ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

DISTILLING KNOWLEDGE OF NEURAL NETWORKS FOR IMAGE ANALYSIS, MODEL COMPRESSION, DATA PROTECTION AND MINIMIZATION

Ph.D. THESIS

Reyhan Kevser KESER

Department of Applied Informatics

Information and Communication Engineering Programme

JULY 2024

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

DISTILLING KNOWLEDGE OF NEURAL NETWORKS FOR IMAGE ANALYSIS, MODEL COMPRESSION, DATA PROTECTION AND MINIMIZATION

Ph.D. THESIS

Reyhan Kevser KESER
(708182005)

Department of Applied Informatics

Information and Communication Engineering Programme

Thesis Advisor: Prof. Dr. Behçet Uğur TÖREYİN

JULY 2024

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

GÖRÜNTÜ ANALİZİ, MODEL SIKIŞTIRMA, VERİ KORUMA VE MİNİMİZASYONU İÇİN YAPAY SİNİR AĞLARININ BİLGİSİNİN DAMITILMASI

DOKTORA TEZİ

Reyhan Kevser KESER
(708182005)

Bilişim Uygulamaları Anabilim Dalı

Bilgi ve Haberleşme Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Behçet Uğur TÖREYİN

TEMMUZ 2024

Reyhan Kevser KESER, a Ph.D. student of ITU Graduate School student ID 708182005 successfully defended the thesis entitled "DISTILLING KNOWLEDGE OF NEURAL NETWORKS FOR IMAGE ANALYSIS, MODEL COMPRESSION, DATA PROTECTION AND MINIMIZATION", which she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Behçet Uğur TÖREYİN** Istanbul Technical University	.....
Jury Members :	**Assoc. Prof. Dr. Nazım Kemal ÜRE** Istanbul Technical University	.....
	Asst. Prof. Dr. Süha TUNA Istanbul Technical University	.....
	Prof. Dr. Maria TROCAN Institut Supérieur d'Électronique de Paris (ISEP)	.....
	Asst. Prof. Dr. Nurullah ÇALIK Medeniyet University	.....

Date of Submission : **16 May 2024**
Date of Defense : **4 July 2024**

To my family,

FOREWORD

I would like to express my heartfelt gratitude to my advisor, Prof. Dr. Behçet Uğur TÖREYİN, for his profound knowledge and invaluable insights during my thesis studies. His unwavering support, understanding, and patience have been crucial on this long and challenging journey.

I am also deeply thankful to the members of my thesis committee, Assoc. Prof. Dr. Nazım Kemal ÜRE and Asst. Prof. Dr. Nurullah ÇALIK, for their time, effort, constructive feedback, and encouraging words. Additionally, I extend my thanks to Prof. Dr. Maria TROCAN and Asst. Prof. Dr. Süha TUNA for participating in the thesis defense. I greatly appreciate my esteemed colleagues and co-authors, Onur Can KOYUN and Dr. Muhammet Sebul BERATOĞLU, whose opinions have been invaluable throughout my doctoral studies.

I would like to thank the organizations that support my studies, which are TUBITAK (The Scientific and Technological Research Council of Turkey) under project no. 121E378, and with 1515 Frontier R&D Laboratories Support Program for BTS Advanced AI Hub: BTS Autonomous Networks and Data Innovation Lab. Project 5239903; TURKCELL with the projects named "Developing Algorithms for Detection of Lip-Speech Synchronization Error in Live Channel Broadcasts using Computer Vision and Machine Learning Techniques" and "Cigarette Detection in Offline Video"; ASELSAN with the projects named "Literature Review and Developing Detection Approaches for Raman Spectroscopy" and "Developing Deep Learning Based Methods for Object Detection in Video"; ITU Vodafone Future Lab with the project named "Neural Embeddings for Social Context in Telecommunications"; Arcelik ITU R&D Center; and ITU BAP with the project named "Data Acquisition, Compression, Analysis and Fusion Methods for New Generation Surveillance Systems", with project PMA-2023-44299, and under grant number MOA-2019-42321.

Last but not least, I want to express my endless gratitude to my family, who have been my biggest supporters throughout my life, always standing by me with their material and moral support.

July 2024

Reyhan Kevser KESER
(Electronics and Communication Engineer)

TABLE OF CONTENTS

ABBREVIATIONS

Acc	: Accuracy
AFD	: Attention Feature Distillation
AP50	: Average Precision at IoU Threshold of 0.5
ATT	: Attention Transfer
CC	: Correlation Congruence
CCA	: Canonical Correlation Analysis
CKA_{linear}	: Centered Kernel Alignment with Linear Kernel
CKA_{rbf}	: Centered Kernel Alignment with RBF Kernel
COCO	: Common Objects in Context Dataset
CPU	: Central Processing Unit
CRD	: Contrastive Representation Distillation
DA	: Domain Adaptor
DN	: DarkNet
DNN	: Deep Neural Network
FT	: Factor Transfer
GPU	: Graphics Processing Unit
HEVC	: High Efficiency Video Coding
HP	: Hint Position
HSIC	: Hilbert-Schmidt Independence Criteria
IP	: Intellectual Property
KD	: Knowledge Distillation
MN	: MobileNetv2
NAS	: Neural Architecture Search
NLP	: Natural Language Processing
ReLU	: Rectified Linear Unit
resnetdxm	: m times wider CIFAR style ResNet-d
resnet32	: CIFAR style ResNet-32
resnet110	: CIFAR style ResNet-110
RN-18	: ImageNet style ResNet-18
RN-50	: ImageNet style ResNet-50
R^2_{CCA}	: Mean Squared CCA
SGD	: Stochastic Gradient Descent
SKD	: Spherical Knowledge Distillation
VID	: Variational Information Distillation
WRN-d-w	: Wide ResNet with depth d and width factor w
WSL	: Weighted Soft Label

SYMBOLS

y	**:** Label
z	**:** Logits of a Network
$\hat{z}$	**:** Adjusted Logits of a Network
λ	**:** Trade-off Between Loss Terms
L_{cls}	**:** Classification Loss
L_{hint}	**:** Hint Loss
L_{task}	**:** Task Loss
L_{feat}	**:** Feature Loss
L_{total}	**:** Total Loss
L_{logit}	**:** Logit Loss
N	**:** Number of Samples
exp	**:** Exponential Function
T	**:** Temperature
p	**:** Predicted Class Probabilities
M	**:** Million
K	**:** Thousand
ms	**:** Milliseconds

LIST OF TABLES

Page

LIST OF FIGURES

Page

DISTILLING KNOWLEDGE OF NEURAL NETWORKS FOR IMAGE ANALYSIS, MODEL COMPRESSION, DATA PROTECTION AND MINIMIZATION

SUMMARY

Knowledge distillation is an effective tool for model training, which refers to the process of knowledge transfer between models. In the context of knowledge distillation, the model to be trained with the injected knowledge is named student, where the teacher refers to the model whose knowledge is acquired. It can be exploited for various aims including improving model performance, accelerating the model, and reducing model parameters. Further, with the advent of diverse distillation schemes, it can be efficiently applied in various scenarios and problems. Thus, it has a wide range of application fields including computer vision and natural language processing.

This thesis comprises the studies conducted on numerous problems of knowledge distillation, as well as the literature review. The first problem we focus on is hint position selection as an essential element in hint distillation, which is transferring features extracted in intermediate layers, namely hints. First, we demonstrate the importance of the determination of the hint positions. Then, we propose an efficient hint point selection methodology based on layer clustering. For this purpose, we exploit the k-means algorithm with specially designed metrics for layer comparison. We validate our approach by conducting comprehensive experiments utilizing various architectures for teacher-student pairs, hint types, and hint distillation methods, on two well-known image classification datasets. The results indicate that the proposed method achieves superior performance compared to the conventional approach.

Another problem focused on in this thesis is model stealing, which refers to acquiring knowledge of a model that is desired to be protected due to the privacy concerns or commercial purposes. Since knowledge distillation can be exploited for model stealing, the concept of the undistillable teacher has been introduced recently, which aims to protect the model from stealing its knowledge via distillation. To contribute to this field, we propose an approach called averager student, whose goal is distilling the undistillable teacher, in this thesis. We evaluate the proposed approach for given teachers which are undistillable or normal. The results suggest that the proposed method outperforms the compared methods whose aim is the same as ours.

The last problem we addressed is cross distillation, which means the distillation process between teacher and student models that operate on different modalities. In this work, we introduce a cross distillation scheme that transfers the compressed domain knowledge to the pixel domain. Further, we employ hint distillation which utilizes our previously proposed hint selection method. We evaluate our approach on two computer vision tasks, that are object detection and recognition. The results demonstrate that compressed domain knowledge can be efficiently exploited in a task in the pixel domain via the proposed approach.

The proposed approaches in the context of the thesis, contribute to studies on image analysis, model compression, data protection, and minimization. First, our study on the selection of efficient hint positions aims to improve model compression performance, although the proposed approach can also be employed for other distillation schemes. The gains of our method in terms of model compression are presented as well as the performance results of the proposed algorithm. Then, our work on model stealing targets to contribute to the literature on model intellectual property (IP) protection and data protection, where we introduce an algorithm to distill a protected model's knowledge. Moreover, our study on cross distillation provides a contribution to data protection and minimization studies, where we propose a distillation methodology that utilizes compressed domain knowledge on pixel domain problems. Our approach demonstrates a technique that expands limited knowledge by employing different modality data instead of more samples. Since we utilize compressed domain images and eliminate the need for more samples to boost performance, we prevent the use of more data that may be personal or sensitive.

GÖRÜNTÜ ANALİZİ, MODEL SIKIŞTIRMA, VERİ KORUMA VE MİNİMİZASYONU İÇİN YAPAY SİNİR AĞLARININ BİLGİSİNİN DAMITILMASI

ÖZET

Bilgi damıtımı, modeller arasında bilgi transferi süreci anlamına gelen ve model eğitimi için kullanılan etkili bir tekniktir. Bilgi damıtımı bağlamında, enjekte edilen bilgi ile eğitilmesi amaçlanan model öğrenci olarak adlandırılırken, bilgiyi aktaran model öğretmen olarak adlandırılır. Model performansını artırmak, modeli hızlandırmak ve model parametrelerini azaltmak gibi çeşitli amaçlar için kullanılabilir. Ayrıca, çeşitli damıtma şemaları ile birlikte, çeşitli senaryo ve problemlerde etkin bir şekilde uygulanabilir. Bu nedenle, bilgisayarlı görü, doğal dil işleme, konuşma tanıma, öneri sistemleri, çekişmeli saldırılar, model sıkıştırma ve hızlandırma, gizli bilgi damıtımı, derin öğrenme modellerinin anlamlandırılması, federe öğrenme, ve nöral mimari arama gibi geniş bir uygulama yelpazesine sahiptir.

Bilgi damıtımı, başarılı bir yaklaşım olmasından dolayı, birçok farklı alanda çalışan araştırmacının dikkatini çekmeyi başarmıştır. Tezde de incelendiği gibi zamanla, Hinton ve arkadaşlarının önerdiği şema geliştirilmiş ve farklı ve karmaşık bilgi damıtma yaklaşımları ortaya çıkmıştır. Bu yaklaşımlar bilgi çeşidi, damtım şeması, uygulama alanları ve algoritmalar gibi farklı açılara göre gruplanabilir. Örneğin, mevcut yöntemler bilgi çeşidine göre üç alt başlıkta toplanabilir: Cevap-tabanlı, öznitelik-tabanlı, ve ilişki-tabanlı bilgi çeşitleri. Cevap-tabanlı yöntemler modelin tahmin ettiği olasılık dağılımını esas alarak bilgi iletimi yaparken, öznitelik-tabanlı yöntemler modelin ara katmanlarda elde ettiği öznitelikleri esas almaktadır. İlişki-tabanlı yöntemler ise modelin öğrendiği, veri kümesindeki örnekler arasındaki ilişkiyi bir şekilde diğer modele iletmektedir. Öte yandan, mevcut bilgi damıtımı yöntemleri kullanılan şema açısından gruplandığında üç başlık ortaya çıkmaktadır: Çevrimdışı, çevrimiçi, ve kendi kendine bilgi damıtımı. Çevrimdışı bilgi damıtımı, klasik yaklaşım olan ve öğretmen modelin önceden eğitildiği durumu ifade eder. Çevrimiçi damıtım ise öğretmen ve öğrenci modelin birlikte eğitildiği senaryoyu tanımlamaktadır. Kendi kendine bilgi damıtımı ise öğretmen ve öğrencinin aynı model olduğu yöntemleri kapsamaktadır. Örneğin, aynı modelin farklı veya aynı katmanları arasındaki bir bilgi aktarımı bu başlık kapsamındadır.

Mevcut bilgi damıtımı yaklaşımlarında kullanılan algoritmalara göz atıldığında çok çeşitli yöntemler karşımıza çıkmaktadır. Örneğin, çekişmeli damıtım, çekişmeli öğrenmenin bilgi damıtımında kullanılmasını ifade eden yaklaşımlar dizisini tanımlar. Çapraz damıtım ise öğretmen ve öğrenci modelin farklı modaliteler üzerinde çalıştığı senaryoları ifade etmektedir. Çoklu öğretmen algoritmaları ise adından anlaşıldığı üzere bilgi damıtımı şemasında birden fazla öğretmen modelin bilgisinden faydalanıldığı durumları tanımlamaktadır. Bunların dışındaki mevcut bilgi damıtım algoritmaları arasında şunlar sayılabilir: Çizge-tabanlı damıtım, dikkat

mekanizması-tabanlı damıtım, verisiz damıtma, ömürboyu damıtım, nöral mimari arama-tabanlı damıtma, ve nicelenmiş damıtma.

Bu tez, çeşitli bilgi damıtımı problemlerine odaklanan çalışmaları ve ilgili literatür taramasını içermektedir. Tez kapsamında yapılan çalışmalar üç başlıkta toplanabilir: İpucu damıtımı için ipucu pozisyonu belirlenmesi, model çalma problemi için ortalamacı öğrenci yöntemi önerilmesi, ve sıkıştırılmış alandan piksel alanına bilgi aktarımı sağlayan bir çapraz damıtma yaklaşımı tasarlanması.

Tez kapsamında ilk olarak, ara katmanlarda çıkarılan özniteliklerin aktarıldığı ve ipucu damıtımı olarak adlandırılan bir tür bilgi damıtımı yaklaşımı için önemli bir unsur olan ipucu pozisyonu seçimine odaklanılmıştır. Bu çalışmada öncelikle, ipucu pozisyonlarının belirlenmesinin bilgi damıtımında önemli bir etken olduğu gösterilmiştir. Sonrasında ise, PURSUhInT olarak adlandırılan ve katman kümeleme tabanlı olan etkili bir ipucu noktası seçim metodolojisi önerilmektedir. Bu amaçla, katmanların kıyaslanabilmesi için yeterli sayıda örnek üzerinden katman temsil matrisleri oluşturulmakta ve katman karşılaştırması için özel olarak tasarlanmış metriklerden faydalanarak k-ortalamalar kümeleme algoritması kullanılmaktadır. Önerilen yöntem, çeşitli öğretmen ve öğrenci mimarileri, ipucu türleri ve ipucu damıtma yöntemleri için ve iki bilindik görüntü sınıflandırma veri kümesi üzerinde kapsamlı deneyler yapılarak değerlendirilmiştir. Elde edilen sonuçlar, önerilen yöntemin geleneksel yaklaşıma göre üstün başarı elde ettiğini göstermektedir. Bununla beraber, önerilen yöntem model sıkıştırma açısından değerlendirilmiş, yöntem sayesinde elde edilen sıkıştırma oranları ve çıkarım süresi kazançları paylaşılmıştır.

Bu tezde odaklanılan bir başka problem ise, ticari amaçlarlar veya gizlilik endişeleri nedeniyle korunmak istenen bir modelin bilgisini ele geçirme olarak tanımlanan model çalma problemidir. Bilgi damıtımı model çalmak için kullanılabileceğinden, yakın zamanlarda sıkıştırılamaz öğretmen kavramı tanıtılmıştır. Bu alana katkıda bulunmak için, bu tezde sıkıştırılamaz öğretmeni sıkıştırmayı amaçlayan ve ortalamacı öğrenci olarak adlandırılan bir yaklaşım önerilmektedir. Bu yaklaşım, öğretmen modelin tahmin ettiği bozulmuş olasılık dağılımlarının, aynı etiketli örnekler için ortalamasını alarak, olasılık dağılımındaki yanıltıcı zirve değerlerini gidermekte ve bu şekilde öğrenci modele iletmektedir. Önerilen yaklaşım, sıkıştırılamaz veya normal olduğu bilinmeyen öğretmen modelleri üzerinde değerlendirilmiştir. Sonuçlar, önerilen yöntemin kıyaslanan yöntemlerden daha iyi performans gösterdiğini belirtmektedir.

Ele alınan son problem, farklı modalitelerde çalışan öğretmen ve öğrenci modeller arasındaki bilgi aktarım süreci olan çapraz damıtmadır. Bu çalışmada, sıkıştırılmış alan bilgisini piksel alanına aktaran bir çapraz damıtma yöntemi önerilmektedir. Çalışma, daha önce önerdiğimiz ipucu noktası seçme yönteminden faydalanan bir ipucu damıtma yaklaşımı içermektedir. Önerilen yaklaşım, nesne tespiti ve tanıma olmak üzere iki bilgisayarlı görü problemi üzerinde değerlendirilmiştir. Elde edilen sonuçlar, sıkıştırılmış alan bilgisinin piksel alanındaki bir görevde önerilen yaklaşım sayesinde etkin bir şekilde kullanılabileceğini göstermektedir. Yapılan ek deneylerde ise, öğretmen ve öğrenci ikilisi için tüm modalite seçenekleri değerlendirilmiş ve farklı modaliteler arasındaki bilgi damıtımının standart bilgi damıtımından daha iyi performans gösterdiği gözlemlenmiştir. Bununla beraber, en iyi performansın,

önerilen yöntemdeki gibi sıkıştırılmış alandan piksel alanına yapılan bilgi aktarımı senaryosunda elde edildiği raporlanmıştır.

Tez kapsamında önerilen yaklaşımlar, görüntü analizi, model sıkıştırma, veri koruma ve minimizasyon çalışmalarına katkı sağlamaktadır. İlk olarak, etkili ipucu pozisyonlarının seçimine ilişkin çalışmamız, model sıkıştırma performansını iyileştirmeyi amaçlamaktadır, ancak önerilen yaklaşım bu alanda sınırlı olmayıp, diğer damıtma şemaları için de kullanılabilir. Yöntemimizin model sıkıştırma açısından kazanımları, önerilen algoritmanın performans sonuçlarının yanı sıra sunulmaktadır. Ardından, model fikri mülkiyet (IP) koruması ve veri koruması literatürüne katkıda bulunmak amacıyla model çalma konusu üzerine yapılan çalışmada, korunan bir modelin bilgisini damıtacak bir algoritma tanıtılmıştır. Bununla beraber, çapraz damıtma üzerine yapılan çalışma veri koruma ve minimizasyon çalışmalarına katkı sağlamaktadır. Bu çalışmada piksel alanı problemlerinde sıkıştırılmış alan bilgisinden yararlanan bir damıtma metodolojisi önerilmiştir. Yaklaşımımız, daha fazla örnek yerine farklı modalite verisi kullanarak sınırlı bilgiyi genişleten bir teknik ortaya koymaktadır. Sıkıştırılmış alan görüntüleri kullanılması ve performansı artırmak için kullanılabilecek daha fazla örnek ihtiyacının ortadan kaldırılması dolayısıyla, kişisel veya hassas olabilecek daha fazla verinin kullanılması önlenmektedir.

1. INTRODUCTION

This chapter comprises the motivation and contributions of the studies conducted in the context of the thesis, as well as the organization of the thesis.

