





# LOSS FUNCTIONS AND METRICS IN DEEP LEARNING


A PREPRINT

**Juan R. Terven**  
CICATA-Qro  
Instituto Politecnico Nacional  
Mexico  
jrtervens@ipn.mx

**Diana M. Cordova-Esparza**  
Facultad de Informática  
Universidad Autónoma de Querétaro  
Mexico  
diana.cordova@uaq.mx

**Alfonso Ramirez-Pedraza**  
Visión Robótica  
Centro de Investigaciones en Óptica A.C.  
Mexico  
pedro.ramirez@cio.mx

**Edgar A. Chavez-Urbiola**  
CICATA-Qro  
Instituto Politecnico Nacional  
Mexico  
eachavezu@ipn.mx

**Julio A. Romero-Gonzalez**  
Facultad de Informática  
Universidad Autónoma de Querétaro  
Mexico  
julio.romero@uaq.mx

August 9, 2024

## ABSTRACT

When training or evaluating deep learning models, two essential parts are picking the proper loss function and deciding on performance metrics. In this paper, we provide a comprehensive overview of the most common loss functions and metrics used across many different types of deep learning tasks, from general tasks such as regression and classification to more specific tasks in Computer Vision and Natural Language Processing. We introduce the formula for each loss and metric, discuss their strengths and limitations, and describe how these methods can be applied to various problems within deep learning. We hope this work serves as a reference for researchers and practitioners in the field, helping them make informed decisions when selecting the most appropriate loss function and performance metrics for their deep learning projects.

**Keywords** Deep learning · loss functions · performance metrics

## Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Loss Functions vs. Performance Metrics</b>	<b>3</b>
2.1	Properties of loss functions	3
<b>3</b>	<b>Basic Tasks</b>	<b>4</b>
3.1	Regression	4
3.1.1	Regression Loss Functions	4
3.1.2	Regression Performance Metrics	9
3.2	Classification	12
3.2.1	Classification Loss Functions	12
3.2.2	Classification Performance Metrics	15

<b>4 Computer Vision Tasks</b>	<b>26</b>
4.1 Image Classification	26
4.1.1 Image Classification Loss Functions	27
4.1.2 Image Classification Metrics	27
4.2 Object Detection	27
4.2.1 Object Detection Loss Functions	27
4.2.2 Object Detection Metrics	31
4.3 Image Segmentation	33
4.3.1 Segmentation Loss Functions	34
4.3.2 Segmentation Metrics	36
4.4 Face Recognition	38
4.4.1 Face Recognition Loss Functions and Metrics	39
4.5 Image Generation	45
4.5.1 Image Generation Loss functions	45
4.5.2 Image Generation Metrics	49
<b>5 Natural Language Processing</b>	<b>52</b>
5.1 Loss Functions used in NLP	53
5.1.1 Cross-Entropy Loss (Token-Level)	53
5.1.2 Hinge Loss	53
5.1.3 Cosine Similarity Loss	54
5.2 Losses for Sequence Generation	55
5.2.1 Minimum Risk Training (MRT)	55
5.2.2 REINFORCE Algorithm	56
5.3 Performance Metrics used in NLP	57
5.3.1 Accuracy	57
5.3.2 Precision, Recall, and F1 Score	58
5.3.3 AUC-ROC	58
5.3.4 BLEU Score	59
5.3.5 ROUGE Score	60
5.3.6 Perplexity	61
5.3.7 Exact Match (EM)	62
<b>6 Discussion</b>	<b>63</b>
<b>7 Conclusion</b>	<b>63</b>
<b>8 Acknowledgments</b>	<b>64</b>

# 1 Introduction

Deep Learning has become the dominant technology for solving problems involving unstructured data, such as images. [1][2][3][4][5][6][7][8], video, audio [9][10][11][12][13], and text [14][15][16][17][18]. One of the critical components of deep learning is the selection of the loss function and performance metrics used for training and evaluating models. Loss functions measure how effectively a model can approximate the desired output, while performance metrics assess the model’s ability to make accurate predictions on unseen data. Choosing the appropriate loss function and performance metric is essential for success in deep learning tasks. However, with many options to choose from, it can be challenging for practitioners to determine the most suitable method for their specific task.

