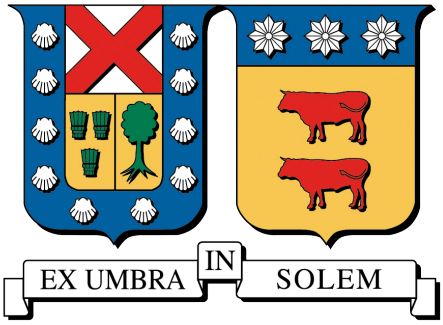
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE ELECTRÓNICA
VALPARAÍSO - CHILE



ADVANCED TRANS-DOMAIN KNOWLEDGE TRANSFER THROUGH TRANSFORMER-BASED DISTILLATION: A NOVEL FRAMEWORK FOR IMAGE-LIDAR INTEGRATION IN AUTONOMOUS SYSTEMS

JESUS EDUARDO ORTIZ SANDOVAL

TESIS PARA OPTAR AL GRADO DE
DOCTOR EN INGENIERÍA ELECTRÓNICA

PROFESOR GUÍA : Dr. Werner Creixell

ENERO 2025

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☐ Memoria o trabajo de título; ☒ Tesis de Postgrado;

Título del trabajo: ADVANCED TRANS-DOMAIN KNOWLEDGE TRANSFER THROUGH TRANSFORMER-BASED DISTILLATION: A NOVEL FRAMEWORK FOR IMAGE-LIDAR INTEGRATION IN AUTONOMOUS SYSTEMS

Nombre del candidato(a): Jesus Eduardo Ortiz Sandoval

Carrera / Grado: Doctorado en Ingenieria Electrónica

Campus: Casa Central Valparaiso **;** **Departamento**: Electrónica

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Werner Creixell, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL

☒ El trabajo **NO contiene información que amerite confidencialidad** y puede ser publicado de inmediato en repositorio con acceso abierto.

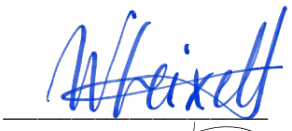
☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (embargo) por:

☐ 6 meses; ☐ 12 meses; ☐ 2 años; ☐ 3 años; ☐ 5 años; ☐ 10 años

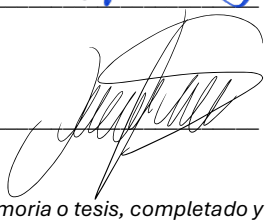
Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 6/10/2025 ; **Firma:** 

Estudiante o Candidato(a):

Fecha: 6/10/2025 ; **Firma:** 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

TÍTULO DE LA TESIS:

ADVANCED TRANS-DOMAIN KNOWLEDGE TRANSFER THROUGH TRANSFORMER-BASED DISTILLATION: A NOVEL FRAMEWORK FOR IMAGE-LIDAR INTEGRATION IN AUTONOMOUS SYSTEMS

AUTOR:

Jesús Eduardo Ortíz Sandoval

TRABAJO DE TESIS, presentado en cumplimiento parcial de los requisitos para el grado de Doctor en Ingeniería Electrónica de la Universidad Técnica Federico Santa María.

Dr. Werner Creixell _____

Dr. Mauricio Araya _____

Dr. Moulay Akhloufi _____

Valparaíso, Enero de 2025.

A mi amada madre, Carmen Sandoval, quien tuvo la fortuna de ser la persona que me enseñó a leer, a escribir, a persistir y a buscar siempre el cumplimiento de mis objetivos sin olvidarme de soñar.

A mi amada Nayli, cuyo apoyo incondicional en los momentos más difíciles me ha ayudado a convertirme en un mejor hombre, un mejor compañero y un mejor profesional.

A mi preciosa hija Luciana, cuya ternura y abrazos recargaban mi corazón cuando estaba completamente exhausto o listo para tirar la toalla.

Un agradecimiento especial a mi tutor, el Dr. Werner Creixell, a quien agradeceré toda la vida por haberme brindado tanto conocimiento, no solo a nivel académico, sino también como persona y ser humano.

Quiero expresar mi más sincero cariño y gratitud a mi compañero de doctorado y de sufrimientos, el Dr. Patricio Olivares, una de las grandes amistades que cultivé durante este proceso, así como a Leonardo Guerrero y Nicolás Olivos del CCT-VAL, tesoros invaluables que encontré en este camino.

También quisiera agradecer al Dr. Jorge Ardila, quien fue un gran apoyo académico y personal, y con quien tuve el privilegio de trabajar en un proyecto extraordinario que se convirtió en la puerta de entrada a mi tesis.

Un saludo especial a mis queridos suegros, Libardo y Anaydu, quienes siempre me han brindado su apoyo incondicional y me han recibido como un hijo más en su familia.

Por último, quiero dedicar unas palabras a mi amado perrito Mateo, su recuerdo vivirá eternamente en mi corazón. Fuiste mi fiel compañero, sé que desde el cielo, tu cola no dejará de moverse al ver que he alcanzado este sueño que una vez compartimos.

ABSTRACT

Recent advances in deep learning have significantly improved the performance of image classification models, yet adapting these models to fundamentally different data types—such as point clouds from Light Detection and Ranging (LiDAR) sensors—remains a challenging task. This thesis addresses that challenge by exploring trans-domain knowledge distillation: transferring capabilities learned from well-established image classification networks to LiDAR point cloud classification. Building on insights gained from earlier research on partial discharge (PD) signal generation using Deep Convolutional Generative Adversarial Networks (DCGANs), our approach leverages adversarial learning principles to preserve domain-specific features during knowledge transfer.

Central to this work is a transformer-based distillation framework that aligns the rich feature representations of teacher models (trained on image datasets) with the unique spatial and structural characteristics of LiDAR point clouds. This transformer architecture employs multi-head attention mechanisms to maintain both global structure and local detail—an insight originally derived from our GAN-based PD signal synthesis, where temporal and spectral fidelity proved essential for realistic data generation. Through rigorous experimental validation on benchmark datasets, our distilled models achieve an F1-score of 90.4

Beyond immediate performance gains, this research underscores the versatility of knowledge distillation techniques for trans-domain adaptation. It illustrates how established models trained on high-fidelity image data can enhance the interpretative power of LiDAR-based classifiers, significantly reducing the reliance on large-scale annotated point cloud datasets. Additionally, the thesis explores the impact of optimizing distillation parameters—such as temperature and weighting factors—and highlights the potential of self-supervised learning for scenarios where annotated teacher data are scarce. The proposed methodology has broad applicability, potentially extending beyond the fusion of image and LiDAR domains to other fields characterized by data disparities in availability, frequency, and richness. Ultimately, this work lays the groundwork for more robust, efficient, and cost-effective perception systems that can accelerate the deployment of intelligent, real-time applications such as autonomous driving and beyond.

Keywords. Knowledge Distillation, Trans-Domain Classification, Point Cloud Classification, LIDAR Data Processing, Deep Learning, Model Compression, Sensor Fusion, Efficient Machine Learning, Autonomous Systems, Cross-Domain Machine Learning

Contents

List of Figures

1 | Introduction

From its earliest moments, artificial intelligence has modeled human behavior, from the simple perceptron inspired by the neuron cell to the reinforcement learning that mirrors human decision-making. These techniques have led to breakthroughs that have generated new frontiers of knowledge. For example, deep learning models have significantly improved image recognition, enabling facial recognition and medical diagnostics applications. In natural language processing, algorithms like transformers have revolutionized machine translation and text generation.

This success has allowed algorithms to approach increasingly complex tasks, such as autonomous driving, real-time language translation, and advanced scientific simulations[1]. Today, artificial intelligence is used in legal[2], social[3], environmental[4], health care[5], and many other areas, in addition to classic applications such as computer vision and natural language processing. With the new application areas of artificial intelligence and the increased available computing capacity, the complexity of artificial intelligence models is growing. This increase in the complexity of artificial intelligence models encourages innovation in data utilization, allowing the extraction and maximization of the value from available information in ways not previously possible. The optimization of existing architectures adapted to new objectives has resulted in new technologies on the market that are becoming more and more impactful. One of the most exciting technologies today is autonomous driving, a demanding and complex task that requires the integration of sensors of diverse nature into the car.[6]

The capabilities of these sensors are enhanced through a process known as sensor fusion[7], which combines different types of sensors that collect various forms of information to create a more efficient and reliable system. For instance, a standard camera, while useful, faces challenges such as occlusion, varying illumination, camouflage, shadows, and the difficulty of accurately measuring distances. Since it captures two-dimensional images, a standard camera lacks depth perception, making it challenging to determine the distance to objects in a scene. This limitation is critical in applications like navigation, obstacle avoidance, and 3D mapping, where precise distance measurements are essential. Integrating a LIDAR (Light Detection And Ranging) sensor with a camera can mitigate some of these limitations[8], resulting in a combined system that is significantly more effective. This fusion approach not only retains the benefits of a camera but also addresses issues related to occlusion and illumination while adding the capability to measure the distance to detected

objects. Such an advanced sensor system is crucial for operating autonomous driving technologies.

Autonomous driving poses significant challenges since it directly involves passengers' safety. This task requires the development of sophisticated technologies, including advanced sensors, control frameworks, and intelligent actuators. The primary focus is on addressing issues related to vehicle safety. This is challenging in many ways since, although complex algorithms are developed for the control, navigation, and automation of processes, the computing capacity installed in the vehicles is limited. Expanding the computation capacity using remote infrastructure, such as cloud computing, is, in many cases, not even feasible since response time restrictions are critical for adequately functioning the entire system. Furthermore, there is a redundancy of equipment, which means that the computing and data transmission capacity increases as new types of sensors are integrated. Sensor fusion generates a new virtual device with high reliability and robustness in its performance. However, these new sensors have a high implementation cost related to the computational power needed to be deployed and the transmission bandwidth needed to communicate data from different domains[9] within the time restrictions imposed by autonomous driving. The proposal presented in this thesis addresses this issue using the sensors directly and realizing the fusion during the training of a deep network that leverages data from different domains through knowledge distillation [10]. Knowledge distillation is a technique for transferring what a large model learns into a simple, smaller model, called model compression. The large model is called "the teacher," and the small one is "the student." Knowledge distillation also allows the student to have more than one teacher network, increasing the student network generalization power for new unseen data by enriching its learning process. Beyond model compression and knowledge transfer, knowledge distillation has been successfully applied in other areas:

- Privacy and Data Security: Knowledge distillation allows training models without direct access to sensitive data, thus protecting privacy and complying with data protection regulations [11].
- Semi-supervised and Unsupervised Learning: It facilitates training models with limited or unlabeled datasets by leveraging information from pre-trained models [12].
- Multi-task and Multi-modal Learning: Knowledge distillation enables a student model to handle multiple tasks or different data types by combining knowledge from several specialized teacher models [13].
- Robustness Against Adversarial Attacks: It improves the student model's resistance to perturbations or attacks by inheriting defenses learned by the teacher model [14].
- Optimization for Mobile and Embedded Devices: Knowledge distillation is crucial for developing AI applications on energy and processing-constrained devices, such as smartphones and IoT systems [15].
- Reduction of Latency in Real-time Applications: Knowledge distillation helps create faster models suitable for real-time systems by simplifying complex models without significant loss in performance [16].

In previous work, the author of this thesis, in collaboration with Ardila-Rey et al. [17], investigated the use of generative adversarial networks (GANs) for synthesizing realistic partial discharge signals. The study demonstrated the potential of adversarial learning in augmenting limited datasets and improving classification models' performance in partial discharge analysis. This work laid the foundation for exploring the application of adversarial learning in tackling the challenges associated with data scarcity and model performance in various domains.

Building upon this prior research experience, the current thesis focuses on extending the application of adversarial learning to the challenge of trans-domain knowledge distillation, specifically in autonomous driving. The insights gained from the successful implementation of GANs in the partial discharge domain have inspired the author to leverage the power of adversarial networks and transformer architectures to bridge the gap between image-based models and LiDAR point cloud classifiers. By adapting the knowledge distillation process to the unique challenges posed by the heterogeneous data modalities in autonomous driving, this work aims to enhance the perception capabilities of self-driving vehicles.

The experience and expertise acquired through the previous study on partial discharge analysis using GANs have equipped the author with a solid understanding of adversarial learning techniques and their potential for addressing data limitations. This knowledge has been instrumental in shaping the current research direction and methodological approach, enabling a seamless transition from the domain of partial discharge analysis to the more complex and demanding field of autonomous driving perception.

By leveraging the insights gained from the prior work and adapting the adversarial learning techniques to the specific requirements of trans-domain knowledge distillation, this thesis aims to significantly contribute to developing robust and efficient perception systems for autonomous vehicles. The successful application of GANs in the previous study is a strong motivation and proof-of-concept for the current research, highlighting the potential of adversarial learning in overcoming data scarcity and enhancing model performance across different domains.

