BABEŞ-BOLYAI UNIVERSITY
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Development of Efficient Methods of Deep Learning in Computer Vision

PhD thesis summary

PhD student: Tudor Alexandru G. ILENI
Scientific supervisor: Prof. dr. Anca M. ANDREICA

2024

Keywords: Computer Vision, Deep learning, machine learning, neural architecture search, dynamic pruning, knowledge distillation

Contents

Thesis table of contents

List of publications

The ranking of publications was performed according to the CNATDCU (National Council for the Recognition of University Degrees, Diplomas and Certificates) standards applicable for doctoral students enrolled after October 1, 2018. All rankings are listed according to the classification of journals[1] and conferences[2] in Computer Science.

Publications in Web of Science - Science Citation Index Expanded

[IDBM22] T.A. Ileni, A.S. Darabant, D.L. Borza, A.I. Marinescu. *DynK-hydra: improved dynamic architecture ensembling for efficient inference*. Complex and Intelligent Systems, pp.1-12, 2022

Rank Q1, 2022 IF=35, 4 points.

[BDIM22] D.L. Borza, A.S. Darabant, T.A. Ileni, A.I. Marinescu. *Effective Online Knowledge Distillation via Attention-Based Model Ensembling*. Mathematics, 10(22), p.4285, 2022

Rank Q1, 2022 IF=2, 4 points.

[NID20] S.C. Nistor, T.A. Ileni, A.S. Dărăbant. *Automatic development of deep learning architectures for image segmentation*. Sustainability, 12(22), p.9707, 2020

Rank Q2, 2020 IF=53, 2 points.

Publications in Web of Science, Conference Proceedings Citation Index

[MID19] A.I. Marinescu, T.A. Ileni, A.S. Darabant. *A versatile 3d face reconstruction from multiple images for face shape classification*. In 2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), pp. 1-6, IEEE, 2019.

Rank B-Conference, 4 points.

[IBD19] T.A. Ileni, D.L. Borza, A.S. Darabant. *Fast In-the-Wild Hair Segmentation and Color Classification*. In VISIGRAPP (4: VISAPP), pp. 59-66, 2019.

Rank B-Conference, 4 points.

[1] https://uefiscdi.ro/premierea-rezultatelor-cercetarii-articole
[2] http://portal.core.edu.au/conf-ranks/

[BID18] D. Borza, T.A. Ileni, A. Darabant. *A deep learning approach to hair segmentation and color extraction from facial images*. In Advanced Concepts for Intelligent Vision Systems: 19th International Conference, ACIVS 2018, Poitiers, France, September 24–27, 2018, Proceedings 19, pp. 438-449, Springer International Publishing.

Rank B-Conference, 4 points.

[MDI20] A.I. Marinescu, A.S. Darabant, T.A. Ileni. *A Fast and Robust, Forehead-Augmented 3D Face Reconstruction from Multiple Images using Geometrical Methods*. In 2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), pp. 1-6, IEEE, 2020.

Rank C-Conference, 2 points.

[MDI21] A.I. Marinescu, A.S. Darabant, T.A. Ileni. *Optimal Stereo Camera Calibration via Genetic Algorithms*. In IJCAI 2021 - International Joint Conference on Artificial Intelligence, Workshop, 2021. ISBN:978-0-9992411-9-6

Rank A-workshop, 1 point.

[TA20] T.A. Ileni. *Efficient iris segmentation and pupil detection for visagisme applications*. In 2020 IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI), pp. 000123-000128, 2020.

Rank C-Conference, 1 point.

Publications score: 26 points.

Chapter 1

Introduction

1.1 Motivation and hypotheses

The work presented in this Thesis follows two tracks of problems. On one hand, we tackle the task of face analysis, in both 2D and 3D space. On the other hand, we approach the task of inference optimization via ensemble dynamic pruning, knowledge distillation, and neural architecture search. Understanding the human face has applications in fields like biometrics (face detection and re-identification), medicine (diagnostics based on eye color evolution), the fashion industry (accessories matching), and human-machine interaction. Modern approaches to face analysis, and Computer Vision in general, employ deep learning models which require large sets of training data, and high computational costs. Hence, creating new datasets and reducing the inference costs are of high interest to us. To sum up, we wanted to validate the following hypotheses from both theoretical and applicative perspectives:

- Is it possible to train a robust and fast face analysis model (hair/face segmentation, pupil detection, hair color classification) and how much data do we need to reach a proper accuracy?
- Starting from a generic 3D face model, can we deform it to create a face of a specific subject in 3D?
- How can we model a genetic algorithm to solve the camera calibration input selection problem?
- By what means can we automatically generate neural network architectures tailored for specific tasks?
- To what degree can we reduce the inference cost in a large network using dynamic pruning during forward pass?
- How much can a lightweight model learn (via Knowledge Distillation) from a larger network and boost its performance?

1.2 Objectives

Starting from both theoretical and applicative motivations, and hypotheses, we draw the following objectives for this Thesis.

- Integration of image processing and deep learning for face analysis tasks. Combine both image processing and deep learning approaches for face analysis. Compare the benefits and drawbacks of image processing methods with deep learning processes in the context of face analysis.

- Developing robust and efficient face analysis models. Train a robust and fast face analysis model for tasks such as hair/face segmentation, pupil detection, and hair color classification and investigate the impact of dataset size on the accuracy of face analysis models.
- Integrating ML models in various tools and applications. Apply face analysis models to specific domains such as biometrics, medicine, and the fashion industry. Explore the deformation of a generic 3D face model to create a face specific to an individual.
- Reducing inference costs in deep learning models. Explore techniques for reducing inference costs in large networks using dynamic pruning during the forward pass. Investigate the extent to which lightweight models can learn from larger networks via knowledge distillation.
- Optimizing searching and computational costs. Design and implement a recurrent network controller for NAS to automatically generate deep learning architecture blocks. Model a genetic algorithm to enhance the camera calibration process.

1.3 Original Contributions

More specifically, in the face analysis area, we developed a hair analysis tool that includes hair segmentation and color classification, a pupil and iris detector, and a pipeline for 3D face reconstruction. To optimize the computational costs we study and develop an ensemble of networks dynamically pruned based on the input, an online distillation mechanism to create accurate lightweight performant *students*, and a recurrent network that predicts suitable block architecture for a specific task.

A great deal of research has focused on face analysis using 2D (planar) images [Kel70, VJ01, KKA+20]. Regardless of the method being used, some problems are present in the context of *face-in-the-wild* 2D captures such as face pose lack of depth data and illumination condition.