In this paper, we present a thorough overview of the most commonly utilized loss functions and performance metrics in the field of deep learning. We analyze the strengths and weaknesses of each method and provide illustrative examples of their application across various deep learning tasks.

First, we delve into the prevalent regression and classification loss functions, such as mean squared error, cross-entropy, and hinge loss, delineating their respective advantages, limitations, and typical use cases. Subsequently, we explore standard tasks in computer vision, such as image classification, object detection, image segmentation, face recognition, and image generation. Finally, we conclude our review by outlining the prevalent loss functions and metrics employed in natural language processing.

This paper is structured as follows. Section 2 outlines the difference between loss functions and performance metrics. Section 3 provides an overview of the most widely used losses and metrics in regression and classification. Section 4 delves into the prevalent computer vision tasks, detailing the associated loss functions and metrics. Section 5 concentrates on the use of losses and metrics in NLP. Section 6 presents a discussion of this research and, finally, Section 7 rounds out the paper with a concluding statement.

## 2 Loss Functions vs. Performance Metrics

Loss functions and performance metrics are two distinct tools for evaluating the performance of a deep learning model and serve different purposes.

During training, a loss function is used to optimize the model’s parameters. Measures the difference between the predicted and expected outputs of the model. The objective of training is to minimize this difference.

In contrast, a performance metric is used to evaluate the model after training. It helps to determine how well the model can generalize to new data and make accurate predictions. Performance metrics also aid in comparing different models or configurations to identify the best-performing one.

The following list details the key differences between loss functions and performance metrics:

- During the training of a deep learning model, loss functions are used to optimize the model’s parameters, whereas performance metrics are used to evaluate the model’s performance after training.
- The choice of loss function typically depends on the model’s architecture and the specific task at hand. In contrast, performance metrics are less dependent on the model’s architecture and can be used to compare different models or configurations of a single model.
- The ultimate goal of training a deep learning model is to minimize the loss function, while evaluating a model aims to maximize the performance metric, with the exception of error performance metrics such as Mean Squared Error.
- Loss functions can be challenging to interpret as their values are often arbitrary and depend on the specific task and data. In contrast, performance metrics are often more interpretable and can be used across different tasks.

### 2.1 Properties of loss functions

Loss functions have a series of properties that need to be considered when selected for a specific task:

1. **Convexity:** A loss function is convex if any local minimum is also the global minimum. Convex loss functions are desirable because they can be easily optimized using gradient-based optimization methods.
2. **Differentiability:** A loss function is differentiable if its derivative with respect to the model parameters exists and is continuous. Differentiability is essential because it allows the use of gradient-based optimization methods.

3. **Robustness:** Loss functions should be able to handle outliers and not be affected by a small number of extreme values.
4. **Smoothness:** A loss function should have a continuous gradient and no sharp transitions or spikes.
5. **Sparsity:** A sparsity-promoting loss function should encourage the model to produce sparse output. This is useful when working with high-dimensional data and when the number of important features is small.
6. **Monotonicity:** A loss function is monotonic if its value decreases as the predicted output approaches the true output. Monotonicity ensures that the optimization process is moving toward the correct solution.

The tables below provide a summary of the loss functions and performance metrics that are discussed in this work. Table 1 outlines the functions and metrics employed in general tasks such as regression, binary classification, and multiclass classification. Table 2 provides an overview of the loss functions and metrics related to computer vision, and table 3 summarizes the work on natural language processing.

Table 1: Loss functions and performance metrics for general tasks.