1.1 Motivation and Contributions

Knowledge distillation (KD) is a training technique that exploits the knowledge learned by some model, to train another model [5,16,17]. It is widely used for enhancing model performance as well as model compression whose aim is obtaining smaller models that operate faster and have less computational cost. The model whose knowledge is obtained is called the teacher, whereas the student refers to the model that receives that knowledge. Since it is an efficient and needed tool for deep learning models, it has drawn the attention of researchers studying numerous fields including computer vision, natural language processing, and speech recognition [11].

In this thesis, we present the studies conducted on various problems of knowledge distillation, which are focused on hint position selection, model stealing, and cross distillation. Hint position is a crucial parameter for hint distillation, which is a research direction of knowledge distillation and aims to transfer knowledge acquired from intermediate layers of the teacher. The features to be transferred, which are obtained from intermediate layers of the teacher are called hints in the distillation terminology. In our study, we demonstrate that hint positions should be selected carefully since the performance of the distillation depends on hint positions as well as the other essential parameters. Further, we propose a hint position selection approach which is based on clustering of the teacher model's layers [18]. For this purpose, we utilize specialized metrics for clustering teacher's layers, by collecting layer representations for a sufficient amount of samples. Our contributions can be listed as follows:

- It is demonstrated that the selection of hint positions affects the student model's performance and utilizing heuristics or rule-of-thumbs for hint positions may degrade the generalization ability of the student.
- A novel method for searching informative hint locations is proposed, which mainly applies clustering on the teacher model's layers and employs cluster centers as the hint locations. This is the first approach that introduces a hint point selection algorithm based on layer clustering for hint distillation methodologies.
- The proposed scheme can be integrated into any hint distillation method that transfers multiple hints. This ability is demonstrated through comprehensive experiments that comprise numerous hint distillation approaches and various architectures for teacher and student models.
- The introduced approach is suitable for any student network and does not require any update during the training, after operating once on a selected teacher network.
- The proposed method is compared with the current state-of-the-art distillation algorithms through experiments on CIFAR-100 and ImageNet datasets. Our approach achieves superior performances on both of these datasets, for compression of numerous teacher networks regardless of having identical or distinct architectures with student models, as the results suggested.

Model stealing is an open research direction, where the studies focus on extracting the knowledge of a model that is desired to be protected since it is a commercial product or trained with private data. On the other hand, model intellectual property (IP) protection aims to protect models from malicious users. With the advent of data-free knowledge distillation methods, KD can be exploited as a tool for model stealing. In this context, a recent study proposed the method called *undistillable teacher* [19], whose aim is to protect its knowledge from malicious student models. On the contrary, to contribute to this field, we introduce an approach called *averager student*, which targets a successful distillation from undistillable teachers, in our study [20]. We can list the contributions of this study as follows:

- We introduce an approach for distilling the undistillable teacher models, which is a simple and effective approach.
- Our method computes the average values of the probability distributions that are predicted by the teacher network, for the samples with the same label.
- We evaluate the proposed approach in the cases of that given teacher model is undistillable or normal.
- Results suggest that our approach outperforms the prior study, whose aim is to transfer the knowledge of the undistillable teacher, too.

Cross distillation refers to the knowledge distillation scheme where teacher and student models operate on different modalities. In our work, we propose a cross distillation scheme to enhance a model operating on the pixel domain, where the teacher is trained on compressed domain images [21]. To train the student model, we utilize hint distillation, which transfers the knowledge of intermediate layers of the teacher model, namely *hints*. With this goal, we use the hint point selection approach we previously introduced. The contributions of this study can be listed as follows:

- We introduce the first cross distillation method that transfers knowledge from the compressed domain to the raw domain, to the best of our knowledge.
- The proposed method enhances the raw domain performance, while compressed domain data are used only in the training step.
- We validate our method through two well-known common computer vision tasks which are object detection and recognition.

The proposed approaches in this thesis contribute to the fields of image analysis, model compression, data protection, and minimization. Firstly, our study on selecting efficient hint positions aims to enhance model compression performance, though this method can also be applied to other distillation techniques. We present the benefits of our method for model compression and the performance results of the proposed algorithm. Secondly, our work on model stealing aims to contribute to the

literature on model intellectual property (IP) and data protection by introducing an algorithm that distills knowledge from a protected model. Additionally, our study on cross-distillation contributes to data protection and minimization by proposing a distillation methodology that uses compressed domain knowledge for pixel domain problems. This work demonstrates a technique for expanding limited knowledge by using data from different modalities rather than increasing the number of samples. By using compressed domain images and reducing the need for additional data, our work prevents the excessive use of potentially personal or sensitive information.

To sum up, knowledge distillation is a vital technique for training models. In this thesis, light is shed on various problems that knowledge distillation can offer solutions to. It is demonstrated that KD is effective for many different problems, datasets, architectures, and purposes. Conducted studies suggest that KD is a multi-functional tool that can be designed as varied frameworks and serves many application areas and tasks.

1.2 Organization of Thesis

The rest of this thesis is organized as follows: In Chapter 2, knowledge distillation is explained in detail. The categorization of knowledge distillation methods, from the viewpoints of knowledge type, distillation scheme, applications, and algorithms, is presented. Further, existing studies in the literature are reviewed.

Chapter 3 presents the studies conducted on the hint position selection problem. After reviewing the related literature, the importance of the selection of hint positions is demonstrated. Further, an efficient approach to determine the hint positions is proposed, which exploits the clustering of the teacher layers by the k-means algorithm via specialized metrics. To validate the proposed approach, comprehensive experiments are conducted, which employ pre-activation and post-activation hints. Moreover, the proposed approach is integrated into state-of-the-art hint distillation methods to compare the performances between the proposed and the conventional approaches. Besides, experiments are conducted for numerous teacher-student pairs, which have similar or distinct architectures. Furthermore, experiments are conducted on two well-known image classification datasets which are CIFAR-100 and ImageNet. Finally, the proposed approach is evaluated from the viewpoint of model compression.

Chapter 4 mentions the model stealing problem and its relationship with knowledge distillation. Further, a recently proposed concept called *undistillable teacher* is described, where the aim is designing undistillable teachers that protect their knowledge from the student models. With the aim of contributing model IP protection literature, a novel approach is proposed, which is called *averager student* and targets to distill the undistillable teachers. The experiments are conducted on CIFAR-100 dataset for the scenarios of that given teacher is undistillable or normal. Results suggest that the proposed method is effective for distilling both types of teacher networks.

Chapter 5 describes the cross distillation framework, where teacher and student networks perform on different modalities. It can be exploited in various scenarios including having insufficient data in the target domain. Thus, a novel methodology is introduced, where the teacher and student operate on the compressed and the pixel domains, respectively. To evaluate the proposed approach, two computer vision tasks are utilized, which are object detection and recognition. Further, experiments are conducted from the viewpoints of domains and results are compared.

Chapter 6 presents the conclusion and possible future work in the context of this thesis.

2. KNOWLEDGE DISTILLATION

2.1 Introduction

Knowledge distillation (KD) refers to the knowledge transfer between models, where the knowledge is mostly transferred from a large model (teacher) to a smaller/lighter model (student). Although KD was first introduced by [16], it is recognized with Hinton and his colleagues' study [5]. The most reputed scheme of KD employs a cost function that comprises the soft targets obtained from the teacher network and hard targets, i.e., labels. The methodologies employing soft targets without hints are referred to as output (logit) distillation [22].

To compute soft targets as in (2.1), training batches are passed through the teacher network, and the softmax layer's output is utilized with a higher temperature value ($T > 1$):

$$p_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \tag{2.1}$$

where p, T, and z represent the softened class probability, temperature value, and logits, respectively. For example, $i, j \in \{0, ..., 99\}$ for a classification task with 100 classes, such as CIFAR-100. To produce both softened logits of teacher and student networks, the same temperature value is utilized, where higher T yields a softer class probability distribution. For training of the student model, a loss function (2.2) is employed, as in the following:

$$L = \lambda L_{logit}(p^S, p^T) + (1-\lambda) L_{cls}(p^S, y) \tag{2.2}$$

where λ represents the balance between classification loss (L_{cls}) and logit distillation loss (L_{logit}), y stands for the label, p^S and p^T are the softened logits of student and teacher networks, respectively. Temperature value is chosen as 1 in L_{cls}. In

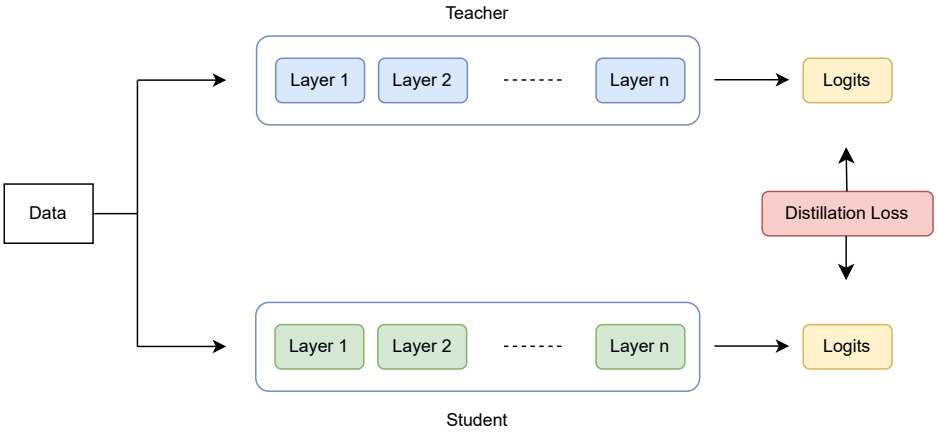


Figure 2.1 : The representative scheme of response-based knowledge distillation [11].

general, Kullback-Leibler and Cross-entropy losses are employed for L_{logit} and L_{cls}, respectively.

KD methods can be categorized in terms of numerous viewpoints including knowledge types, distillation schemes, applications, and algorithms [11].

2.2 Knowledge Types in KD

Regarding the knowledge type, KD studies can be divided into three classes, which are response-based, feature-based, and relation-based knowledge.

2.2.1 Response-based KD

Response-based knowledge distillation refers to transferring output knowledge, which is generally obtained from logits, where logit is the output of the softmax layer in a classifier model. For different tasks, the content of the knowledge to be transferred may be altered. For instance, logits are transferred as well as the bounding box parameters in an object detection task [23]. Further, heatmaps for the landmarks can be transferred as the teacher network's output in a semantic landmark localization task [24]. Nevertheless, the most exploited response-based knowledge is knowledge obtained from logits, where classical knowledge distillation proposed by Hinton et al. [5], utilizes logits. Thus, in this thesis, *logit distillation* term is utilized for the response-based KD, which is visualized in Figure 2.1.

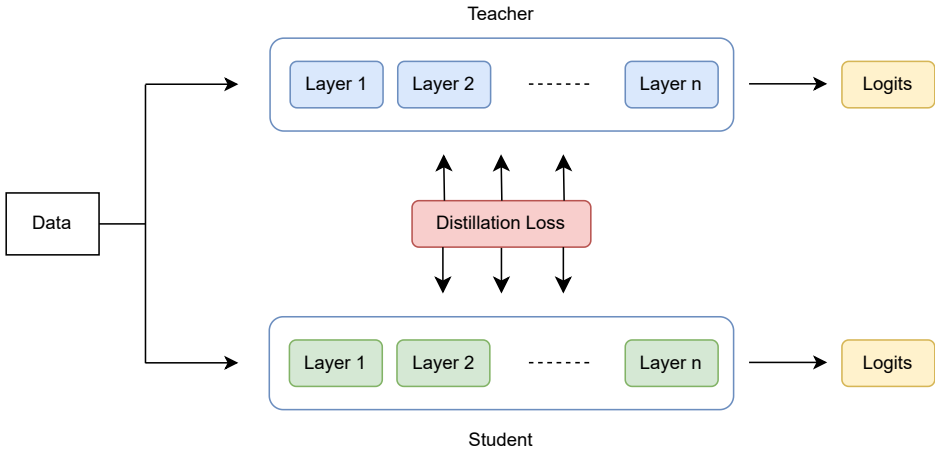


Figure 2.2 : The representative scheme of feature-based knowledge distillation [11].

After Hinton et al.'s seminal study, numerous approaches were introduced to improve logit distillation. For example, [25] proposes an approach to adjust logits that may be incorrect or uncertain. [26] proposes to employ a third network called *teacher assistant*, which receives knowledge from the teacher model and transfers the knowledge to the student model. [27] introduces a method called *Inverse Probability Weighting Distillation* which applies a weighting scheme to highlight under-represented samples in the machine domain, that refers to the domain of teacher model outputs.

Although response-based knowledge distillation is an open research direction, these methods don't exploit the knowledge acquired from intermediate layers of the teacher network, where that knowledge is a crucial source of learned representations. Moreover, logit distillation is only applicable to supervised learning, which limits the usage of these approaches [11].

2.2.2 Feature-based KD

Since neural networks extract features from the input layer by layer, these features/learned representations can be employed as a supervision source in a knowledge distillation scheme. Thus, feature-based knowledge distillation refers to the process of transferring intermediate representations to the student model, as shown in Figure 2.2. It can be also called hint distillation since the features to be transferred are named hints [28].

First, it was proposed by Romero et al. [28], where the feature maps in the middle layers of teacher and student models, are forced to match. After that study, numerous approaches are introduced, where most of them alter the feature representation style by applying some transformation. For instance, [1] proposes to transfer the attention maps of the teacher network to the student network. [10] introduces an approach that paraphrases the knowledge obtained from the teacher to ease the mimicking process of the student model. [29] proposes to transfer the teacher's knowledge acquired from the layer before the activation function. [22] analyzes previous studies from the viewpoint of teacher transformation, student transformation, feature position, and distance function. Then, it designs a hint distillation scheme so that these components operate in harmony.

Moreover, [30] combines self-distillation and online distillation schemes to introduce a novel hint distillation method. In another study [31], a hint distillation scheme is proposed, where the issue of teacher and student having different-sized feature maps is mitigated in a setup of low-resolution object recognition problem. Further, a self-distillation approach is suggested by [32], where the knowledge of the deepest layer is transferred to the shallower layers in a multi-exit neural network.

Existing hint distillation studies mostly focus on improving the representation of features. Nevertheless, the selection of hint layers is commonly overlooked [11], which is an important parameter for the distillation scheme, that affects the performance of the distillation [18].

2.2.3 Relation-based KD

On the contrary to exploiting the outputs of intermediate or last layers as in feature- and response-based KD, the knowledge of relationships between layers or the samples is the source of the supervision in the relation-based KD scheme, where an exemplary scheme is demonstrated in Figure 2.3. For instance, [4] proposes to transfer the *flow of solving procedure (FSP)* matrix of the teacher, which is obtained by the inner product between features of different layers. Further, [33] introduces a method that transfers the instance relation-based knowledge. Moreover, [34] employs singular value decomposition (SVD) to extract the essential information in feature maps.

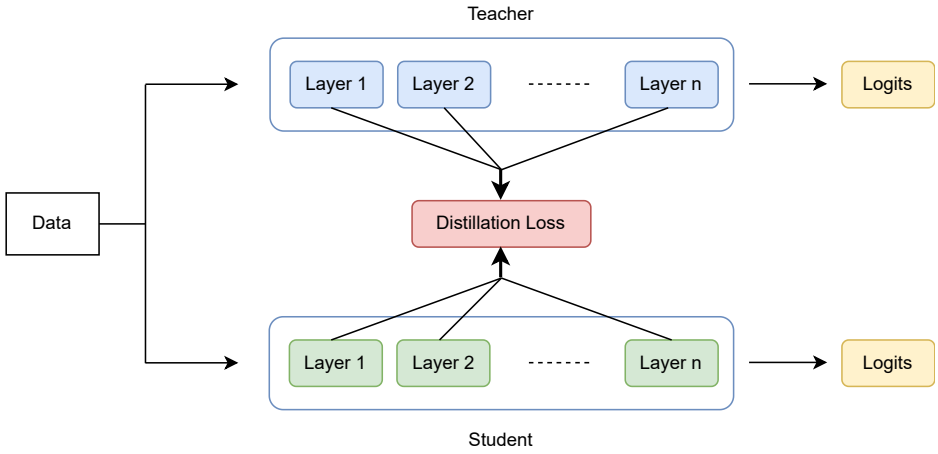


Figure 2.3 : An exemplary scheme of relation-based knowledge distillation that utilizes the knowledge of relationships between the layers [11].

[6] introduces a similarity-preserving distillation scheme, where pairwise similarity matrices obtained from the activation maps are transferred between the models. Thus, the student tries to learn the similarity between the data points, instead of mimicking the representation space of the teacher. Besides, [35] proposes to transfer the *instance relationship graph*, which comprises instance features, instance relationships, and the feature space transformation across layers.

2.3 KD Schemes

From the viewpoint of learning schemes, knowledge distillation approaches can be categorized into three groups: Offline distillation, online distillation, and self-distillation. Figure 2.4 demonstrates the knowledge distillation schemes.

2.3.1 Offline distillation

Classical knowledge distillation [5] comprises two steps. First, the teacher model is trained on a training set. Second, the extracted features from the trained teacher network are transferred to the student network. The first step is not deliberated mostly, since the teacher network is considered to be pre-trained [11]. Recently, this scheme adopting a pre-trained teacher has been named offline distillation, with the introduction of other schemes, including online distillation. Previously mentioned approaches including [5], [28], [1], [26], and [36], adopt an offline distillation scheme.

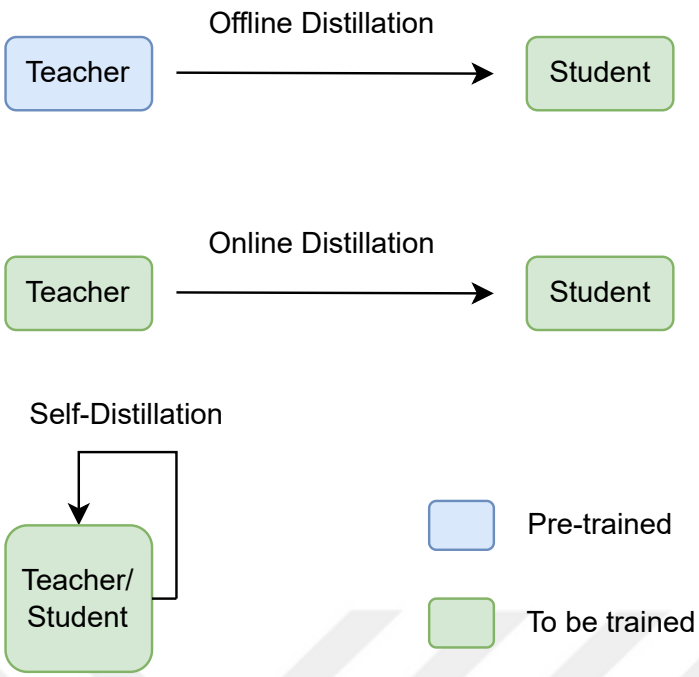


Figure 2.4 : Knowledge distillation schemes [11].

2.3.2 Online distillation

Online distillation is a technique introduced to address the constraints of offline distillation, where it suggests training teacher and student networks simultaneously. This scheme is particularly beneficial when a teacher with high performance is unavailable. Numerous studies have adopted this scheme including [37], [38], and [39]. For instance, [37] proposes a method called adversarial feature map distillation, where it employs two discriminators to make each model mimic the other one at the feature map level. Moreover, [38] introduces a scheme that exploits a group of student networks where each network transfers its knowledge to other networks for training. That study points out that there is no need for a pre-trained high-performing teacher. Further, [39] proposes an online distillation method that employs *m* networks, where *m-1* peer networks are trained using each other's knowledge, and then, a group leader network is trained with the knowledge of the peer networks.