Face analysis tasks are a subsection of Computer Vision (CV), in which the classification algorithms are adapted for human faces, for extracting or classifying several attributes like hair color, face shape, eye color, face ID, head pose, identity, etc.

For the face analysis work, both in 2D and 3D space, the original contributions of the Thesis are:

- Manually annotate and make publicly available 3.5k images containing the face and hair mask, 20k annotation of hair color, and 300 samples of the eye contour [BID18, IBD19, TA20].
- Evaluate face/hair segmentation deep neural networks analyze the results using multiple classifiers, and compare the performances with state-of-the-art models [TA20].
- Develop and fine-tune an iris and pupil detection algorithm, which deals with multiple types of noise, using both Image Processing and machine learning techniques, and compare the results with similar methods from literature.
- Train a hair color classifier, using the hair area only, on a balanced dataset with the following taxonomy: red, brown, grey, black, and blonde. Analyze the impact of the used colorspace, and compare the results with the current state of the art, obtaining comparable or better results [BID18, IBD19].
- Develop a 3D face reconstruction pipeline model that reconstructs the entire face area, not limited to the forehead, as most of the state-of-the-art models [MDI20, MID19].

In what follows, we explain in more detail the contributions. For the face analysis task, we first train a head segmentation Deep Neural Network, which infers the area of the face and the hair. Because the research community lacks a large annotated dataset we manually annotate more than 3.5k images to feed our predictor. For the segmentation task, we achieved a mean pixel accuracy of 92.1% for frontal face pictures, which is higher than similar works (at the moment of publishing 91.5% for Figaro1k dataset [MSLB18]).

Using the hair area we feed another predictor, this time a three-layer Artificial Neural Network (ANN) to classify the hair color of the subject. The best results were obtained when we fed the ANN with the histograms of color computed on the LAB colorspace. For each of the three input channels, we compute a histogram of size 256/8=32, that is concatenated and normalized. The hair taxonomy includes black, brown, blond, grey, and red classes. To enlarge the available hair color dataset we annotate more than 20k images. For hair color classification our accuracy (of 89.6%) is slightly better compared with the state-of-the-art of 88.6% at the time of publishing. As inconvenient most similar works also have unbalanced datasets.

While analyzing the face, we are also interested in the eyes area. For this, we develop and fine-tune a detection algorithm for the iris and pupil. The pipeline starts with a regressor (trained with 300 manually annotated samples) which predicts the corners of the eyes, the pupil, the eyelids, and a point on the iris border. Using those points we tune an Image Processing algorithm that identifies the pupil as the darkest area of the eye, and the iris using a threshold-based binary image.

To alleviate the previously mentioned problems in 2D space as the lack of depth data and illumination conditions, the researchers and the industrial applications adopt 3D face models in their analysis. In the broader context of 3D capturing, the following techniques are used: stereo-vision systems (requiring calibrated or non-calibrated cameras) [ARL+10], RGB-D cameras (which also recover the depth information), 3D laser scans [LTW95], 3D-from-2D methods (usually in non calibrated camera scenarios). In our work, for 3D face reconstruction, we are proposing a statistical model-fitting algorithm based on Structure from Motion and a deformable model.

The reconstruction pipeline starts with detecting the face contour using a state-of-the-art face alignment regressor, to which we add five more points from the upper area of the forehead (extracted using our face segmentation network). Most of the researchers limit the 3D face reconstruction up to the eyebrow line or the middle of the forehead. Using multiple acquisitions of the subject we construct a 3D point cloud of the face landmarks (using Structure from Motion) and then we deform a base 3D face model to obtain the final result. Compared with other work we obtain similar results, having the benefit of a fully 3D face reconstruction.

Still, in the 3D modeling field, we are developing a parallel genetic algorithm that aims to enhance the camera calibration process by selecting the most appropriate subset of acquisitions of a pattern (usually a chess board) required as input for the calibration algorithm.

In this endeavor, we model the chromosome as a binary array having a length equal to the number of acquisitions. If an element is true, the corresponding photograph is used as input to the calibration algorithm. During the population evolution, we use mutation and crossover operators.

For the second track (optimization methods in deep learning) of our research we tackle the following problems: neural architecture search, knowledge distillation, and dynamic pruning. We list the original contributions in the field:

- Improve the robustness of the camera calibration process, by designing a parallel genetic algorithm to select the optimal input for the calibration algorithm (captures of a pattern). Compared to state-of-the-art models we get improved performances [MDI21].

- Designing and training a self-pruning deep learning architecture that reduces the inference

costs by more than 2x compared to similar work and more than 5x compared to classical deep networks [IDBM22].

- In the field of Neural Architecture Search we implement a recurrent network controller that predicts architecture blocks for specific Computer Vision tasks. The experiments were performed to study the energy consumption impact in the process of searching for a suitable architecture. [NID20].

- Boost the performances of light networks through a process of Online Knowledge Distillation, by using a cohort of students trained in parallel. We obtain a gain in accuracy of more than 2% while keeping the same number of parameters [BDIM22].

We propose a neural architecture search (NAS) controller to automatically create deep learning architecture blocks, to alleviate the problem of manually choosing a suitable model. The research community has proposed multiple ways to touch these problems including combining multiple weak learners (boosting [DCJ+94], bagging [Bre96], stacking algorithm), training multiple models on the same dataset, and combining/selecting their predictions to get a better result [HLP+17, BWHY05, FHL19] (i.e. Ensemble Learning), training a larger model and pruning it for specialized inference [CGW+19], dynamic inference using a slice of the network [MMSF18], dynamically creating a task-specialized architecture (i.e. Neural Architecture Search) [EMH18, BGNR16, ZL16], knowledge distillation to create a lightweight model [HVD15, LZG18, GYMT20].

We found the Neural Architecture Search (NAS) approach very promising. We developed a recurrent predictor to generate cells (group of layers) that are further stacked into deeper templates. We test the network discovery pipeline to an instance segmentation task, and we get a high IoU score.

Optimizing the computational cost (especially for inference) is a research topic of high interest nowadays. With the trade-off of needing more resources during training (time and memory), tasks such as Knowledge Distillation, dynamic neural networks, or a Mixture of Experts are trying to achieve this goal. Our approach is an online attention-based distillation where a cohort of students are trained together and their output is dynamically combined to create a more powerful teacher. This output is further distilled to the individual components. The results (expressed in accuracy gain) exceed or are at least comparable to the current state-of-the-art approaches.