Deep Learning Task	Loss Functions	Performance Metrics
Regression	MSE (3.1.1)	MSE (3.1.1)
	MAE (3.1.1)	MAE (3.1.1)
	Huber loss (3.1.1)	RMSE (3.1.2)
	Log-Cosh (3.1.1)	MAPE (3.1.2)
	Quantile loss (3.1.1)	SMAPE (3.1.2)
	Poisson loss (3.1.1)	$R^2$ (3.1.2)
		Adjusted $R^2$ (3.1.2)
Binary Classification	BCE (3.2.1)	Accuracy (3.2.2)
	Hinge loss (3.2.1)	Precision (3.2.2)
	Focal loss (4.2.1)	Recall or TPR (3.2.2)
		F1-Score (3.2.2)
		AUC-ROC (3.2.2)
		PR Curve (3.2.2)
Multi-Class Classification	CCE (3.2.1)	Accuracy (3.2.2)
	Sarse CCE (3.2.1)	Precision (3.2.2)
	CCE w/label smoothing (3.2.1)	Recall or TPR (3.2.2)
	Focal loss (4.2.1)	F1-Score (3.2.2)
	PolyLoss (3.2.1)	PR Curve (3.2.2)
	Hinge loss (3.2.1)	

### 3 Basic Tasks

In the following sections, we dive into two essential tasks in machine learning: regression and classification. We examine the loss functions and performance metrics used and offer practical guidance on when to use each, as well as real-world applications.

#### 3.1 Regression

Regression is a supervised learning problem in machine learning that aims to predict a continuous output value based on one or more input features. Regression is used in various domains, including finance, healthcare, social sciences, sports, and engineering. Some practical applications include house price prediction [19], energy consumption forecasting [20], healthcare and disease prediction [21], stock price forecasting [22], and customer lifetime value prediction [23].

In the following subsections, we review the most common lost functions and performance metrics used for regression.

##### 3.1.1 Regression Loss Functions

Table 4 shows the common loss functions used for regression and their applications.

The following subsections describe each of these loss functions in more detail.

Table 2: Loss functions and performance metrics used in Computer Vision.

Deep Learning Task	Loss Functions	Performance Metrics
Object Detection	Smooth L1 (4.2.1)	AP (4.2.2)
	IoU loss (4.2.1)	AR (4.2.2)
	Focal loss (4.2.1)	
	YOLO loss (4.2.1)	
Semantic Segmentation	CCE	IoU (4.2.1),
	IoU loss (4.2.1)	Pixel Accuracy (4.3.2),
	Dice Loss (4.3.1)	AP (4.2.2)
	Tversky loss (4.3.1)	BF (4.3.2)
	Lovasz loss (4.3.1)	
Instance Segmentation	CCE (3.2.1)	AP (4.2.2)
	IoU loss (4.3.1)	
	Smooth L1 (4.2.1)	
Panoptic Segmentation	CCE (3.2.1)	PQ (4.3.2)
	Dice Loss (4.3.1)	
Face Recognition	A-Softmax (4.4.1)	Accuracy (3.2.2)
	Center loss (4.4.1)	Precision (3.2.2)
	CosFace (4.4.1)	Recall (3.2.2)
	ArcFace (4.4.1)	F1-Score (3.2.2)
	Triplet loss (4.4.1)	
	Contrastive loss (4.4.1)	
	Circle loss (4.4.1)	
Image Generation	Adversarial Loss (4.5.1)	PSNR (4.5.2)
	Reconstruction loss (4.5.1)	SSIM (4.5.2)
	KL Divergence (4.5.1)	IS (4.5.2),
	Wasserstein Loss (4.5.1)	FID (4.5.2)
	Contrastive Divergence (4.5.1)	

**Mean Squared Error (MSE)** The MSE measures the average of the squared differences between the predicted values and the true values [24]. The MSE loss function can be defined mathematically as

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (1)$$

where  $n$  is the number of samples,  $y_i$  is the true value of the  $i^{th}$  sample and  $\hat{y}_i$  is the predicted value of the  $i^{th}$  sample.