2.3.3 Self-distillation

Self-distillation refers to employing the same networks for both teacher and student networks. This scheme can be considered a special case of online distillation [11]. For

example, [40] proposes a scheme where deeper layers transfer their knowledge to the shallower layers. That study claims that self-distillation helps the networks solve the vanishing gradient problem. Similarly, [41] introduces a scheme that transfers attention maps of layers from deeper layers to shallower layers, for the lane detection problem. Further, [42] proposes a method that utilizes the student model as the teacher model after some epochs of training. Again, after some epochs of training, the student model becomes the new teacher model. This process repeats until the training is done.

2.4 KD Algorithms

Various algorithms have been introduced to enhance the distillation process over time, by designing more complicated schemes. For instance, adversarial distillation refers to the knowledge distillation process in an adversarial setting [43]–[45]. For example, [43] introduces a method called KDGAN which employs a discriminator, a teacher, and a student. In addition to the distillation loss, an adversarial loss is utilized, where the student learns to mimic the teacher by trying to fool the discriminator. Further, some studies exploit a multi-teacher setting [46]–[48]. For instance, [48] introduces a multi-teacher distillation scheme that transfers all of the teachers' knowledge to the student network at the feature-level, by employing non-linear transformations. Moreover, another scheme of KD is cross-modal distillation, where the knowledge is transferred between different modalities [49]–[51]. For example, [49] introduces a cross-modal distillation scheme, where the teacher network operates on RGB images. On the other hand, the student network operates on depth or optical flow images, which are unlabeled and paired with the teacher's samples. With this scheme, the teacher's intermediate representations obtained by using labeled data are transferred to the student model that has no labeled data on a different modality.

Graph-based distillation can be another example of KD algorithms [33,52,53]. The primary concepts behind graph-based distillation approaches include exploiting the graph as a courier for the teacher's knowledge and controlling the message passing of the teacher's knowledge through the graph [11]. For instance, [52] designs a graph distillation method, where two graphs are obtained from logits and intermediate features of the teachers for knowledge transfer in a multi-teacher scheme. Besides,

some approaches utilize the attention mechanism, where they can be categorized as attention-based distillation methods [1,54,55]. For example, [55] proposes to match the distributions of neuron selectivity patterns between the models for knowledge transfer. That method operates by minimizing the maximum mean discrepancy (MMD) metric by these distributions, where *Attention Transfer* [1] can be classified as a special case of that approach. In addition, some methods aim to eliminate the difficulties arising from the unavailable data, which are named as data-free KD methods [44,56,57]. For instance, [57] exploits the layer activation statistics to reconstruct the original dataset to train the student model.

Other KD algorithms include quantized distillation, lifelong distillation, and NAS-based distillation [11]. Quantized distillation refers to the KD methods that employ quantization [58,59]. For instance, [58] introduces a KD scheme comprising a full-precision teacher and a low-precision student model, where the feature maps of the full-precision teacher are quantized before transferring them to the quantized student. Moreover, lifelong distillation refers to the KD methods that utilize lifelong learning, whose goal is to perform the learning similarly to humans [60,61]. For example, [60] employs a meta-learning scheme to specify the content and the location of the knowledge to be transferred. Furthermore, some approaches exploit neural architecture search (NAS) in a KD framework, to determine the suitable student architecture [62], whereas some studies utilize KD in a NAS framework, to enhance the performance of the NAS method [63].

2.5 KD Applications

Since knowledge distillation is an efficient tool for training models with various aims including model compression, enhancing the model's performance, and reducing the inference time, it has a wide range of usage areas for numerous tasks. The primary fields of artificial intelligence that exploit KD, can be categorized as follows [11]:

Visual Recognition: KD is often utilized in various visual recognition applications including image classification [18], image or video segmentation [64], object detection [23], video classification [52], visual question answering [65], and anomaly detection [66].

Natural Language Processing (NLP): Since conventional language models are very cumbersome, KD is a popular approach that enables training lightweight models, in the field of NLP. Some of the applications in that field can be listed as text generation [67], document retrieval [68], text recognition [69], and neural machine translation [70].

Speech Recognition: Numerous studies utilize KD to improve the recognition accuracy or reduce the response time of the deep acoustic models for speech recognition, which have a range of applications including audio classification [71], speech enhancement [72], spoken language identification [73], acoustic event detection [74], and speech synthesis [75].

Other Applications: KD can be integrated into the training of the recommendation models, to enhance the effectiveness of these models [76]. Further, KD is a powerful tool for solving adversarial attacks or perturbations of deep models, as well as model compression and acceleration [77]. Moreover, private knowledge distillation is an open research direction, whose goal is addressing privacy concerns of models [78]. Other applications of KD include interpretability of DNNs [79], federated learning [80], and neural architecture search [63].

3. HINT POSITION SELECTION

Hint distillation is one of the most powerful approaches for model compression, which refers to training a student model by injecting knowledge (hints) from intermediate layers of the teacher model [81,82]. Even though choosing different hint positions/points can yield quite different compression performances, this fact is ignored in conventional knowledge distillation studies, where the same hint positions in the early studies are utilized. Hence, in this work, we introduce a hint selection approach, which is based on clustering. Specifically, the teacher model's layers are clustered utilizing specially designed metrics, and the center positions of the clusters are employed as the hint positions. Since the proposed approach mainly operates on the teacher model, it is valid for any student network, once it is performed on the selected teacher model. The proposed method's performance is assessed through comprehensive experiments on CIFAR-100 and ImageNet datasets, employing numerous pairs of teacher and student models, and various hint distillation schemes. The hint points determined by the proposed approach yield superior compression performance compared to the state-of-the-art knowledge distillation methods on the same setup of student models and datasets, as the results suggested.

3.1 Introduction

In recent years, there has been a serious surge in employing Deep Neural Networks (DNNs) to succeed in numerous tasks of computer vision, such as classification [83,84], object detection [85], segmentation [86], face recognition [87] and tracking [88]. Further, implementing these models on edge devices is a requirement for various applications [89]. Nevertheless, deploying large-scale deep learning models is challenging due to their huge memory consumption and high computational load. In this work, we aim to push the state-of-the-art in compressing large-scale neural networks to enhance their inference speed while keeping the original performance

as much as possible. With this aim, we introduce a novel knowledge distillation methodology that exploits ideas from clustering and hint distillation, which is proven to be the superior algorithm for the compression of numerous models performing various tasks.

The development of deep learning-driven computer vision models gained momentum with the introduction of the ILSVRC challenge [90]. Over the years, at the expense of an important increase in model size and computational load, model performance is enhanced with the launch of VGGNet [91] and ResNet [84], after AlexNet model [83]. Alongside these advancements, researchers also put extensive efforts into model compression [92], where numerous algorithms were proposed to change the large models into smaller ones that perform comparably with the original models. Knowledge distillation [5] is a proper methodology among these algorithms since the compressed model can be determined to conform to the requirements and limitations of the target hardware, which is essential to deploy the models on edge devices. Among knowledge distillation methods, hint distillation algorithms [1,28,30,31] particularly have been proven successful in model compression due to numerous design choices of the scheme, where the original and large model is called as teacher, and the compressed and small model is called student networks. The effectiveness of these methods arises from incorporating intermediate feature maps/hints into the distillation process, where the student model is forced to learn, not only the logit response but also the response of intermediate layers of the teacher model, for a given input. The optimization of the hint distillation scheme and its performance is still an open research direction.

The main contribution of this work is a novel hint distillation methodology that focuses on the optimization of the hint locations utilizing a clustering operation of the teacher model's layers via specially designed metrics. We can summarize the contributions of this work as follows:

- It is demonstrated that the selection of hint positions affects the student model's performance and utilizing heuristics or rule-of-thumbs for hint positions may degrade the generalization ability of the student.

- A novel method for searching informative hint locations is proposed, which mainly applies clustering on the teacher model's layers and employs cluster centers as the hint locations. This is the first approach that introduces a hint point selection algorithm based on layer clustering for hint distillation methodologies.
- The proposed scheme can be integrated into any hint distillation method that transfers multiple hints. This ability is demonstrated through comprehensive experiments that comprise numerous hint distillation approaches and various architectures for teacher and student models.
- The introduced approach is suitable for any student network and does not require any update during the training, after operating once on a selected teacher network.
- The proposed method is compared with the current state-of-the-art distillation algorithms through experiments on CIFAR-100 and ImageNet datasets. Our approach achieves superior performances on both of these datasets, for compression of numerous teacher networks regardless of having identical or distinct architectures with student models, as the results suggested.

3.2 Related Work

Knowledge distillation is a methodology that aims to transfer knowledge between models. Hinton et al. proposed to transfer the knowledge of logits, which are the outputs of the softmax layers of teacher and student models [5]. After the recent studies employing hints in the transfer scheme, Hinton et al.'s method is categorized as a logit/output distillation method.

Among the numerous approaches introduced after the seminal paper of Hinton et al., Spherical Knowledge Distillation (SKD) [93] was recently proposed to improve KD, where it applies a projection to all logits of the teacher and the student on a sphere by normalization. To be specific, the SKD approach scales the logits into a unit vector and multiplies them with the average teacher norm, to recover its norm to the original level. Further, [94] proposes a novel distillation methodology, where it reformulates the conventional logit distillation. Moreover, in [95], a logit distillation scheme is

introduced to train a transformer student by employing lightweight teacher models that have distinct architectures.

Logit distillation is an open research direction, where various schemes are employed including data-free distillation [96], online distillation [8], ensemble distillation [97] or federated learning [98]. Besides, hint distillation targets to transfer much more knowledge to the student model via employing intermediate layers of the teacher in the transfer process [1,22,28,31], which is also termed as feature matching and constitutes the base of our introduced approach. It offers more advantages compared to the logit distillation, by diminishing the capacity gap between teacher and student models and being applicable to a wide range of tasks including regression and low-level vision problems [99]. Thus, this section analyzes existing studies on model compression and hint distillation.

3.2.1 Model compression

Considerable progress recently has been made in computer vision thanks to employing deep neural networks. Nevertheless, many state-of-the-art models can not be utilized for real-world applications, since employing these models is computationally costly and requires high memory spaces. Thus, numerous model compression methodologies have been introduced to address this problem, to compress especially deep neural networks. These methods, namely model compression approaches, can be categorized into four groups, which are parameter pruning and quantization, low-rank factorization, compact convolutional filters, and knowledge distillation [100].

Parameter pruning techniques [101] explore discarding unnecessary weights based on thresholding. Conversely, quantization-based methods aim to represent parameters with fewer bits. Although parameter pruning methods lower the model size, they typically do not lead to a reduction in inference time, which is a concern for real-world applications on edge devices. On the other hand, low-rank factorization methods are difficult to implement and computationally costly due to the decomposition process. Further, compact convolutional filters are introduced in [102] and modified by SqueezeNet [103] and MobileNet [104]. Employing compact convolutional filters usually yields a decrease in computational load. Nevertheless, the performance of

the models may be weakened in relatively complicated tasks due to a substantially reduced number of parameters [105]. So, in this study, we concentrate on knowledge distillation.

3.2.2 Hint distillation

In KD [5] and SKD [93], the soft targets consist of only logits of the teacher model, for training of the student model. After the foundation of KD [5], approaches that exploit intermediate feature maps of the teacher, have been introduced to enhance the logit (output) distillation's performance [28,30,106]–[109].

In [28], the knowledge obtained from the middle layer of the teacher is transferred to the middle layer of the student model. This methodology is named as FitNet and comprises a two-stage training process. In the first step, the student model is trained up to the guided layer which is the middle layer. After that, conventional logit distillation is employed to train the whole student network, in the second step. It should be highlighted that the classical KD [5] transfers the softened class probabilities of the teacher model, which are employed in the cost function along with the cross entropy loss between predictions and the labels. On the contrary, FitNet [28] transfers the knowledge obtained from the middle layer of the teacher along with the standard task loss. Further, [1] introduces activation-based and gradient-based Attention Transfer (AT) methodology, that transfers knowledge from a heavy teacher model to a compact student model. To employ AT in the knowledge distillation scheme, the distance between l_2-normalized attention maps of teacher-student layer pairs is minimized by the cost function, as well as the standard cross entropy loss. Another study [106] proposes a distillation scheme called Variational Information Distillation (VID), that aims to maximize the mutual information of the teacher-student layer pairs to train the student network. In [22], transferring knowledge obtained from the layer before the activation function is proposed, as in [29]. Furthermore, this approach exploits a loss function and a teacher transform, which are consistent with the feature position. It revises the feature position and, however, retains the conventional grouping scheme.

In another study [110] focused on hint distillation, namely IAKD, models are partitioned into smaller groups compared to the conventional grouping scheme. That

study states that transferring hints from all of the breakpoints yields a degraded performance compared to the conventional grouping. Further, [111] notes that results from experiments exploiting numerous hint points exhibit poor performance.

Lately, few studies have addressed the hint position problem, which is called feature linking by some studies [107]. For instance, [60] introduces a meta-learning methodology for transfer learning to choose hint positions along with the weights for hints. Nevertheless, this method deals with only the last layers of blocks of a model as the potential hint points and implementation of it is computationally costly. With the same purpose, [107] introduces an attention-based methodology, named AFD. However, this method repeats the hint point search process for each iteration. This is a limitation of the algorithm since it results in challenges in implementing the algorithm for each potential hint point for large models such as resnet110. On the contrary, our methodology offers a solution that is ready to apply to any student model and valid for the whole training stage, which is obtained using a selected teacher model.

Moreover, transferring knowledge between different stages of teacher and student networks is firstly suggested in [112]. Nevertheless, that method is based on the conventional grouping of layers, too. [113] proposes two methodologies for hint point selection on some problems of natural language processing, which are selecting every k^{th} layers and selecting the last k layers as the hint points. This method exploits a stationary rule for hint points which does not consider the information that the network layers own. [114] establishes a meta-learning scheme to determine hint points and hint transformations which are utilized for the alignment of hints between teacher and student. However, their search space is narrow and consists of only the last layers of conventional groups for hint points. On the contrary, our method redesigns the grouping of layers regarding the knowledge that layers maintain, where it allows all layers to be selected as hint points. Further, [115] proposes a non-conventional, random-selection-based hint point selection scheme for distillation. [116] aims to address the elimination of redundant knowledge in the transfer scheme, as in our proposed method. For this purpose, a selection is made among the samples to transfer the selected samples' knowledge. Nevertheless, our proposed method performs on hint points of the teacher model. In [117], an attention-based method is employed, where

natural language models are primarily concentrated. The knowledge obtained from different layers of the teacher model is fused to transfer, in that study.

To summarize, conventional hint distillation methods acquire intermediate feature maps by grouping layers of the teacher model, and then transfer the knowledge to the student model. This is achieved by dividing layers considering the spatial sizes of the feature maps. Opposed to employing a pre-determined scheme for hint point selection utilized in the state-of-the-art knowledge distillation approaches, in this study, we introduce a layer clustering approach as a systematized way to select more informative hint positions.

3.3 Method

In this study, the effect of the selection of hint positions on the performance of distillation is demonstrated. And, a layer clustering-based selection approach for informative hint points, namely PURSUhInT, is introduced. To be specific, pre-trained teacher networks are employed and a clustering process is applied to their sub-blocks. To this end, layer representations of each sub-block are collected. Then, k-means [118] clustering algorithm is applied on these representation matrices to acquire clusters of sub-blocks. This process results in clusters of similar sub-blocks in terms of the predetermined layer similarity metric. As a result, one hint point per cluster is selected for the distillation of the teacher model's knowledge in a comprehensively precise manner. Figure 3.1 visualizes the stages of the proposed approach, which will be elaborated in this section.

3.3.1 Layer representation

Sub-blocks should be expressed in a suitable style, to cluster the sub-blocks of models. In this work, the usually utilized setup for representing layers [119]–[121] is followed, which will be explained in detail, in this section.

Let H_i be the feature map of i^{th} sub-block in a teacher model that comprises blocks such as ResNet. The size of H_i is $N \times C \times H \times W$, where N represents the number of samples, C shows the number of channels, H and W stand for the height and width of the feature map, respectively. The channel dimension holds more information because spatial

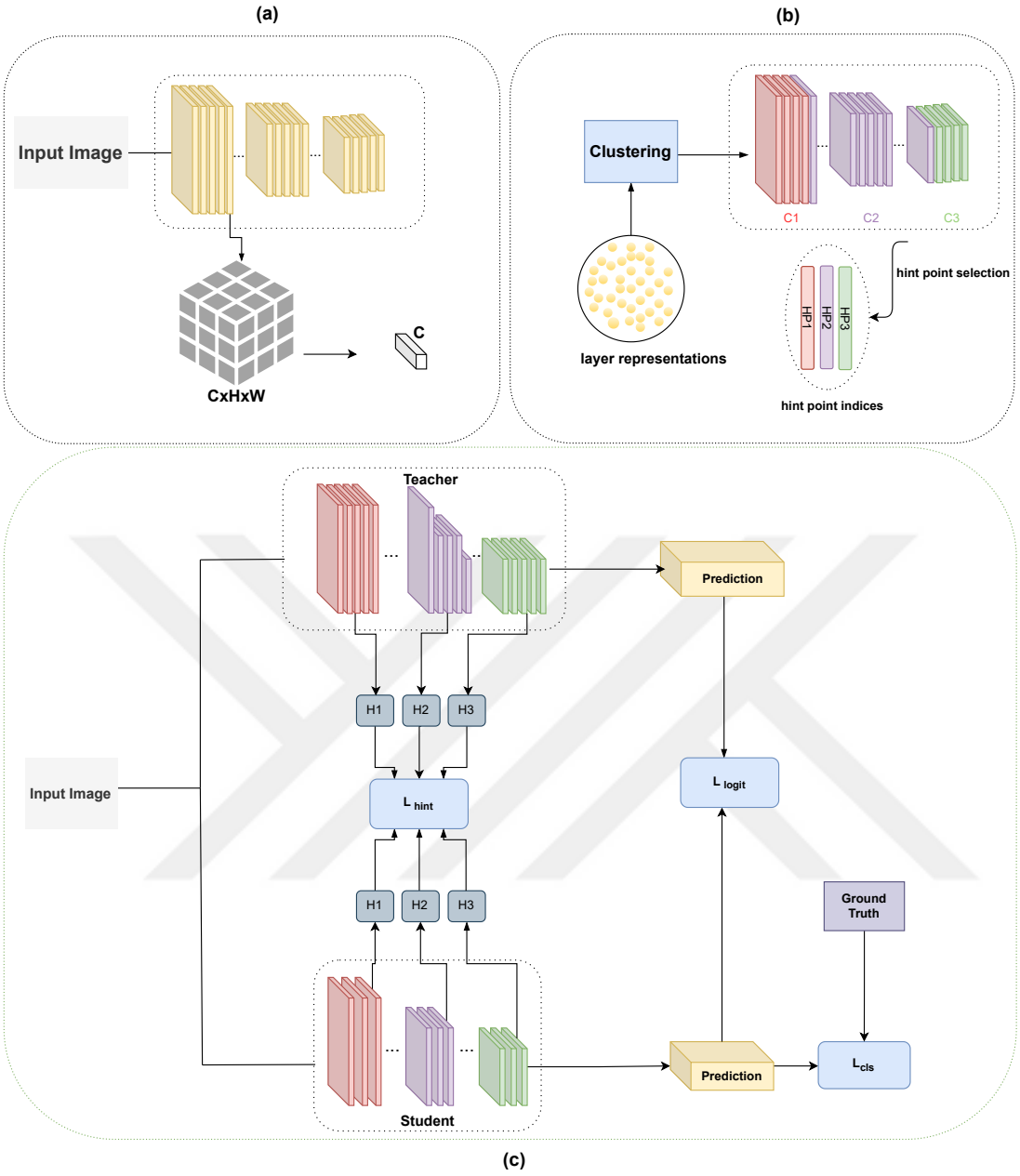


Figure 3.1 : The workflow of PURSUhInT comprises three stages: (a) The acquisition of layer representations. To this end, feature maps for a certain number of samples (N) are collected from all of the sub-blocks of the pre-trained teacher and averaged in dimensions of height and width to form matrices with the size of $N \times C$. (b) Clustering layer representations to obtain clusters of sub-blocks utilizing k-means. We select one hint position for each cluster, specifically the center positions of the clusters, to eliminate the redundancy. (c) Distillation with the determined hint positions in the previous step. In the diagrams, HP stands for the index of the hint point, H represents the transformation that determines the hint type, and CL depicts the determined cluster. Best viewed in color.

dimensions share parameters. Thus, the feature maps are averaged in dimensions of height and width to acquire representation matrices with the size of $N \times C$. For the clustering step, feature maps of the sub-blocks are collected for 10^4 samples from the training set.