On the other hand, dynamic networks (or dynamic pruning) [HHS+21] are an improvement of the classical artificial neural networks, where only a part of the network is executing, based on some halting/scoring conditions, or only a part of the input is considered. In this work, we are creating an ensemble, composed of multiple (convolutional neural network) branches, having a common stem and a gater, which decides which branches to activate for each input sample dynamically. In this manner, the number of inference flops (floating-point operations per second) is considerably reduced.

In this process, we divide the input space into multiple partitions and assign a branch model (from the ensemble) to each partition. This *assignment* is done via a loss function which encourages each branch model to be specialized on the specific input space.

Chapter 2

Background

In this Chapter we are providing the general context and background of the notions we approached in the Thesis. Our discussion is about the traditional and Deep Learning methods used in computer vision, especially on the task of face analysis, and optimization in Deep Learning as Neural Architecture Search, Knowledge Distillation and Dynamic pruning.

2.1 Traditional Image Processing Methods and Deep Learning Techniques in Computer Vision

Traditional image processing algorithms in face analysis are developed for various tasks like face detection, recognition, hair segmentation, and iris detection. These methods involve using specific algorithms for each task. For instance, in face recognition [Kel70], methods like background subtraction, template matching, dynamic threshold smoothing, and edge detection are employed. The Haar Feature-based Cascade Classifiers [VJ04] are a notable example, which combine image processing techniques and machine learning. This method constructs integral images, where each pixel's value is the sum of all pixels above and to its left, to rapidly compute features like Haar features. Feature selection is typically achieved through machine learning algorithms like AdaBoost [Sch13], which selects relevant features from a larger set to create weak learners that work collectively for tasks like face detection.

Deep learning approaches address some limitations of traditional image processing methods, such as the need for predefining and tuning algorithms for specific tasks and constraints related to image location and external hardware requirements. Deep learning methods typically involve selecting a mathematical model, usually a Convolutional Neural Network (CNN), gathering labeled data, and running optimization algorithms guided by a loss function. These methods automatically incorporate constraints like uniform texture distribution or the need for specialized devices. For instance, deep learning methods for hair segmentation [PN17, LCP+18] and hairstyle classification in unconstrained environments use CNNs to construct hair probability maps from image patches, categorized by classifiers like Random Forest. Other deep learning applications in head segmentation, hair color classification, and iris segmentation [WZL+19] leverage the ability of CNNs to learn features from data without explicit programming for each task. This approach offers more flexibility and adaptability in diverse and dynamic environments.

2.2 Common Approaches in 3D (Face) Reconstruction and Camera Calibration

In 3D face reconstruction, several techniques are employed, including stereo-vision systems [ARL+10], RGB-D cameras, 3D laser scans [LTW95], and 3D-from-2D methods. These methods can be further categorized into statistical model fitting, photometric systems, and deep learning approaches [MPS21].

The most common technique in statistical model fitting is the use of 3D Morphable Models (3DMM). In this approach, a generic face model, composed of geometry (vertices), albedo (reflection property), and texture (color), is altered based on the subject's facial characteristics to reconstruct a new 3D face that best fits the given photographs. This method requires a pre-defined 3D face model which is adapted to fit individual facial features.

Photometric models assume a Lambertian reflection model, where the brightness of a surface remains consistent regardless of the observer's viewpoint. These models deconstruct an image or set of images into normals, albedo, and color.

Deep Learning models for 3D-from-2D reconstruction use a single facial photograph to reconstruct the face. These models employ a learning volumetric model trained to infer the depth characteristics of the subject. While effective and straightforward, challenges arise when parts of the face are not present in the picture due to occlusions, non-frontal face angles, poor illumination, etc. The model then needs to "invent" the missing information, such as geometry and texture.

Camera calibration is a fundamental process in computer vision, especially in the context of 3D reconstruction. It involves determining the camera's internal (intrinsic) and external (extrinsic) parameters to accurately capture the geometry of the environment.

Among the first camera calibration methods [Bro71, Fai75] used an object having multiple 3D orthogonal planes, which are accurately represented in a 3D (digital) space. To eliminate the need for the expensive calibration device, self-calibration techniques appeared. This framework requires a moving camera around a static scene. Each captured frame provides two constraints on the camera intrinsic parameters. Al least three images are enough to recover both intrinsic and extrinsic parameters [Har94, LF97].

2.3 Overview of Neural Architecture Search

Neural Architecture Search (NAS) is pivotal in the evolution of machine learning algorithms, especially as data becomes increasingly varied and abundant. Traditional machine learning algorithms [KSH12, SZ14] are typically developed and fine-tuned by humans for specific tasks. However, as the need to apply these algorithms across various domains grows, there's a pressing need for an automated framework capable of creating and constructing these algorithms. This is where NAS plays a crucial role.

NAS involves three fundamental elements: the search space, the search strategy, and the performance estimation strategy. The search space includes all possible architectures that the NAS method considers. For example, conventional Convolutional Neural Network (CNN) architectures are constructed by layering convolutional and pooling layers to achieve the desired dimensions, balancing latency and accuracy. Modern NAS approaches, however, focus on creating smaller modules or cells [LCY13, SLJ+15] that serve as building blocks for the full CNN architecture, tailored to specific tasks.

The NAS search space can become extensive, encompassing various operations or layers and their

connections. Efficiently traversing this space is crucial. Random searches can be time-consuming and may not consider previously acquired information. To address these challenges, several NAS algorithms have been proposed, including evolutionary algorithms, reinforcement learning, fully-differentiable methods, and methods designed for fast inference, some considering real-device latency.

2.4 Ensemble Learning and Gating Mechanisms Optimization

Ensemble learning has proven its superiority over single predictors in both theoretical and empirical aspects for various tasks, including regression and classification. The concept of ensemble diversity is crucial, with ensembles categorized into four levels [SS97] based on the number of errors. These levels range from no coincident error (Level 1), where majority voting is always correct, to situations where no member is correct (Level 4). The diversity within an ensemble is key, and insights suggest that members of a Level 2 or Level 3 ensemble can be adjusted to improve its overall classification accuracy.

2.5 Model Compressing via Attention-Based Knowledge Distillation

Knowledge Distillation (KD) involves a smaller model (the student) learning from a larger model (the teacher) [BCNM06]. The student is trained and penalized based on how its outputs differ from the teacher's. A key advancement in KD was the introduction of a "soften" process in the teacher's output, using a temperature scaling factor on the softmax activation [HVD15]. This approach allows the student to learn the relationship between the actual class and similar ones. The aim is to distill the "dark-knowledge" from the teacher to the student, ensuring that the student model learns intricate relationships and nuances captured by the larger model.

