRD [17], ReviewKD [18], SimKD [19], FGFI [27] and CAT-KD [20], WKD [22].
- *Logits distillation*: KD [3], TAKD [28], DKD [7], CTKD [8], OFA [23], WKD [22], LS [5], MLKD [4] , EAKD [10] and LDLKD [13].

Evaluation Metrics: We evaluate the models based on standard metrics: Classification performance is assessed using top-1 accuracy on CIFAR-100 and ImageNet. Object detection performance is evaluated using Average Precision (AP, AP_{50}, AP_{75}) on COCO2017.

A. Classification

Training Details. For CIFAR-100. we follow the settings as works [4], [5], [7], [18].CNN models are trained for 240 epochs, with the learning rate decayed by 0.1 at the 150-th, 180-th, and 210-th epochs. The initial learning rate is set as 0.05, while 0.01 for MobileNet and ShuffleNet. The batch size is set as 64, and the weight decay is set as 0.0005

TABLE I: Knowledge distillation performance comparison between homogenous architectures. It reports Top-1 test accuracy on CIFAR-100. Performance comparison between different knowledge distillation methods.

Type	Teacher	ResNet32×4	VGG13	WRN-40-2	WRN-40-2	ResNet56	ResNet110	ResNet110
		79.42	74.64	75.61	75.61	72.34	74.31	74.31
	Student	ResNet8×4	VGG8	WRN-40-1	WRN-16-2	ResNet20	ResNet32	ResNet20
		72.50	70.36	71.98	73.26	69.06	71.14	69.06
Feature	FitNet [14]	73.50	71.02	72.24	73.58	69.21	71.06	68.99
	AT [15]	73.44	71.43	72.77	74.08	70.55	72.31	70.65
	RKD [16]	71.90	71.48	72.22	73.35	69.61	71.82	69.25
	CRD [17]	75.51	73.94	74.14	75.48	71.16	73.48	71.46
	ReviewKD [18]	75.63	74.84	75.09	76.12	71.89	73.89	71.34
	SimKD [19]	78.08	74.89	74.53	75.53	71.05	73.92	71.06
	CAT-KD [20]	76.91	74.65	74.82	75.60	71.62	73.62	71.37
Logit	KD [3]	73.33	72.98	73.54	74.92	70.66	73.08	70.67
	DKD [7]	76.32	74.68	74.81	76.24	71.97	74.11	71.06
	CTKD [8]	73.39	73.52	73.93	75.45	71.19	73.52	70.99
	KD+LS [5]	76.62	74.36	74.37	76.11	71.43	74.17	71.48
	EA+KD [10]	75.46	74.08	74.38	-	-	-	-
	MLKD [4]	77.08	75.18	75.35	76.63	72.19	74.11	71.89
	LDLKD [13]	77.20	75.06	74.98	76.35	77.20	74.16	71.98
	Ours	**77.85**	**76.25**	**75.75**	**77.04**	**72.25**	**74.55**	**72.29**

TABLE II: Knowledge distillation performance comparison between heterogeneous architectures. It reports Top-1 test accuracy on CIFAR-100. Performance comparison between different knowledge distillation methods.