3.3.2 Clustering of layers

Layer clustering is the major methodology exploited in the proposed knowledge distillation approach. Clustering methods require a similarity or distance metric to be able to perform, as known. To compute the similarity between distinct layers of an artificial neural network, existing approaches utilize numerous attributes of the model, including activation maps [119,120], weight matrices [122], and the layers with the same order at different models [123]. For instance, [123] suggests exploiting correlation as a similarity metric to assess the obtained knowledge of distinct models that share an identical architecture. In that study, it is reported that correlation performs similarly with mutual information, where they are utilized for comparison. Further, [120] employs canonical correlation analysis (CCA) to assess the similarity between layers. In addition, [121] suggests exploiting a weighting scheme to enhance this metric. After that, [119] introduces to employ centered kernel alignment (CKA) as a similarity index, whose linear case is reconciled with CCA. In this study, CKA and mean squared CCA (R^2_{CCA}) are utilized as the similarity metrics for layers of a neural network, where R^2_{CCA} is a statistic that represents the convergence of CCA [119]. Further, R^2_{CCA} is recognized as Yanai's GCD measure [124].

Mean squared CCA can be computed by (3.1) as in the following:

$$R^2_{CCA}(X,Y) = \frac{\sum_{i=1}^{p_1} \rho_i^2}{p_1} = \frac{\left\| Q_Y^T Q_X \right\|_F^2}{p_1} \tag{3.1}$$

where p_1 represents the minimum of shapes of the components, ρ_i stands for the i^{th} canonical correlation coefficient, Q_Y and Q_X are the orthonormal bases for the columns of Y and X, respectively.

Moreover, CKA is calculated by (3.2) as in the following:

$$CKA(K,L) = \frac{HSIC(K,L)}{\sqrt{HSIC(K,K)HSIC(L,L)}} \tag{3.2}$$

where HSIC represents Hilbert-Schmidt independence criteria. The empirical estimator of HSIC is computed in (3.3) as:

$$HSIC(K,L) = \frac{1}{(n-1)^2} tr(KHLH) \tag{3.3}$$

where H represents the centering matrix, tr is the trace of the matrix and n stands for the number of samples. $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$ where k and l are two kernels. In this study, we utilize linear kernels unless it is explicitly stated that RBF kernel is employed.

Hints should have succinct information, to distill the teacher's knowledge accurately. Thus, we acquire clusters of similar layers in terms of the selected metrics, namely R^2_{CCA} and CKA, which are exploited to group the layers of the teacher model. Then, optimum hint positions comprising only the necessary information, are determined by choosing one hint position per cluster. To this end, $1-R^2_{CCA}$ and $1-CKA$ are employed as the distance metrics for the k-means algorithm where k corresponds to the number of hints.

K-means is a clustering approach whose goal is splitting the n observations into k clusters. While k-means searches the best centroids for each cluster, it repeats two steps until the assignments are converged, where these steps are referred to as assignment and update, and are explained as in the following:

1) Assignment: Assign the data points based on the labels of the current nearest centroids:

$$y_i = \arg\min_j D(x_i, \mu_j), \tag{3.4}$$

where μ_j stands for the centroid of j^{th} cluster. x_i and y_i are the features and assigned cluster of the i^{th} data point, respectively. D is the selected distance metric, which is determined experimentally.

2) Update: Determine the centroids based on the current assignment of data points for each cluster:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i, \tag{3.5}$$

where C_j represents the j^{th} cluster.

In this study, we employ k-means as the clustering methodology, because of its simplicity and effectiveness. Even though it has disadvantages for other applications, these are not a limitation for our problem, since the number of clusters is fixed in advance, which is the number of hints and our data does not possess high dimensionality. Further, our scheme eliminates the seeding problem by selecting three layers farthest from each other according to their order in the network. To be specific, we initialize the seeds as the first, center, and last layers (points) of a network.

The execution times for the first two steps of our workflow, namely, the acquisition of layer representations and clustering of these representations, are exhibited in Table 3.1, for the teacher models used in experiments on CIFAR-100. It should be noted that executing the first two steps once is adequate. Further, it is observed that clustering takes a shorter time when R^2_{CCA} is employed as the similarity metric, compared to employing CKA.

Table 3.1 : Execution times for Step 1 and Step 2 in our workflow, on CIFAR-100 dataset. Step 2 is accomplished by k-means with the stated similarity metrics. It should be highlighted that the execution times for the clustering step are measured by using only CPU, while this step can be accelerated by employing GPU.

Teacher model	Number of layers	Similarity metric	Execution time (s)	
			Step 1	Step 2
WRN-40-2	18	R^2_{CCA}	272	180
resnet32x4	15	R^2_{CCA}	239	220
resnet110	54	R^2_{CCA}	941	336
		CKA		4797

3.3.3 Hint distillation with hint point search

In general, hint distillation approaches are employed to enhance logit distillation, where the teacher sends more information to the student via the knowledge learned in intermediate layers, namely hints. Thus, the loss function comprises multiple terms

to supervise numerous layers of the student network [7,125]. Hence, the total loss function, L_{total}, utilized to train the student network can be stated as in the following:

$$L_{total} = \gamma L_{cls}(p^S, y) + \alpha L_{logit}(p^S, p^T) + \beta L_{hint}(f^S, f^T) \tag{3.6}$$

where L_{logit} and L_{cls} are the output (logit) distillation loss and classification loss, respectively. y, p, and f represent labels, logits, and hints obtained from networks, respectively. Moreover, T and S superscripts denote teacher and student networks, respectively. Further, γ, α, and β stand for the weights for the losses of classification, logit distillation, and hint distillation, respectively.

Hint distillation loss L_{hint} can be defined as in the following:

$$L_{hint} = \sum_{i}^{k} L(F(S_i), F(T_i)), \tag{3.7}$$

where k denotes the number of hints, L represents the predetermined loss function for hint transfer, F stands for the transformation which determines the hint type, T_i and S_i are the features from the i^{th} group of layers in teacher and student networks, respectively. It should be highlighted that, contrary to most of the prior studies focusing on the transformation F, our approach revises and reformulates the grouping approach.

The groups of layers that have the same spatial size, are employed in the studies done so far. Nevertheless, in our analyses, it is monitored that some of these conventional hint points belong to the same cluster which is obtained by clustering via particularly designed metrics for layer similarity (See Figure 3.2). Our analyses indicate that determining hint points has a significant effect on the distillation performance. Thus, we suggest constructing the groups of layers by clustering utilizing suitable metrics, to achieve the efficient hint points. Our approach avoids utilizing unnecessary information acquired from similar layers, for distillation.

Further, we suggest exploiting the center points as the representative point of the layer groups, instead of the last points as in the conventional method. Thus, we analyze the performances corresponding to these points with distinct metrics as shown in Table 3.2. In these experiments, the first points in the clusters are excluded, because selecting the first points/layers in clusters results in transferring knowledge from the first layer of a

model for the first hint. This would yield transferring uninformative knowledge since the features learned in the first layer of networks are generic and usually irrelevant with the objective function [126].

Table 3.2 : Top-1 accuracy (%) results of distilled resnet20 for different hint positions in baseline grouping and clusters acquired with different metrics, on CIFAR-100. Attention Transfer is employed as the distillation type in these experiments. * denotes that normalization is applied before clustering. Although R^2_{CCA} results in different clusters for the normalized and non-normalized scenarios, the center points are identical for these scenarios.

Metric type	Position on cluster	Accuracy
R^2_{CCA}*	last	71.22 ± 0.13
	center	71.59 ± 0.32
R^2_{CCA}	last	70.99 ± 0.16
	center	71.59 ± 0.32
CKA_{linear} or CKA_{rbf}	last	71.20 ± 0.17
	center	71.50 ± 0.14
None (Baseline)	last	71.35 ± 0.43
	center	71.29 ± 0.15

The results indicate that employing center points results in better distillation performance than the last points of clusters we obtained. Moreover, the experiments highlight that altering hint points as the center points without the suggested grouping, results in a decrease in the distillation performance instead of an improvement.

3.4 Experiments

To assess the performance of the proposed method, we conduct experiments on two datasets, namely CIFAR-100 and ImageNet, which are usually employed datasets in knowledge distillation studies. Furthermore, we exploit a wide collection of distillation approaches and architectures for teacher and student networks including ResNet [84], Wide ResNet [127], and ShuffleNet [128]. In addition, we analyze our method on the scenarios of pre-activation and post-activation hints.

Our approach comprises three stages, as mentioned in previous sections. Following the acquirement of layer representations of candidate hint points, we apply clustering on these representations utilizing the k-means method. To conform with the conventional number of hints, we select $k = 3$ for the experiments on CIFAR-100 and $k = 4$ for

the experiments on ImageNet. Following the clustering step, we determine the center points of the clusters as the proposed hint points. Table 3.3 points out the distinctness between the conventional hint positions and hint positions determined by our method. In addition, the classification performances of the teacher networks employed in our experiments are demonstrated. It is monitored that some of the conventional hint positions may be settled in the same clusters determined by our approach, as presented in Figure 3.2. This is the main reason for the inefficiency in transferring knowledge from the teacher network to the student network in the conventional methods. The last stage of our scheme is the distillation via the proposed hint points, which will be described in the following sections.

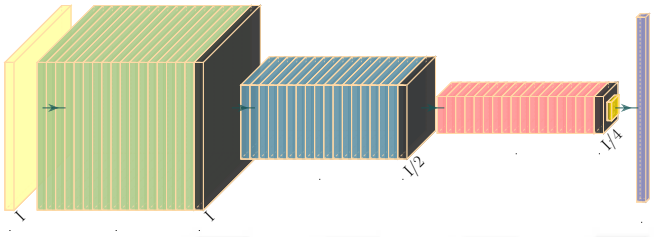
Table 3.3 : Comparison of baseline and proposed hint positions (HPs) for teacher networks employed in our experiments. Proposed HPs are determined by our scheme which exploits k-means with particular metrics as the distance functions. Further, the datasets utilized to test the teacher models and the accuracies achieved by the teacher networks are demonstrated. AT stands for the Attention Transfer [1] as the hint type.

Teacher Model	Dataset	Accuracy	Baseline HPs	Proposed HPs
Pre-activation hints:				
resnet110 (for AT)	CIFAR-100	74.49	18, 36, 54	6,27,49
resnet110 (for others)	CIFAR-100	74.49	18, 36, 54	8,29,49
Post-activation hints:				
resnet110	CIFAR-100	74.31	18, 36, 54	7, 29, 49
resnet32x4	CIFAR-100	79.42	5, 10, 15	3, 8, 13
WRN-40-2	CIFAR-100	75.61	6, 12, 18	1, 8, 16
ResNet-34	ImageNet	73.31	3, 7, 13, 16	2, 6, 11, 15

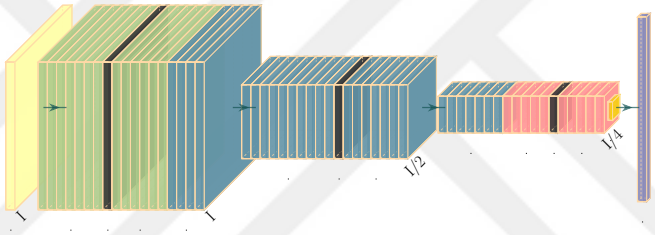
3.4.1 Experimental settings

The similarity metrics employed in the clustering process are chosen taking into account the distillation performances, in our experiments. To implement the hint types, the setup of [3] is followed, except FitNets [28]. In this work, three hints are utilized for FitNets, for a fair comparison with other hint types. Further, weights for the loss terms are selected via Bayesian search [129], for all methods. Besides, the query-key dimension for AFD method [107] is selected similarly.

SGD is employed as the optimizer in all experiments. For experiments on pre-activation hints, pre-ReLU hints are utilized for all methods. The maximum



((a))



((b))

Figure 3.2 : Characteristic structure of ResNet models with hint positions which are colored black. The labels at the bottom right of the blocks stand for the spatial sizes of feature maps, where *I* represents the spatial size of the input image. (a) Common method to group layers of ResNets which only considers the spatial size of the feature maps. The last points of these groups are determined as the hint positions. (b) Our proposed method which employs clustering to focus on layer similarity. This figure demonstrates the clusters determined by our approach with *CKA* metric on resnet110 sub-blocks, for the CIFAR-100 dataset. We select the center points of each cluster as the hint positions for the distillation of teacher models. It is visualized using [12] and best viewed in color.

number of epochs and the batch size are selected as 350 and 64, respectively. Further, the initial learning rate is 0.05 and it is decayed by 0.1 every 50 epochs after the first 200 epochs. For experiments on post-activation hints, post-ReLU hints are utilized for all methods. The maximum number of epochs is chosen as 240 to compare fairly the proposed method with other methodologies. The remaining hyperparameters are selected via Bayesian search, including temperature, learning rate, and batch size.

For ImageNet experiments, Attention Transfer is utilized as the hint type with $p = 1$ [1]. The maximum number of iterations is chosen as 100 epochs, batch size as 256, and temperature as 4, by following [3]. SGD is employed as the optimizer. Further, the initial learning rate is 0.1 and it is decayed by 0.1 at epochs of 30, 60, 80, and 90.

3.4.2 Results on CIFAR-100

To assess the performance of the proposed method on the CIFAR-100 dataset, experiments are conducted for pre-activation and post-activation hints, which are the hints acquired before and after the last ReLU function in a sub-block, respectively.

3.4.2.1 Pre-activation hints

For experiments on pre-activation hints, three hint distillation methods are selected to compare our findings, which are FitNets [28], Attention Transfer (AT) [1] and Variational Information Distillation (VID) [106]. Two logit distillation methods are chosen to employ along hint distillation methods, which are Hinton's Knowledge Distillation (KD) [5] and Spherical Knowledge Distillation (SKD) [93]. Hence, 6 experiments are conducted for each teacher-student pair. To evaluate our performance, proposed hint positions are compared with the baseline hint positions, which are the last layers before the downsampling. CIFAR-style ResNets are utilized as the teacher and student networks for distillation. Further, the R^2_{CCA} metric is exploited for the AT (Attention Transfer) type of distillation, and the *CKA* metric is exploited for two other types of hint distillation, in these experiments. Baseline and proposed hint positions are demonstrated in Table 3.3.

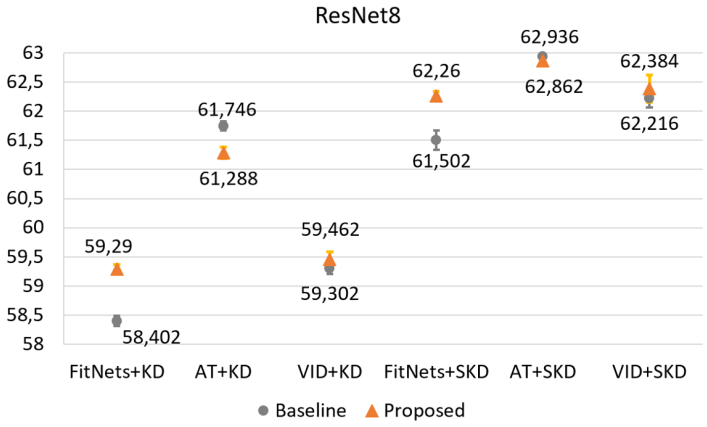
The proposed hint positions are applicable for any student network with the same number of hints since the proposed approach suggests efficient hint positions based

on the given teacher model. Thus, we conduct experiments for three student networks, which are resnet8, resnet20, and resnet32, by exploiting the hint positions obtained on the teacher network, namely resnet110. The results of distillation from resnet110 to these student networks are demonstrated in Figure 3.3, for the scenarios of employing the proposed hint positions and the baseline hint positions.

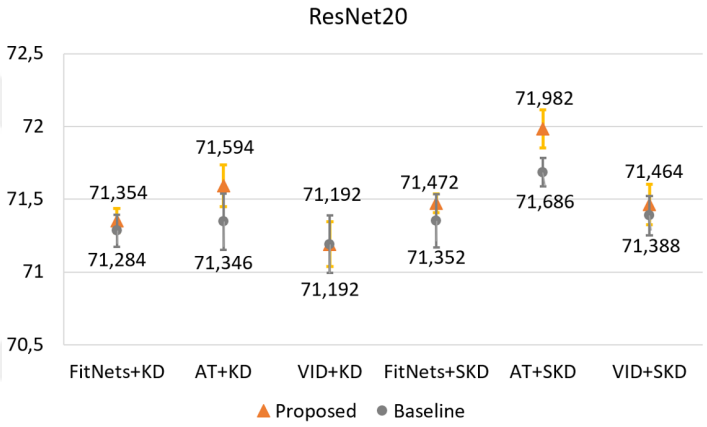
The proposed approach outperforms the conventional method for all hint types for resnet20, as the results suggested. Further, our approach results in the best distillation performance for resnet32. Moreover, it performs better than the baseline method except AT type of distillation for resnet8, while it obtains a very close performance to the best result among all hint types. Besides, the proposed approach substantially enhances the distillation performance of FitNets for resnet8.

3.4.2.2 Post-activation hints

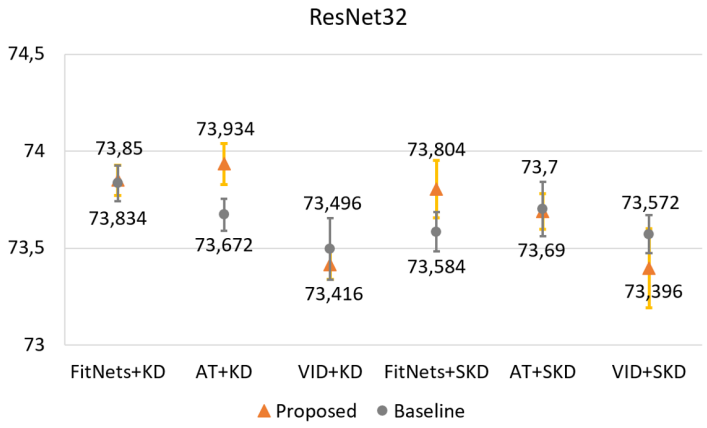
Six teacher-student pairs are employed for the experiments on post-activation hints. Our approach is implemented on Attention Transfer (AT) along with the conventional KD, which is referred to as AT+KD, where baseline and proposed hint positions are exploited for AT hints. Further, our method is applied to state-of-the-art distillation methods to boost their classification performance. With this aim, two recently introduced approaches, namely, Weighted Soft Label (WSL) [130] and Attention-based Feature Distillation (AFD) [107] are selected. WSL is a recent logit distillation approach that performs a sample-wise weighting on the objective function. This approach is enhanced by combining it with AT, to conduct experiments on hint position. To combine these approaches, the loss function in (3.6) is employed, where L_{hint} stands for the loss for AT hints and other terms refer to the loss of the WSL approach. After that, the results of WSL+AT are acquired utilizing the baseline and proposed hint positions for transferring AT hints. Another approach leveraged for the experiments is AFD, which searches informative hint positions between teacher and student networks, by exploiting an attention mechanism. It performs a knowledge transfer between these specified positions as well as the logits and repeats the search process along the training step. All intermediate output positions are assumed as potential hint positions for making connections between networks, in that approach.