Online Knowledge Distillation differs from offline distillation in that there is no pre-trained teacher model [ZXHL18, LZG18]. Instead, a cohort of students simultaneously trains and shares knowledge, each acting as both student and teacher to the others. This method involves multiple learners (students) sharing and distilling knowledge, which adds complexity but can lead to more robust learning outcomes.

Attention mechanisms in knowledge distillation are inspired by the human sensory system's ability to focus on essential elements. In the context of KD, attention mechanisms are used to redistribute the weights of a feature map. This approach emphasizes the most significant features and channels, directing the student model's learning towards these crucial aspects. It is particularly useful in computer vision tasks, where focusing on specific features or areas can significantly enhance performance.

Chapter 3

Facial Feature Extraction And Analysis

3.1 Face analysis method description

This chapter is structured into two main sections: Image Processing Methods and Deep Learning Approaches.

Image Processing Methods: This section details techniques for eye segmentation including pupil detection, eyelid and eye corner recognition, and iris segmentation. It addresses challenges such as eyelid coverage, non-uniform iris color, pupil dilation, and noise due to eyelashes and light reflections.

Deep Learning Approaches: This part covers advanced methods for hair segmentation and color classification using machine learning. It introduces fully convolutional neural networks (FCN) inspired by U-Net and VGG architectures for hair segmentation, and methods for hair color classification using both artificial neural networks (ANN) and Random Forest classifiers. This section also discusses the detection of baldness as an extension of hair segmentation.

In the first section, we detail an innovative approach for eye segmentation, focusing particularly on the detection and segmentation of the pupil, iris, eyelids, and eye corners. The techniques leverage a combination of traditional image processing methods and machine learning algorithms. Key steps include initial face and eye region detection, landmark detection for eye features, and refinement of pupil and iris segmentation. The chapter also addresses challenges like varying eyelid coverage, non-uniform iris color, pupil dilation, and interference from eyelashes and light reflections. These methods are significant for applications in facial feature analysis, especially in areas requiring precise eye feature detection. The section emphasizes the importance of accuracy and robustness in these techniques due to their wide range of applications, from biometric identification to emotional analysis.

The second section focuses on two methods for hair segmentation and hair color classification in facial images using machine learning techniques. It introduces two main deep learning methods for hair segmentation, exploring fully convolutional neural networks inspired by U-Net and VGG architectures. Additionally, the section covers hair color classification, presenting approaches that involve using hair color histograms with artificial neural networks and super-pixel classification with Random Forest Classifiers. By extending the hair segmentation architecture with a fully connected branch we also perform the classification of baldness. This section is crucial for understanding advanced deep-learning applications in facial feature analysis, specifically targeting hair attributes.

3.2 Face analysis numerical results

In this section, we provide a summary of the results obtained on the facial analysis tasks.

The method used for iris and pupil detection involved manual annotation of 200 images from the

Chicago face dataset. The evaluation results were reported in terms of mean square error and mean absolute error for each landmark, combining points for both the left and right eye. For detecting the eye contour we get the following errors expressed in terms of mean absolute error Pupil: 1.24 Radius: 2.46 Outer corner: 2.46, inner corner: 3.11, eye top: 3.65, eye bottom: 2.68. Considering the accuracy of point detections, calculated using a threshold of 0.05 for the worst cumulative error we have a value above 95%. This means that the error is less than the pupil size.

For hair segmentation, the two deep learning methods on fully convolutional neural networks (FCN) were inspired by U-Net and VGG architectures. For hair segmentation using FCN, a pixel accuracy of 84.75% was achieved on a subset of the Figaro 1k database, and an average pixel accuracy of 90.77% with a confidence interval (C.I.) of [86.43, 95.11], which is at least at par with similar method on the literature [MSLB18]. To train and evaluate the segmentation models we manually annotate over 3.5k images and make them publicly available.

A probabilistic model was proposed for baldness detection and finetuned on over 6500 images from the CelebA dataset. The model's overall accuracy for the baldness detection task was reported to be 93.33%, with other metrics like precision, recall, and f1-score all at 93.50. The confidence interval at 95% significance level is [91.3, 95.3].

The performance of the ANN-based hair color classification module for different colorspaces and feature vector sizes was reported, with the best results obtained using the LAB colorspace with 8 bin sizes. For instance, the accuracy percentages for different colorspaces with bin size 111 were: RGB - 87.80%, HSV - 88.10%, and LAB - 88.30%. In a non-black hair scenario, the method attained a hair color recognition accuracy of 89.6% with a confidence interval of [87.16%, 92.04%], surpassing other works in the literature [KPR+14]. For this task, we manually annotate 20k images for training and 2k images for testing.

Chapter 4

Automatic and Dynamic Development of Deep Learning Networks

In this Chapter we present our fundamental research on Deep Learning optimization. The proposed methods attack different stages of working with DL models, from architecture search, to training, and inference. The numerical results are promising, surpassing the current State of the Art approaches.

4.1 Neural Architecture Search for Optimal Specialized Network

NAS is an automated process for designing neural network architectures. The main objective of NAS in this context is to identify the most efficient and effective architecture for a given task. This approach is particularly relevant for tasks that require specialized network configurations. The goal is to optimize the architecture to enhance performance metrics like accuracy, while also considering computational efficiency.

This section introduces a new method for automatically searching and optimizing CNN architectures for specific tasks, like eyeglasses segmentation. The approach uses a Recurrent Neural Network (RNN) based on reinforcement learning to generate efficient cell configurations for CNNs. These cells form the building blocks of the final CNN architecture. The structure of these cells is represented as a directed acyclic graph, optimizing for performance with varying operations like convolution and dilated convolution.

The process involves training the RNN with reinforcement learning to maximize a reward function, which is generally related to the performance of the generated architectures on a validation dataset. Over time, the RNN learns to propose increasingly effective architectures. After the search process, the best-performing architecture is trained from scratch to evaluate its performance on the target task.

The primary advantage of NAS is its ability to automate the design of neural network architectures, which can be a highly complex and time-consuming task when done manually. For example for 8 possible operations, and cell size of 5 nodes we may have 327.680 possible combination to create a cell. By automating this process, NAS can potentially discover novel and highly efficient architectures that might not be apparent through manual exploration. Additionally, NAS can be tailored to optimize networks not just for accuracy, but also for other constraints like computational efficiency, making it suitable for deployment in resource-constrained environments. For a specific task of eyeglasses segmentation our generated cell, instantiated in a larger network template, perform at 0.96 IoU.