The MSE loss function has the following properties:

- **Non-negative:** MSE is always non-negative because the differences between the predicted and actual values are squared. A value of 0 indicates a perfect fit, while larger values correspond to higher discrepancies between predictions and actual values.
- **Sensitive to outliers:** MSE is a quadratic function of the prediction errors, which means it places more emphasis on larger errors than smaller ones. This property makes it sensitive to outliers and can lead to models that prioritize reducing large errors over smaller ones.
- **Differentiable:** MSE is a smooth and continuous function for the model parameters. This property allows for the efficient computation of gradients, which is essential for optimization algorithms such as gradient descent.
- **Convex:** MSE is a convex function, which means it has a unique global minimum. This property simplifies the optimization process, as gradient-based optimization techniques can converge to the global minimum without getting trapped in local minima. However, for deep neural networks, the error landscape is generally non-convex due to the multiple layers of non-linear activation functions, leading to a complex and highly non-linear optimization problem.
- **Scale-dependent:** The value of MSE depends on the scale of the target variable, making it difficult to compare the performance of models across different problems or target variable scales. For this purpose, researchers often use the root mean squared error (RMSE) or mean squared percentage error (MSPE) for computing the performance of the model.

Table 3: Loss functions and performance metrics used in Natural Language Processing.

Deep Learning Task	Loss Functions	Performance Metrics
Text Classification	CE (3.2.1)	Accuracy (5.3.1)
	Hinge loss (5.1.2)	P/R/F1 (5.3.2)
		AUC-ROC (5.3.3)
Language Modeling	T-CE (5.1.1)	Perplexity (5.3.6)
		BLEU (5.3.4)
		ROUGE (5.3.5)
Machine Translation	T-CE (5.1.1)	BLEU (5.3.4)
	MRT (5.2.1)	ROUGE (5.3.5)
	REINFORCE (5.2.2)	Perplexity (5.3.6)
		Exact match (5.3.7)
Name Entity Recognition	T-CE (5.1.1)	Accuracy (5.3.1)
		P/R/F1 (5.3.2)
		Exact match (5.3.7)
Part-of-Speech Tagging	T-CE (5.1.1)	Accuracy (5.3.1)
		P/R/F1 (5.3.2)
Sentiment Analysis	CE (3.2.1)	P/R/F1 (5.3.2)
	Hinge loss (5.1.2)	AUC-ROC (5.3.3)
	Cosine Similarity (5.1.3)	
	Adj. Circle (4.4.1)	
Text Summarization	T-CE (5.1.1)	BLEU (5.3.4)
	MRT (5.2.1)	ROUGE (5.3.5)
	REINFORCE (5.2.2)	Exact match (5.3.7)
Question Answering	CE (3.2.1)	P/R/F1 (5.3.2)
	Hinge loss (5.1.2)	Exact match (5.3.7)
	Cosine Similarity	
Language Detection	CE (3.2.1)	Accuracy (5.3.1)
	Hinge loss (5.1.2)	P/R/F1 (5.3.2)

Table 4: Loss Functions and their applications in regression tasks.

Loss Function	Applications
Mean Squared Error (MSE)	Linear Regression, Ridge Regression, Lasso Regression, Neural Networks, Support Vector Regression, Decision Trees, Random Forests, Gradient Boosting
Mean Absolute Error (MAE)	Quantile Regression, Robust Regression, L1 Regression, Neural Networks, Decision Trees, Random Forests, Gradient Boosting
Huber Loss	Robust Linear Regression, Robust Neural Networks, Gradient Boosting, Random Forests
Log-Cosh Loss	Robust Regression, Neural Networks, Gradient Boosting
Quantile Loss	Quantile Regression, Distributional Regression, Extreme Value Prediction
Poisson Loss	Poisson Regression, Count Data Prediction, Generalized Linear Models, Neural Networks, Gradient Boosting

The Mean Squared Error (MSE), which is also known as L2 loss, is a simple and widely used method for calculating the average of the squared differences between predicted and actual values. It is not only utilized as a loss function during the training of models but also serves as a performance metric to evaluate the accuracy of predictions. However, it is important to note that the use of the square of the error term can make it susceptible to outliers in the data. In such cases, it is recommended to use other loss functions such as Mean Absolute Error (MAE) or Huber Loss, which are more robust to outliers.