This process mirrors how humans can quickly learn complex concepts from a small number of samples, even across different categories. The work presented in this thesis exploits the flexibility of knowledge distillation in a scenario where the teacher and student models learn from data of a very different nature, called trans-domain knowledge distillation. The proposed method introduces a novel transformer-based distillation mechanism that adapts knowledge from the image classification domain to the LiDAR point cloud domain. Unlike traditional sensor fusion methods that require extensive retraining and large-scale datasets, the approach presented in this thesis leverages image models' rich, pre-trained features to enhance LiDAR data interpretation without compromising performance. To the author's knowledge, this is the first time a transformer architecture is employed for knowledge distillation between images and point clouds. By doing so, this work offers a more efficient and cost-effective alternative to conventional sensor fusion techniques, preserving the essential reliability and robustness required for autonomous driving technologies while decreasing implementation costs and computational demands.

The main contributions of this thesis are as follows:

1. A novel knowledge distillation approach based on adversarial networks is proposed for transferring knowledge from image-based models to LiDAR point cloud classifiers. The proposed method leverages the power of transformers to bridge the gap between the image and LiDAR domains, enabling the distillation of rich feature representations.

2. A comprehensive framework that integrates knowledge distillation and adversarial learning is developed, allowing for effective knowledge transfer across domains. The framework includes a teacher network trained on image data, a student network operating on LiDAR point clouds, and an adversarial component that enhances the alignment between the two domains.

3. Extensive experiments on benchmark datasets are conducted to validate the effectiveness of the proposed approach. The performance of the proposed method is evaluated in various object detection and classification tasks, comparing it to state-of-the-art techniques in both the image and LiDAR domains. The results demonstrate significant improvements in LiDAR-based object classification, showcasing the potential of the proposed approach in enhancing the perception capabilities of autonomous vehicles.

4. Insights into the interplay between competitive and collaborative interactions in adversarial learning are provided, highlighting the importance of balancing these dynamics for effective knowledge transfer. The impact of different adversarial objectives and training strategies on the framework's performance is explored, shedding light on the key factors contributing to successful trans-domain classification.

5. A comprehensive analysis of the learned representations is presented, visualizing the feature spaces and examining the alignment between the image and LiDAR domains. The analysis provides a deeper understanding of how knowledge distillation and adversarial learning enable the transfer of discriminative features across modalities, facilitating the adaptation of image-based models to the LiDAR domain.

The remainder of this thesis is organized as follows: Chapter 3 presents a comprehensive review of the relevant literature, covering topics such as deep learning, knowledge distillation, adversarial networks, and cross-domain adaptation. The state-of-the-art techniques in these areas are discussed, and the gaps this work aims to address are highlighted. Chapter 4 describes the proposed methodology in detail, presenting the architecture of the knowledge distillation and adversarial learning framework. A step-by-step explanation of the training process and the key components of the approach is provided. Chapter 5 outlines the experimental setup, including the datasets, evaluation metrics, and implementation details. The baselines and state-of-the-art methods that the proposed approach is compared against are also discussed. Chapter 6 presents the experiments' results, showcasing the proposed approach's performance in various object detection and classification tasks. A thorough analysis of the results is provided, comparing the proposed method to existing techniques and discussing the implications of the findings. Finally, Chapter 7 concludes the thesis,

summarizing the contributions and outlining potential directions for future research. By advancing the state of the art in trans-domain knowledge distillation, this thesis contributes to developing more robust and efficient perception systems for autonomous vehicles. The proposed approach has the potential to significantly reduce the reliance on expensive and time-consuming data collection and annotation processes, ultimately paving the way for safer and more intelligent transportation solutions. Furthermore, this work opens up new avenues for exploring the interplay between competitive and collaborative interactions in adversarial learning, providing insights that can be applied to a wide range of domain adaptation and transfer learning problems.

2 | Research Method

2.1 Hypothesis

The main research hypothesis can be formulated as: "Knowledge distillation techniques can effectively bridge the domain gap between image classification and LIDAR point cloud data, enabling the transfer of learned features while maintaining or improving classification performance compared to traditional point cloud classification methods."

2.2 Objectives

The main goal of this thesis is to develop novel methodologies for trans-domain knowledge transfer between different sensor modalities in autonomous systems.

1) Design and implement a transformer-based knowledge distillation architecture capable of transferring learning from image domain to point cloud domain.
2) Develop robust validation methodologies to evaluate the effectiveness of trans-domain knowledge transfer.
3) Quantify and compare the performance improvements achieved through trans-domain knowledge distillation versus traditional single-domain approaches.

2.3 Methodology

The methodology to achieve these objectives consists of:

1. **Architecture Design**:
    - Development of a novel transformer-based distillation mechanism
    - Implementation of teacher-student network architecture
    - Integration of domain-specific processing modules for image and point cloud data

2. **Training Framework**:
    - Three-phase training approach: preprocessing, model training, and generation
    - Implementation of temperature and alpha parameter tuning
    - Optimization of attention mechanisms for cross-domain feature transfer
3. **Validation Strategy**:
    - Model architecture validation using k-fold cross-validation
    - Hyperparameter optimization through grid search
    - Domain adaptation validation through ablation studies
4. **Experimental Evaluation**:
    - Comparative analysis with state-of-the-art point cloud classification methods
    - Performance evaluation across multiple metrics (precision, recall, F1-score)

3 | Related work

Knowledge distillation, a technique widely explored for its potential in model compression, acceleration, and transfer learning, has garnered considerable attention over recent years [18]. This surge of interest is mainly due to the increasing need to implement complex deep-learning models in environments with limited resources.

The timeline shown in Figure 3.1 illustrates how several research streams have converged to enable modern trans-domain knowledge transfer. Beginning with Goodfellow et al. [19]'s introduction of GANs in 2014; the field established fundamental principles for generating and adapting data across domains. This work laid fundamental groundwork for understanding how neural networks could learn and transfer domain-specific characteristics.

The evolution towards practical multi-domain applications gained momentum through Fung et al. [20]'s comprehensive analysis of sensor fusion methods in 2017. Their work established essential frameworks for integrating data from heterogeneous sensors, which are particularly relevant for combining camera and LIDAR data in autonomous systems. This contribution highlighted the challenges and opportunities in working with different sensor modalities, setting the stage for future advances in cross-domain learning.

Significant breakthroughs in bridging domain gaps emerged in 2018, with Wei et al. [21] and Wang et al. [22] demonstrating how GANs could enhance knowledge distillation across domains. These works showed the potential of combining adversarial learning with knowledge transfer techniques, introducing new approaches for handling domain adaptation challenges.

A practical demonstration of these principles emerged in our previous work [23], where we showed how DCGANs could be effectively applied to generate and process complex signal data. This research validated the potential of adversarial learning for domain adaptation in sensor processing applications, providing insights that would later prove valuable for trans-domain knowledge transfer.

Recent developments have focused on specific challenges in autonomous systems, with Tan et al. [24] establishing new benchmarks for processing LIDAR data through their work on 3D point cloud classification. Their subsequent research on transform domain learning [25] introduced novel approaches for handling multi-modal data, particularly relevant for autonomous vehicle perception systems.

This historical progression demonstrates how the field has evolved from separate research threads in gener-

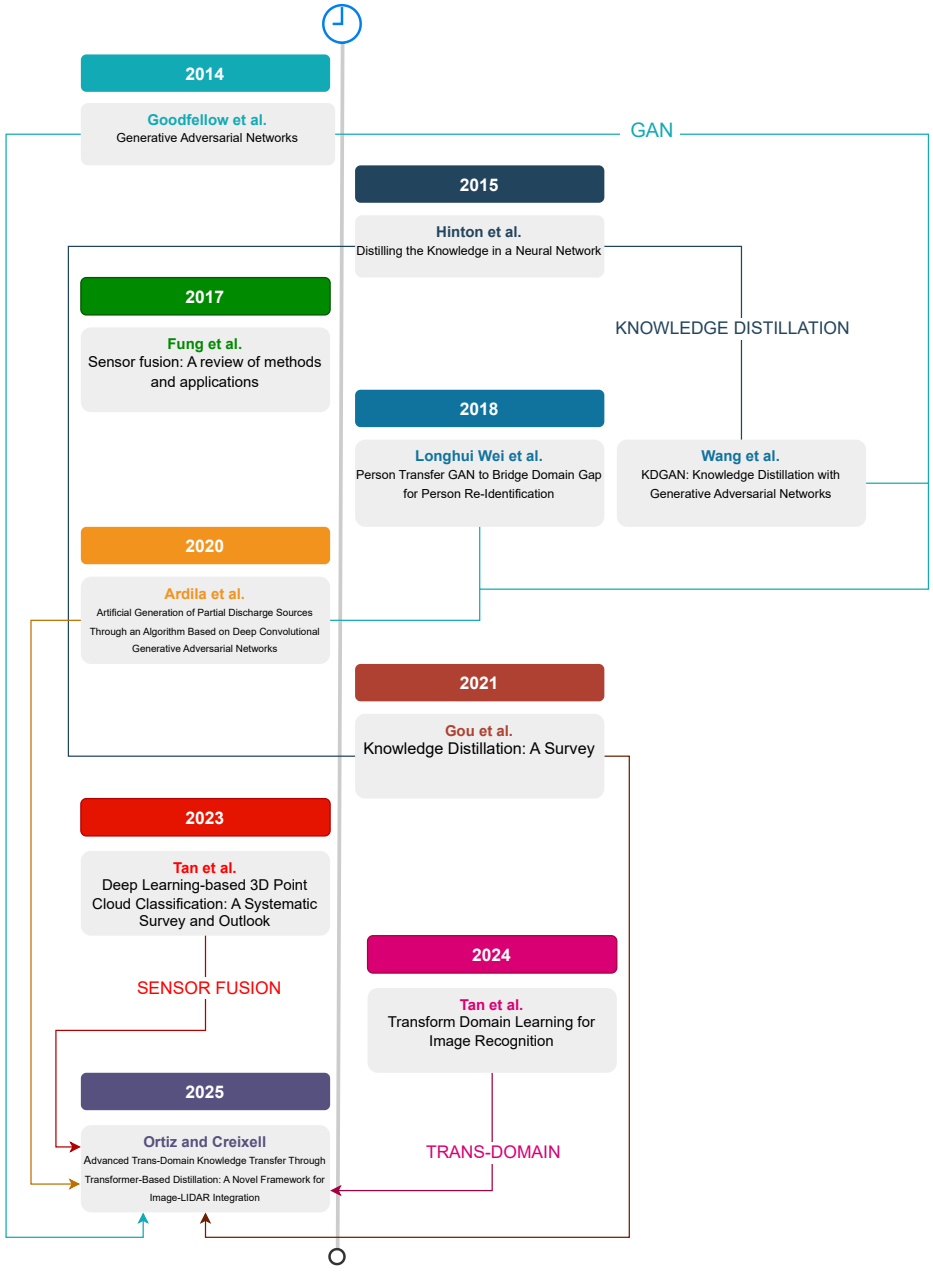


Figure 3.1: Timeline of the State of the Art. This figure illustrates the chronological evolution of key contributions in Generative Adversarial Networks (GAN), knowledge distillation, sensor fusion, and transform-based domain learning from 2014 to 2025. Starting with Goodfellow et al. (2014) on GANs and Hinton et al. (2015) on knowledge distillation highlights significant milestones, such as Fung et al. (2017) on sensor fusion, Gou et al. (2021) on distillation surveys, and Tan et al. (2023, 2024) on 3D point, cloud classification, and transform domain learning, culminating in Ortiz & Creixell (2025), who propose advanced trans-domain knowledge transfer using transformer-based distillation.

ative models, knowledge distillation, and sensor fusion towards integrated approaches that can effectively bridge different data domains. Our current work [26] builds upon these foundations to propose a novel

transformer-based approach for knowledge distillation between image and LIDAR domains. Building on these foundations, modern approaches to knowledge distillation have emerged as powerful tools for addressing the challenges of deep neural networks, as discussed in the following sections.

Deep neural networks are powerful tools known for their impressive predictive capabilities and adaptability[27]. However, they come with a downside - their inherent architectural complexity and depth demand extensive computational resources. You need substantial computing power to train and make inferences with these networks. Storing these models also requires considerable space[28]. These factors pose significant challenges, especially when aiming to deploy these models in environments with limited resources, like mobile phones, Internet of Things (IoT) devices, or other embedded systems.

Knowledge distillation provides a promising solution to these challenges. This process allows for the transfer of knowledge from a more significant, resource-intensive "teacher" model to a smaller, more efficient "student" model [29]. During this process, the 'student' model learns to mimic the behavior of the 'teacher,' which results in a model that performs at a comparable level but is significantly more resource-efficient.