4.2 Dynamic Pruning in Ensembles using Gating Mechanism

Dynamic pruning in ensembles is an advanced technique in neural network optimization. Its primary purpose is to enhance the computational efficiency and accuracy of deep learning models. This is achieved by selectively activating only relevant parts of an ensemble of neural networks based on the input. The technique is particularly useful in situations where different network branches are specialized for different sub-tasks or classes.

The thesis proposes the DynK-Hydra framework, which enhances deep neural networks for specific tasks by dividing classes into clusters and specializing network branches for these subtasks. A gating mechanism dynamically chooses the most relevant branches for each input, improving both accuracy and speed. This framework is tested on several datasets, showing promising results in terms of efficiency and accuracy.

The core of this approach is the DynK-Hydra framework, which divides the overall task into clusters, with each cluster being addressed by a specialized branch of the network. This framework is designed to dynamically choose the most relevant branch for each input during inference, thereby reducing unnecessary computations that are typical in large, monolithic models.

At the heart of this dynamic pruning approach is the gating mechanism. This mechanism functions as a smart switch that determines which branches of the ensemble should be activated for a given input. The decision is based on the characteristics of the input data, ensuring that only the most relevant and specialized segments of the network are employed. This process leads to a significant reduction in computational overhead, as only a fraction of the entire ensemble is active at any given time.

This framework is tested on several datasets, demonstrating its effectiveness. These results highlight the efficiency gains in terms of speed and reduced computational requirements, along with improvements or maintenance of high accuracy in task performance. For example, for CIFAR-100 [KH+09] DynK-Hydra achieved approximately 74% accuracy with 139M FLOPs, demonstrating a 2.7x times improvement over HydraRes [MMSF18], which required 378M FLOPs for similar accuracy. In comparison, the largest ResNet architecture achieved 73.56% accuracy with 767M FLOPs, making DynK-Hydra approximately 5.5 times more efficient in terms of inference time. On the Tiny-ImageNet [LY15] dataset, a 1.43% gain in accuracy was achieved using the proposed framework. The confidence interval widths for vanilla and KD experiments were [51.94, 53.90] and [53.37, 55.33], respectively.

4.3 Knowledge Distillation via Attention Based Learning

This section introduces an innovative approach for optimizing deep neural networks. This approach aims to create a lightweight model that mimics the behavior and performance of a larger model. It utilizes an online distillation mechanism where the supervisor is a weighted combination of peer students and the knowledge (supervisor output) is distilled back to the students. After training, only one (more accurate) student is used, thus achieving high accuracy while maintaining a low memory footprint and execution time.

The online KD framework treats multiple models as students, learning from each other's predictions. The attention mechanism dynamically weights each student's output in the final ensemble. The experiments were conducted using various well-known network architectures and on several image classification datasets. The classical cross-entropy loss was used for training the individual "vanilla" models before training the same network architecture with the KD framework.

The online KD framework treats multiple models as students, learning from each other's predictions. The attention mechanism dynamically weights each student's output in the final ensemble. The experiments were conducted using various well-known network architectures and on several image classification datasets. The classical cross-entropy loss was used for training the individual "vanilla" models before training the same network architecture with the KD framework.

The framework utilizes the Convolutional Block Attention Module (CBAM) to compute attention maps across channel and spatial dimensions. This mechanism focuses on the model's most discriminative features using pooling operations and a shared multi-layer perceptron. The channel attention mechanism calculates the weights used in ensembling the students' predictions.

The KD framework was evaluated on CIFAR-10 [HP18], CIFAR-100, [KH+09] and TinyImageNet [LY15] image classification benchmarks. The "vanilla" version of the model was first trained and evaluated independently, then the same architecture was used in the ensemble, and the best student was selected after training. The improvement in accuracy (i.e. gain) was a critical metric for evaluating the effectiveness of the knowledge distillation process. A numerical comparison, using the ResNet-32 architecture [HZRS16], with similar approaches form the literature is presented in Table 4.1 for CIFAR-10 and in Table 4.2 for CIFAR-100.

Table 4.1: Comparison with state of the art works on CIFAR-10 database using ResNet-32 [HZRS16] as student network.

Method	Vanilla	KD	KD Gain
ONE [LZG18]	93.07%	94.01%	0.94%
CLCNN [SC18]	93.17%	94.14%	0.97%
OKDDip [CMW+20] net.	93.66%	94.38%	0.72%
OKDDip [CMW+20] br.	93.66%	94.42%	0.76%
PCL [WG21]	93.26%	94.33%	1.07%
Proposed	92.77%	93.88%	**1.11%**

Table 4.2: Comparison with state-of-the-art works for ResNet-32 architecture trained on CIFAR-100 dataset.

Method	Vanilla	KD	KD Gain
DML [ZXHL18]	68.99%	71.19%	2.20%
KDCL [GWW+20] [1]	71.28%	73.76%	2.48%
OKDDip net. [CMW+20]	71.24%	74.60%	**3.36%**
OKDDip br. [CMW+20]	71.24%	74.37%	3.13%
SAD [JHP21]	75.32%	77.47%	2.15%
PCL [WG21]	71.28%	74.14%	2.86%
Proposed	69.6%	72.76%	3.16%

Chapter 5

3D Scene and Face Reconstruction

This Chapter discusses advancements in 3D reconstruction, focusing on overcoming limitations posed by 2D image analysis, particularly in non-frontal views and depth information. The chapter introduces a camera calibration optimization using a genetic algorithm, an extension of the forehead landmarks for complete face geometry, and two 3D Morphable Model (3DMM) reconstruction models utilizing 73 face landmarks.

5.1 Camera Calibration using an Evolutionary Approach

We propose an innovative approach using a genetic algorithm for optimizing camera calibration. Camera calibration is the process of finding the camera parameters. The intrinsic parameters relate to the camera's internal characteristics like focal length, principal point, and aspect ratio, while the extrinsic parameters describe the camera's position and orientation in space. We didn't attack the camera calibration algorithm, we use the well established OpenCV implementation for the algorithm [**?**]. Our method involves selecting the best captures from a set of images to minimize calibration errors. The calibration process uses evolutionary techniques to manage a large solution space effectively, employing mutation and crossover operators. The methodology is tested through various experiments, demonstrating its superiority in accuracy and robustness compared to traditional methods, especially in stereo calibration for 3D measurements.