((a))



((b))



((c))

Figure 3.3 : Top-1 accuracy (%) results of (a) resnet8, (b) resnet20, and (c) resnet32 student models for 6 distillation schemes. The accuracy of the teacher network is 74.490. Each score is computed over 5 runs. (Best viewed in color.)

Although it normally searches informative positions among all intermediate output positions, we integrate our method into this approach by limiting its potential hint points of the teacher as our proposed three hint points.

Table 3.4 demonstrates the results on state-of-the-art approaches, while the performances of the teacher models employed in the experiments are presented in Table 3.3. *CKA* metric is exploited for resnet110 and R^2_{CCA} metric is employed for resnet32x4 and WRN-40-2 teacher networks, to achieve the proposed hint points. The proposed hint position approach performs better than the baseline in most scenarios, as the results suggested. Further, it achieves an increase in accuracy up to 1.5 % compared to the baseline hint points. In addition, it consistently enhances the distillation performance for resnet32x4, resnet20, and resnet8x4 networks where the first one is employed as a teacher and/or the others as student networks. Though the search space of AFD is constrained to three proposed hint positions, the proposed approach mostly results in better distillation performances than the baseline. To be specific, the proposed method enhances the WSL+AT approach except for the resnet32 student network and the AFD approach except for one teacher-student pair which is WRN-40-2 - ShuffleNetV1. Moreover, it results in a 6.482 % increase in accuracy for ShuffleNetV1 student, compared to the vanilla training.

Our approach is compared with numerous distillation methods as in [3], for further assessment. To this end, the results of our approach are acquired by integrating it with four distillation methods, utilizing the same pre-trained teachers as other methods. The results acquired on the last epochs for a fair comparison with [3], are presented in Table 3.5.

Our method results in superior performance compared to the numerous methods, except for one teacher-student pair, as the results suggested. It is monitored that the best performances are achieved mostly by AT+KD and CRD+AT+KD methods with the proposed hint points. Moreover, integrating CRD to AT+KD results in a decrease in classification performance for the resnet32x4 teacher, while it enhances distillation performances of the two teacher-student pairs which are resnet110 - resnet20 and WRN-40-2 - WRN-16-2. Though the AFD approach with the proposed

Table 3.4 : Top-1 accuracy (%) results of student networks utilizing state-of-the-art distillation approaches on the CIFAR-100 dataset. The results are compared between the proposed and baseline hint points for each distillation scheme. Values in bold stand for the best classification accuracies for each distillation approach. Each score is computed over 5 runs. The column marked by * represents the reported results in [2].

Teacher	Student	Vanilla	AT + KD	
			Baseline	Proposed
resnet110	resnet20	69.06	70.95	**71.06**
	resnet32	71.14	**73.59**	73.48
resnet32x4	ShuffleNetV1	70.50	75.49	**76.98**
	resnet8x4	72.50	75.06	**76.07**
WRN-40-2	WRN-16-2	73.26	**75.64**	75.45
	ShuffleNetV1	70.50	76.27	**76.78**
			WSL + AT	
		WSL	Baseline	Proposed
resnet110	resnet20	71.24	71.45	**71.63**
	resnet32	73.49	**73.87**	73.73
resnet32x4	ShuffleNetV1	74.33	75.95	**76.98**
	resnet8x4	75.24	75.04	**75.69**
WRN-40-2	WRN-16-2	75.56	75.49	**75.60**
	ShuffleNetV1	75.29	75.84	**76.52**
			AFD	
			Baseline*	Proposed
resnet110	resnet20		71.20	**71.49**
	resnet32		73.46	**74.05**
resnet32x4	ShuffleNetV1		75.08	**75.11**
	resnet8x4		74.72	**75.80**
WRN-40-2	WRN-16-2		75.41	**75.60**
	ShuffleNetV1		**75.63**	75.11

Table 3.5 : Top-1 accuracy (%) results of networks utilizing numerous distillation approaches on the CIFAR-100 dataset. Each score is computed over 5 runs. Values in bold represent the best classification accuracies for each distillation method. It should be highlighted that reported results are based on the last epoch for a fair comparison with [3] where the results of other approaches are quoted from. Further, some results for the FSP [4] approach cannot be acquired since it is suitable for only student-teacher pairs having identical architectures.

Teacher Student	WRN-40-2 WRN-16-2	resnet110 resnet20	resnet110 resnet32	resnet32x4 resnet8x4	resnet32x4 ShuffleNetV1	WRN-40-2 ShuffleNetV1
Vanilla	75.61 73.26	74.31 69.06	74.31 71.14	79.42 72.50	79.42 70.50	75.61 70.50
KD [5]	74.92	70.67	73.08	73.33	74.07	74.83
FitNet [28]	73.58	68.99	71.06	73.50	73.59	73.73
AT [1]	74.08	70.22	72.31	73.44	71.73	73.32
SP [6]	73.83	70.04	72.69	72.94	73.48	74.52
CC [7]	73.56	69.48	71.48	72.97	71.14	71.38
VID [106]	74.11	70.16	72.61	73.09	73.38	73.61
RKD [33]	73.35	69.25	71.82	71.90	72.28	72.21
PKT [36]	74.54	70.25	72.61	73.64	74.10	73.89
AB [29]	72.50	69.53	70.98	73.17	73.55	73.34
FT [10]	73.25	70.22	72.37	72.86	71.75	72.03
FSP [4]	72.91	70.11	71.89	72.62	n/a	n/a
NST [55]	73.68	69.53	71.96	73.30	74.12	74.89
CRD [3]	75.48	**71.46**	73.48	75.51	75.11	76.05
Ours (AT+KD)	75.16	70.86	73.12	**75.84**	**76.74**	76.55
Ours (WSL+AT)	75.39	71.43	73.44	75.49	76.70	76.36
Ours (AFD)	75.33	71.22	**73.79**	75.52	74.94	74.94
Ours (CRD+AT+KD)	**75.82**	71.32	73.19	74.20	74.88	**76.66**

hint points, does not perform as well as others for ShuffleNetV1 student, it results in the best accuracy for the resnet110 - resnet32 pair. It should be noted that our method on AT+KD results in a 1.6 % improvement over the second-best approach for teacher-student pair of resnet32x4 - ShuffleNetV1.

3.4.3 Results on ImageNet

For further assessment of the performance of our method, the ImageNet dataset is employed, which is a well-known large-scale dataset. With this aim, Attention Transfer (AT) is exploited as the hint type to be transferred, and KD is employed for logit distillation, where this experiment is referred to as AT+KD. In these experiments, ImageNet-style ResNet-34 and ResNet-18 architectures are employed as teacher and

Table 3.6 : Top-1 and Top-5 error rates on ImageNet validation set. ResNet-34 and ResNet-18 are employed as the teacher and student networks, respectively. Results of other methods are quoted from [3], where other methods comprises AT [1], KD [5], SP [6], CC [7], Online KD [8] and CRD [3].

	Teacher	Student	AT	KD	SP	CC	Online KD	CRD	CRD+KD	Ours (AT+KD)
Top-1	26.69	30.25	29.30	29.34	29.38	30.04	29.45	28.83	28.62	**28.51**
Top-5	8.58	10.93	10.00	10.12	10.20	10.83	10.41	9.87	**9.51**	9.83

student networks as in [3], respectively. The *CKA* metric is exploited for clustering the layers of the ResNet-34 teacher network to acquire the proposed hint points for AT hints. The results compared with the other knowledge distillation approaches are demonstrated in Table 3.6, where the results of other approaches are quoted from [3]. Further, selected hint points by our approach are presented in Table 3.3.

Results suggest that the proposed hint position approach results in the best Top-1 accuracy among the compared distillation methods on ImageNet, which is one of the most challenging datasets in computer vision. Furthermore, it is observed that only altering hint points for a typical hint type may result in superior results compared to the recent approaches employing contrastive learning.

The proposed approach enhances the hint distillation schemes which play a significant role in knowledge distillation for model compression, as the results suggested. The gains of our approach in terms of accuracy among inference time and model size, are demonstrated in Table 3.7. Results point out that the inference time of resnet110 may be enhanced by a factor of 2.99 with only a Top-1 accuracy loss of 0.26%. Moreover, resnet32x4 can be compressed into resnet8x4 with a memory gain of 83.4 % and a speed-up in inference time of 2.34. Further, WRN40-2 can be compressed into ShuffleNetV1 with a memory gain of 57.9 % and an increase in accuracy of 1.27%. In addition, WRN40-2 can be compressed into WRN-16-2 with a memory gain of 68.8 % and an increase in accuracy of 0.4%. Furthermore, results on ImageNet indicate that ResNet-34 can be compressed into ResNet-18 with a speed-up in inference time of 1.55 and memory gain of 46.4 %.

Table 3.7 : Number of parameters, inference time per sample and Top-1 accuracy(%) results for teacher and student models. Student models are compared with teacher models in the viewpoints of compression ratio and achieved speed-up, where student networks are trained by exploiting the proposed hint points. It should be highlighted that the classification accuracies of ResNet-34 and ResNet-18 networks are acquired on ImageNet.

Networks	**#Parameters**	**Compression ratio**	**Inference time**	**Speed-up**	**Accuracy**
Teacher: resnet110	1.74 M	-	24.66 ms	-	74.31
resnet20	278.32 K	84.0 %	5.84 ms	x4.22	71.63
resnet32	472.76 K	72.8 %	8.25 ms	x2.99	74.05
Teacher: resnet32x4	7.43 M	-	8.00 ms	-	79.42
ShuffleNetV1	949.26 K	87.2 %	13.37 ms	x0.60	76.98
resnet8x4	1.23 M	83.4 %	3.42 ms	x2.34	76.07
Teacher: WRN-40-2	2.26 M	-	10.07 ms	-	75.61
WRN-16-2	703.28 K	68.8 %	4.71 ms	x2.14	76.01
ShuffleNetV1	949.26 K	57.9 %	13.37 ms	x0.75	76.88
Teacher: ResNet-34	21.80 M	-	3.30 ms	-	73.31
ResNet-18	11.69 M	46.4 %	2.13 ms	x1.55	71.49

3.5 Conclusion

In this study, the grouping problem on teacher models is addressed to select the hint points among the model's sub-blocks. With this aim, the k-means algorithm with metrics designed for layer similarity is exploited for clustering these sub-blocks. The proposed method is suitable for any hint distillation scenario employing at least two hints, in an offline distillation setup. Moreover, it is valid for any student network since it results in efficient hint positions based on the teacher model.

To assess the performance of the proposed method, we integrate our approach to state-of-the-art distillation approaches with the comparison of the conventional method and the proposed method. Further, a comprehensive comparison among numerous distillation methods and approaches employing the proposed hint points is demonstrated. The proposed method outperforms the state-of-the-art algorithms for various architectures on CIFAR-100 and ImageNet datasets, as the experimental results suggested. In addition, the proposed approach performs successfully in the viewpoint of model compression, where it may result in a high compression ratio, speed-up in inference time, and a gain in accuracy, at the same time.

4. MODEL STEALING

Today, some black-box models are released by companies as a service for users, where users can obtain the output of the model which is fed by their input. However, these models are vulnerable to stealing via knowledge distillation by malicious users, who aim to access the operating model which is a commercial product [131,132]. For example, a malicious user can collect the model's predictions for the inputs submitted, which are designed to extract information from the model. Then this information is utilized to train a substitute model. The malicious user can now offer this model as a commercial product, which is obtained by only queries of this user. Further, the malicious user can generate adversarial examples to make the original model make wrong predictions, using the substitute model. To address this issue, the concept of an undistillable teacher [19] has been introduced recently to prevent knowledge leakage. An undistillable teacher is a specially trained model that nearly performs as a normal teacher while it degrades the student's performance which is trained by knowledge distillation from the undistillable teacher. Hence, aiming to contribute to solutions for protecting model intellectual property (IP), we propose a novel method that enhances the distillation from an undistillable teacher, in this chapter. Our approach contributes to the fields of model protection and data protection since protecting the model's intellectual property covers protecting the data used to train the model.

4.1 Introduction

Protecting the intellectual property (IP) of models is crucial due to the potential negative consequences. Firstly, a stolen model could be exploited for commercial gain, posing a threat to the original model's owner company. Additionally, malicious users might leverage the stolen model to generate adversarial examples. Moreover, the extraction of a model could lead to the exposure of sensitive training data used in its training step [133].

While knowledge distillation offers many benefits across various applications, it also enables a risk of IP stealing by malicious actors [19]. For instance, data-free knowledge distillation methods can be employed to obtain the functionality and training data of the original model [57]. In the literature, the methods to obtain the model's IP can be categorized as model stealing or model extraction [134,135].

4.2 Related Work

In the context of the model stealing problem, a recent study [19] introduced an approach termed *self-undermining knowledge distillation* to protect models against stealing via knowledge distillation. The proposed method in that study, namely, *undistillable teacher*, is obtained with the training process via self-undermining knowledge distillation. An undistillable teacher hardens the training of a student that is trained with knowledge distillation from that teacher, while it keeps a similar performance with normally trained teachers. Building on this research, [9] introduced a method called *skeptical student* which aims to mitigate the performance degradation caused by undistillable teachers. Their method changes the position of distillation to train their student model, along with employing auxiliary classifiers and a self-distillation approach.

4.3 Recap: Knowledge Distillation

Knowledge distillation (KD) is a method for training a model, which defines the process of transferring knowledge between models via minimizing some sort of distance function between the outputs of chosen layers of the models. Soft targets define the knowledge obtained from the teacher model, where they are generally transferred in combination with labels, namely, hard targets. We can define the loss function for the student model as in (4.1), to clarify the difference between the conventional knowledge distillation and the proposed approach:

$$L_S = \lambda L_{KD}(z^S, z^T) + (1-\lambda) L_{cls}(z^S, y) \tag{4.1}$$

where λ represents the balance between distillation loss L_{KD} and classification loss L_{cls}, y is the label, z^S and z^T are the logits of student and teacher networks, respectively. Kullback-Leibler and cross-entropy losses are commonly utilized for L_{KD} and L_{cls}, respectively. The distillation loss, L_{KD}, can be calculated as in (4.2):

$$L_{KD} = \sum_b \sum_i L(z_i^T, z_i^S) \tag{4.2}$$

where L is the loss function, b represents the batches, i shows the samples in a batch, z_i^T and z_i^S indicate the logits correspond to sample i obtained from teacher and student, respectively.

4.4 Proposed Method

In both [9] and [19], it is demonstrated that multiple peaks are exhibited in undistillable teacher's softmax responses, which are then transferred to the student models. These peaks are believed to be the major cause of misguidance in the student models. Thus, in this chapter, we propose an approach to mitigate the impact of these multiple peaks in the softmax responses of the teachers, where we introduce transferring the mean of features with identical labels, as the soft labels. To achieve this, the features of the teacher are extracted in batches and the mean values of them are computed. Without loss of generality, we utilize logits as the features throughout this section, corresponding to the scenario of logit distillation [5]. The loss function employed to train the student model is presented in (4.3) as follows:

$$L_S = \sum_b \sum_i L(\widehat{z}_i^T, z_i^S) \tag{4.3}$$

where $\widehat{z}$ is the adjusted logits in a batch, which consists of the mean values of logits obtained from samples with the same labels. The mean values of logits in a batch for a label y, namely $\widehat{z}_y$, is computed as in (4.4):

$$\widehat{z}_y = \frac{1}{N} \sum_x z_{x_y} \tag{4.4}$$

where x_y represent the samples with the label y, N is the number of samples that have the label y, in a batch. $\widehat{z}_y$ is the mean of logits correspond to x_y.

4.5 Experiments and Results

To evaluate the performance of our proposed approach, we follow the experimental setup of the prior work [9]. Thus, we obtain distillation results from undistillable teachers, using the same architectures for teacher and student models as in [9]. To be precise, we utilize ResNet-18 and ResNet-50 as the teacher models on CIFAR-100 dataset. We utilize ResNet-18, MobileNetV2, and ResNet-50 as the student models.

Moreover, we obtain results for the ensemble scenario as in the prior work. The results are presented in Table 4.1. Results point out that our approach always outperforms [9] for all scenarios.

Table 4.1 : Results on CIFAR-100 for the undistillable teachers. KD and Skeptical represent [5]'s and [9]'s approach, respectively. Bold values indicate the best results for each scenario. (RN: ResNet, MN: MobileNetV2, Ens: Ensemble)

Teacher		**Student**		**KD**	**Skeptical**	**Ours**	**Skeptical -Ens**	**Ours -Ens**
Model	**Acc. (%)**	**Model**	**Acc. (%)**					
RN-18	77.55	RN-18	77.55	75.00	77.33	**77.39**	76.38	**79.69**
		MN	69.24	7.13	66.62	**71.80**	64.26	**74.92**
RN-50	76.57	RN-18	77.55	72.28	77.25	**77.53**	75.48	**79.62**
		RN-50	78.04	74.14	78.65	**79.18**	77.61	**81.60**
		MN	69.24	7.72	66.38	**70.73**	62.93	**74.98**

In a real scenario, it is not clear whether the teacher is a normally trained or an undistillable model. Thus, our approach is also evaluated on normal teachers, where distillation performances are given in Table 4.2.

Results point out that our method mostly outperforms [9]. We achieve an improvement in accuracy up to 2.2 % on CIFAR-100, for the scenario of a normal teacher.

4.6 Conclusion

In this chapter, a novel approach is introduced to improve the distillation performance of a student from an undistillable teacher. The proposed approach mitigates the incorrect supervision caused by the undistillable teacher by employing a simple yet

Table 4.2 : Results on CIFAR-100 for the normal teachers. KD and Skeptical represent [5]'s and [9]'s approach, respectively. Bold values indicate the best results for each scenario. (RN: ResNet, MN: MobileNetV2, Ens: Ensemble)

Teacher		**Student**		**KD**	**Skeptical**	**Ours**	**Skeptical -Ens**	**Ours -Ens**
Model	**Acc. (%)**	**Model**	**Acc. (%)**					
RN-18	77.55	RN-18	77.55	78.96	78.79	**78.98**	79.68	**79.95**
		MN	69.24	**75.12**	71.63	73.83	**75.45**	75.345
RN-50	78.04	RN-18	77.55	79.21	78.51	**79.39**	79.86	**80.205**
		RN-50	78.04	79.56	**80.66**	80.53	81.96	**82.41**
		MN	69.24	**75.28**	71.76	73.42	**76.32**	75.455

effective technique. Results show that our method outperforms the compared methods. We believe that our approach will contribute to the solutions for the field of model and data protection.

5. CROSS DISTILLATION

Data play a crucial role in the training of a model and directly impacts its performance. Nevertheless, there might be inadequate data for a model to achieve sufficient performance on some tasks [136,137]. To address this challenge, we propose a knowledge distillation-based approach that aims to alleviate the limitations caused by data scarcity. To be precise, our approach enhances the pixel domain performance of a model by leveraging compressed domain knowledge via cross distillation between these two modalities. We validate our method through experiments conducted on two well-known computer vision tasks: object detection and recognition. The results point out that compressed domain knowledge can benefit a pixel domain task via the proposed method, especially when available data is limited due to privacy or copyright concerns.

To sum up, the proposed methodology contributes to the literature on data protection and minimization. Our work demonstrates a technique that expands limited knowledge by exploiting different modality data instead of more samples. Since we use compressed domain data and eliminate the need for more samples to enhance the performance, we prevent the use of more data that may be personal or sensitive.