However, knowledge distillation is about more than just model compression and acceleration. It also plays a vital role in trans-domain learning[30] and transfer learning[31]. In these scenarios, the 'student' model learns to apply the knowledge gained from the 'teacher' model to new domains. This ability to adapt and generalize makes knowledge distillation even more valuable in today's fast-evolving machine-learning landscape.

The importance of enhancing the efficacy of distillation techniques cannot be overstated. Optimizing the distillation process not only results in improved performance of the student models but also can bring significant savings in computational resources and storage space. Furthermore, as deep learning advances, it is crucial to identify and develop techniques that can effectively deal with the increased complexity of emerging models.

Applying transformer architectures to knowledge distillation represents a significant advancement in addressing cross-domain challenges [32]. Traditional knowledge distillation methods, while effective within single domains [27], often struggle when faced with fundamental differences between sensing modalities such as cameras and LIDAR systems. Recent work by Liu et al. [33] has demonstrated the effectiveness of transformer-based architectures in bridging the semantic gap between image and point cloud data. Adapting transformer architectures for sensor fusion applications has required significant innovations in architectural design and training methodologies. Chen et al. [34] introduced several key modifications to the standard transformer architecture to handle the specific challenges of multi-modal sensor data. Their work demonstrates that careful attention to the design of positional encodings and the integration of sensor-specific preprocessing layers can significantly improve the transformer's ability to handle different data types [35].

The theoretical foundations supporting the effectiveness of transformer-based knowledge distillation in cross-domain applications have been substantially strengthened by recent research [36]. Their work establishes that the success of transformer-based approaches can be attributed to several key factors: the ability to learn hierarchical representations, the flexibility to model complex relationships through multiple layers of self-attention, and the capacity to maintain separate but interrelated representations for different sensor

modalities.

The theoretical foundations supporting the effectiveness of transformer-based knowledge distillation in cross-domain applications have been substantially strengthened by recent research [36]. Their work establishes that the success of transformer-based approaches can be attributed to several key factors: the ability to learn hierarchical representations, the flexibility to model complex relationships through multiple layers of self-attention, and the capacity to maintain separate but interrelated representations for different sensor modalities.

The broader field of cross-modal and multi-modal learning provides essential context for understanding trans-domain knowledge transfer, representing a fundamental paradigm shift from traditional single-modal approaches toward leveraging complementary information across different data modalities. Early pioneering work by Ngiam et al. [37] demonstrated the potential of deep learning architectures for audio-visual speech recognition, establishing foundational principles for multi-modal representation learning. Their approach showed that jointly training on multiple modalities could significantly improve performance compared to single-modal baselines, particularly in noisy environments where one modality might be compromised, laying the groundwork for understanding how different data types could be effectively combined for enhanced performance.

The field expanded significantly with advances in vision-language understanding, where researchers began exploring more sophisticated methods for aligning visual and textual representations. Kiros et al. [38] introduced skip-thought vectors for learning distributed sentence representations that could be effectively aligned with visual features, establishing foundational techniques for cross-modal retrieval and understanding. This work laid crucial groundwork for numerous applications in image captioning, visual question answering, and cross-modal retrieval tasks that would follow. Building on these foundations, more recent developments have focused on attention-based mechanisms for cross-modal alignment, with Anderson et al. [39] proposing bottom-up and top-down attention mechanisms for image captioning, demonstrating how spatial attention over image regions could be effectively combined with linguistic representations. Similarly, Lu et al. [40] introduced ViLBERT, a transformer-based architecture for learning joint representations of image content and natural language, showcasing the power of attention mechanisms in bridging different modalities.

In the domain of audio-visual learning, researchers have explored self-supervised approaches that leverage the natural synchronization between different sensory modalities. Owens et al. [41] explored self-supervised learning by predicting ambient sounds from visual scenes, demonstrating how one modality could provide supervisory signals for learning representations in another without requiring explicit annotations. This approach highlighted the potential for cross-modal supervision, where the natural correspondence between modalities could be exploited for representation learning. Zhao et al. [42] developed methods for sound source localization in videos, further demonstrating the potential for learning meaningful correspondences between audio and visual information. These approaches share conceptual similarities with knowledge distillation in that they both involve transferring information from one domain to enhance learning in another.

Recent work has extended cross-modal learning to more challenging scenarios involving significant domain gaps between different types of structured data. Peng et al. [43] proposed cross-modal learning for 3D shape recognition using 2D images, addressing the fundamental challenge of learning meaningful correspondences between 2D visual information and 3D geometric structures. Wang et al. [44] developed methods for cross-modal hashing between images and text, enabling efficient retrieval across different modalities. These approaches demonstrate the evolution of cross-modal learning toward handling increasingly complex domain gaps, moving beyond natural correspondences like audio-visual synchronization toward learning alignments between fundamentally different data representations.

Our approach to knowledge distillation between images and LIDAR point clouds extends this cross-modal learning paradigm to the specific challenge of autonomous driving perception, where the domain gap between 2D images and 3D point clouds presents unique challenges. While previous cross-modal work has primarily focused on vision-language combinations or audio-visual scenarios where some natural correspondence exists, our method addresses the geometric and semantic differences between 2D RGB images and sparse 3D point cloud representations. This requires novel architectural innovations and training strategies specifically designed to bridge this particular domain gap, leveraging the rich feature representations learned from high-resolution image data to enhance the interpretation of geometrically structured but semantically sparse LIDAR data.

The paper by Li[45] excellently demonstrates this endeavor. The authors' innovative approach to self-distillation, incorporating a selective feature fusion module, represents a promising direction for improving distillation techniques. Moreover, the authors' finding that integrating fused features into the network can further enhance its performance and stresses the potential for future explorations in this direction. Such advancements in distillation techniques will be vital in deploying powerful deep learning models in resource-constrained environments, thereby expanding the reach and impact of deep learning technologies.

Moreover, the survey by Xiaojun Chang [46] thoroughly examines knowledge distillation applied to object detection (OD), a core field of focus within computer vision. The authors underline that despite substantial performance enhancements achieved by deep neural network-based object detection models, these models depend heavily on intricate network architectures and large-scale parameters. Such dependency often translates into high computational and storage requirements, posing challenges for real-time or edge-device applications [47].

This critical role of knowledge distillation becomes increasingly significant when we look beyond object detection [48] and into other tasks such as image classification[49]. Like OD, image classification also benefits from knowledge distillation in model compression and performance enhancement, thus making it a widely applicable strategy across machine learning tasks. This thorough exploration of knowledge distillation, especially its refinement and evolution, highlights the method's central position in machine learning and its expected role in driving future advancements.

Recent work by Kim et al. [50] has focused on optimizing the computational efficiency of sensor fusion

algorithms while maintaining robustness. Their research establishes important benchmarks for real-time performance in resource-constrained environments, providing frameworks for evaluating the practical deployability of fusion systems in autonomous vehicles.

4 | Methodology

4.1 Proposal

This investigation delves deeper into a specific application of knowledge distillation beyond its traditional confines. While the broader field continues to explore various novel areas, this study specifically utilizes knowledge distillation not for model compression but for trans-domain learning[51]. The focus here is training a model for image classification that subsequently facilitates teaching a student network to classify objects in point clouds. This approach not only underscores the adaptability of knowledge distillation but also showcases its relevance in bridging distinct domains of machine learning.

By applying knowledge distillation principles within trans-domain machine learning tasks, our research directly tackles the pressing issues of model efficiency, generalization, and adaptability across varied data domains. This approach not only enhances model functionality beyond the traditional confines of model compression and transfer learning but also rigorously tests the bounds of these models in novel, cross-disciplinary settings. Our findings significantly advance the dialogue within the machine learning community, promoting a more rigorous and systematic exploration of knowledge distillation. This methodology facilitates the development of models that adeptly handle modern technological ecosystems' complexities and dynamic nature, particularly in their application to disparate data types and domains.

The architecture employs a teacher-student model for KD tailored for trans-domain applications. Unlike typical distillation, which focuses on model compression, this approach maintains network size parity, aiming to bridge the domain gap between image processing and LIDAR point cloud classification. This novel training strategy eliminates constraints when training new architectures, departing from conventional practices.

The teacher network corresponds to AlexNet architecture, chosen specifically for its balance between complexity and experimental tractability. Since we intended to integrate pre-trained YOLO models, AlexNet's relatively straightforward architecture facilitates better experimentation with this novel knowledge distillation technique compared to more complex networks. This deliberate choice allows us to focus on validating the trans-domain transfer mechanism without the added complexity of extremely deep architectures. The network consists of five convolution down-sampling layers, a dense block, and spatial pyramid pooling layers[52]. This structure extracts a feature set from image data, integrating features at various abstraction levels. The

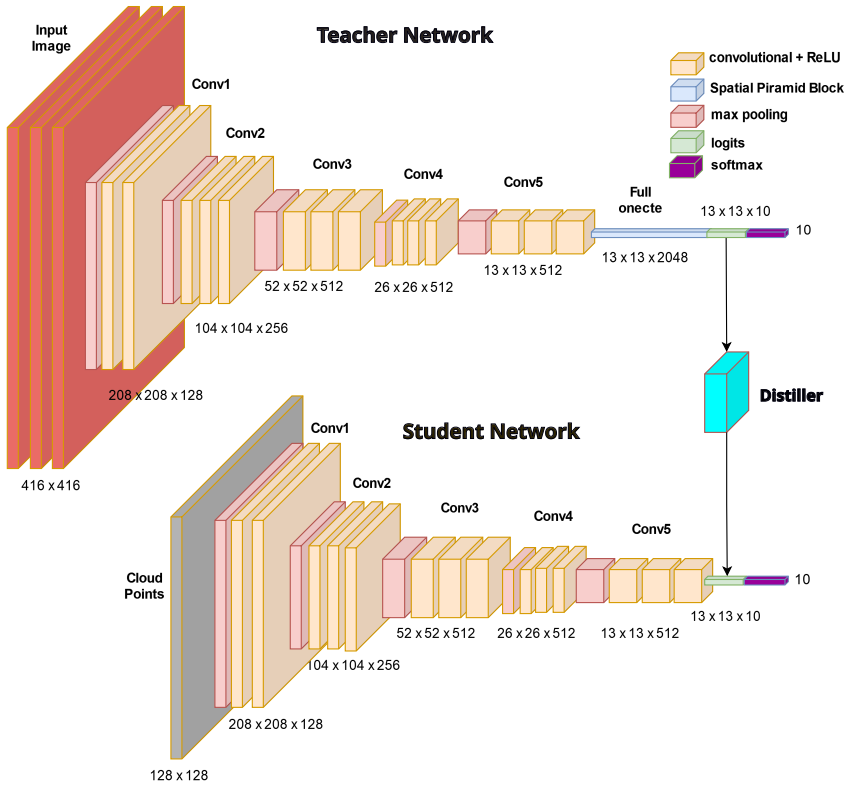


Figure 4.1: Architecture Overview of Teacher and Student Networks in Knowledge Distillation. The diagram illustrates the pre-trained teacher network processing input images through convolutional downsampling followed by dense and spatial pyramid blocks to generate logits for object detection. These logits are then passed to a transformer for distillation. The student network, also pre-trained, similarly processes cloud points and uses the distilled knowledge from the teacher to improve its object detection capabilities.

spatial pyramid pooling further refines the feature map, preparing it for the final classification layers.

In the distiller model, knowledge distillation is driven by the dynamic adjustment of the temperature and alpha hyperparameters. These adjustments are critical for both modulating the softening of outputs and balancing the emphasis between hard and soft targets during the knowledge transfer process. This fine-tuning occurs within the distiller block, explicitly incorporating these two parameters to enhance control over the model's learning dynamics. We conduct a grid search to optimize temperature and alpha, aiming to improve essential metrics such as Mean Average Precision Score (MAPS), accuracy, mean accuracy, and F1 score. These metrics are crucial for our object classification task, ensuring that our model's performance is evaluated comprehensively.

The core of the distiller incorporates a novel use of a transformer model[53], structured with an encoder-decoder framework, to adeptly bridge the gap between distinct data domains—specifically transitioning from image data to the classification of point clouds obtained from LIDAR. This transdomain knowledge distillation is needed because of the nature of the input domains, where traditional distillation techniques fall short. Our transformer model variably employs 8 to 32 attention heads to manage the complexity and diversity of the data involved effectively. By transforming image domain logits to align with LIDAR classification

logits, the model ensures that the distilled knowledge is preserved and aptly adapted to meet the demands of this particular cross-domain application.

In the architecture of our distiller, two key hyperparameters, temperature and alpha, play a central role. The temperature parameter primarily controls the smoothing of the outputs of the teacher network by multiplying the logits fed to the distiller. Higher temperatures result in smooth probability distributions over classes, facilitating a more effective knowledge transfer by emphasizing the relational information between classes rather than absolute predictions. This procedure is particularly beneficial in complex classification tasks with significant nuanced differences between categories. The alpha parameter, on the other hand, balances the relative importance of matching the soft targets provided by the teacher with the challenging targets of the proper labels. It allows the student model to learn the teacher's exact outputs and refine them based on the ground truth, fostering a more robust learning environment.