The numerical results and comparisons section details the experimental outcomes of the camera calibration process. It includes stereo and mono calibration results, comparing the proposed genetic algorithm-based approach against established methods. The experiments demonstrate sub-millimeter accuracy in 3D measurements for stereo point pairs. For mono calibration, the proposed method outperforms existing techniques as *CalWiz* [PS19] by a significant margin in terms of root mean square error. In our best experiment we get an error of 0.357px compared to *CalWiz* best experiment of 0.631px.

5.2 New Pipeline for 3D Face Modeling

This Chapter introduces two innovative approaches for reconstructing 3D faces. Both utilize Structure From Motion (SFM) and Radial Basis Function (RBF) techniques on high-density vertex models, starting from various facial images at different poses, mostly frontal. The first approach employs a generic 3D model retrained with additional forehead landmarks, while the second approach, using a deformable model, integrates pose estimation and forehead points inferred from a hair mask. Analyzing current approaches based on Deep Learning [FWS+18], they are reconstructing the 3D face

model from a single camera capture. Their system is based on a statistical model, which was trained on the publicly available dataset. Those methods are susceptible to outliers, as those subjects are far apart from the pre-determined statistical face model, of faces with variate poses. In our experiments we emphasizes the improvement in capturing facial details, especially in the forehead area, and outlines the process from landmark detection to detailed 3D reconstruction.

Our process begins with a generic, textured 3D model of a human face that represents a midpoint between male and female characteristics. A crucial step is establishing a mapping between the 73 2D landmarks and corresponding 3D vertices on this model. To achieve this, we render the generic 3D model into a 2D image and then apply the facial landmark detector. Subsequently, we employ a ray-casting technique originating from the system's origin point to the detected landmarks. This method is used to map each 2D landmark to a specific vertex index on the 3D model. The mapping is conducted based on the nearest triangle-vertex proximity, ensuring accurate alignment between the 2D and 3D representations.

The first approach leverages the BU-3DFE 3D model [YWS+06a] and an extended DLIB [Kin09] regressor for forehead landmark prediction. The provided extension include the 5 top forehead points. The second approach uses the Basel model [GFB+18] and infers forehead points from a hair mask, integrating a pose estimation network. Both approaches involve using Structure From Motion (SFM) and Radial Basis Function (RBF) on a high-density vertex model.

To measure the accuracy of our proposed method, we computed the percentage of landmarks that fall within an $\epsilon = 10^{-2}$. We use the 3D face models from BU-3DFE Database [YWS+06a], which contains 100 subjects (56 female and 44 male). We obtain an accuracy of 87.34% on the female data set and 91.25% on the male data set. The overall accuracy over the face contour shape only (without inner face depth features) is 98%.

Chapter 6

Conclusions and Future Work

In this thesis, we have contributed to the multifaceted domain of Computer Vision, focusing on enhancing facial analysis through advanced methods in hair segmentation and color classification, pupil and iris detection, dynamic pruning, attention-based ensemble knowledge distillation, and 3D face reconstruction. Our work synergizes traditional computer vision techniques with modern machine learning approaches, leading to innovative solutions and significant advancements in the field.

Each segment of this work is not a standalone contribution but part of an integrated whole, aiming to push the boundaries of how we understand and process visual information in the digital age. More specifically, we start from the intricate details of 2D face analysis, advance into the expansive 3D space for a more holistic understanding, and culminate in the refinement of deep learning models through innovative architecture search and inference optimization.

6.1 Unified Approach of Face Analysis in the 2D Space

Our exploration in 2D face analysis, involving pupil, iris detection [TA20], hair segmentation, and color classification [BID18, IBD19] leverages a combination of regression trees, morphological operations, and neural network-based classification. The developed methods have exhibited high accuracy and efficiency, demonstrating their practical applicability.

Our methods are not constrained by any hardware devices and can accurately segment and detect the visible region of the iris, face, and hair. The pupil and the iris are precisely detected by employing an iterative algorithm that determines the darkest part of the eye. The experimental results show that the worst cumulative error is less than 0.05, meaning that the predicted pupil location is within the interior of the real pupil.

Moving further in analyzing the face, we proposed a pipeline of algorithms starting with hair and face segmentation and then color classification. Eventually, we move beyond 2D space by proposing a 3D face reconstruction process. More specifically, the hair and face areas are determined using a state-of-the-art CNN. Additional morphological operators are applied to the hair mask to fill in the eventual gaps in the hair area. The hair pixels are further analyzed either by a Random Forest Classifier based on superpixel features [BDD16]) or by a classical artificial neural network [IBD19] to determine the hair color. To train and test the proposed algorithm, we annotated more than 4000 images from an existing database with the hair color.

The Thesis also presents comparative studies of different techniques used in face analysis. Using the numerical results, one can decide which color space or network architecture fits best for similar tasks and provides a good intuition for estimating the trade-off between latency and accuracy.

In future work, we plan to refine the current approaches by making them more robust to illumi-

nation conditions and face pose. We also plan to extend the face analysis tasks to include face shape and eye color detection. One hot topic nowadays is generating synthetic (realistic) images using generative methods and using them to train supervised models. Furthermore, we will investigate the direction where the input for hair color classification is the entire image. For this, we will use a deep learning architecture and potentially more input data.

6.2 Beyond 2D Analysis using 3D Face Reconstruction

2D Face Analysis has its limitations, especially for tasks where the subject should be in a frontal view (face shape detection) or when depth-related measures on the face are required. Recognizing the limitations of 2D analysis, our work in 3D face reconstruction marks a significant step forward. In the context of the 3D space, we made the following contributions: we designed a genetic algorithm for enhancing camera calibration (which selects a better subset of input acquisitions) and developed two pipelines for 3DMM face reconstruction, including the forehead area.

Regarding the camera calibration procedure, we introduced a method based on genetic algorithms to automatically choose the most suitable stereo calibration images from a larger collection of acquisitions. In contrast to the compared approach, our pipeline is entirely integrated and autonomous, eliminating the need for any user involvement other than capturing multiple image snapshots of a calibration pattern from various distinctive angles. We consider this approach more suitable for calibrating cameras for several reasons. Firstly, it uses all available data in an evolutionary approach rather than in a greedy manner like similar works. Secondly, the system is plug-and-play, runs using parallel computing, and can achieve excellent calibration for any number of cameras, including mono and stereo calibration.