5.1 Introduction

Recent advancements in computer vision have led to significant improvements in model performance. And, it is well known that these models often require large amounts of data. Nevertheless, acquiring sufficient data for some tasks may be challenging. For instance, collecting medical and multi-modal data poses difficulties. Moreover, processing personal data is hindered by tight regulations, which limits the available data to utilize in the training process of a model. These concerns are related to the concept of data minimization, which aims to keep and process only the necessary data for the operations [138]. It is an essential requirement imposed by data

regulations including The European General Data Protection Regulation (GDPR) and The California Privacy Rights Act (CPRA) [139,140]. Besides, knowledge distillation aligns with the goals of data minimization since the student is limited to being fed by the teacher's feedback in some cases including unsupervised schemes and data-free KD [141].

Given that data scarcity appears on numerous tasks due to various reasons, there is growing research interest aimed at enhancing knowledge extracted from limited data. On the other hand, compressed domain data is attractive data to be processed since it requires less computational complexity and enables faster processing time [142,143]. Further, operating on the compressed domain is a more conservative approach for processing sensitive data, compared to the pixel domain. For instance, the approach introduced in [144], generates images from the HEVC compression scheme where it limits the information in images for unintended tasks such as text recognition.

Thus, in this chapter, we introduce an approach to address the concerns of data scarcity, protection, and minimization. It leverages the compressed domain knowledge, while the task is in the pixel domain, on the contrary. Our method consists of a cross distillation scheme that performs on two image domains, as illustrated in Figure 5.1. Our approach contributes to data protection by enhancing model performance through the utilization of compressed domain data rather than relying on additional pixel domain data. Similarly, it addresses the concerns of data minimization, since it prevents the need for more pixel domain samples that may be sensitive, to boost the performance.

In this study, knowledge from the compressed domain is primarily employed to address tasks in the pixel domain. To achieve this objective, images in the HEVC domain are exploited, where HEVC (High Efficiency Video Coding) is a video compression standard developed jointly by the ITU-T Video Coding Experts Group and the ISO/IEC Moving Picture Experts Group. HEVC enables significantly improved compression efficiency compared to its predecessor, H.264/MPEG-4 AVC. It utilizes a hierarchical block structure, dividing a frame into coding tree units, coding units, transform units, and prediction units for processing. The prediction mechanism in HEVC can be

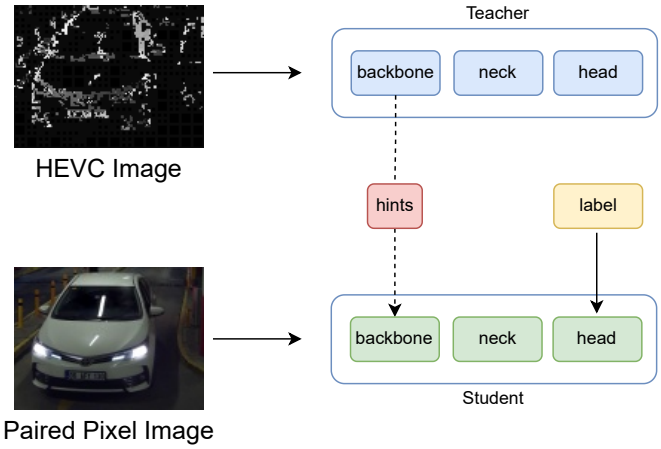


Figure 5.1 : The workflow of the proposed approach. The task loss is utilized to train the student model, as well as the cross distillation loss, which is computed between hints obtained from the pre-trained teacher model operating on the compressed domain and features of the student model operating on the pixel domain. A prediction unit (PU) based attribute of the HEVC legacy codec is applied on the "Paired Pixel Image", to obtain "HEVC Image".

either intra or inter, which uses information within the same frame to reduce spatial correlation or utilizes data from previous frames to mitigate temporal correlation, respectively.

The HEVC images utilized in this work, are derived from the prediction-based technique introduced by Beratoglu et al. (2021) [144]. The prediction units (PUs) in the HEVC bitstream are exploited in this method, which predicts the pixel values in the current block based on neighboring blocks. This approach involves a function that assigns varying intensity levels depending on the size of the coding blocks and particular regions within the image. It also employs a linear transformation of prediction unit values, spanning from 0 to 34, to the pixel intensity range of 0-255. This conversion procedure yields distinct gray-level regions in the ensuing image corresponding to different prediction unit values. Thus, this technique preserves the critical scene features while notably reducing image complexity.

Knowledge distillation is frequently employed to transfer knowledge from the pixel domain to the compressed domain, in existing studies on the compressed domain. Specifically, optical flow information is transferred into the models that perform on the compressed domain, since acquiring optical flow at inference time is a costly

computation. On the other hand, our proposed method employs compressed domain knowledge to enhance the models' performance on the pixel domain via knowledge distillation. Therefore, we establish the direction of knowledge transfer from the compressed domain to the pixel domain, contrary to the previous studies.

To summarize, we propose a method that leverages the compressed domain knowledge for computer vision tasks on the pixel domain, to expand the knowledge obtained from limited data. Thus, our contributions can be listed as follows:

- We introduce the first cross distillation method that transfers knowledge from the compressed domain to the pixel domain, to the best of our knowledge.
- The proposed method enhances the pixel domain performance, while compressed domain data are used only in the training step.
- We validate our method through two well-known common computer vision tasks which are object detection and recognition.

5.2 Related Work

In this section, we review the literature from the viewpoint of two concepts, that are cross distillation and knowledge distillation on the compressed domain.

5.2.1 Cross distillation

Knowledge distillation (KD) [5] mainly comprises the transfer of knowledge between models, and this field has broadened to encompass a variety of subtopics including hint distillation [145], self-distillation [40] and, online distillation [39]. One of these subtopics, namely *cross distillation* [49], denotes the transfer of knowledge across various modalities, such as RGB and depth images, or, pixel and compressed domain images. It can find applications in numerous tasks including human pose estimation, emotion recognition, object detection, and action recognition [49,146]–[149]. For instance, cross distillation is utilized for the action recognition task in [146], where the teacher model is fed by RGB videos and the student model is fed by paired RGB videos and sequences of 3D human poses. Further, a cross distillation scheme is employed in [49] to transfer knowledge between the modalities of RGB images

and depth images. In that study, obtained hints are transferred to the student model. In another study [147], knowledge between the domains of RF heatmaps and RGB frames is transferred to estimate human poses. Moreover, [149] employs an online distillation scheme, where cross distillation is leveraged between modalities of RGB and DensePose sequences. These mentioned studies utilize the RGB domain as the source of auxiliary knowledge. On the contrary, our method employs RGB and compressed domains as the main and auxiliary domains, whereas the RGB domain includes more rich knowledge.

5.2.2 Knowledge distillation on the compressed domain

Knowledge distillation (KD) has been employed with various aims on the compressed domain, in recent studies. It is mostly utilized to transfer the pixel domain knowledge to models performing on the compressed domain in the literature. Since the computation of optical flow is a costly operation, KD is used to transfer the optical flow information to the models that are fed by compressed domain features in [150], [151], [152], [153], [154] and [155]. For instance, optical flow information is transferred from a pre-trained teacher model in [156], where a multi-modal student is employed which takes motion vectors as the input, as well as the residuals. In [157], a KD scheme is introduced to train a compressed domain student which imitates a pixel domain teacher. This KD approach includes a hint distillation scheme between the models, where I-frames and augmented residual frames are the inputs of the student model, whose task is action recognition. Likewise, a cross distillation scheme is proposed to transfer the pixel domain knowledge to the compressed domain in [158], where mostly the last layers of the models are chosen as the distillation position. On the other hand, the knowledge is transferred in a reverse direction in our proposed approach, i.e., from the compressed domain to the pixel domain.

Furthermore, some studies employ KD for transferring knowledge or model compression on the compressed domain, where both the teacher and student models perform on the compressed domain. For example, a multi-teacher scheme is introduced in [159], where each teacher model receives input from a single modality as in [160] and the student receives I-frames as the input. Moreover, KD is employed to transfer

spatial information to the student model that performs in the frequency domain, where both the teacher and student receive I-frames as the input, in [161]. On the contrary, the student model undergoes the training and inference steps within the pixel domain, in our workflow.

5.3 Method

In this study, a cross distillation method is introduced where the pixel domain performance is enhanced by leveraging the compressed domain knowledge. To achieve this goal, a pair of teacher-student models is used, where HEVC domain images are utilized to pre-train the teacher model. To train the student model, the student receives the pixel domain images as the input, as well as the hints obtained from the teacher whose inputs are the paired compressed domain images, as presented in Figure 5.1.

5.3.1 Training of teacher model

The HEVC domain images produced by the prediction-based method, are employed to train the teacher model, which is trained from scratch, i.e., only task loss is utilized.

5.3.2 Training of student model

The set of image representations of the teacher model, namely Φ_T, is defined as follows:

$$\Phi_T = \{\phi^i \ \forall i \in \{1,...,k\}\} \tag{5.1}$$

where i is the index of the hint and k represents the number of hints. The representation of the i^{th} hint layer, ϕ^i, can be denoted as:

$$\phi^i : X_{comp} \rightarrow \mathbb{R}^{(H \times W \times C)_i} \tag{5.2}$$

A compressed domain image is mapped to a feature map with the size of $(H \times W \times C)_i$, by ϕ^i. It should be noted that i stands for the index of the hint layer, instead of a layer in a neural network.

Likewise, the set of image representations of the student model, namely Φ_S, can be defined as follows:

$$\Phi_S = \{\psi^i \ \forall i \in \{1,...,k\}\} \tag{5.3}$$

where ψ^i stands for the representation of the i^{th} hint layer of the student model. A pixel domain image is mapped to a feature map with the size of $(H \times W \times C)_i$, by ψ^i:

$$\psi^i : X_{pix} \rightarrow \mathbb{R}^{(H \times W \times C)_i} \tag{5.4}$$

The following loss function is employed to train the student model:

$$L_{total} = \lambda L_{hint} + L_{task} \tag{5.5}$$

where, λ represents the balance between loss terms and L_{task} is the task loss, which is object detection or object recognition loss, in our work. Hint loss L_{hint} can be defined as in the following:

$$L_{hint} = \sum_{i}^{k} L_{feat}\ (\phi^i(I_{comp}),\ DA(\psi^i(I_{pix}))) \tag{5.6}$$

where $(I_{pix}, I_{comp}) \in P$. P stands for the paired dataset which comprises the pixel domain images and their corresponding compressed domain images. Moreover, DA and L_{feat} are the domain adaptor module and feature loss that is determined by hint type, respectively. Further, a convolutional layer is employed as the domain adaptor, DA. We utilize a domain adaptor to diminish the domain gap between the representations of teacher and student.

The neural networks in this study are characterized as a combination of three components which are backbone, neck, and head, where the definition in [162] is followed. To be specific, the proposed cross modal distillation scheme is established between the backbones of teacher and student models, where the hint layers of the teacher model are determined following the grouping and selection approach proposed in our previous study [18].

5.4 Experiments

In this section, conducted experiments will be explained, which are based on two computer vision tasks. First, experiments are conducted for all scenarios in terms of relative sizes of teacher-student model pairs, for evaluation of the proposed approach. Further, various hint types in the proposed distillation scheme are applied for additional assessment.

5.4.1 Datasets

In this work, a cross distillation scheme to transfer the knowledge from the compressed domain to the pixel domain is introduced. To achieve this goal, two compressed domain datasets are exploited that are CD-LP [13] and [15], which comprises the compressed domain images of BIT-Vehicle dataset [14]. [13] and [15] are employed for the problems of license plate detection and vehicle classification, respectively.

CD-LP dataset which comprises paired 2400 images for pixel and HEVC domains, is employed for the object detection problem. In [144], two types of images are produced for the HEVC domain, by two methods, namely block partition and prediction-based approaches. The training and test sets involve 1800 and 600 images, respectively. All images in the dataset are the identical size, which is 128×96. It should be highlighted that we employ the HEVC images produced by the prediction-based method, in this work.

The dataset in [15] is exploited for the object recognition problem, which comprises 9850 vehicle images for pixel and HEVC domains. This dataset was generated based on the BIT-Vehicle dataset [14]. The employed dataset comprises six classes of vehicles, which are bus, sedan, sport-utility vehicle (SUV), microbus, truck, and minivan. The pixel domain images in the dataset can be in two sizes which are 1600×1200 and 1920×1080, while all HEVC domain images are 128×96 in size. Nevertheless, the pixel domain images are resized to 128×96 pixels in the conducted experiments. The dataset is divided into a training and a test set with a ratio of 4:1, following the prior study. Further, the images containing more than one vehicle are removed and the HEVC images produced by the prediction-based method, are

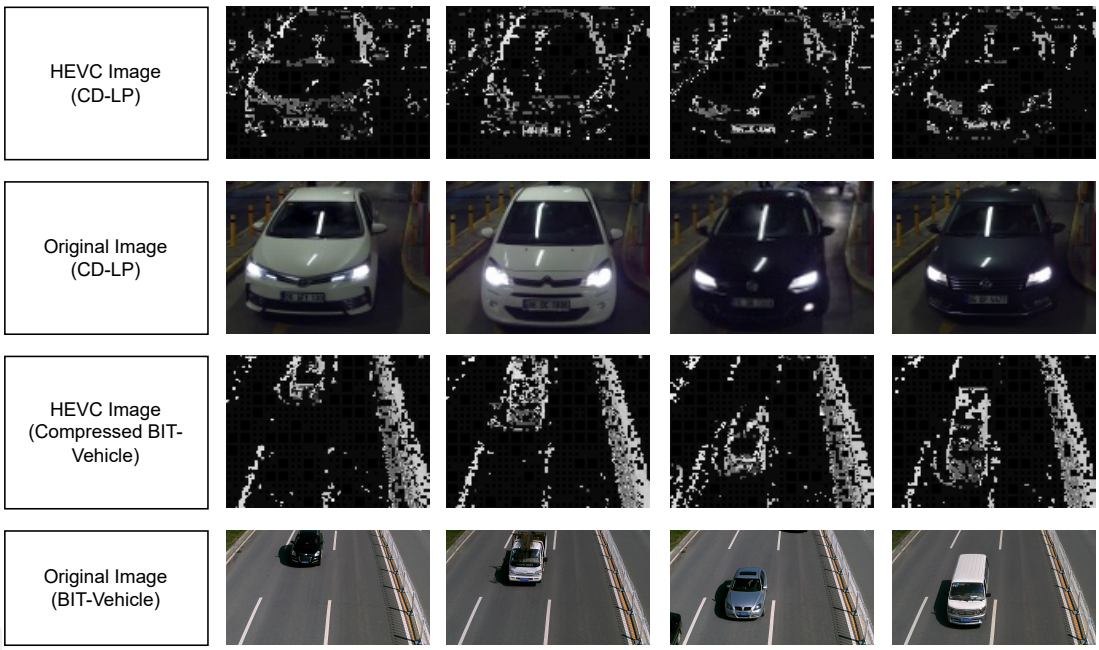


Figure 5.2 : Samples of compressed and pixel domain images from the license plate detection dataset, namely CD-LP [13], and the vehicle classification datasets, namely BIT-Vehicle [14] and compressed BIT-Vehicle [15].

exploited in this work. Some exemplary images of these datasets are presented in Figure 5.2.

5.4.2 Implementation details and results

5.4.2.1 Object detection experiments

First, experiments are conducted on the license plate detection problem, to assess the performance of the proposed approach. To achieve this goal, the YOLOv3 model [163] is exploited as the object detector. Two types of backbones are employed to acquire different-sized pairs of teacher and student models, which are DarkNet-53 [163] and MobileNetv2 [104]. For these experiments, *Attention* [1] is utilized as the hint type to be transferred, where the number of hints is 3. AdamW [164] is employed as the optimizer. Furthermore, the maximum number of iterations is chosen as 300 epochs and the batch size is set as 64. Moreover, the learning rate is chosen as 0.01 and the learning rate multiplier for the backbone is set as 0.1 or 1.

Table 5.1 : Student and teacher models' performance in terms of AP50 (%) score, on license plate detection problem. The row at the top stands for the backbones of YOLOv3 teacher-student model pairs where DN is for DarkNet-53 and MN is for MobileNetv2. Vanilla represents the training from scratch, i.e., training without distillation. It should be highlighted that the performances of teacher and student models are assessed on compressed and pixel domains, respectively. Each student model's score is computed over three runs.

T-S pairs	DN $\rightarrow$ MN	MN $\rightarrow$ MN	MN $\rightarrow$ DN
Teacher	97.70	97.60	97.60
Vanilla	98.70	98.70	97.30
Proposed	**99.03**	**98.77**	**97.97**

The performance scores of the teacher and student models on the CD-LP dataset are presented in Table 5.1, where the obtained results for teacher and student are on the compressed and pixel domains, respectively. The performances are measured in terms of the AP50 score which is the average precision when the Intersection over Union (IoU) threshold is set to 0.50. The proposed method achieves an increase in AP50 score up to 0.67 (%) compared to the baseline method, as the results suggested.

5.4.2.2 Object recognition experiments

Secondly, the proposed method is validated on a vehicle classification problem. To achieve this goal, ResNet-18 [165] is chosen as the teacher and student models. In these experiments, distinct hint types [1,7,10] are applied in the proposed distillation scheme, to further assess the proposed method, where the number of hints is chosen as 4. SGD [166] is employed as the optimizer. Furthermore, the maximum number of iterations is set as 300 epochs, the batch size is chosen as 64 and the learning rate is set as 0.1. The obtained results for the vehicle classification problem on the BIT-Vehicle dataset are presented in Table 5.2.

The proposed approach boosts the student's performance regardless of the hint type to be transferred, as the results suggested. Moreover, the proposed method yields an increase in accuracy up to 0.40 (%) compared to the student model trained from scratch.

Table 5.2 : Accuracy (%) scores of teacher and student models for vehicle classification task on BIT-Vehicle dataset. Vanilla stands for training from scratch. Numerous hint types, which are FT [10], ATT [1] and CC [7], are applied in our proposed distillation scheme. It should be highlighted that teachers' and students' performances are computed on compressed and pixel domains, respectively. Each student's score is calculated over three runs.

	Teacher	**Vanilla**	**Ours**		
		Student	ATT	CC	FT
Accuracy	91.68	95.45	95.62	95.76	**95.85**

5.4.3 Ablation study

Besides these experiments on cross distillation, experiments on classical distillation and reverse distillation, are conducted, where the first refers to employing teacher and student models performing on the same domain, and the latter refers to employing student performing on compressed domain and teacher performing on pixel domain. The results of classical and reverse distillation are presented in Table 5.3, which are computed on the license plate detection task.

Student models that are performing on pixel domain, are better in detection than models that are performing on compressed domain, as the results suggested, where it is an expected outcome. Furthermore, from a student's perspective, utilizing a teacher in a different domain is always better than that in the same domain. This proves that a cross distillation scheme performs better than a classical distillation scheme. Moreover, the proposed cross distillation scheme outperforms the reverse distillation,

Table 5.3 : AP50 (%) scores of the student models on the license plate detection task, for the distillation cases in the perspective of the domains that student and teacher models are performing on. DarkNet and MobileNetv2 are used as the backbones of YOLOv3 models for teacher and student models, respectively. Each score is computed over three runs.

		Teacher	
		Pixel	HEVC
Student	Pixel	98.63	**99.03**
	HEVC	98.17	97.97

which highlights the significance of determining convenient domains for the teacher and student models.

5.5 Conclusion

In this work, a cross distillation method is proposed to transfer knowledge between different domains, to eliminate challenges of scarce or not fully available data in the target domain, due to privacy or copyright concerns. Our method is validated through experiments on two computer vision tasks, that are object detection and recognition. The obtained results point out that the proposed cross distillation scheme between different modalities boosts the performance of a uni-modal computer vision task. In conclusion, our study is a step in the related literature that demonstrates the potential of utilizing compressed domain knowledge in pixel domain tasks. Further, our work contributes to the literature on data protection and minimization, by expanding limited knowledge via exploiting different modality data instead of more samples. Moreover, employing compressed domain images helps to avoid the use of more data that may be personal or sensitive.