The transformer-based distillation process in our architecture can be formally represented through a series of transformations. As shown in Figure 2, the process begins with the teacher and student logits being processed through parallel streams with temperature and alpha modulation, respectively. The core transformations can be expressed as:

$$z_t = \frac{\text{logits}_{\text{teacher}}}{\tau} \tag{4.1}$$

Where τ is the temperature parameter that controls the softness of the probability distribution. For the multi-head attention blocks, each transformation follows:

$$\text{MHA}(Q, K, V) = \text{Concat}(h_1, ..., h_n)W^O \tag{4.2}$$

where each head h_i is computed as:

$$h_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{4.3}$$

The Add & Norm layers implement a residual connection followed by layer normalization:

$$\text{AddNorm}(x) = \text{LayerNorm}(x + \text{Sublayer}(x)) \tag{4.4}$$

The final logistic output is computed as follows:

$$\text{output} = \text{Softmax}(\text{Linear}(\alpha \cdot \text{AddNorm}(\text{FeedForward}(\text{AddNorm}(\text{MHA}(z_t, z_s)))))) \tag{4.5}$$

Where z_t represents the teacher logits processed through temperature scaling, z_s represents the student logits, and α is the balancing parameter that controls the contribution of the transformed knowledge in the final prediction. The linear transformation followed by Softmax produces the final probability distribution for the target classes.

Using a transformer within our distiller is pivotal for addressing the challenges posed by cross-domain data

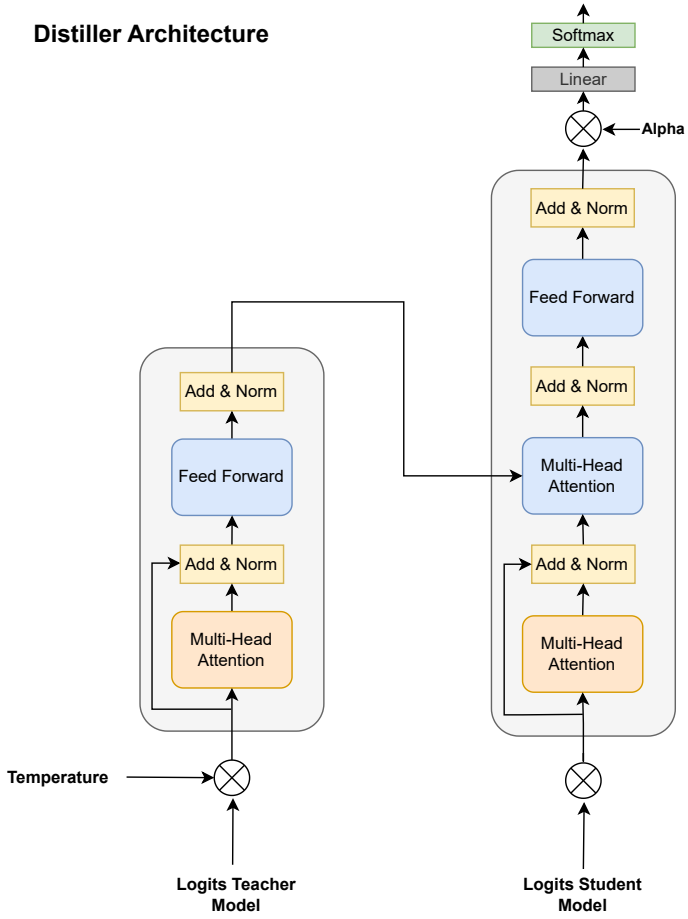


Figure 4.2: Architecture of the Distillation Enhancer in KD model This diagram outlines the sophisticated structure of the distillation process within teacher and student logistic models, highlighting the pivotal roles of alpha and temperature adjustments. It features multi-head attention and feed-forward layers, coupled with recurrent 'Add & Norm' stages for model regulation. These adjustments are crucial for modulating the knowledge transfer, refining logistic outputs, and enhancing the student model's learning precision.

characteristics.

Traditional distillation methods typically assume that the teacher and student models operate within the same input domain. However, our task involves transferring knowledge from image data domains to the radically different domains of LIDAR point cloud classifications. The teacher network is trained on COCO dataset [54] images, which typically have a resolution of approximately 640x480 pixels in RGB space, with the image capture frequency at 30 frames per second. In contrast, the LIDAR data from the NuScenes dataset [55] [56] comprises around 300,000 points per scan with a capture frequency of 20 frames per second. Transformers, composed of layers of self-attention mechanisms, are particularly suited to this task due to their capacity to handle high-dimensional sequential input data from disparate sources such as images and LIDAR point clouds. Additionally, the latent spaces they develop enhance their ability to model various data relationships. These latent spaces effectively capture the underlying patterns and nuances of the data, facilitating more

robust and adaptable cross-domain knowledge transfer.

The transformer employs an encoder-decoder structure. The encoder processes the logits from the image domain, mapping these into a 512-dimensional space that captures the latent dynamics of the process, underlying patterns, and structures essential for classification tasks. This representation is then passed to the decoder, which translates these abstract features into a form suitable for predicting classes in the point cloud data from LIDAR. The ability of the transformer to modify its attention focus across different parts of the input sequence allows it to adaptively learn how to represent best the knowledge necessary for classification in a different domain.

Choosing 8 to 32 attention heads in the transformer's architecture further refines its capability. Although increasing the number of attention heads can enhance the model's ability to understand complex relationships within the data by allowing each head to focus on different segments of the input sequence or represent diverse aspects of the input, this approach is not without limitations. Merely adding more attention heads does not guarantee improved performance. There is a point of diminishing returns where additional heads may contribute to overfitting or computational inefficiency. Moreover, the effective adaptation of logits from one domain to another also depends on the alignment and calibration of these heads. It is essential to optimize their number and configuration to ensure that the student network can accurately interpret and utilize the distilled knowledge from the teacher network.

The performance of the trained student network in classifying point cloud data is benchmarked against established systems to assess our knowledge distillation strategy's effectiveness in overcoming the ground truth scarcity in LIDAR data.

4.2 Validation Methodology

The validation process for our trans-domain knowledge distillation framework demands a rigorous and systematic approach to ensure the architecture and methodology's robustness and reliability. Our comprehensive validation strategy addresses the unique challenges presented by cross-domain knowledge transfer between image and point cloud data, focusing on establishing the effectiveness and stability of our proposed approach. This validation framework is structured around four principal components: a general validation process that establishes the foundational evaluation criteria, architectural validation that examines the structural integrity and performance of our model, hyperparameter optimization that ensures optimal knowledge transfer, and domain adaptation validation that verifies the effectiveness of cross-domain feature transformation. Each component is designed to provide specific insights while contributing to a comprehensive understanding of the system's capabilities and limitations.

4.2.1 General Validation Process

The foundation of our validation methodology lies in establishing robust evaluation protocols that account for the unique challenges of cross-domain knowledge transfer. This process begins with implementing a comprehensive data management strategy, ensuring that our validation results accurately reflect the model's true capabilities. We establish rigorous criteria for assessing both the quantitative performance metrics and the qualitative aspects of knowledge transfer between domains.

The validation process incorporates multiple assessment levels, from individual component evaluation to system-level performance analysis. This hierarchical approach ensures that each aspect of the architecture is thoroughly validated while focusing on the overall objective of effective knowledge transfer. Performance metrics are carefully selected to reflect the accuracy of classification outcomes and the quality of knowledge transfer between domains.

4.2.2 Architectural Validation

The architectural validation phase focuses on verifying our proposed model's structural integrity and performance characteristics. This phase begins with strategically partitioning the dataset, allocating 70% for training purposes while equally dividing the remaining data between validation and testing sets, with 15% each. This distribution ensures sufficient data for training while maintaining statistically significant sets for validation and testing purposes.

We implement a k-fold cross-validation strategy with k=5 to strengthen the validation process, providing deep insights into the model's stability and generalization capabilities across different data distributions. This approach is particularly crucial in our context, where the model must maintain consistent performance across varying data characteristics in both image and point cloud domains.

The architectural validation process incorporates sophisticated monitoring and control mechanisms throughout the training phase. An early stopping protocol with a patience parameter of 10 epochs prevents overfitting while ensuring optimal convergence, which is particularly important given the complex nature of cross-domain knowledge transfer. Monitoring key performance metrics provides comprehensive insights into the model's learning dynamics and adaptation capabilities.

4.2.3 Hyperparameter Optimization

The optimization of hyperparameters represents a crucial aspect of our validation methodology, particularly given the complex nature of trans-domain knowledge distillation. The effectiveness of knowledge transfer between image and point cloud domains heavily depends on the precise tuning of several key parameters that govern the distillation process and the transformer architecture. This optimization process requires careful consideration of each parameter's theoretical foundations while balancing the practical implementation

constraints.

The temperature parameter t plays a fundamental role in knowledge distillation, controlling the softness of probability distributions generated by the teacher network. Theoretically, this parameter modifies the logits of the teacher network according to the relationship $z_t = logits_teacher/t$, where higher temperatures produce softer probability distributions across classes. This softening process is crucial as it reveals the inter-class relationships learned by the teacher network, providing richer training signals for the student network. Our comprehensive grid search validation spans the interval $[0.1, 10]$, with particular emphasis on the range $[1.0, 3.5]$. The selection of this range is theoretically motivated by the need to balance two competing factors: temperatures below 1.0 tend to approach hard labels, potentially losing valuable relational information, while temperatures above 3.5 risk over-smoothing the distributions, diluting the discriminative power of the teacher's knowledge.

The alpha parameter α , which controls the relative weighting between teacher and student predictions, requires equally careful optimization. This parameter operates within the $[0.01, 1.0]$ range. It determines the balance between two competing objectives: maintaining fidelity to the teacher's knowledge and allowing the student to adapt to the point cloud domain's specific characteristics. The theoretical importance of α lies in its role in the loss function, where it modulates the contribution of the distillation loss relative to the task-specific loss. Values closer to 0 emphasize the student's direct learning from ground truth labels, while values closer to 1 prioritize matching the teacher's outputs. Our sequential validation process explores this range systematically, with finer granularity in regions where performance shows high sensitivity to parameter changes. In our transformer-based architecture, the configuration of attention heads represents a critical hyperparameter that significantly impacts model performance and computational efficiency. The number of attention heads affects the model's capacity to capture different aspects of the relationship between image and point cloud representations. Our validation explores configurations with 8, 16, and 32 heads, offering different trade-offs between model capacity and computational requirements. The theoretical basis for these choices stems from the transformer's ability to perform multiple attention operations in parallel, with each head potentially specializing in different aspects of the cross-domain relationship. The learning rate schedule, another crucial hyperparameter, requires careful optimization to ensure stable convergence during the knowledge transfer. We implement a warm-up strategy followed by a decay schedule, with the base learning rate varying within $[0.001, 0.01]$. This approach is fundamental in our context, as the early stages of training require careful parameter updates to establish proper cross-domain mappings. The warm-up period, spanning the first 10,000 steps, allows the model to establish stable gradient flows before reaching the full learning rate.

The search space spans from 16 to 128, focusing on the $[32, 96]$ range. This parameter significantly impacts both the quality of gradient estimates and the effectiveness of batch normalization layers, which are crucial for stabilizing the training of deep networks. The theoretical motivation for our chosen range comes from the need to maintain sufficient statistical diversity within each batch while ensuring effective utilization of

computational resources.

Beyond individual parameter optimization, we also consider the interaction effects between hyperparameters. For instance, the effectiveness of temperature scaling can be influenced by the learning rate and batch size, requiring a holistic optimization approach. Our validation process, therefore, includes multi-parameter optimization phases where we explore combinations of parameters within their respective ranges to identify optimal operating points that consider these interactions.

Additionally, we implement adaptive hyperparameter schedules for certain parameters, particularly the temperature and alpha values. This approach allows these parameters to evolve during training, adapting to the changing needs of the knowledge transfer process as the student network progresses in its learning. The schedules are designed based on theoretical considerations of the knowledge distillation process, with parameters adjusted gradually to maintain stable training while optimizing performance.

This comprehensive hyperparameter optimization process ensures that our architecture operates at its optimal point for effective cross-domain knowledge transfer. The careful balance of theoretical considerations with practical constraints enables our system to achieve robust performance while maintaining computational efficiency.

4.2.4 Domain Adaptation Validation

The validation of domain adaptation capabilities addresses the fundamental challenge of our research: ensuring effective knowledge transfer between the substantially different domains of image and point cloud data. This validation phase employs specialized approaches to verify the effectiveness of our cross-domain knowledge transfer methodology.

We implement comparative analyses of the student network's performance with and without knowledge distillation, providing crucial insights into the value added by our approach. The validation process includes a detailed examination of gradient stability during training, an assessment of feature space alignment between domains, and an evaluation of the preservation of spatial relationships during knowledge transfer.