Type	Teacher	ResNet32×4	ResNet32×4	ResNet32×4	WRN-40-2	WRN-40-2	VGG13	ResNet50
		79.42	79.42	79.42	75.61	75.61	74.64	79.34
	Student	SHN-V2	WRN-16-2	WRN-40-2	ResNet8×4	MN-V2	MN-V2	MN-V2
		71.82	73.26	75.61	72.50	64.60	64.60	64.60
Feature	FitNet [14]	73.54	74.70	77.69	74.61	68.64	64.16	63.16
	AT [15]	72.73	73.91	77.43	74.11	60.78	59.40	58.58
	RKD [16]	73.21	74.86	77.82	75.26	69.27	64.52	64.43
	CRD [17]	75.65	75.65	78.15	75.24	70.28	69.73	69.11
	ReviewKD [18]	77.78	76.11	78.96	74.34	71.28	70.37	69.89
	SimKD [19]	78.39	77.17	79.29	75.29	70.10	69.44	69.97
	CAT-KD [20]	78.41	76.97	78.59	75.38	70.24	69.13	71.36
Logit	KD [3]	74.45	74.90	77.70	73.97	68.36	67.37	67.35
	DKD [7]	77.07	75.70	78.46	75.56	69.28	69.71	70.35
	CTKD [8]	75.37	74.57	77.66	74.61	68.34	68.50	68.67
	KD+LS [5]	75.56	75.26	77.92	77.11	69.23	68.61	69.02
	EA+KD [10]	75.91	-	-	-	-	69.17	69.67
	MLKD [4]	78.44	76.52	79.26	77.33	70.78	70.57	71.04
	LDLKD [13]	77.33	-	-	-	-	70.11	70.74
	Ours	**79.17**	**77.50**	**80.06**	**77.69**	**71.41**	**71.37**	**71.57**

for all models. For distillation experiments involving vision transformers, we adopt the settings of [22], [23] and conduct experiments across Convolutional Neural Networks (CNNs) and vision transformers. For ImageNet, we use the mainstream training process [5], [7] that trains the model for 100 epochs, with the learning rate decayed by 0.1 at the 30-th, 60-th, and 90-th epochs. The initial learning rate is set as 0.1, the batch size is set as 512, and the weight decay is set as 0.0001.

Results on CIFAR-100. The results on CIFAR-100 demonstrate the effectiveness of our method in improving classification accuracy, as shown in Table I, II, III. We observe that EDLKD substantially improves distillation performance. On CNNs, EDLKD outperforms standard KD by 4.72% in top-1 accuracy when using ResNet-32×4 as the teacher and SHN-V2 as the student. For experiments involving vision transformers, EDLKD yields a 1.79% in top-1 accuracy gain with Swin-T as teacher and RN18 as student. The decoupling of correct and incorrect predictions contributes to a more robust learning process, resulting in better generalization on the test set. Compared to other distillation techniques, EDLKD consistently achieves higher accuracy across various student-teacher pairs.

Results on ImageNet. EDLKD continues to show superior performance, as shown in Table V. The method also performs competitively against more complex distillation techniques, highlighting its effectiveness even in large-scale scenarios.

B. Object Detection

Training Details. All the models used in this section are provided by Detectron2 [30]. Student models are trained with an initial learning weight of 0.01 for 18000 steps, with learning decayed by 0.1 at 12000-th and 16000-th steps.

Results on COCO2017. EDLKD achieves a notable increase in AP, AP_{50}, and AP_{75} compared to standard KD,

TABLE III: Knowledge distillation performance involving vision transformers. It reports Top-1 test accuracy on CIFAR-100 across CNNs and Transformers.

	TRANSFER	TRANSFORMER→CNN			CNN→TRANSFORMER	
TYPE	TEACHER	SWIN-T	VIT-S	VIT-S	CONVNEXT-T	CONVNEXT-T
		89.26	92.04	92.04	88.41	88.41
	STUDENT	RN18	RN18	MN-V2	DEIT-T	SWIN-P
		74.01	74.01	73.68	68.00	72.63
FEATURE	FITNET [14]	78.87	77.71	73.54	60.78	24.06
	CC [21]	74.19	74.26	70.67	68.01	72.63
	RKD [16]	74.11	73.72	68.46	69.79	71.73
	CRD [17]	77.63	76.60	78.14	65.94	67.09
	WKD [22]	81.57	81.12	79.11	73.27	74.80
LOGIT	KD [3]	78.74	77.26	72.77	72.99	76.44
	DKD [7]	80.26	78.10	69.8	74.60	76.8
	OFA [23]	80.54	80.15	78.45	75.76	78.32
	WKD [22]	81.42	80.81	79.04	76.11	78.94
	OURS	**83.21**	**80.85**	**80.04**	**76.26**	**79.52**