**Mean Absolute Error (MAE)** The MAE is another commonly used loss function in regression problems. It measures the average of the absolute differences between the predicted values and the true values [25]. The MAE loss can be defined as

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (2)$$

where  $n$  is the number of samples,  $y_i$  and  $\hat{y}_i$  are the true and predicted value of the  $i^{th}$  sample.

The MAE loss function has the following properties:

- **Non-negative:** Like MSE, MAE is always non-negative because it takes the absolute value of the differences between predicted and actual values. A value of 0 indicates a perfect fit, while larger values correspond to higher discrepancies between predictions and actual values.
- **Robust to outliers:** MAE is a linear function of the prediction errors, which treats all errors equally regardless of their magnitude. This property makes MAE less sensitive to outliers than MSE, as it does not disproportionately emphasize large errors. MAE is considered a more robust loss function than MSE. This makes it suitable for applications where the presence of outliers is expected or the distribution of errors is not symmetric.
- **Non-differentiable:** Although MAE is continuous, it is not differentiable when the prediction error is zero due to the absolute value function. This property can complicate the optimization process for specific algorithms, particularly those relying on gradient-based techniques. However, subgradient methods [26, 27, 28, 29] can be employed to overcome this issue.
- **Convex:** MAE is a convex function, which means it has a unique global minimum. This property simplifies the optimization process, as gradient-based optimization techniques can converge to the global minimum without getting trapped in local minima. Like the MSE, the MAE is non-convex for Deep neural networks due to the multiple layers with non-linear activation functions.

The MAE called L1 loss, is often used as an evaluation metric. It is computationally simple and easy to understand, but it does not have the smooth and differentiable property of the MSE and is not sensitive to outliers. However, like MSE, the value of MAE depends on the scale of the target variable, making it difficult to compare the performance of models across different problems or target variable scales. To address this issue, it is often use scale-invariant metrics such as mean absolute percentage error (MAPE) or normalized mean absolute error (NMAE) to compare models across different scales or units.

**Huber Loss** The Huber loss combines the properties of both Mean Squared Error (MSE) and Mean Absolute Error (MAE). Huber loss is designed to be more robust to outliers than MSE while maintaining smoothness and differentiability [30]. The Huber loss function is defined as

$$L(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{for } |y - \hat{y}| \leq \delta \\ \delta(|y - \hat{y}| - \frac{1}{2}\delta) & \text{otherwise,} \end{cases} \quad (3)$$

where  $y$  is the true value,  $\hat{y}$  is the predicted value, and  $\delta$  is a user-specified threshold value.

When the error is small, the Huber loss function behaves like the MSE loss function, and when the error is large, the Huber loss function behaves like the MAE loss function. This property makes the Huber loss function more robust to outliers than the MSE loss function, as it is less sensitive to large errors.

The Huber loss function is differentiable, which makes it suitable for use in gradient-based optimization algorithms such as stochastic gradient descent (SGD). It is commonly used in linear regression and time series forecasting, as it can handle outliers and noise in the data. It is also used in robust optimization problems where the data may contain outliers or noise.

The threshold  $\delta$  can be chosen empirically by trying different values and evaluating the model's performance. However, common practice is to set  $\delta$  to a small value if the data has a lot of noise and to a large value if the data has outliers.