Through carefully designed ablation studies, we validate the contribution of each architectural component to the overall system performance. These studies systematically examine the impact of different architectural choices, from the basic network structure to sophisticated feature transformation mechanisms. This comprehensive component analysis provides crucial insights into the relative importance of different aspects of our design.

The domain adaptation validation process pays particular attention to the quality of knowledge transfer, examining classification accuracy and the preservation of semantic relationships and geometric properties across domains. That procedure ensures that our model effectively bridges the gap between image and point cloud representations while maintaining the essential characteristics of each domain.

5 | Experimental setup

This study's experimental framework was designed to explore the efficacy of knowledge distillation techniques in transferring the classification capabilities from RGB image models to LIDAR point cloud classifiers. Our setup involved three distinct phases: configuration of the teacher-student model architecture, selection of appropriate datasets and hyperparameter tuning to optimize model performance across diverse metrics. For the first phase, teacher-student model configuration, we tried three different teacher networks in order to test the system sensitivity to the Teacher implementation; the architectures tested were YoloV4 [57], EfficientNET-1 [58], and YoloV7 [59]. The best-performing architecture was used to compare it against the point cloud classification methods. The nuScenes dataset was selected for this study due to its comprehensive coverage of urban driving scenarios, which include a wide range of environmental conditions and object interactions. This dataset uniquely contains synchronized point cloud data captured by LIDAR and corresponding high-resolution RGB images, allowing for direct comparison and knowledge transfer between these two modalities. The availability of both data types in the same temporal and spatial context is crucial for the effectiveness of the knowledge distillation approach, ensuring a robust and versatile test in real-world settings. Although other datasets such as KITTI [60] and Waymo Open Dataset [61] also provide point clouds and image data, nuScenes offers a higher frequency of data capture and richer annotation across a more varied array of urban scenarios. These features make nuScenes particularly well-suited for exploring the complex dynamics of urban environments and testing the adaptability of the models to different and challenging real-world conditions. The teacher model used in our experiments was a pre-trained state-of-the-art convolutional neural network.

The experimental setup for grid search in transformer model training was crucial in understanding the learning process of the model. We conducted a grid search to tune the hyperparameters of the transformer model during training. We focused on three key hyperparameters: alpha, temperature, and the number of attention heads in the transformer's architecture. We created a grid of parameter values for each hyperparameter to conduct the grid search. For the alpha hyperparameter, we examined values ranging from 0.01 to 1, with increments of 0.05. Similarly, we explored values ranging from 0.1 to 10 for the temperature hyperparameter, with increments of 0.5. We considered values from 8 to 32 for the number of attention heads, with increments of 4. Configurations of 8, 16, and 32 attention heads were tested, finding that 16 heads offered the best balance

between performance and computational cost. Increasing to 32 heads resulted in only marginal accuracy gains (0.5%) but increased training time by 20. In each experiment of the grid search, we modified the size of the student network to observe its impact on performance. We conducted experiments with varying alpha, temperature, and attention head combinations. This grid search aimed to find the optimal values of these hyperparameters that would improve the transformer model's performance in our specific task of transferring knowledge from the image domain to LIDAR point cloud classifications. In each experiment of the grid search, we trained the transformer model using the selected hyperparameter values and computational budgets. We employed a learning rate schedule with a warm-up period of 40,000 steps and decayed the learning rate with the inverse square root of the number of training steps after the warm-up. The grid search process was essential for finding the optimal combination of hyperparameters that would maximize the performance of the transformer model.

Therefore, two fixed-length arrays represent each input protein, one the average vector of size 512 and the 512×512 covariance matrix. They are then processed by a linear layer followed by a dropout of value 0.1, adapting them to the exact shape (1×512). After that, the two vectors are concatenated in a single vector of shape 1×1024, which is used as input of the hyperbolic generator network.

5.1 The nuScenes Dataset

The experimental framework of this study relies fundamentally on the nuScenes dataset [62], a large-scale multi-modal dataset designed explicitly for autonomous driving research. Developed by Motional (formerly nuTonomy), nuScenes represents a significant advancement over previous datasets in the field, being the first to provide synchronized data from a complete sensor suite of an autonomous vehicle. While other datasets like KITTI focus primarily on camera-based detection, nuScenes offers comprehensive coverage across multiple sensor modalities. It suits our knowledge distillation approach between RGB images and LIDAR point clouds. The dataset encompasses 1000 carefully selected driving scenes collected across Boston and Singapore, two cities chosen for their dense traffic patterns and challenging driving conditions. Each scene spans 20 seconds and captures diverse driving maneuvers, traffic situations, and environmental conditions. This geographical and environmental diversity proves crucial for developing robust autonomous driving systems, as it allows for evaluating algorithm performance across different traffic rules (left versus right-hand driving), weather conditions, and urban environments. The data collection vehicle, a modified Renault Zoe, was equipped with a comprehensive sensor suite as detailed in Table 5.1. This setup included a 32-beam LIDAR providing 360-degree coverage, six cameras ensuring complete visual coverage around the vehicle, and five RADAR sensors for additional object detection and velocity measurement capabilities.

Figure 5.1: NuScenes Sensor Data (Night). Example of nighttime multi-sensor data from NuScenes, including several camera viewpoints along with RADAR and LiDAR top-down overlays. Bounding boxes (orange/blue) indicate detected objects in low-light conditions.

5.1.1 Dataset Characteristics and Sensor Setup

The data collection vehicle, a modified Renault Zoe, was equipped with a comprehensive sensor suite as detailed in Table 5.1. This setup included a 32-beam LIDAR providing 360-degree coverage, six cameras ensuring complete visual coverage around the vehicle, and five RADAR sensors for additional object detection and velocity measurement capabilities. As illustrated in Figure 5.3, the sensors are strategically positioned to provide optimal coverage around the vehicle. The six cameras are positioned to ensure complete 360-degree

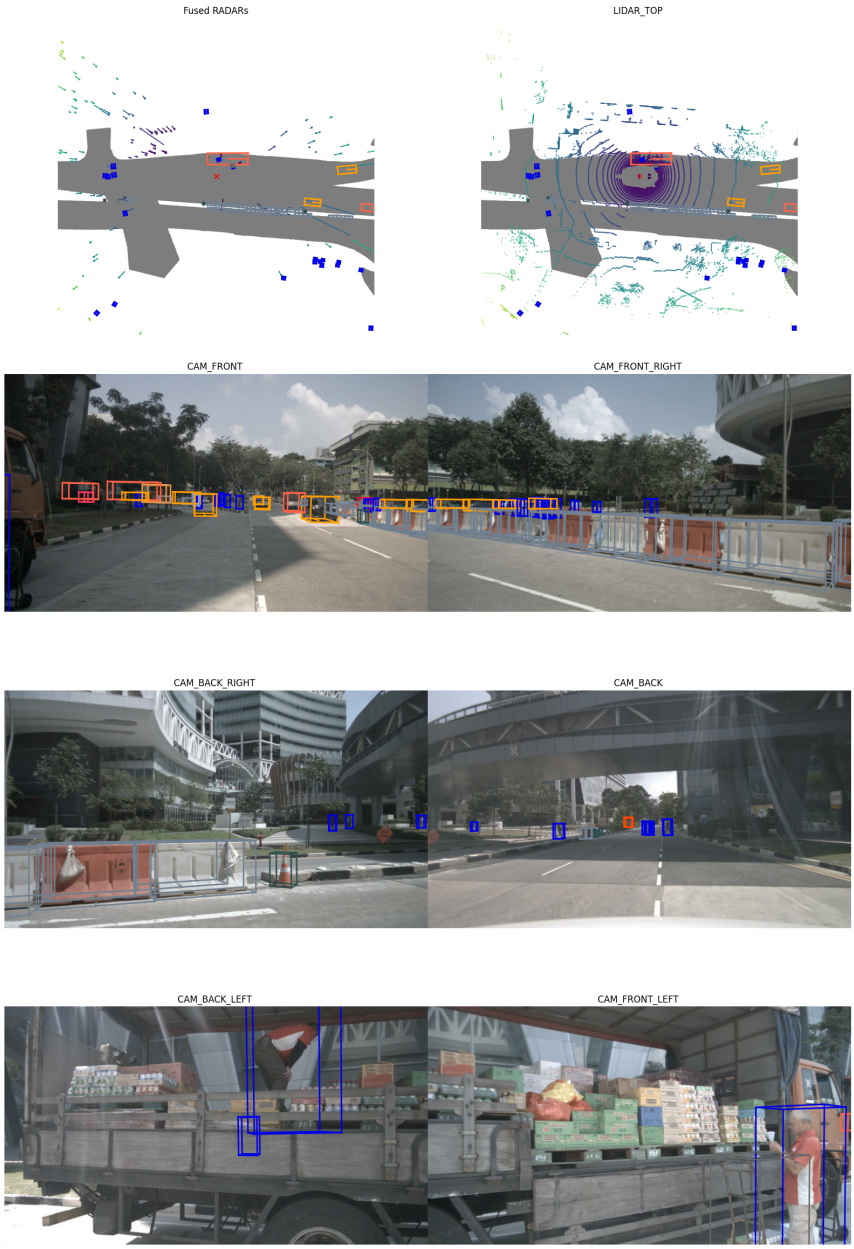


Figure 5.2: NuScenes Sensor Data (Day). Daytime multi-sensor visualization from NuScenes (cameras, RADAR, and top-down LiDAR). Multiple bounding boxes (in different colors) show vehicles, pedestrians, and other objects in well-lit urban conditions.

visual coverage, with front, side, and rear views. The RADAR sensors are placed at the corners and front of the vehicle to maximize detection coverage. In contrast, the LIDAR sensor is mounted on the roof to provide unobstructed 360-degree point cloud data. The dataset's scale and comprehensiveness are evident in its statistics: approximately 1.4M camera images, 390k LIDAR sweeps, and 1.4M RADAR sweeps, all synchronized and annotated. Table 5.2 presents the key dataset statistics, highlighting its significant scale compared to previous autonomous driving datasets.

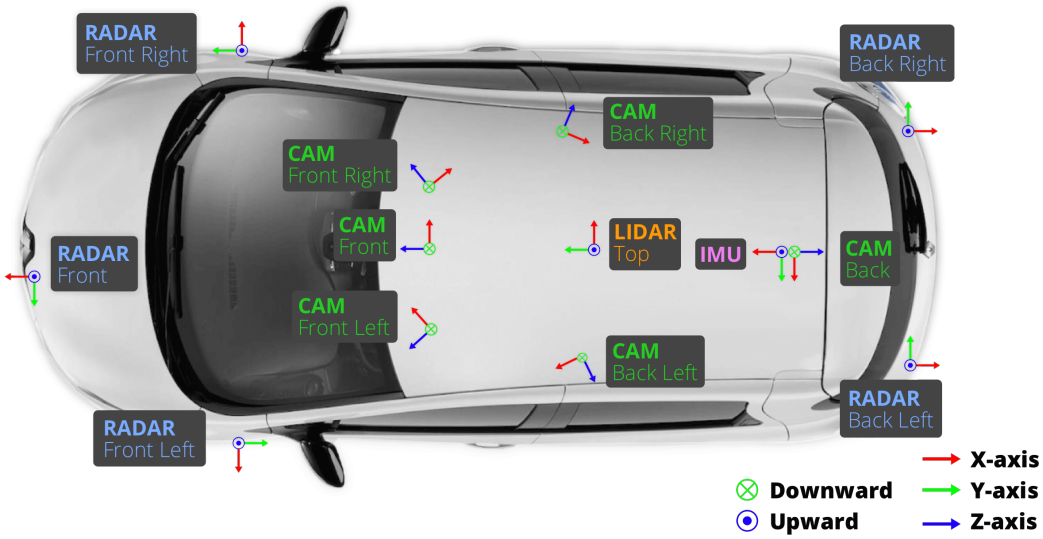


Figure 5.3: NuScenes Sensor Configuration. Top-down view of the sensor setup on the Renault Zoe vehicle. The configuration includes six cameras (CAM, green), 5 RADAR sensors (blue), 1 LIDAR sensor (orange), and an IMU (purple). Coordinate systems for each sensor are indicated with colored arrows (X-axis in red, Y-axis in green, Z-axis in blue).