For 3D facial reconstruction, we propose a pipeline composed of Structure From Motion followed by Radial Basis Functions applied to a high-density vertex model to yield a suitable approximation of the subject's face in three dimensions. The first step is employed to create a 73-point cloud of 3D landmarks, while the second step deforms a generic 3D model to achieve the final reconstruction. In the two proposed approaches, we use different generic models such as Basel [GFB+18] and BU-3DFE [YWS+06b]. We extract the 68 2D facial landmarks using the DLIB regressor [Kin09], while the upper forehead landmarks are extracted using a retrained version of the [Kin09] regressor or from a face mask, that we previously detected.

Our methods have been tested in real-world scenarios and have successfully reconstructed 3D facial models of random individuals. While the alternative approach of one-shot reconstruction using Deep Learning may be tempting, it has the limitation that occluded areas should be "invented" by the network.

Regarding future work, we plan to increase the accuracy of the model by expanding the number of facial landmarks used. One approach is to generate synthetic faces with well-known positions of facial landmarks. Additionally, we will focus on the sensitive subject of face shape classification and determine various internal facial metrics (such as interpupillary distance, eye size, etc.) in the hope of developing an accurate eyewear recommendation system that takes into account enhanced 3D metrics.

6.3 Enhance Architecture Search, Training and Inference in Deep Learning Models

As we previously mentioned, our research and experiments are mostly based on Deep Learning models. When trying to deploy the models to fit real-world scenarios, we encounter problems in terms of latency and memory usage. To address these issues we consider the following solutions automatic Neural Architecture Search, dynamic pruning, and knowledge distillation.

Designing and implementing a general-purpose algorithm that processes all this data may be challenging, and thus various types and variations of algorithms need to be created. Solving each problem individually requires significant human resources and may take longer time. One solution is to use Neural Architecture Search algorithms to find the required algorithms, which reduces the need for human resources. We developed a neural architecture search approach that generates a convolutional neural network using a recurrent neural network as a proposer. Our framework generates cells (groups of layers) that are individually evaluated on a smaller dataset. The best cells are further instantiated in larger templates to evaluate them on the target dataset.

We chose to test our generated cell on two semantic segmentation tasks. The first experiment was performed on a self-made dataset of eyeglasses, and the second on the public dataset Pascal VOC 2012 [EVGW+10]. The mean intersection-over-union on the testing set of our dataset is 0.9683, and for the general Pascal VOC dataset, it is 0.3289. The cell used for evaluation was discovered on a smaller eyeglasses dataset, and this may be a reason for the low performance on the general-purpose Pascal VOC dataset. For future work, we plan to discover a specific cell for each task.

For network optimization, we proposed two methods (1) training dynamic networks (i.e., gating ensembles) and (2) boosting a small network using knowledge distillation.

For the former optimization approach, our dynamic ensemble called *DynK-Hydra* is targeted at reducing the inference time of classification tasks with a medium to a large number of classes while preserving overall accuracy. We show improvements in the inference time on the order of 2-5.5 times compared to baseline ResNet networks [HZRS16], and a marginal accuracy improvement of 1.2% against similar work such as HydraRes [MMSF18]. We apply dynamic sparse execution and show that a significant reduction of inference time is still possible compared to HydraRes. As the proposed process involves only the training phase (making it more complex), we consider it justifies the inference time gains.

The latter optimization mechanism involves a cohort of students that are trained simultaneously. Their output is dynamically concatenated as a weighted sum performed by an attention-based mechanism. The final output represents the *teacher's* knowledge, which is distilled back to the components. The proposed method was tested on multiple benchmark datasets, using well-known deep learning architectures. In all the training scenarios, the knowledge-distilled student is more powerful than a *vanilla* independently trained student. Compared to similar state-of-the-art approaches our mechanism obtains a better accuracy gain or is at least comparable.

Yet, there are many inference optimization processes, and all of them usually imply an overhead during training. As an overview, if very large datasets are used and the system benefits from large memory capacity, a dynamic ensemble is a good fit. If the task at hand is not generic, and one is looking for a suitable, relatively small architecture, a neural architecture search can be employed. If we want to use a small architecture but benefit from high training capacity, one can gather many light students and boost their performance by training them in online knowledge distillation.

Generating tailored deep neural network architecture and reducing inference costs (both memory footprint and latency) are of high interest to us. In future work, we plan to enhance the capabilities of all the proposed methods: generative framework, dynamic ensemble, and knowledge distillation

process. To enhance the generative (architecture) framework and make it even faster, we plan to introduce a surrogate evaluation function that can evaluate the intermediary cells without training them. For the knowledge distillation process, we are looking for ways to distill the knowledge at the level of the feature maps. Moreover, we are investigating various extra features to be used for the attention-based mechanism to ensure that we distill only the relevant knowledge. In the dynamic ensemble, we intend to explore the loss landscape (both visually and numerically), especially for the individual branches, to better understand and guide the training process. Additionally, we will focus on improving the training procedure to reduce the number of activated branches to 1 (currently, the mean activation branches are greater than 2.5).

Bibliography

[ARL+10] Oleg Alexander, Mike Rogers, William Lambeth, Jen-Yuan Chiang, Wan-Chun Ma, Chuan-Chang Wang, and Paul Debevec. The digital emily project: Achieving a photorealistic digital actor. *IEEE Computer Graphics and Applications*, 30(4):20–31, 2010.

[BCNM06] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

[BDD16] Diana Borza, Adrian Darabant, and Radu Danescu. Real-time detection and measurement of eye features from color images. *Sensors*, 16(7):1105, 2016.

[BDIM22] Diana-Laura Borza, Adrian Sergiu Darabant, Tudor Alexandru Ileni, and Alexandru-Ion Marinescu. Effective online knowledge distillation via attention-based model ensembling. *Mathematics*, 10(22):4285, 2022.

[BGNR16] Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. Designing neural network architectures using reinforcement learning. *arXiv preprint arXiv:1611.02167*, 2016.

[BID18] Diana Borza, Tudor Ileni, and Adrian Darabant. A deep learning approach to hair segmentation and color extraction from facial images. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 438–449. Springer, 2018.

[Bre96] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[Bro71] D. C. Brown. Close-Range Camera Calibration. *Photogrammetric Engineering and Remote Sensing*, 37(8):855–866, 1971.

[BWHY05] Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information fusion*, 6(1):5–20, 2005.

[CGW+19] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791*, 2019.

[CMW+20] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(04), pages 3430–3437, 2020.

[DCJ+94] Harris Drucker, Corinna Cortes, Lawrence D Jackel, Yann LeCun, and Vladimir Vapnik. Boosting and other ensemble methods. *Neural Computation*, 6(6):1289–1301, 1994.

[EMH18] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *arXiv preprint arXiv:1808.05377*, 2018.