6. CONCLUSIONS

In this thesis, knowledge distillation is explained in detail, which is a powerful tool for training models, and related literature is reviewed. Further, studies conducted on numerous problems of knowledge distillation, including hint position selection, model stealing, and cross distillation, are presented. For the problem of hint position selection, a novel hint point selection approach is introduced, which employs clustering of the teacher model's layers. It is applicable for any student architecture and hint distillation scheme that uses two hints at least. Results suggest that the proposed method outperforms the conventional approach, which only considers the spatial size of the feature maps. Possible future work comprises the following:

- Evaluating the proposed approach on different tasks such as object detection on the COCO dataset.
- Integration of the proposed approach into different distillation schemes including online distillation and self-distillation.
- Enhancing clustering step by employing different metrics and layer representation styles including optimized tensor decomposition approaches.

For the model stealing problem, an approach called *averager student* is proposed to distill an undistillable teacher, whose aim is to protect itself from student networks. It is a simple yet effective method that mainly exploits averaging operation on predictions of the teacher network. Results demonstrate that the proposed method outperforms the compared approaches. Possible future work for the model stealing problem includes the following:

- Analyzing the undistillable teacher's predictions by paying attention to the training process of the undistillable teacher.

- Improving the proposed approach by employing a sample-wise process after the averaging operation.
- Enhancing the proposed method by analyzing the outlier probability distributions of the teacher.

For the cross distillation problem, a novel scheme is introduced where a teacher network trained on the compressed domain, is employed to train the student network which operates on the pixel domain. The proposed approach comprises a hint distillation scheme that utilizes our previously proposed hint point selection method. Results indicate that the proposed cross distillation framework between different modalities can help to boost the performance of a model on a uni-modal computer vision problem. Possible future work comprises the following:

- Exploiting other compressed domain attributes in the proposed cross distillation scheme.
- Extension of the proposed method by employing other knowledge types including the response-based knowledge, in the cross distillation scheme.
- Analyzing the effects of the hyperparameters of the proposed scheme including hint types and hint positions, on the performance of distillation.

The introduced methodologies in this thesis contribute to studies in image analysis, model compression, data protection, and minimization. First, our research on selecting efficient hint positions aims to enhance model compression performance, though the proposed method can also be applied to other distillation schemes. We present the benefits of our method for model compression and the performance results of the proposed algorithm. Additionally, our work on model stealing aims to advance the literature on model intellectual property (IP) and data protection by introducing an algorithm to distill knowledge from a protected model. Furthermore, our study on cross-distillation contributes to data protection and minimization by proposing a distillation methodology that exploits compressed domain knowledge for pixel domain problems. This work introduces a technique for expanding limited knowledge using

different modality data rather than more samples. By using compressed domain images and reducing the need for additional samples to boost performance, we avoid the extra usage of pixel domain data which is potentially personal or sensitive.

To conclude, knowledge distillation is a crucial methodology for training models. This thesis explores various problems that knowledge distillation can address, demonstrating its effectiveness across different problems, datasets, architectures, and purposes. The studies conducted point out that knowledge distillation is a versatile tool that serves numerous application areas and tasks thanks to being adaptable enough to be constructed as diverse frameworks.

REFERENCES

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

[2] **Shao, B. and Chen, Y.** (2021). Multi-granularity for knowledge distillation, *Image and Vision Computing*, *115*, 104286.

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

[4] **Yim, J.**, **Joo, D.**, **Bae, J. and Kim, J.** (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4133–4141.

[5] **Hinton, G.**, **Vinyals, O. and Dean, J.** (2015). Distilling the knowledge in a neural network, *arXiv preprint arXiv:1503.02531*.

[6] **Tung, F. and Mori, G.** (2019). Similarity-preserving knowledge distillation, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.1365–1374.

[7] **Peng, B.**, **Jin, X.**, **Liu, J.**, **Li, D.**, **Wu, Y.**, **Liu, Y.**, **Zhou, S. and Zhang, Z.** (2019). Correlation congruence for knowledge distillation, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.5007–5016.

[8] **Lan, X.**, **Zhu, X. and Gong, S.** (2018). Knowledge distillation by on-the-fly native ensemble, *Advances in neural information processing systems*, *31*.

[9] **Kundu, S.**, **Sun, Q.**, **Fu, Y.**, **Pedram, M. and Beerel, P.** (2021). Analyzing the Confidentiality of Undistillable Teachers in Knowledge Distillation, *Advances in Neural Information Processing Systems*, *34*.

[10] **Kim, J.**, **Park, S. and Kwak, N.** (2018). Paraphrasing complex network: Network compression via factor transfer, *Advances in neural information processing systems*, *31*.

[11] **Gou, J.**, **Yu, B.**, **Maybank, S.J. and Tao, D.** (2021). Knowledge distillation: A survey, *International Journal of Computer Vision*, *129*(6), 1789–1819.

[12] **Iqbal, H.** (2018). *HarisIqbal88/PlotNeuralNet v1.0.0*, `https://doi.org/10.5281/zenodo.2526396`.

[13] **Beratoglu, M.S. and Toreyin, B.U.** (2020). *CD-LP: Compressed Domain License Plate Detection Database*, https://dx.doi.org/10.21227/2511-ym76.

[14] **Dong, Z.**, **Wu, Y.**, **Pei, M. and Jia, Y.** (2015). Vehicle type classification using a semisupervised convolutional neural network, *IEEE transactions on intelligent transportation systems*, *16*(4), 2247–2256.

[15] **Beratoğlu, M.S. and Töreyin, B.U.** (2023). Vehicle Detection and Classification without Residual Calculation: Accelerating HEVC Image Decoding with Random Perturbation Injection, *arXiv preprint arXiv:2305.08265*.

[16] **Buciluǎ, C.**, **Caruana, R. and Niculescu-Mizil, A.** (2006). Model compression, *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.535–541.

[17] **Schmidhuber, J.** (1992). Learning complex, extended sequences using the principle of history compression, *Neural computation*, *4*(2), 234–242.

[18] **Keser, R.K.**, **Ayanzadeh, A.**, **Aghdam, O.A.**, **Kilcioglu, C.**, **Toreyin, B.U. and Ure, N.K.** (2023). PURSUhInT: In Search of Informative Hint Points Based on Layer Clustering for Knowledge Distillation, *Expert Systems with Applications*, *213*, 119040.

[19] **Ma, H.**, **Chen, T.**, **Hu, T.K.**, **You, C.**, **Xie, X. and Wang, Z.** (2021). Undistillable: Making a nasty teacher that cannot teach students, *arXiv preprint arXiv:2105.07381*.

[20] **Keser, R.K. and Töreyin, B.U.** (2023). Averager Student: Distillation from Undistillable Teacher.

[21] **Keser, R.K.**, **Beratoğlu, M.S. and Töreyin, B.U.** (2024). Generated Compressed Domain Images to the Rescue: Cross Distillation from Compressed Domain to Pixel Domain, *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, pp.1–5.

[22] **Heo, B.**, **Kim, J.**, **Yun, S.**, **Park, H.**, **Kwak, N. and Choi, J.Y.** (2019). A comprehensive overhaul of feature distillation, *Proceedings of the IEEE International Conference on Computer Vision*, pp.1921–1930.

[23] **Chen, G.**, **Choi, W.**, **Yu, X.**, **Han, T. and Chandraker, M.** (2017). Learning efficient object detection models with knowledge distillation, *Advances in neural information processing systems*, *30*.

[24] **Zhang, F.**, **Zhu, X. and Ye, M.** (2019). Fast human pose estimation, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp.3517–3526.

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

[26] **Mirzadeh, S.I.**, **Farajtabar, M.**, **Li, A.**, **Levine, N.**, **Matsukawa, A. and Ghasemzadeh, H.** (2020). Improved knowledge distillation via teacher assistant, *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp.5191–5198.

[27] **Niu, Y.**, **Chen, L.**, **Zhou, C. and Zhang, H.** (2022). Respecting transfer gap in knowledge distillation, *Advances in Neural Information Processing Systems*, *35*, 21933–21947.

[28] **Romero, A.**, **Ballas, N.**, **Kahou, S.E.**, **Chassang, A.**, **Gatta, C. and Bengio, Y.** (2015). Fitnets: Hints for thin deep nets, *Proc. ICLR*, *2*.

[29] **Heo, B.**, **Lee, M.**, **Yun, S. and Choi, J.Y.** (2019). Knowledge transfer via distillation of activation boundaries formed by hidden neurons, *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp.3779–3787.

[30] **Li, S.**, **Lin, M.**, **Wang, Y.**, **Wu, Y.**, **Tian, Y.**, **Shao, L. and Ji, R.** (2022). Distilling a powerful student model via online knowledge distillation, *IEEE Transactions on Neural Networks and Learning Systems*.

[31] **Huang, Z.**, **Yang, S.**, **Zhou, M.C.**, **Li, Z.**, **Gong, Z. and Chen, Y.** (2022). Feature Map Distillation of Thin Nets for Low-resolution Object Recognition, *IEEE Transactions on Image Processing*.

[32] **Zhang, L.**, **Bao, C. and Ma, K.** (2021). Self-distillation: Towards efficient and compact neural networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(8), 4388–4403.

[33] **Park, W.**, **Kim, D.**, **Lu, Y. and Cho, M.** (2019). Relational knowledge distillation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.3967–3976.

[34] **Lee, S.H.**, **Kim, D.H. and Song, B.C.** (2018). Self-supervised knowledge distillation using singular value decomposition, *Proceedings of the European conference on computer vision (ECCV)*, pp.335–350.

[35] **Liu, Y.**, **Cao, J.**, **Li, B.**, **Yuan, C.**, **Hu, W.**, **Li, Y. and Duan, Y.** (2019). Knowledge distillation via instance relationship graph, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7096–7104.

[36] **Passalis, N. and Tefas, A.** (2018). Learning deep representations with probabilistic knowledge transfer, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.268–284.

[37] **Chung, I.**, **Park, S.**, **Kim, J. and Kwak, N.** (2020). Feature-map-level online adversarial knowledge distillation, *International Conference on Machine Learning*, PMLR, pp.2006–2015.

[38] **Zhang, Y.**, **Xiang, T.**, **Hospedales, T.M. and Lu, H.** (2018). Deep mutual learning, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.4320–4328.

[39] **Chen, D.**, **Mei, J.P.**, **Wang, C.**, **Feng, Y. and Chen, C.** (2020). Online knowledge distillation with diverse peers, *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp.3430–3437.

[40] **Zhang, L.**, **Song, J.**, **Gao, A.**, **Chen, J.**, **Bao, C. and Ma, K.** (2019). Be your own teacher: Improve the performance of convolutional neural networks via self distillation, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.3713–3722.

[41] **Hou, Y.**, **Ma, Z.**, **Liu, C. and Loy, C.C.** (2019). Learning lightweight lane detection cnns by self attention distillation, *Proceedings of the IEEE/CVF international conference on computer vision*, pp.1013–1021.

[42] **Yang, C.**, **Xie, L.**, **Su, C. and Yuille, A.L.** (2019). Snapshot distillation: Teacher-student optimization in one generation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.2859–2868.

[43] **Wang, X.**, **Zhang, R.**, **Sun, Y. and Qi, J.** (2018). Kdgan: Knowledge distillation with generative adversarial networks, *Advances in neural information processing systems*, *31*.

[44] **Micaelli, P. and Storkey, A.J.** (2019). Zero-shot knowledge transfer via adversarial belief matching, *Advances in Neural Information Processing Systems*, *32*.

[45] **Belagiannis, V.**, **Farshad, A. and Galasso, F.** (2018). Adversarial network compression, *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp.0–0.

[46] **Chen, X.**, **Su, J. and Zhang, J.** (2019). A two-teacher framework for knowledge distillation, *Advances in Neural Networks–ISNN 2019: 16th International Symposium on Neural Networks, ISNN 2019, Moscow, Russia, July 10–12, 2019, Proceedings, Part I 16*, Springer, pp.58–66.

[47] **Papernot, N.**, **Abadi, M.**, **Erlingsson, U.**, **Goodfellow, I. and Talwar, K.** (2016). Semi-supervised knowledge transfer for deep learning from private training data, *arXiv preprint arXiv:1610.05755*.

[48] **Park, S. and Kwak, N.**, (2020). Feature-level ensemble knowledge distillation for aggregating knowledge from multiple networks, ECAI 2020, IOS Press, pp.1411–1418.

[49] **Gupta, S.**, **Hoffman, J. and Malik, J.** (2016). Cross modal distillation for supervision transfer, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2827–2836.

[50] **Zhao, L.**, **Peng, X.**, **Chen, Y.**, **Kapadia, M. and Metaxas, D.N.** (2020). Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.6528–6537.

[51] **Garcia, N.C.**, **Morerio, P. and Murino, V.** (2018). Modality distillation with multiple stream networks for action recognition, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.103–118.

[52] **Zhang, C. and Peng, Y.** (2018). Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification, *arXiv preprint arXiv:1804.10069*.

[53] **Lee, S. and Song, B.C.** (2019). Graph-based knowledge distillation by multi-head attention network, *arXiv preprint arXiv:1907.02226*.

[54] **Srinivas, S. and Fleuret, F.** (2018). Knowledge transfer with jacobian matching, *International Conference on Machine Learning*, PMLR, pp.4723–4731.

[55] **Huang, Z. and Wang, N.** (2017). Like what you like: Knowledge distill via neuron selectivity transfer, *arXiv preprint arXiv:1707.01219*.

[56] **Chen, H.**, **Wang, Y.**, **Xu, C.**, **Yang, Z.**, **Liu, C.**, **Shi, B.**, **Xu, C.**, **Xu, C. and Tian, Q.** (2019). Data-free learning of student networks, *Proceedings of the IEEE/CVF international conference on computer vision*, pp.3514–3522.

[57] **Lopes, R.G.**, **Fenu, S. and Starner, T.** (2017). Data-free knowledge distillation for deep neural networks, *arXiv preprint arXiv:1710.07535*.

[58] **Wei, Y.**, **Pan, X.**, **Qin, H.**, **Ouyang, W. and Yan, J.** (2018). Quantization mimic: Towards very tiny cnn for object detection, *Proceedings of the European conference on computer vision (ECCV)*, pp.267–283.

[59] **Polino, A.**, **Pascanu, R. and Alistarh, D.** (2018). Model compression via distillation and quantization, *arXiv preprint arXiv:1802.05668*.

[60] **Jang, Y.**, **Lee, H.**, **Hwang, S.J. and Shin, J.** (2019). Learning what and where to transfer, *International Conference on Machine Learning*, PMLR, pp.3030–3039.

[61] **Li, Z. and Hoiem, D.** (2017). Learning without forgetting, *IEEE transactions on pattern analysis and machine intelligence*, *40*(12), 2935–2947.

[62] **Kang, M.**, **Mun, J. and Han, B.** (2020). Towards oracle knowledge distillation with neural architecture search, *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp.4404–4411.

[63] **Macko, V.**, **Weill, C.**, **Mazzawi, H. and Gonzalvo, J.** (2019). Improving neural architecture search image classifiers via ensemble learning, *arXiv preprint arXiv:1903.06236*.

[64] **He, T.**, **Shen, C.**, **Tian, Z.**, **Gong, D.**, **Sun, C. and Yan, Y.** (2019). Knowledge adaptation for efficient semantic segmentation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.578–587.

[65] **Mun, J.**, **Lee, K.**, **Shin, J. and Han, B.** (2018). Learning to specialize with knowledge distillation for visual question answering, *Advances in neural information processing systems*, *31*.

[66] **Bergmann, P.**, **Fauser, M.**, **Sattlegger, D. and Steger, C.** (2020). Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp.4183–4192.

[67] **Chen, Y.C.**, **Gan, Z.**, **Cheng, Y.**, **Liu, J. and Liu, J.** (2019). Distilling knowledge learned in BERT for text generation, *arXiv preprint arXiv:1911.03829*.

[68] **Shakeri, S.**, **Sethy, A. and Cheng, C.** (2019). Knowledge distillation in document retrieval, *arXiv preprint arXiv:1911.11065*.

[69] **Wang, Z.R. and Du, J.** (2021). Joint architecture and knowledge distillation in CNN for Chinese text recognition, *Pattern Recognition*, *111*, 107722.

[70] **Hahn, S. and Choi, H.** (2019). Self-knowledge distillation in natural language processing, *arXiv preprint arXiv:1908.01851*.

[71] **Gao, L.**, **Mi, H.**, **Zhu, B.**, **Feng, D.**, **Li, Y. and Peng, Y.** (2019). An adversarial feature distillation method for audio classification, *IEEE Access*, *7*, 105319–105330.

[72] **Watanabe, S.**, **Hori, T.**, **Le Roux, J. and Hershey, J.R.** (2017). Student-teacher network learning with enhanced features, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.5275–5279.

[73] **Shen, P.**, **Lu, X.**, **Li, S. and Kawai, H.** (2018). Feature Representation of Short Utterances Based on Knowledge Distillation for Spoken Language Identification., *Interspeech*, pp.1813–1817.

[74] **Price, R.**, **Iso, K.i. and Shinoda, K.** (2016). Wise teachers train better DNN acoustic models, *EURASIP Journal on Audio, Speech, and Music Processing*, *2016*, 1–19.

[75] **Oord, A.**, **Li, Y.**, **Babuschkin, I.**, **Simonyan, K.**, **Vinyals, O.**, **Kavukcuoglu, K.**, **Driessche, G.**, **Lockhart, E.**, **Cobo, L.**, **Stimberg, F.** *et al.* (2018). Parallel wavenet: Fast high-fidelity speech synthesis, *International conference on machine learning*, PMLR, pp.3918–3926.

[76] **Chen, X.**, **Zhang, Y.**, **Xu, H.**, **Qin, Z. and Zha, H.** (2018). Adversarial distillation for efficient recommendation with external knowledge, *ACM Transactions on Information Systems (TOIS)*, *37*(1), 1–28.

[77] **Papernot, N.**, **McDaniel, P.**, **Wu, X.**, **Jha, S. and Swami, A.** (2016). Distillation as a defense to adversarial perturbations against deep neural networks, *2016 IEEE symposium on security and privacy (SP)*, IEEE, pp.582–597.

[78] **Wang, J.**, **Bao, W.**, **Sun, L.**, **Zhu, X.**, **Cao, B. and Philip, S.Y.** (2019). Private model compression via knowledge distillation, *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp.1190–1197.

[79] **Liu, X.**, **Wang, X. and Matwin, S.** (2018). Improving the interpretability of deep neural networks with knowledge distillation, *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, IEEE, pp.905–912.

[80] **He, C.**, **Annavaram, M. and Avestimehr, S.** (2020). Group knowledge transfer: Federated learning of large cnns at the edge, *Advances in Neural Information Processing Systems*, *33*, 14068–14080.

[81] **Hao, Z.**, **Guo, J.**, **Han, K.**, **Tang, Y.**, **Hu, H.**, **Wang, Y. and Xu, C.** (2024). One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation, *Advances in Neural Information Processing Systems*, *36*.

[82] **Ye, X.**, **Jiang, R.**, **Tian, X.**, **Zhang, R. and Chen, Y.** (2024). Knowledge Distillation via Multi-Teacher Feature Ensemble, *IEEE Signal Processing Letters*.

[83] **Krizhevsky, A.**, **Sutskever, I. and Hinton, G.E.** (2017). Imagenet classification with deep convolutional neural networks, *Communications of the ACM*, *60*(6), 84–90.

[84] **He, K.**, **Zhang, X.**, **Ren, S. and Sun, J.** (2016). Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.770–778.

[85] **Redmon, J. and Farhadi, A.** (2018). Yolov3: An incremental improvement, *arXiv preprint arXiv:1804.02767*.

[86] **Long, J.**, **Shelhamer, E. and Darrell, T.** (2015). Fully convolutional networks for semantic segmentation, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.3431–3440.