Table 5.1: NuScenes Sensor Suite Specifications

Sensor Type	Specifications
LIDAR	Velodyne HDL32E: 20Hz capture rate, 32 beams, 360° Horizontal FOV, +10° to -30° Vertical FOV, Effective range up to 70m, ±2cm accuracy
Cameras	6x Basler acA1600-60gc: 12Hz capture rate, 1600x900 resolution, Auto exposure (max 20ms), Complete 360° coverage
RADAR	5x Continental ARS 408-21: 13Hz capture rate, 77GHz frequency, Range up to 250m, Velocity accuracy of ±0.1 km/h
Localization	GPS/IMU with 20mm position accuracy, 0.1° roll/pitch accuracy

5.1.2 Data Annotation and Object Categories

The annotation process follows a rigorous protocol, providing 3D bounding boxes and attribute labels for 23 object classes at a frequency of 2Hz. The dataset includes comprehensive annotations for common and rare object categories, with particular attention paid to challenging scenarios. Table 5.3 presents the distribution of annotations across major object categories, demonstrating the dataset's rich diversity. A unique feature of nuScenes is its nuScenes-lidarseg extension, which provides point-wise semantic labels for the LIDAR data. This includes 32 semantic classes (23 foreground "things" and nine background "stuff" classes), with over 1.4 billion annotated points. This rich semantic segmentation of point clouds makes the dataset particularly valuable for our research into knowledge transfer between image and point cloud domains. The dataset's temporal continuity, with 20-second scenes sampled at regular intervals, enables the development and evaluation of tracking algorithms alongside detection tasks. Furthermore, the synchronized nature of the sensor data allows for direct comparison and correlation between different sensor modalities, making it an

Table 5.2: NuScenes Dataset Statistics and Comparison

Characteristic	nuScenes	KITTI	Waymo
Total Scenes	1,000	22	1,000+
Annotated Frames	40,000	15,000	200,000
3D Boxes	1.4M	200K	12M
Object Classes	23	8	4
Sensor Suite	Full	Partial	Full

Table 5.3: Distribution of 3D Object Annotations in nuScenes

Category	Number of Instances	Percentage
Vehicle (Car)	493,322	42.30%
Pedestrian (Adult)	208,240	17.86%
Movable Object (Barrier)	152,087	13.04%
Traffic Cone	97,959	8.40%
Truck	88,519	7.59%
Bus	16,321	1.40%
Trailer	24,860	2.13%
Motorcycle	12,617	1.08%
Bicycle	11,859	1.02%
Others	60,403	5.18%
Total	1,166,187	100.00%

ideal testbed for our knowledge distillation experiments between RGB images and LIDAR point clouds. This comprehensive and well-structured dataset forms the foundation of our experimental framework, providing the necessary data diversity and annotation quality to evaluate our proposed knowledge distillation approach effectively. The subsequent sections detail how we leverage this dataset within our Azure-based infrastructure to implement and validate our methodology.

5.2 Azure Infrastructure and System Architecture

The experimental framework was implemented on Microsoft Azure's Machine Learning Studio platform, leveraging its comprehensive cloud infrastructure for deep learning experimentation. As depicted in Figure 5.4, the system architecture was designed to ensure robust data processing, efficient model training, and comprehensive performance monitoring. The core computational infrastructure consisted of a virtual machine equipped with an NVIDIA Tesla V100 GPU, 32GB of RAM, and a 2TB SSD storage solution, specifically configured to handle the intensive computational demands of deep learning workloads and large-scale data processing.

5.2.1 Computational Infrastructure

All experiments were conducted on Microsoft Azure cloud infrastructure using Standard_NC6s_v3 virtual machines equipped with NVIDIA Tesla V100 GPUs. The computational specifications included:

- **GPU:** 1x NVIDIA Tesla V100 with 16GB VRAM

- **CPU:** 6 vCPUs (Intel Xeon E5-2690 v4)
- **RAM:** 32 GB system memory
- **Storage:** 2TB Premium SSD for dataset storage
- **Operating System:** Ubuntu 18.04 LTS
- **Framework:** PyTorch 1.8.1 with CUDA 11.1

Training sessions typically required 3-4 hours for the complete nuScenes dataset, with early stopping implemented at 10 epochs of patience to prevent overfitting. All experiments were conducted with identical random seeds (seed=42) to ensure reproducibility.

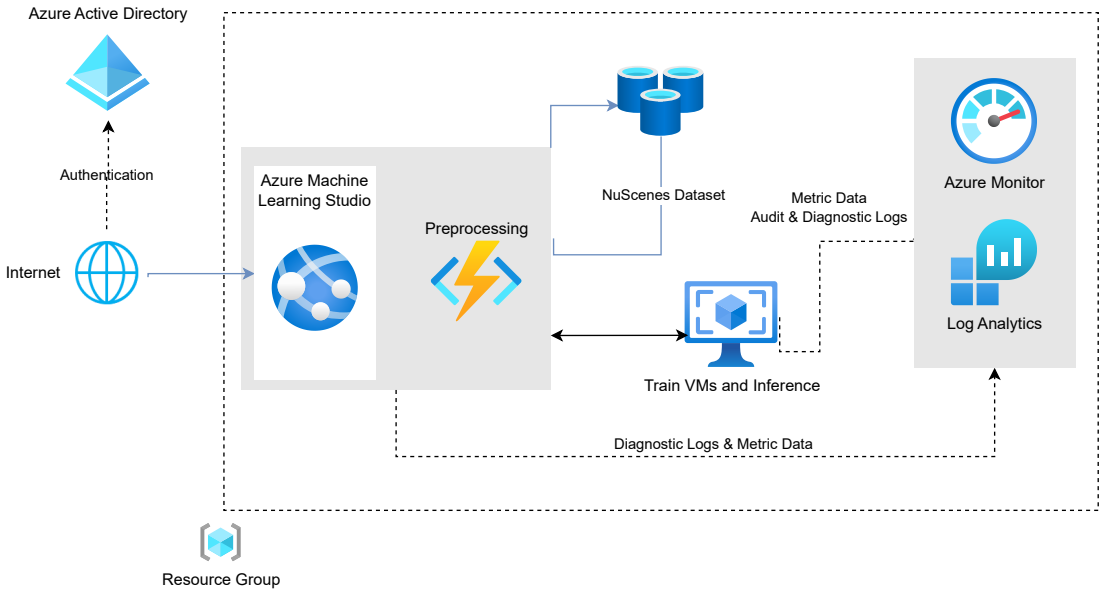


Figure 5.4: Overview of the Cloud-Based ML Pipeline. This diagram illustrates an end-to-end solution using Azure Machine Learning Studio, where the NuScenes dataset is stored and preprocessed. Training and inference tasks run on dedicated VMs within a single resource group. Azure Active Directory handles authentication from the internet while Azure Monitor and Log Analytics collect metric data, diagnostic logs, and audit information for real-time monitoring and analysis.

As illustrated in Figure 5.4, the architecture implements a multi-layered approach to data processing and model training. At the infrastructure level, Azure Active Directory provides secure authentication and access control, ensuring that all computational resources and data assets are adequately protected. The Azure Machine Learning Studio environment is the central orchestration platform, managing the interaction between various system components and executing training workflows. Data processing follows a structured pipeline within the Azure environment. The nuScenes dataset, housed in dedicated storage, undergoes preprocessing steps that prepare both RGB images and LIDAR point cloud data for model training. This preprocessing stage is crucial for maintaining data quality and ensuring efficient training processes. The system implements dedicated processing workflows for training and inference phases, with specific optimizations for handling

the multi-modal nature of our dataset. The monitoring infrastructure comprises two key components: Azure Monitor and Log Analytics, both illustrated in the right portion of Figure 5.4. Azure Monitor provides comprehensive system-level metrics, tracking resource utilization, computational efficiency, and overall system health. Log Analytics complements this by offering detailed insights into the training process, capturing performance metrics, and enabling in-depth analysis of model behavior. This dual monitoring approach ensures both system stability and optimal model training progression. All components are integrated within the Resource Group boundary shown in the diagram to form a cohesive experimental environment. The system's design emphasizes scalability and reproducibility, enabling systematic evaluation of our knowledge distillation approach across different model configurations and dataset scenarios. The infrastructure supports parallel processing capabilities, allowing for efficient handling of the initial nuScenes-mini validation phase and subsequent complete dataset experiments. The diagram also illustrates the implementation of diagnostic logging and metric collection paths, represented by the feedback loops connecting various components to the monitoring systems. This comprehensive logging infrastructure proves invaluable for tracking model training progress, identifying potential bottlenecks, and optimizing system performance. The integration with TensorBoard, while not explicitly shown in the architecture diagram, provides additional capabilities for visualizing training metrics and model behavior in real time. As indicated by the directional arrows in Figure 5.4, network connectivity and data flow patterns are optimized to minimize latency and ensure efficient data transfer between components. The system's architecture supports both synchronous and asynchronous processing patterns, enabling flexible execution of training and evaluation workflows while maintaining data consistency and processing efficiency.

5.3 Network Architecture Analysis

Building upon the established Azure infrastructure, selecting and configuring appropriate neural network architectures formed a critical component of our experimental framework. The study evaluated three distinct teacher networks to assess system sensitivity and optimize knowledge transfer capabilities. Each architecture was deployed and evaluated within the Azure Machine Learning environment, leveraging the computational resources and monitoring infrastructure described in the previous section. The YOLOv4 architecture served as our initial baseline, implementing its CSPDarknet53 backbone to achieve an optimal balance between accuracy and processing speed. This architecture's implementation proved particularly effective in handling complex urban environments, benefiting from its spatial pyramid pooling and path aggregation neck for multi-scale feature fusion. Integrating Azure's GPU resources enabled efficient processing of high-resolution input images while maintaining real-time performance capabilities. EfficientNET-1 provided an alternative approach through its compound scaling method, which systematically balanced network depth, width, and resolution. This architecture's implementation demonstrated particular advantages in resource utilization, aligning well with our Azure infrastructure's computational constraints while maintaining competitive accuracy levels.

The model's efficient scaling properties proved valuable when processing varying input resolutions from the nuScenes dataset, adapting well to daytime and nighttime scenarios. YOLOv7, representing our study's most recent architectural advancement, incorporated improvements, including E-ELAN (Extended Efficient Layer Aggregation Network), for enhanced feature extraction. This architecture demonstrated superior performance in challenging lighting conditions, leveraging its improved feature aggregation mechanisms and expanded receptive field. The implementation within our Azure framework allowed for a comprehensive evaluation of its advanced features, particularly in processing complex urban scenarios captured in the nuScenes dataset.

5.4 Dataset Integration and Processing

Integrating the nuScenes dataset within our Azure-based experimental framework required careful consideration of data handling and processing strategies. Initial validation using nuScenes-mini, comprising 404 scenes, allowed for rapid iteration and system optimization while minimizing computational resource consumption. The Azure infrastructure proved particularly effective in handling this initial validation phase, processing the complete mini dataset in approximately 40 minutes while maintaining comprehensive logging and monitoring capabilities. As illustrated in Figures 5.1 and 5.2, the dataset encompasses various environmental conditions, from well-lit daytime scenarios to challenging nighttime situations. The availability of synchronized LIDAR point clouds and high-resolution RGB images provided an ideal testing ground for our knowledge distillation approach. The 2TB SSD storage solution in our Azure configuration ensured efficient data access and processing capabilities, while the Tesla V100 GPU enabled rapid image and point cloud data processing.

5.5 Training Process and Optimization

The training process leveraged Azure's computational resources to implement a sophisticated learning rate schedule to optimize model convergence. Implementing a 40,000-step warm-up period, followed by an inverse square root decay of the learning rate, proved effective in managing the learning dynamics across different training phases. This approach monitored through Azure's logging infrastructure and TensorBoard integration, helped mitigate early training instabilities while ensuring effective learning rate adaptation throughout the training process. The grid search for optimal hyperparameters revealed interesting relationships between model performance and architectural choices. The investigation of attention head configurations demonstrated that while increasing from 8 to 16 heads provided substantial performance improvements, further expansion to 32 heads yielded only marginal benefits (0.5% accuracy improvement) at the cost of significantly increased computational overhead. These findings, captured through our comprehensive monitoring infrastructure, guided our final architectural decisions. Temperature parameter tuning proved particularly crucial for effective knowledge distillation. Exploring temperature values ranging from 0.1 to 10.0 revealed that lower

temperature values (below 1.0) generally resulted in sharper probability distributions. In comparison, higher values produced softer distributions that better facilitated knowledge transfer between modalities. The alpha parameter, controlling the balance between hard and soft targets, showed optimal performance in the range of 0.3 to 0.5, suggesting the importance of maintaining a balance between teacher signal and ground truth supervision. These optimization processes were continuously monitored and logged through Azure's analytics infrastructure, enabling detailed analysis of the relationship between hyperparameter settings and model performance.