**Log-Cosh Loss** The Log-Cosh loss function is smooth and differentiable. It is commonly used in regression problems where the data may contain outliers or noise [31]. The Log-Cosh loss is defined as

$$L(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n \log(\cosh(y_i - \hat{y}_i)), \quad (4)$$

where  $y$  is the true value,  $\hat{y}$  is the predicted value and  $n$  is the number of samples.



One of the advantages of the log-cosh loss function is that it is less sensitive to outliers than the mean squared error (MSE), as it is not affected by extreme data values. However, it is more sensitive to small errors than the Huber loss.

**Huber Loss vs Log-Cosh Loss** Huber loss and Log-Cosh loss are both loss functions that aim to improve robustness to outliers compared to the standard Mean Squared Error (MSE). They do so by blending properties of MSE and Mean Absolute Error (MAE), but each has distinct characteristics and advantages.

We use Huber loss when we need robustness to outliers and prefer a loss function that has a controllable transition point (via the  $\delta$  parameter) between quadratic and linear behavior. It is useful when we have a reason to define a specific point where the loss function should switch from quadratic to linear, depending on the noise characteristics of the data.

We use Log-Cosh Loss when we want a smooth and more universally applicable loss function that is always differentiable, making it suitable for optimization algorithms that benefit from smooth gradients. It can be a better choice when we do not have clear reasons to manually set a transition threshold as in Huber loss.

**Quantile Loss** Also known as quantile regression loss, this function is often used for predicting an interval instead of a single value [32]. If we denote the quantile as  $q$  where  $0 < q < 1$ , and the predicted and actual values as  $\hat{y}$  and  $y$  respectively, then the quantile loss is given by

$$L(y, \hat{y}) = q \cdot \max(y - \hat{y}, 0) + (1 - q) \cdot \max(\hat{y} - y, 0), \quad (5)$$

where  $\max(a, b)$  represents the maximum of  $a$  and  $b$ . The expression  $y - \hat{y}$  is used when the prediction underestimates, and  $\hat{y} - y$  is used when the prediction overestimates. The loss is scaled by  $q$  for underestimations and  $(1 - q)$  for overestimations.

Note that when  $q = 0.5$ , the quantile loss is equivalent to the Mean Absolute Error (MAE), making it a generalization of MAE that allows for asymmetric penalties for underestimations and overestimations.

Overestimation occurs when a model's prediction exceeds the actual value. Underestimation is the opposite of overestimation. It occurs when a model's prediction is lower than the actual value.

Practical examples of quantile regression include:

**Financial Risk Management:** To estimate Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR), which are measures of financial risk used in risk management. These quantile-based measures help to understand the potential for extreme losses [33].

**Supply Chain and Inventory Management:** Predicting demand for products can benefit from quantile loss as it can give a range of potential demand rather than a single point, which can help manage inventory and reduce stockouts or overstock situations [34].

**Energy Production:** To predict power output, having a range of potential outputs to manage grid stability [35].

**Economic Forecasting:** Predicting economic indicators can use quantile regression to give a range of possible values, which can help planning and policy-making [36].

**Weather Forecasting:** Can be useful for predicting variables like temperature or rainfall, where providing a range can be more informative than a single-point estimate [37] [38].

**Real Estate Pricing:** Predicting the price of a property within a range can be more useful than predicting a single price [39].

**Healthcare:** Quantile regression can predict a range of possible patient outcomes based on a set of features, which can assist doctors in making more informed decisions [40].

**Poisson Loss** Poisson loss is used in regression tasks when the target variable represents count data and is assumed to follow a Poisson distribution. The Poisson loss is derived from the negative log-likelihood of the Poisson distribution. It maximizes the likelihood of observing the count data given the predicted values [41]. It is defined as

$$L(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i \log(\hat{y}_i)), \quad (6)$$

where  $y_i$  represents the actual target value,  $\hat{y}_i$  is the predicted value, and  $n$  is the number of samples.