TABLE IV: Results on COCO2017 based on Faster-RCNN [29]-FPN [1]. Performance comparison between different knowledge distillation methods.

		mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
TYPE	TEACHER	RESNET101			RESNET101			RESNET50		
		42.04	62.48	45.88	42.04	62.48	45.88	40.22	61.02	43.81
	STUDENT	RESNET18			RESNET50			MN-V2		
		33.26	53.61	35.26	37.93	58.84	41.05	29.47	48.87	30.90
FEATURE	FITNET [14]	34.13	54.16	36.71	38.76	59.62	41.80	30.20	49.80	31.69
	FGFI [27]	35.44	55.51	38.17	39.44	60.27	43.04	31.16	50.68	32.92
	REVIEWKD [18]	36.75	56.72	34.00	40.36	60.97	44.08	33.71	53.15	36.13
LOGIT	KD [3]	33.97	54.66	36.62	38.35	59.41	41.71	30.13	50.28	31.35
	TAKD [28]	34.59	55.35	37.12	39.01	60.32	43.10	31.26	51.03	33.46
	DKD [7]	35.05	56.60	37.54	39.25	60.90	42.73	32.34	53.77	34.01
	CTKD [8]	33.26	53.61	35.26	-	-	-	31.39	52.34	33.10
	MLKD [4]	36.03	57.28	38.51	40.15	61.67	**44.57**	33.83	54.01	35.22
	OURS	**36.54**	**57.85**	**39.17**	**40.61**	**62.14**	44.23	**34.40**	**55.63**	**36.60**

TABLE V: Top-1 accuracy on ImageNet. Performance comparison between different knowledge distillation methods.

Type	Teacher	ResNet34	ResNet50
		73.31	76.16
	Student	**ResNet18**	**MN-V2**
		69.75	68.87
Feature	AT [15]	70.69	69.56
	CRD [17]	71.17	71.37
	ReviewKD [18]	71.61	72.56
	SimKD [19]	71.59	72.25
	CAT-KD [20]	71.26	72.24
Logit	KD [3]	71.03	70.50
	DKD [7]	71.70	72.05
	CTKD [8]	71.38	71.16
	KD+LS [5]	71.42	72.18
	MLKD [4]	71.90	73.01
	Ours	**71.95**	**73.11**

especially with smaller student models like MN-V2, as shown in Table IV. This method demonstrates effectiveness in transferring knowledge relevant to object detection, leading to improved performance.

C. Ablation Studies

To better understand the contribution of each component in EDLKD, we conduct ablation studies, as shown in Table VI:

Effect of Distance Loss: We compare cosine distance with KL divergence, finding that cosine distance better preserves the angular relationships in logits, leading to more accurate knowledge transfer.

Impact of Error Decoupling: Removing the error decoupling step, i.e., skipping the decoupling of logits into correct and incorrect predictions, leads to a significant performance drop, underscoring its crucial role in the distillation process.

Effect of Data Augmentation (**DA**): DA has been shown to improve the effectiveness of knowledge distillation [4], [31], [32]. To ensure a fair comparison, we adopt the same experimental setup as the MLKD [4] method. To further validate the effectiveness of our method, we also evaluate it without data augmentation. The results, shown in Table VII (with additional results available in the Appendix), indicate that our approach outperforms the best previous methods that do not use data augmentation. Additionally, our method trained for only 240 epochs with the same data augmentation strategy as MLKD achieves superior performance compared to MLKD trained for 480 epochs on the CIFAR-100 dataset.

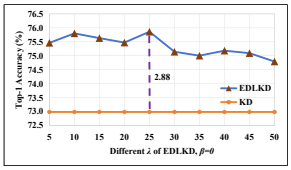

Fig. 3: Impact of the hyperparameter λ for VGG8 & VGG-13 on CIFAR-100 ($\beta = 0$).