6 | Results and discusion

6.1 Experimental Results

The experimental evaluation shows the effectiveness of our transformer-based knowledge distillation approach for cross-domain object classification. The results presented in table 6.1 shows that the proposed KDNet architecture has a high performance through all the tested variants: KD-Y4, KD-EFF, and KD-Y7. KD-Y4 establishes a strong baseline with consistent performance, achieving 90.6% precision, 87.6% recall, and 88.9% F1-score. This model demonstrates particularly robust performance in complex categories such as Person (91.6% F1-score) and Car (91.0% F1-score), validating the effectiveness of our trans-domain approach. KD-EFF, while showing comparable performance woth KD-Y4 in specific categories like Person (92.3% F1-score) and Motorcycle (92.1% F1-score), achieved lower overall metrics with 80.2% precision, 86.5% recall, and 82.6% F1-score. This performance difference highlights the importance of teacher network selection in knowledge distillation. The model maintained strong performance in everyday urban objects but showed decreased effectiveness in challenging categories like Fire Hydrant (60.4% F1-score) and Parking Meter (59.6% F1-score).

KD-Y7, shows the best overall performance among the knowledge distillation models, achieving 89.0% precision, 92.1% recall, and 90.4% F1-score. This model shows remarkable improvements in traditionally challenging categories, with Fire Hydrant and Parking Meter classifications improving precision from 51.2% to 76.5% and from 49.8% to 74.2%, respectively. The model excels particularly in Person detection (95.5% F1-score) and Car classification (94.3% F1-score), suggesting effective transfer of features from the image domain to point cloud classification The comparison with traditional point cloud classification methods, detailed in table 6.2, reveals significant advantages of our approach. PointNet++, the best-performing traditional method, achieves 74.1% precision, 79.3% recall, and 75.6% F1-score. DGCNN shows lower performance with 38.0% precision, 71.0% recall, and 50.5% F1-score, while the baseline PointNet achieves 14.92% precision, 72.2% recall, and 27.2% F1-score. Our benchmark comparison, presented in table 6.3, directly contrasts the best-performing models from each approach. KD-Y7 (90.4% F1-score) significantly outperforms PointNet++ (75.6% F1-score), demonstrating a 19.6% relative improvement in classification

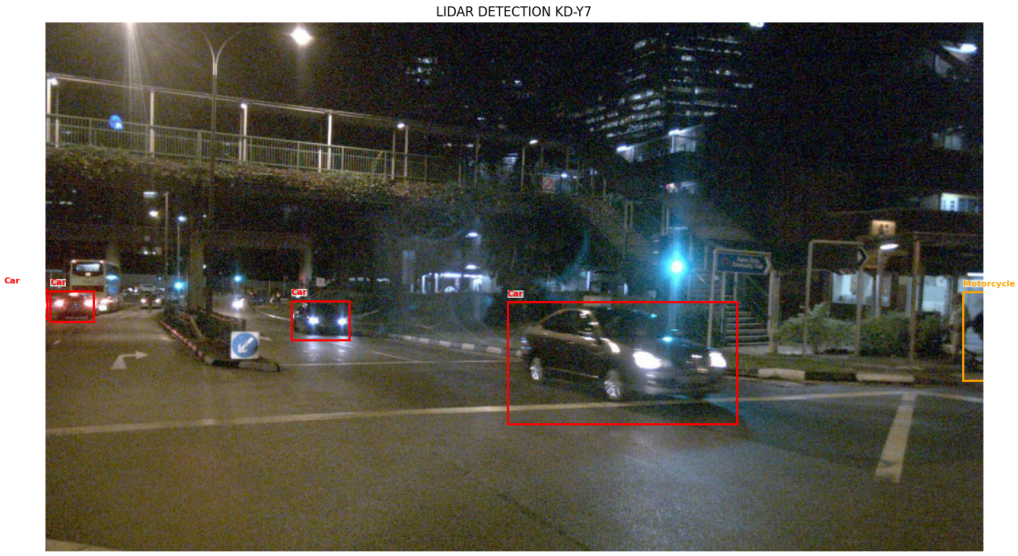


Figure 6.1: Nighttime LiDAR Detection Using KD-Y7. A single-view output illustrating cars, motorcycles, and other vehicles identified by the KD-Y7 model at night.

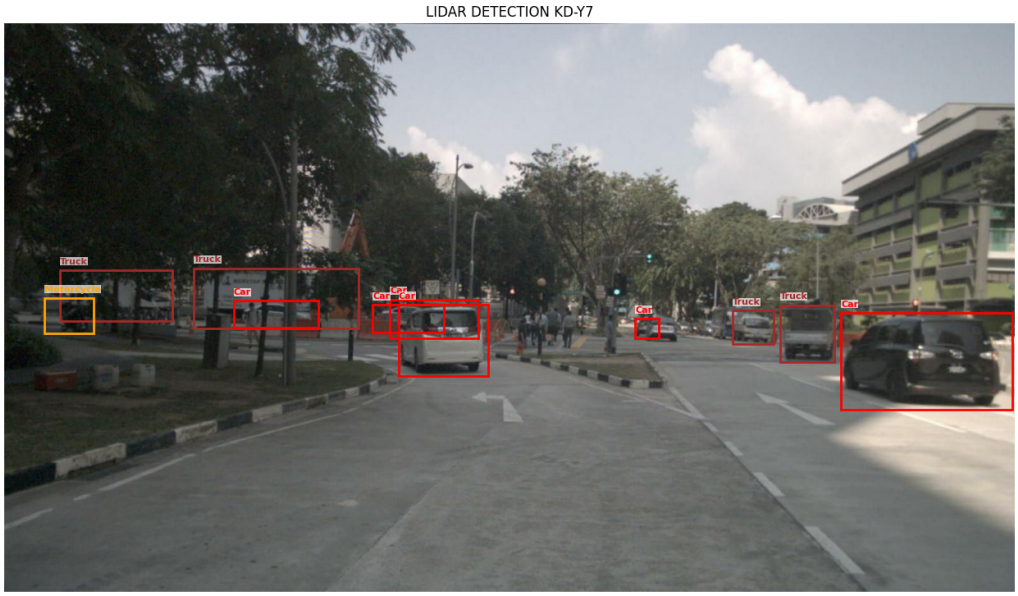


Figure 6.2: Daytime LiDAR Detection Using KD-Y7. Example of the KD-Y7 model's detection output in daylight, highlighting cars, trucks, and motorcycles.

performance. This substantial performance gap validates the effectiveness of our trans-domain knowledge distillation strategy and suggests its potential for broader applications in autonomous driving and robotics applications.

These results show that our transformer-based knowledge distillation approach effectively bridges the domain gap between image-based classification and point cloud data, maintaining high performance across diverse object categories while addressing the challenges inherent in trans-domain learning.

Analysis of failure cases revealed specific challenges in classifying particular urban objects. The most

Table 6.1: Performance of Knowledge Distillation for three teacher network: KDNet-YoloV4 (KD-Y4), KDNet-EfficientNet (KD-EFF), and KDNet-YoloV7 (KD-Y7)

Class	KD-Y4			KD-EFF			KD-Y7		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Person	92.5	90.7	91.6	93.5	91.2	92.3	96.2	94.8	95.5
Bicycle	88.3	87.2	87.7	87.1	86.5	86.8	91.5	90.2	90.8
Car	92.3	89.8	91.0	91.5	88.2	89.8	95.2	93.5	94.3
Motorcycle	91.5	90.9	91.2	92.8	91.5	92.1	94.5	93.8	94.1
Bus	91.2	89.4	90.3	89.5	88.9	89.2	93.8	92.5	93.1
Truck	85.3	86.7	86.0	84.8	85.9	85.3	89.5	90.8	90.1
Traffic Light	88.6	88.2	88.4	87.2	87.5	87.3	91.8	92.5	92.1
Fire Hydrant	51.2	78.9	62.0	49.5	77.8	60.4	76.5	89.5	82.5
Stop Sign	79.5	89.5	84.3	78.2	88.1	82.9	86.8	93.2	89.9
Parking Meter	49.8	80.3	61.5	47.9	79.5	59.6	74.2	90.5	81.5
Average	**90.6**	**87.6**	**88.9**	**80.2**	**86.5**	**82.6**	**89.0**	**92.1**	**90.4**

Table 6.2: Performance of Point Cloud Classification Methods

Class	PointNet++			DGCNN			PointNet			Houghes Net		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Person	85.6	82.8	83.8	44.4	75.1	57.2	21.9	75.3	33.9	67.07	82.1	73.5
Bicycle	81.4	79.3	79.9	37.3	72.7	50.6	18.5	72.1	29.7	53.85	79.8	64.6
Car	85.4	81.9	83.2	41.3	79.2	55.5	12.5	80.6	21.7	60.33	85.4	70.2
Motorcycle	84.6	83.0	83.4	47.0	76.6	59.6	32.2	75.4	44.6	62.24	84.7	71.8
Bus	84.3	81.5	82.5	50.1	77.9	62.3	32.8	76.3	45.6	70.11	86.5	77.5
Truck	78.4	78.8	78.2	39.1	74.1	52.5	12.8	73.5	21.7	65.58	81.2	72.3
Traffic Light	81.7	80.3	80.6	35.0	68.4	47.5	2.5	65.7	5.3	69.85	77.3	73.4
Fire Hydrant	44.3	71.0	54.2	20.1	55.6	30.8	6.7	62.0	12.1	32.09	54.6	40.5
Stop Sign	72.6	81.6	76.5	47.0	76.6	59.6	42.5	81.2	55.7	54.04	78.3	64.2
Parking Meter	42.9	72.4	53.7	18.9	53.9	29.3	4.1	60.1	7.7	32.79	52.7	40.4
Average	**74.1**	**79.3**	**75.6**	**38.0**	**71.0**	**50.5**	**14.92**	**72.2**	**27.2**	**63.01**	**78.1**	**68.1**

notable difficulties were detecting and classifying Fire Hydrants and Parking Meters, where KD-Y4 achieved only 51.2% and 49.8% precision respectively. These low-performance metrics can be attributed to several factors: the relatively small size of these objects in the point cloud data, their sparse point representation due to occlusion by other urban elements, and their structural similarity to other street furniture. While KD-Y7 significantly improved the classification of these objects, they remain the most challenging categories in our classification task. This difficulty persists across all tested architectures, suggesting an inherent challenge in distinguishing these urban elements from point cloud data alone. In these cases, the improvement shown by KD-Y7 demonstrates that knowledge distillation from the image domain can help overcome some limitations by leveraging the rich feature representations learned from high-resolution image data. Regarding computational performance, both models were evaluated on an Azure Machine Learning instance with 4 CPU cores, 32GB RAM, and NVIDIA T4 GPU acceleration. The KD-Y7 model required 3 hours and 35 minutes for training on the complete nuScenes dataset, while PointNet++ required 4 hours and 15 minutes. Although the training process involves additional computational steps due to the knowledge distillation mechanism, KD-Y7 demonstrates superior inference efficiency, processing the complete test set in 35 seconds, compared to 50 seconds for PointNet++. This represents a 30% improvement in inference speed. The enhanced

Table 6.3: Benchmark Comparison of Best Models

Best Models	P (%)	R (%)	F1 (%)
KD-Y7	89.0	92.1	90.4
PointNet++ (Traditional)	74.1	79.3	75.6

inference performance, combined with the superior classification accuracy shown in 6.3, establishes our approach as a viable solution for real-time applications where accuracy and processing speed are critical factors.

6.2 Limitations and Negative Results

Throughout development, we conducted an extensive series of experiments to determine whether knowledge distillation could compress the teacher network while preserving detection performance. These tests explored various architectures, optimizers, and hyperparameters, and many configurations produced unsatisfactory results, illustrating important limitations of our approach.

Early trials with large models. Our first experiments used a YOLOv4 teacher with $\sim$ 65 million trainable parameters and a student network of comparable size ($\sim$ 55 million parameters). The student was trained exclusively on the teacher's logits (with the teacher frozen). Despite the substantial model size and high learning rate, the student's accuracy stagnated below 50%, and training curves showed oscillatory or divergent losses. Attempts to reduce the student's complexity (from 55 million down to 44 million, 50 million, and even 10 million parameters) consistently degraded performance, highlighting that naïvely shrinking the model was not effective.

Exploration of compact architectures. We then designed a lightweight student with $\sim$ 2.24 million parameters. This network required over 100 epochs to converge and still achieved only $\sim$ 16.3% accuracy on the test set. The distillation loss remained high and oscillatory, suggesting that the student was unable to extract meaningful features from the teacher. These results underscore the difficulty of transferring rich representations to very shallow students.

Hyper-parameter tuning and optimizer search. Subsequent experiments varied the teacher's training duration (100 vs. 200 epochs), batch size, learning rate, and momentum. For example, increasing the teacher's epochs to 200 with a learning rate of 0.1 and momentum of 0.9 yielded a teacher accuracy of 82.3%, yet the corresponding student accuracy remained far lower. A grid search over optimizers (Adam, Adagrad, Adadelta), dropout rates, and temperature/α settings for distillation revealed a narrow band of configurations where the student accuracy rose modestly. In particular, tuning the distillation temperature and the weight α on the distillation loss increased the student's accuracy from $\sim$ 43.3% to $\sim$ 47.5%, and finally to $\sim$ 51.6%. Nonetheless, even the best configuration fell well short of the teacher's performance, and many trials resulted in "random features" with low accuracy.