When applying the Poisson loss function to model count data, we must ensure that the predicted values are non-negative since negative counts are not meaningful in real-world scenarios. To achieve this, it is common to use a link function that transforms the linear combination of input features to a non-negative output, which can then be interpreted as the expected count.

A link function is a mapping from the linear predictor to the predicted value. In the context of Poisson regression, the exponential function is a common choice for the link function because it guarantees non-negative outputs. The exponential function has the following form:

$$\hat{y}_i = \exp(\mathbf{w}^\top \mathbf{x}_i + b), \quad (7)$$

where  $\mathbf{w}$  is a vector of weights,  $\mathbf{x}_i$  is a vector of input features for the  $i$ -th observation, and  $b$  is the bias term.

Using the exponential function as a link function, we ensure that the predicted values  $\hat{y}_i$  are always non-negative. In this case, the Poisson loss function can be written as

$$L(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n (\exp(\mathbf{w}^\top \mathbf{x}_i + b) - y_i \log(\exp(\mathbf{w}^\top \mathbf{x}_i + b))) \quad (8)$$

The Poisson distribution is typically used for modeling the number of times an event occurred in an interval. Here are some examples of applications where Poisson loss can be useful.

**Traffic Modeling:** Poisson regression can predict the number of cars that pass through a toll booth during a given time interval based on factors like the time of day, day of the week, and weather conditions [42].

**Healthcare:** Epidemiology can predict the number of disease cases in different regions based on variables like population density, vaccination rates, and social behavior patterns [43].

**Insurance:** In the insurance industry, it can be used to model claim counts for certain types of insurance policies [44].

**Customer Service:** Poisson regression can be used to predict the number of calls that a call center receives during different times of the day, to aid in staff scheduling [45].

**Internet Usage:** It can be used to model the number of website visits or clicks on an ad during a given time interval to help understand user behavior and optimize ad placement [46].

**Manufacturing:** It can predict the number of defects or failures in a manufacturing process, helping in quality control and maintenance planning [47].

**Crime Analysis:** Poisson regression can be used to model the number of occurrences of certain types of crimes in different areas to help in police resource allocation and crime prevention strategies [48].

### 3.1.2 Regression Performance Metrics

Table 5 shows the most common metrics used in regression tasks. The following sections delve into more details on each of these metrics skipping the mean square error (MSE) and the mean absolute error (MAE) because they are the same discussed previously as loss functions.

**Root Mean Squared Error (RMSE)** The RMSE is the square root of the mean squared error (MSE) defined as

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (9)$$

where  $y_i$  is the true value,  $\hat{y}_i$  is the predicted value, and  $n$  is the number of samples.

The RMSE measures the average deviation of the predictions from the true values. This metric is easy to interpret because it is in the same units as the data. However, it is sensitive to outliers. Lower RMSE values indicate better model performance, representing smaller differences between predicted and actual values.

The RMSE can be interpreted geometrically as the Euclidean distance between the vector of observed values and the vector of predictions, which is analogous to the length of a hypotenuse in a right triangle, representing the disparity between predicted and actual values.

Table 5: Common performance metrics used in regression.

Performance Metric	Applications
Mean Squared Error (MSE)	General-purpose regression, model selection, optimization, linear regression, neural networks
Root Mean Squared Error (RMSE)	General-purpose regression, model selection, optimization, linear regression, neural networks
Mean Absolute Error (MAE)	General-purpose regression, model selection, optimization, robustness to outliers, time series analysis
R-squared ( $R^2$ )	Model evaluation, goodness-of-fit, linear regression, multiple regression
Adjusted R-squared	Model evaluation, goodness-of-fit, linear regression, multiple regression with many predictors
Mean Squared Logarithmic Error (MSLE)	Forecasting, model evaluation, skewed target distributions, finance, sales prediction
Mean Absolute Percentage Error (MAPE)	Forecasting, model evaluation, time series analysis, business analytics, supply chain optimization

RMSE is widely used in fields where accurate error magnitude is critical, such as in financial forecasting, energy load predictions, and more generally in regression analysis and machine learning. This is due to its ability to penalize large errors more severely, which can be crucial for some applications.