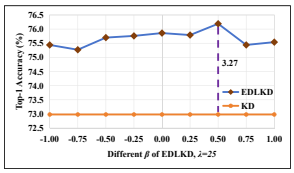

Fig. 4: Impact of the hyperparameter β for VGG8 & VGG-13 on CIFAR-100 ($\lambda = 25$).

We evaluate the impact of the hyperparameters λ and β on the performance of knowledge distillation. In our experiments, we first set β to 0 to avoid introducing the influence of incorrect predictions from the teacher model during the distillation process. We analyze the model's performance by varying λ in the range from 5 to 50 at equal intervals of 5 (Fig. 3).

To investigate the impact of incorrect knowledge on the distillation process, we vary β from -1.00 to 1.00 in steps of 0.25 $Fig.$ 4. When $\beta > 0$, the model consistently shows improved performance, reaching a peak accuracy of 76.25% at $\beta = 0.50$. This indicates that moderately suppressing the contribution of incorrect teacher predictions helps distillation: suppression increases the relative weight of the correct-prediction term L_c during optimization, allowing *correct knowledge to dominate* and thereby improving convergence stability and final accuracy.

By contrast, when $\beta < 0$ the model's performance deteriorates progressively relative to the $\beta = 0$ baseline, reaching a minimum at $\beta = -0.75$. A negative β effectively inverts the student's alignment with the teacher's distribution on the erroneous class, pushing the student away from the relative structure that the teacher encodes across the entire output space. When the teacher is "top-1 wrong but reasonably ranked" (i.e., the soft targets still convey useful ordering among classes), this form of "negative distillation" can destroy beneficial ranking information and the boundary-smoothing effect, thereby harming generalization. A preferable strategy is to attenuate the weight of the teacher's incorrect outputs via positive re-weighting rather than subtracting them outright—that is, reduce the influence of erroneous teacher predictions while preserving useful relational cues.

Importantly, even for some negative values of β, our method still outperforms the standard distillation setting without error decoupling, further validating the robustness of the proposed approach. Adjusting β therefore provides a practical mechanism to regulate the balance between correct and incor-

TABLE VI: Ablation Study. The experients are conducted on CIFAR-100, with VGG13 as the teacher model, VGG8 as the student model, and Top-1 accuracy as the evaluation metric.

KL Divergence	Cosine Distance	Error Decoupling	Data Augmentation	Acc
✓				72.98
✓		✓		73.68
✓			✓	74.69
✓		✓	✓	74.71
	✓			74.59
	✓	✓		74.95
	✓		✓	75.43
	✓	✓	✓	**76.25**

TABLE VII: Top-1 accuracy comparison on CIFAR-100: Effectiveness of the proposed method with and without data augmentation, compared to other knowledge distillation methods.

Type	Teacher	ResNet32×4	VGG13	WRN-40-2	WRN-40-2	ResNet56	ResNet110	ResNet110
		79.42	74.64	75.61	75.61	72.34	74.31	74.31
	Student	ResNet8×4	VGG8	WRN-40-1	WRN-16-2	ResNet20	ResNet32	ResNet20
		72.50	70.36	71.98	73.26	69.06	71.14	69.06
Logit	KD [3]	73.33	72.98	73.54	74.92	70.66	73.08	70.67
	DKD [7]	76.32	74.68	74.81	76.24	71.97	74.11	71.06
	CTKD [8]	73.39	73.52	73.93	75.45	71.19	73.52	70.99
	KD+LS [5]	76.62	74.36	74.37	76.11	71.43	74.17	71.48
	w/o DA	76.65	74.95	74.99	76.29	72.03	74.38	71.72
	w/ DA	**77.85**	**76.25**	**75.75**	**77.04**	**72.25**	**74.55**	**72.29**

rect knowledge: positive β attenuates harmful teacher errors and raises the effective signal-to-noise ratio of transferred knowledge, while negative β has the opposite effect. The reported hyperparameter choices were determined via preliminary experiments to strike a balance between preserving useful relational information and suppressing detrimental erroneous predictions.