[87] **Parkhi, O.M.**, **Vedaldi, A. and Zisserman, A.** (2015). Deep face recognition.

[88] **Ciaparrone, G.**, **Sánchez, F.L.**, **Tabik, S.**, **Troiano, L.**, **Tagliaferri, R. and Herrera, F.** (2020). Deep learning in video multi-object tracking: A survey, *Neurocomputing*, *381*, 61–88.

[89] **Sze, V.**, **Chen, Y.H.**, **Yang, T.J. and Emer, J.S.** (2017). Efficient processing of deep neural networks: A tutorial and survey, *Proceedings of the IEEE*, *105*(12), 2295–2329.

[90] **Russakovsky, O.**, **Deng, J.**, **Su, H.**, **Krause, J.**, **Satheesh, S.**, **Ma, S.**, **Huang, Z.**, **Karpathy, A.**, **Khosla, A.**, **Bernstein, M.**, **Berg, A.C. and Fei-Fei, L.** (2015). ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision (IJCV)*, *115*(3), 211–252.

[91] **Simonyan, K. and Zisserman, A.** (2014). Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*.

[92] **Choudhary, T.**, **Mishra, V.**, **Goswami, A. and Sarangapani, J.** (2020). A comprehensive survey on model compression and acceleration, *Artificial Intelligence Review*, 1–43.

[93] **Guo, J.**, **Chen, M.**, **Hu, Y.**, **Zhu, C.**, **He, X. and Cai, D.** (2020). Spherical Knowledge Distillation, *arXiv preprint arXiv:2010.07485*.

[94] **Zhao, B.**, **Cui, Q.**, **Song, R.**, **Qiu, Y. and Liang, J.** (2022). Decoupled Knowledge Distillation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.11953–11962.

[95] **Ren, S.**, **Gao, Z.**, **Hua, T.**, **Xue, Z.**, **Tian, Y.**, **He, S. and Zhao, H.** (2022). Co-advise: Cross inductive bias distillation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.16773–16782.

[96] **Kang, M. and Kang, S.** (2021). Data-free knowledge distillation in neural networks for regression, *Expert Systems with Applications*, *175*, 114813.

[97] **Nam, G.**, **Lee, H.**, **Heo, B. and Lee, J.** (2022). Improving Ensemble Distillation With Weight Averaging and Diversifying Perturbation, *International Conference on Machine Learning*, PMLR, pp.16353–16367.

[98] **Ni, X.**, **Shen, X. and Zhao, H.** (2022). Federated optimization via knowledge codistillation, *Expert Systems with Applications*, *191*, 116310.

[99] **Wang, L. and Yoon, K.J.** (2021). Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[100] **Cheng, Y.**, **Wang, D.**, **Zhou, P. and Zhang, T.** (2017). A survey of model compression and acceleration for deep neural networks, *arXiv preprint arXiv:1710.09282*.

[101] **Han, S.**, **Mao, H. and Dally, W.J.** (2016). Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding, *International Conference on Learning Representations (ICLR)*.

[102] **Szegedy, C.**, **Ioffe, S.**, **Vanhoucke, V. and Alemi, A.A.** (2017). Inception-v4, inception-resnet and the impact of residual connections on learning, *Thirty-first AAAI conference on artificial intelligence*.

[103] **Iandola, F.N.**, **Han, S.**, **Moskewicz, M.W.**, **Ashraf, K.**, **Dally, W.J. and Keutzer, K.** (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size, *arXiv:1602.07360*.

[104] **Sandler, M.**, **Howard, A.**, **Zhu, M.**, **Zhmoginov, A. and Chen, L.C.** (2018). Mobilenetv2: Inverted residuals and linear bottlenecks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4510–4520.

[105] **Howard, A.**, **Sandler, M.**, **Chu, G.**, **Chen, L.C.**, **Chen, B.**, **Tan, M.**, **Wang, W.**, **Zhu, Y.**, **Pang, R.**, **Vasudevan, V.** *et al.* (2019). Searching for mobilenetv3, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.1314–1324.

[106] **Ahn, S.**, **Hu, S.X.**, **Damianou, A.**, **Lawrence, N.D. and Dai, Z.** (2019). Variational information distillation for knowledge transfer, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.9163–9171.

[107] **Ji, M.**, **Heo, B. and Park, S.** (2021). Show, Attend and Distill: Knowledge Distillation via Attention-based Feature Matching.

[108] **Zhang, K.**, **Zhanga, C.**, **Li, S.**, **Zeng, D. and Ge, S.** (2021). Student network learning via evolutionary knowledge distillation, *IEEE Transactions on Circuits and Systems for Video Technology*.

[109] **Wu, A. and Deng, C.** (2022). Single-Domain Generalized Object Detection in Urban Scene via Cyclic-Disentangled Self-Distillation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.847–856.

[110] **Fu, S.**, **Li, Z.**, **Liu, Z. and Yang, X.** (2021). Interactive knowledge distillation for image classification, *Neurocomputing*, *449*, 411–421.

[111] **Ruffy, F. and Chahal, K.** (2019). The State of Knowledge Distillation for Classification, *arXiv preprint arXiv:1912.10850*.

[112] **Chen, P.**, **Liu, S.**, **Zhao, H. and Jia, J.** (2021). Distilling knowledge via knowledge review, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.5008–5017.

[113] **Sun, S.**, **Cheng, Y.**, **Gan, Z. and Liu, J.** (2019). Patient Knowledge Distillation for BERT Model Compression, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp.4323–4332.

[114] **Deng, X.**, **Sun, D.**, **Newsam, S. and Wang, P.** (2022). DistPro: Searching A Fast Knowledge Distillation Process via Meta Optimization, *arXiv preprint arXiv:2204.05547*.

[115] **Haidar, M.A.**, **Anchuri, N.**, **Rezagholizadeh, M.**, **Ghaddar, A.**, **Langlais, P. and Poupart, P.** (2021). RAIL-KD: RAndom Intermediate Layer Mapping for Knowledge Distillation, *arXiv preprint arXiv:2109.10164*.

[116] **Li, C.**, **Lin, M.**, **Ding, Z.**, **Lin, N.**, **Zhuang, Y.**, **Huang, Y.**, **Ding, X. and Cao, L.** (2022). Knowledge Condensation Distillation, *arXiv preprint arXiv:2207.05409*.

[117] **Passban, P.**, **Wu, Y.**, **Rezagholizadeh, M. and Liu, Q.** (2021). Alp-kd: Attention-based layer projection for knowledge distillation, *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp.13657–13665.

[118] **Jain, A.K.** (2010). Data clustering: 50 years beyond K-means, *Pattern recognition letters*, *31*(8), 651–666.

[119] **Kornblith, S.**, **Norouzi, M.**, **Lee, H. and Hinton, G.** (2019). Similarity of neural network representations revisited, *International Conference on Machine Learning*, PMLR, pp.3519–3529.

[120] **Raghu, M.**, **Gilmer, J.**, **Yosinski, J. and Sohl-Dickstein, J.** (2017). Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability, *Advances in neural information processing systems*, pp.6076–6085.

[121] **Morcos, A.**, **Raghu, M. and Bengio, S.** (2018). Insights on representational similarity in neural networks with canonical correlation, *Advances in Neural Information Processing Systems*, *31*, 5727–5736.

[122] **Neill, J.O.**, **Steeg, G.V. and Galstyan, A.** (2020). Compressing Deep Neural Networks via Layer Fusion, *arXiv preprint arXiv:2007.14917*.

[123] **Li, Y.**, **Yosinski, J.**, **Clune, J.**, **Lipson, H. and Hopcroft, J.E.** (2015). Convergent learning: Do different neural networks learn the same representations?, *FE@ NIPS*, pp.196–212.

[124] **Ramsay, J.**, **ten Berge, J. and Styan, G.** (1984). Matrix correlation, *Psychometrika*, *49*(3), 403–423.

[125] **Yang, D.**, **Zhou, Y.**, **Zhang, A.**, **Sun, X.**, **Wu, D.**, **Wang, W. and Ye, Q.** (2022). Multi-view correlation distillation for incremental object detection, *Pattern Recognition*, *131*, 108863.

[126] **Yosinski, J.**, **Clune, J.**, **Bengio, Y. and Lipson, H.** (2014). How transferable are features in deep neural networks?, *Advances in neural information processing systems*, *27*.

[127] **Zagoruyko, S. and Komodakis, N.** (2016). Wide Residual Networks, *British Machine Vision Conference 2016*, British Machine Vision Association.

[128] **Zhang, X.**, **Zhou, X.**, **Lin, M. and Sun, J.** (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.6848–6856.

[129] **Biewald, L.** (2020). *Experiment Tracking with Weights and Biases*, `https://www.wandb.com/`, software available from wandb.com.

[130] **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, *arXiv preprint arXiv:2102.00650*.

[131] **Liang, J.**, **Pang, R.**, **Li, C. and Wang, T.** (2024). Model extraction attacks revisited, *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*, pp.1231–1245.

[132] **Beetham, J.**, **Kardan, N.**, **Mian, A.S. and Shah, M.** Dual Student Networks for Data-Free Model Stealing, *The Eleventh International Conference on Learning Representations*.

[133] **Kariyappa, S. and Qureshi, M.K.** (2020). Defending against model stealing attacks with adaptive misinformation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.770–778.

[134] **Tramèr, F.**, **Zhang, F.**, **Juels, A.**, **Reiter, M.K. and Ristenpart, T.** (2016). Stealing Machine Learning Models via Prediction {APIs}, *25th USENIX security symposium (USENIX Security 16)*, pp.601–618.

[135] **Chandrasekaran, V.**, **Chaudhuri, K.**, **Giacomelli, I.**, **Jha, S. and Yan, S.** (2020). Exploring connections between active learning and model extraction, *29th USENIX Security Symposium (USENIX Security 20)*, pp.1309–1326.

[136] **Ding, K.**, **Nouri, E.**, **Zheng, G.**, **Liu, H. and White, R.** (2024). Toward robust graph semi-supervised learning against extreme data scarcity, *IEEE Transactions on Neural Networks and Learning Systems*.

[137] **Chen, Z.**, **Gan, W.**, **Wu, J.**, **Hu, K. and Lin, H.** (2024). Data Scarcity in Recommendation Systems: A Survey, *ACM Transactions on Recommender Systems*.

[138] **Goldsteen, A.**, **Ezov, G.**, **Shmelkin, R.**, **Moffie, M. and Farkash, A.** (2022). Data minimization for GDPR compliance in machine learning models, *AI and Ethics*, *2*(3), 477–491.

[139] *The California Privacy Rights Act (CPRA)*, `https://www.caprivacy.org/annotated-cpra-text-with-ccpa-changes/`, accessed: 2024-07-12.

[140] *The European General Data Protection Regulation (GDPR)*, https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en, accessed: 2024-07-12.

[141] **Jagielski, M.**, **Nasr, M.**, **Lee, K.**, **Choquette-Choo, C.A.**, **Carlini, N. and Tramer, F.** (2024). Students parrot their teachers: Membership inference on model distillation, *Advances in Neural Information Processing Systems*, *36*.

[142] **Chen, L.**, **Sun, H.**, **Katto, J.**, **Zeng, X. and Fan, Y.** (2021). Fast object detection in hevc intra compressed domain, *2021 29th European Signal Processing Conference (EUSIPCO)*, IEEE, pp.756–760.

[143] **Jaballah, S. and Larabi, M.C.** (2019). Fast object detection in H264/AVC and HEVC compressed domains for video surveillance, *2019 8th European Workshop on Visual Information Processing (EUVIP)*, IEEE, pp.123–128.

[144] **Beratoğlu, M.S. and Töreyin, B.U.** (2021). Vehicle license plate detector in compressed domain, *IEEE Access*, *9*, 95087–95096.

[145] **Passalis, N.**, **Tzelepi, M. and Tefas, A.** (2020). Multilayer probabilistic knowledge transfer for learning image representations, *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, pp.1–5.

[146] **Thoker, F.M. and Gall, J.** (2019). Cross-modal knowledge distillation for action recognition, *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, pp.6–10.

[147] **Zhao, M.**, **Li, T.**, **Abu Alsheikh, M.**, **Tian, Y.**, **Zhao, H.**, **Torralba, A. and Katabi, D.** (2018). Through-wall human pose estimation using radio signals, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.7356–7365.

[148] **Albanie, S.**, **Nagrani, A.**, **Vedaldi, A. and Zisserman, A.** (2018). Emotion recognition in speech using cross-modal transfer in the wild, *Proceedings of the 26th ACM International Conference on Multimedia*, pp.292–301.

[149] **Xu, C.**, **Wu, X.**, **Li, Y.**, **Jin, Y.**, **Wang, M. and Liu, Y.** (2021). Cross-modality online distillation for multi-view action recognition, *Neurocomputing*, *456*, 384–393.

[150] **Zhang, B.**, **Wang, L.**, **Wang, Z.**, **Qiao, Y. and Wang, H.** (2016). Real-time action recognition with enhanced motion vector CNNs, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2718–2726.

[151] **Lu, H. and Deng, Z.** (2021). A Boundary-aware Distillation Network for Compressed Video Semantic Segmentation, *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, pp.5354–5359.

[152] **Li, H.**, **Jiang, X.**, **Guan, B.**, **Tan, R.R.M.**, **Wang, R. and Thalmann, N.M.** (2021). Joint feature optimization and fusion for compressed action recognition, *IEEE Transactions on Image Processing*, *30*, 7926–7937.

[153] **Zhang, J.**, **Guo, J. and Ma, M.** (2022). Multi-Knowledge Attention Transfer Framework for Action Recognition, *Artificial Neural Networks and Machine Learning–ICANN 2022: 31st International Conference on Artificial Neural Networks, Bristol, UK, September 6–9, 2022, Proceedings, Part I*, Springer, pp.231–242.

[154] **Han, H.**, **Yang, G.**, **Huo, Y.**, **Lu, Z. and Wen, J.R.** (2021). Complex Action Segmentation in Compressed Videos, *2021 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, pp.1–6.

[155] **Huo, Y.**, **Xu, X.**, **Lu, Y.**, **Niu, Y.**, **Ding, M.**, **Lu, Z.**, **Xiang, T. and Wen, J.r.** (2020). Lightweight action recognition in compressed videos, *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, Springer, pp.337–352.

[156] **Huang, S.**, **Lin, X.**, **Karaman, S. and Chang, S.F.** (2019). Flow-distilled IP two-stream networks for compressed video action recognition, *arXiv preprint arXiv:1912.04462*.

[157] **Battash, B.**, **Barad, H.**, **Tang, H. and Bleiweiss, A.** (2020). Mimic the raw domain: Accelerating action recognition in the compressed domain, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, pp.684–685.

[158] **Tu, H.**, **Li, L.**, **Zhou, W. and Li, H.** (2022). Learning in Compressed Domain via Knowledge Transfer.

[159] **Wu, M.C.**, **Chiu, C.T. and Wu, K.H.** (2019). Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks, *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.2202–2206.

[160] **Wu, C.Y.**, **Zaheer, M.**, **Hu, H.**, **Manmatha, R.**, **Smola, A.J. and Krähenbühl, P.** (2018). Compressed video action recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.6026–6035.

[161] **Xiong, L.**, **Jia, X.**, **Ming, Y.**, **Zhou, J.**, **Feng, F. and Hu, N.** (2021). Faster-FCoViAR: Faster Frequency-Domain Compressed Video Action Recognition.

[162] **Contributors, M.** (2021). *OpenMMLab Model Compression Toolbox and Benchmark*, `https://github.com/open-mmlab/mmrazor`.

[163] **Redmon, J. and Farhadi, A.** (2018). Yolov3: An incremental improvement, *arXiv preprint arXiv:1804.02767*.

[164] **Loshchilov, I. and Hutter, F.** (2017). Decoupled weight decay regularization, *arXiv preprint arXiv:1711.05101*.

[165] **He, K.**, **Zhang, X.**, **Ren, S. and Sun, J.** (2016). Deep residual learning for image recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770–778.

[166] **Sutskever, I.**, **Martens, J.**, **Dahl, G. and Hinton, G.** (2013). On the importance of initialization and momentum in deep learning, *International Conference on Machine Learning*, PMLR, pp.1139–1147.

CURRICULUM VITAE

Name SURNAME: Reyhan Kevser KESER

EDUCATION:

- **B.Sc.:** 2016, Istanbul Technical University, Faculty of Electrical and Electronics Engineering, Department of Electronics and Communication Engineering
- **M.Sc.:** 2018, Istanbul Technical University, Informatics Institute, Applied Informatics Programme

PROFESSIONAL EXPERIENCE AND PROJECTS PARTICIPATED:

- 2017 - ...: Research and Teaching Assistant at Istanbul Technical University
- 2023 – 2023: (Researcher) Developing Algorithms for Detection of Lip-Speech Synchronization Error in Live Channel Broadcasts using Computer Vision and Machine Learning Techniques - TURKCELL
- 2022 – 2023: (Researcher) Literature Review and Developing Detection Approaches for Raman Spectroscopy - ASELSAN
- 2022 – 2022: (Researcher) Cigarette Detection in Offline Video - TURKCELL
- 2021 - 2023: (Researcher) Video Understanding in Compressed Domain – TUBITAK 1001 Project
- 2020 – 2021: (Researcher) Developing Deep Learning Based Methods for Object Detection in Video - ASELSAN
- 2018 - 2021: (Researcher) Neural Embeddings for Social Context in Telecommunications – ITU Vodafone Future Lab
- 2017 – 2019: (Researcher) Data Acquisition, Compression, Analysis and Fusion Methods for New Generation Surveillance Systems – ITU Research Project

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Keser, R. K.**, Beratoğlu M. S., & Töreyin B., U. (2024). Generated Compressed Domain Images to the Rescue: Cross Distillation from Compressed Domain to Pixel Domain. *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE.

- **Keser, R. K.**, & Töreyin, B. U. (2023). Averager Student: Distillation from Undistillable Teacher. *Tiny Papers @ ICLR 2023*.

- **Keser, R. K.**, Ayanzadeh, A., Aghdam, O. A., Kilcioglu, C., Toreyin, B. U., & Ure, N. K. (2023). PURSUhInT: In Search of Informative Hint Points Based on Layer Clustering for Knowledge Distillation. *Expert Systems with Applications*, 213, 119040.

OTHER PUBLICATIONS, PRESENTATIONS AND PATENTS:

- Koyun, O. C., **Keser, R. K.**, Şahin, S. O., Bulut, D., Yorulmaz, M., Yücesoy, V., & Töreyin, B. U. (2024). RamanFormer: A Transformer-Based Quantification Approach for Raman Mixture Components. *ACS Omega*.

- Koyun, O. C., **Keser, R. K.**, Akkaya, İ. B., & Töreyin, B. U. (2022). Focus-and-Detect: A small object detection framework for aerial images. *Signal Processing: Image Communication*, 104, 116675.

- Nallbani, I., Ayanzadeh, A., **Keser, R. K.**, Çalık, N., & Töreyin, B. U. (2021). Representation Learning using Graph Autoencoders with Residual Connections. arXiv preprint arXiv:2105.00695.

- **Keser, R. K.**, Nallbani, I., Çalik, N., Ayanzadeh, A., & Töreyin, B. U. (2020, October). Graph Embedding For Link Prediction Using Residual Variational Graph Autoencoders. *2020 28th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

- **Keser, R. K.**, & Töreyin, B. U. (2019, November). Autoencoder Based Dimensionality Reduction of Feature Vectors for Object Recognition. *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)* (pp. 577-584). IEEE.

- **Keser, R. K.**, Ergün, E., & Töreyin, B. U. (2018). Vehicle Logo Recognition with Reduced-Dimension SIFT Vectors Using Autoencoders. *Multidisciplinary Digital Publishing Institute Proceedings* (Vol. 2, No. 2, p. 92).