[EVGW+10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[Fai75] W. Faig. Close-Range Camera Calibration:Mathematical Formulation. *Photogrammetric Engineering and Remote Sensing*, 41(12):1479–1486, 1975.

[FHL19] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[FWS+18] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018.

[GFB+18] Thomas Gerig, Andreas Forster, Clemens Blumer, Bernhard Egger, Marcel Lüthi, Sandro Schönborn, and Thomas Vetter. Morphable face models - an open framework. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82, 2018.

[GWW+20] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11020–11029, 2020.

[GYMT20] Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge distillation: A survey. *arXiv preprint arXiv:2006.05525*, 2020.

[Har94] Hartley. An algorithm for self calibration from several views. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 908–912, 1994.

[HHS+21] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[HLP+17] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

[HP18] Tien Ho-Phuoc. Cifar10 to compare visual recognition performance between deep neural networks and humans. *arXiv preprint arXiv:1811.07270*, 2018.

[HVD15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[IBD19] Tudor Alexandru Ileni, Diana Laura Borza, and Adrian Sergiu Darabant. Fast in-the-wild hair segmentation and color classification. In *VISIGRAPP (4: VISAPP)*, pages 59–66, 2019.

[IDBM22] Tudor Alexandru Ileni, Adrian Sergiu Darabant, Diana Laura Borza, and Alexandru Ion Marinescu. Dynk-: improved dynamic architecture ensembling for efficient inference. *Complex & Intelligent Systems*, pages 1–12, 2022.

[JHP21] Mingi Ji, Byeongho Heo, and Sungrae Park. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7945–7952, 2021.

[Kel70] Michael David Kelly. *Visual identification of people by computer*. Number 130 in -. Department of Computer Science, Stanford University., 1970.

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *arXiv*, 2009.

[Kin09] Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[KKA+20] Khalil Khan, Rehan Ullah Khan, Kashif Ahmad, Farman Ali, and Kyung-Sup Kwak. Face segmentation: A journey from classical to deep learning paradigm, approaches, trends, and directions. *IEEE Access*, 8:58683–58699, 2020.

[KPR+14] A Krupka, J Prinosil, K Riha, J Minar, and M Dutta. Hair segmentation for color estimation in surveillance systems. In *Proc. 6th Int. Conf. Adv. Multimedia*, pages 102–107, 2014.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[LCP+18] Alex Levinshtein, Cheng Chang, Edmund Phung, Irina Kezele, Wenzhangzhi Guo, and Parham Aarabi. Real-time deep hair matting on mobile devices. In *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 1–7. IEEE, 2018.

[LCY13] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[LF97] Q.T. Luong and O. D. Faugeras. Self-Calibration of a Moving Camera from Point-Correspondences and Fundamental Matrices. *Int. Journal of Computer Vision*, 22(3):261–289, 1997.

[LTW95] Yuencheng Lee, Demetri Terzopoulos, and Keith Waters. Realistic modeling for facial animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 55–62, 1995.

[LY15] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7:7, 2015.

[LZG18] Xu Lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble. *arXiv preprint arXiv:1806.04606*, 2018.

[MDI20] Alexandru Ion Marinescu, Adrian Sergiu Darabant, and Tudor Alexandru Ileni. A fast and robust, forehead-augmented 3d face reconstruction from multiple images using geometrical methods. In *2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pages 1–6. IEEE, 2020.

[MDI21] Alexandru Ion Marinescu, Adrian Sergiu Darabant, and Tudor Alexandru Ileni. Optimal stereo camera calibration via genetic algorithms. In *2021 Workshop of Artificial Intelligence for Autonomous Driving (IJCAI)*, pages 1–1. IEEE, 2021.

[MID19] Alexandru Ion Marinescu, Tudor Alexandru Ileni, and Adrian Sergiu Darabant. A versatile 3d face reconstruction from multiple images for face shape classification. In *2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pages 1–6. IEEE, 2019.

[MMSF18] Ravi Teja Mullapudi, William R Mark, Noam Shazeer, and Kayvon Fatahalian. Hydranets: Specialized dynamic architectures for efficient inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8080–8089, 2018.

[MPS21] Araceli Morales, Gemma Piella, and Federico M Sukno. Survey on 3d face reconstruction from uncalibrated images. *Computer Science Review*, 40:100400, 2021.

[MSLB18] Umar Riaz Muhammad, Michele Svanera, Riccardo Leonardi, and Sergio Benini. Hair detection, segmentation, and hairstyle classification in the wild. *Image and Vision Computing*, 71:25–37, 2018.

[NID20] Sergiu Cosmin Nistor, Tudor Alexandru Ileni, and Adrian Sergiu Dărăbant. Automatic development of deep learning architectures for image segmentation. *Sustainability*, 12(22):9707, 2020.

[PN17] Hugo Proença and João C Neves. Soft biometrics: Globally coherent solutions for hair segmentation and style recognition based on hierarchical mrfs. *IEEE Transactions on Information Forensics and Security*, 12(7):1637–1645, 2017.

[PS19] Songyou Peng and Peter Sturm. Calibration Wizard: A guidance system for camera calibration based on modelling geometric and corner uncertainty. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1497–1505, 2019.

[SC18] Guocong Song and Wei Chai. Collaborative learning for deep neural networks. *arXiv preprint arXiv:1805.11761*, 2018.

[Sch13] Robert E Schapire. Explaining adaboost. In *Empirical inference*, pages 37–52. Springer, 2013.

[SLJ+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[SS97] Amanda JC Sharkey and Noel E Sharkey. Combining diverse neural nets. *The Knowledge Engineering Review*, 12(3):231–247, 1997.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[TA20] Ileni Tudor-Alexandru. Efficient iris segmentation and pupil detection for visagisme applications. In *2020 IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pages 000123–000128. IEEE, 2020.

[VJ01] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001.

[VJ04] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[WG21] Guile Wu and Shaogang Gong. Peer collaborative learning for online knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35(12), pages 10302–10310, 2021.

[WZL+19] Caiyong Wang, Yuhao Zhu, Yunfan Liu, Ran He, and Zhenan Sun. Joint iris segmentation and localization using deep multi-task learning framework. *arXiv preprint arXiv:1901.11195*, 2019.

[YWS+06a] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. A 3d facial expression database for facial behavior research. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, FGR '06, page 211–216, USA, 2006. IEEE Computer Society.

[YWS+06b] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. A 3d facial expression database for facial behavior research. *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 211–216, 2006.

[ZL16] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

[ZXHL18] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.