IV. Conclusion and Future Work

In this paper, we theoretically analyze the negative impact of a teacher's erroneous predictions on the distillation process and propose Error-Decoupled Learning for Knowledge Distillation, which improves the distillation process by decoupling correct and incorrect predictions from the teacher model. Experimental results show that EDLKD outperforms existing mainstream logit-based methods in both classification and object detection tasks. In addition, similar to mainstream distillation approaches, EDLKD is designed for supervised learning settings. Extending this framework to self-supervised or semi-supervised tasks remains an open challenge. A promising direction is to explore whether "erroneous" representations identified through threshold-based binning or confidence-aware grouping could still convey transferable relational structures under self-supervised objectives. Such an extension could provide new insights into the role of error information in representation learning and broaden the applicability of error-decoupled distillation beyond traditional supervised paradigms.

References

[1] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[2] M. Farina, M. Mancini, E. Cunegatti, G. Liu, G. Iacca, and E. Ricci, "Multiflow: Shifting towards task-agnostic vision-language pruning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 185–16 195.

[3] G. Hinton, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[4] Y. Jin, J. Wang, and D. Lin, "Multi-level logit distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 276–24 285.

[5] S. Sun, W. Ren, J. Li, R. Wang, and X. Cao, "Logit standardization in knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 731–15 740.

[6] R. Zhang, J. Shen, T. Liu, J. Liu, M. Bendersky, M. Najork, and C. Zhang, "Do not blindly imitate the teacher: Using perturbed loss for knowledge distillation," *arXiv preprint arXiv:2305.05010*, 2023.

[7] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 11 953–11 962.

[8] Z. Li, X. Li, L. Yang, B. Zhao, R. Song, L. Luo, J. Li, and J. Yang, "Curriculum temperature for knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1504–1512.

[9] T. Wen, S. Lai, and X. Qian, "Preparing lessons: Improve knowledge distillation with better supervision," *Neurocomputing*, vol. 454, pp. 25–33, 2021.

[10] C.-P. Su, C.-H. Tseng, B. Pu, L. Zhao, J. Yang, Z. Chen, and S.-J. Lee, "Ea-kd: Entropy-based adaptive knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 731–740.

[11] N. Giakoumoglou and T. Stathaki, "Relational representation distillation," *arXiv preprint arXiv:2407.12073*, 2024.

[12] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967–3976.

[13] L. Xu, K. Liu, J. Liu, L. Wang, L. Xu, and J. Cheng, "Local dense logit relations for enhanced knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 4539–4549.

[14] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[15] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[16] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967–3976.

[17] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.

[18] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5008–5017.

[19] D. Chen, J.-P. Mei, H. Zhang, C. Wang, Y. Feng, and C. Chen, "Knowledge distillation with the reused teacher classifier," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 933–11 942.

[20] Z. Guo, H. Yan, H. Li, and X. Lin, "Class attention transfer based knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 868–11 877.

[21] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5007–5016.

[22] J. Lv, H. Yang, and P. Li, "Wasserstein distance rivals kullback-leibler divergence for knowledge distillation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 65 445–65 475, 2024.

[23] Z. Hao, J. Guo, K. Han, Y. Tang, H. Hu, Y. Wang, and C. Xu, "One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 79 570–79 582, 2023.

[24] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[27] T. Wang, L. Yuan, X. Zhang, and J. Feng, "Distilling object detectors with fine-grained feature imitation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4933–4942.

[28] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 5191–5198.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[30] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[31] W. Cui and S. Yan, "Isotonic data augmentation for knowledge distillation," *arXiv preprint arXiv:2107.01412*, 2021.

[32] H. Wang, S. Lohit, M. N. Jones, and Y. Fu, "What makes a" good" data augmentation in knowledge distillation-a statistical perspective," *Advances in Neural Information Processing Systems*, vol. 35, pp. 13 456–13 469, 2022.