Sensitivity to distillation parameters. Changing the distillation temperature and α affected convergence significantly. Low temperatures produced erratic training curves and negligible improvements, while higher temperatures sometimes increased the student loss. Only by carefully balancing α and temperature did we achieve a slight performance gain. This sensitivity suggests that knowledge distillation can be brittle: without careful tuning, the student may learn noise or overly smooth distributions instead of informative soft targets.

Overall, these negative results highlight that compressing large object-detection models via knowledge distillation is non-trivial. Many naive or small-student configurations failed, and only a subset of hyper-parameter combinations yielded modest improvements. These findings motivated us to refine the architecture and training strategy described in the main results section.

6.3 Experimental Fairness and Reproducibility

All experiments were conducted under consistent conditions to allow a fair comparison across models and distillation strategies. We used the same training/validation/test splits for every trial, identical data augmentation and preprocessing pipelines, and evaluated performance with the same metrics (accuracy, precision, recall, and F1-score). The hardware environment (GPU model and memory) and software stack (Python version, deep-learning framework, random seeds) were kept fixed. Only one hyper-parameter (e.g., optimizer or temperature) was varied at a time while holding all others constant, ensuring that any differences in outcome were attributable to the change under study. These controlled conditions support the conclusion that the observed variability stems from the models and distillation parameters themselves rather than from external factors.

7 | Conclusions

The evolution of this research, from signal generation through adversarial networks to transformer-based knowledge distillation, demonstrates the effectiveness of deep learning architectures in addressing cross-domain challenges. Our initial work with Deep Convolutional Generative Adversarial Networks (DCGANs) for partial discharge signal synthesis established crucial principles that directly informed our current approach to knowledge distillation. The success in maintaining temporal and spectral characteristics during signal generation provided valuable insights into preserving domain-specific features during knowledge transfer between image and LIDAR domains.

The implementation of knowledge distillation techniques in our current work builds upon several key discoveries from our GAN research. The ability of adversarial architectures to capture and reproduce complex signal characteristics translates directly to the challenge of cross-domain feature alignment. Our experience with GAN-based signal generation revealed that successful domain adaptation requires careful attention to both the global structure and local details of the data, a principle we incorporated into our transformer-based distillation approach through multi-head attention mechanisms.

Validation methodologies developed for assessing GAN-generated signals proved invaluable in designing evaluation frameworks for cross-domain knowledge transfer. The spectral power clustering technique (SPCT) used to validate synthetic PD signals informed our approach to evaluating feature preservation in LIDAR point cloud classification. This methodological transfer resulted in more robust validation procedures, as evidenced by our comprehensive performance metrics across different object categories and environmental conditions. The architectural insights gained from implementing DCGANs significantly influenced our current transformer-based design. The importance of maintaining proper gradient flow and feature hierarchy in GANs directly parallels the challenges in cross-domain knowledge distillation. Our success in achieving 90.4% F1-score in object detection and 30% improvement in inference speed builds upon lessons learned about efficient feature extraction and representation learning from our GAN research.

The implementation of knowledge distillation techniques in this study has validated their effectiveness in reducing the computational demands required to deploy complex models and confirmed their capacity to maintain performance comparably to the original models trained on high-quality data sources. This

efficient utilization of computational resources is pivotal, particularly in real-time systems where processing power could be a constraint. Moreover, the ability to transfer knowledge from high-quality image data to comparatively lower-quality LIDAR point clouds requires a significant improvement in model performance and, as a result, enhancement of the utility of less detailed data sources.

This research particularly emphasizes the trans-domain adaptability of knowledge distillation techniques. By successfully applying these techniques between the domains of RGB images and LIDAR point clouds, the study introduces a novel approach to leveraging the strengths of existing technologies in new, diverse settings. This broadens the scope of applications for established machine learning models and fosters the development of innovative applications in fields hampered by data type disparities.

In addition to the demonstrated effectiveness of knowledge distillation techniques in leveraging diverse data sources, it is important to note the complexity of experimental setup across trans-domain differences. By benchmarking the complexity of trans-domain adaptation, it becomes evident that knowledge distillation not only enhances model performance but also facilitates the integration of machine learning models across disparate data types.

While the impact of knowledge distillation on diverse data sources has been evident, the optimization of parameters such as alpha and temperature can be significantly improved. With different alternatives to grid search for these specific parameters, the adaptability of machine learning models across diverse domains can be further optimized. This improved grid search methodology could not only enhance the performance of models but also streamline the process of integrating machine learning models across disparate data types. The integration of an advanced grid search for parameters related to trans-domain complexity should provide a more robust framework for leveraging knowledge distillation techniques in diverse settings. Finally, the trans-domain potential of this architecture extends beyond LIDAR and image classification, potentially impacting fields facing similar disparities in data availability, frequency, richness, and diversity.

A limitation of our approach is the reliance on high-quality image datasets, pre-trained image models, or any other high-quality teacher network dataset. Future work will focus on adapting the method for environments with fewer annotated datasets for the teacher network, like the images in this study, and further exploring self-supervised learning techniques to enhance model robustness.

Bibliography

[1] Johann Laconte, Abderrahim Kasmi, Romuald Aufrère, Maxime Vaidis, and Roland Chapuis. A survey of localization methods for autonomous vehicles in highway scenarios. *Sensors*, 22(1), 2022. 1

[2] Sayash Kapoor, Peter Henderson, and Arvind Narayanan. Promises and pitfalls of artificial intelligence for legal applications, 2024. 1

[3] Nicole Gross. What chatgpt tells us about gender: A cautionary tale about performativity and gender biases in ai. *Social Sciences*, 12(8), 2023. 1

[4] Sun Park, Chan-Su Yang, and JongWon Kim. Design of vessel data lakehouse with big data and ai analysis technology for vessel monitoring system. *Electronics*, 12(8), 2023. 1

[5] Simrat K Gill, Andreas Karwath, Hae-Won Uh, Victor Roth Cardoso, Zhujie Gu, Andrey Barsky, Luke Slater, Animesh Acharjee, Jinming Duan, Lorenzo Dall'Olio, Said el Bouhaddani, Saisakul Chernbumroong, Mary Stanbury, Sandra Haynes, Folkert W Asselbergs, Diederick E Grobbee, Marinus J C Eijkemans, Georgios V Gkoutos, Dipak Kotecha, BigData@Heart Consortium, and the cardAIc group. Artificial intelligence to enhance clinical value across the spectrum of cardiovascular healthcare. *European Heart Journal*, 44(9):713–725, 01 2023. 1

[6] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023. 1

[7] Greg Welch and Gary Bishop. An introduction to the kalman filter. Technical Report TR 95-041, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 1995. 1

[8] Ninad Mehendale and Srushti Neoge. Review on lidar technology. *SSRN Electronic Journal*, 2020. 1

[9] Man Lok Fung, Michael Z. Q. Chen, and Yong Hua Chen. Sensor fusion: A review of methods and applications. In *2017 29th Chinese Control And Decision Conference (CCDC)*, pages 3853–3860, 2017. 1

[10] Youngmoon Kim, Sungpill Jang, Taesu Kim, and Sungroh Lee. Knowledge distillation for small-footprint speech recognition models using teacher-student training. *IEEE Signal Processing Letters*, 27:845–849, 2020. 1

[11] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016. 1

[12] Ilija Radosavovic, Piotr Doll'ar, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4119–4128, 2018. 1

[13] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1285–1294, 2017. 1

[14] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3996–4003, 2020. 1

[15] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 1

[16] Quanquan Li, Zhengxia Wen, and Bin He. Mimicking very efficient network for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6356–6364, 2017. 1

[17] J. A. Ardila-Rey, J. E. Ortiz, W. Creixell, F. Muhammad-Sukki, and N. A. Bani. Artificial generation of partial discharge sources through an algorithm based on deep convolutional generative adversarial networks. *IEEE Access*, 8:24561–24575, 2020. 1

[18] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, March 2021. 3

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2014. 3

[20] Man Lok Fung, Michael ZQ Chen, and Yong Hua Chen. Sensor fusion: A review of methods and applications. In *2017 29th Chinese Control And Decision Conference (CCDC)*, pages 3853–3860. IEEE, 2017. 3

[21] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018. 3

[22] Xiaojie Wang, Fisher Yu, Lisa Chen, and Gonzalo Gallego. Kdgan: Knowledge distillation with generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 783–794, 2018. 3

[23] Jorge A Ardila-Rey, Jesus Ortiz, Werner Creixell, Firdaus Muhammad-Sukki, and Nurul Aini Bani. Artificial generation of partial discharge sources through an algorithm based on deep convolutional generative adversarial networks. *IEEE Access*, 8:24561–24575, 2020. 3

[24] Wei Tan, Xiangang Qin, Lei Yu, Zhixiang Wu, and Hao Yue. Deep learning-based 3d point cloud classification: A systematic survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[25] Wei Tan, Lei Zhang, Yang Liu, and Tao Wang. Transform domain learning for image recognition. *IEEE Transactions on Image Processing*, 2024. 3

[26] Jesus Ortiz and Werner Creixell. Advanced trans-domain knowledge transfer through transformer-based distillation: A novel framework for image-lidar integration in autonomous systems. *IEEE Access*, 0:1–10, 2025. 3

[27] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 3

[28] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online, November 2020. Association for Computational Linguistics. 3

[29] Taohua Zhou, Kun Jiang, Zhongyang Xiao, Chunlei Yu, and Diange Yang. Object detection using multi-sensor fusion based on deep learning. In *CICTP 2019*. American Society of Civil Engineers, July 2019. 3

[30] Song Wu, Yicheng Liu, and Junsong Yuan. Deep sensor fusion for multi-modal human action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[31] Abdolmaged Alkhulaifi, Fahad Alsahli, and Irfan Ahmad. Knowledge distillation in deep learning and its applications. *PeerJ Computer Science*, 7:e474, April 2021. 3

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 3

[33] R. Liu, K. Zhang, and J. Wang. Multi-modal knowledge distillation for point cloud classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[34] L. Chen, R. Wang, and Y. Zhang. Adaptive sensor fusion through knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[35] S. Park, J. Kim, and H. Lee. Temporal-aware knowledge distillation for sensor fusion. *IEEE Transactions on Intelligent Transportation Systems*, 24(5), 2023. 3

[36] J. Kim, S. Park, and H. Lee. Cross-modal knowledge transfer in autonomous systems. *IEEE Transactions on Robotics*, 39(4), 2023. 3

[37] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 689–696. Omnipress, 2011. 3

[38] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 3

[39] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018. 3

[40] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13–23, 2019. 3

[41] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–816, 2016. 3

[42] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 570–586, 2018. 3

[43] Songyou Peng, Sebastian Hamann, Thomas Stieglitz, and Alexander Meder. Cross-modal learning for 3d shape recognition. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 1–9. IEEE, 2019. 3

[44] Di Wang, Quan Wang, Le An, Bing Xu, and Yiming Tang. Cross-modal hashing for multimedia search. *Pattern Recognition*, 100:107111, 2020. 3

[45] Linfeng Li, Weixing Su, Fang Liu, Maowei He, and Xiaodan Liang. Knowledge fusion distillation: Improving distillation with multi-scale attention mechanisms. *Neural Processing Letters*, 55(5):6165–6180, January 2023. 3

[46] Zhihui Li, Pengfei Xu, Xiaojun Chang, Luyao Yang, Yuanyuan Zhang, Lina Yao, and Xiaojiang Chen. When object detection meets knowledge distillation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10555–10579, 2023. 3

[47] Dengyu Xiao, Yixiang Huang, Chengjin Qin, Zhiyu Liu, Yanming Li, and Chengliang Liu. Transfer learning with convolutional neural networks for small sample size problem in machinery fault diagnosis. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 233(14):5131–5143, March 2019. 3

[48] C.P. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 555–562, 1998. 3

[49] D. Lu and Q. Weng. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, 28(5):823–870, March 2007. 3

[50] J. Kim, S. Park, and H. Lee. Real-time multi-modal perception for autonomous vehicles. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 3

[51] Markus J. Buehler. Unsupervised cross-domain translation via deep learning and adversarial attention neural networks and application to music-inspired protein designs. *Patterns*, 4(3):100692, March 2023. 4.1

[52] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017. 4.1

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 4.1

[54] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 4.1

[55] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 4.1

[56] Zhenwei Shen, Yi He, Xianming Du, Jing Yu, Hong Wang, and Yuxiao Wang. Ycanet: Target detection for complex traffic scenes based on camera-lidar fusion. *IEEE Sensors Journal*, 2024. 4.1

[57] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 5

[58] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *International conference on machine learning*, pages 6105–6114, 2019. 5

[59] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. 5

[60] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018. 5

[61] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurélien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2443–2451, 2020. 5

[62] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 5.1