The limitations of the RMSE are two-fold: on the one hand, it is sensitive to outliers; on the other hand, RMSE does not distinguish between types of errors (systematic and random). The scale and size of the dataset also influences it:

$$\text{Scale Influence: } RMSE \propto \sigma_y, \quad (10)$$

where  $\sigma_y$  is the standard deviation of the observed values.

To make RMSE comparable across different datasets or models, it can be normalized:

$$\text{Normalized RMSE} = \frac{RMSE}{\sigma_y}, \quad (11)$$

which standardizes the RMSE by the variability of the dataset, thereby removing scale effects.

Interpreting RMSE involves comparing it to the standard deviation of the dependent variable. A lower RMSE relative to the standard deviation suggests a model with a better fit:

$$\text{Relative Error} = \frac{RMSE}{\sigma_y}, \quad (12)$$

which helps assess model accuracy relative to the inherent variability of the data.

**Mean Absolute Percentage Error (MAPE)** The MAPE measures the average percentage error of the model's predictions compared to the true values. It is defined as

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \times 100, \quad (13)$$

where  $y_i$  is the true value,  $\hat{y}_i$  is the predicted value, and  $n$  is the number of samples.

One of the advantages of using MAPE is that it is easy to interpret, as it is expressed in percentage terms. It is also scale-independent, which allows comparison of forecast accuracy across different scales of the target variable. However, MAPE has two main limitations:

- It can produce undefined results when  $y_i = 0$ , which may require modifying the dataset or the formula to handle such cases.
- It is sensitive to outliers because large errors in the denominator can disproportionately influence the overall percentage error.

MAPE is particularly useful in scenarios where the relative error is more important than the absolute error magnitude. It is extensively used in industries like finance and retail where percentage errors provide a more intuitive sense of prediction errors relative to scales. Moreover, it is beneficial in comparative model assessments across datasets where absolute scales of targets vary widely.

**Symmetric Mean Absolute Percentage Error (SMAPE)** The SMAPE is a variation of the Mean Absolute Percentage Error (MAPE) commonly used to evaluate the accuracy of predictions in time series forecasting [49]. SMAPE is defined as

$$SMAPE = \frac{2}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|} * 100, \quad (14)$$

where  $y_i$  is the true value,  $\hat{y}_i$  is the predicted value, and  $n$  is the number of samples.

One of the advantages of using SMAPE is that it is symmetric, which means that it gives equal weight to over-predictions and under-predictions. This is particularly useful when working with time series data, where over-predictions and under-predictions may have different implications, and SMAPE helps to ensure that the model is equally penalized for both types of errors, leading to better overall performance in terms of how well it meets the business needs or objectives. However, SMAPE has some limitations; for example, it can produce undefined results when both  $y_i$  and  $\hat{y}_i$  are zero and can be sensitive to outliers.

The implications of over-predictions and under-predictions varied depending on the application. In the following, we discuss real-world examples.

**Inventory Management:** Over-predicting demand can lead to excess inventory, which ties up capital and can result in waste if products expire or become obsolete. Under-predicting demand can lead to stockouts, lost sales, and damage to customer relationships [50]. A symmetric error measure like SMAPE penalizes both cases because over-prediction and under-prediction have costly implications.

**Energy Demand Forecasting:** Over-prediction of energy demand can cause unnecessary production, leading to waste and increased costs. Under-prediction can lead to insufficient power generation, resulting in blackouts or the need for expensive on-demand power generation [51].

**Financial Markets:** In financial markets, over-prediction of a stock price might lead to unwarranted investments resulting in financial loss, while under-prediction might result in missed opportunities for gains [52].

**Sales Forecasting:** Over-prediction of sales could lead to overstaffing, overproduction, and increased costs, while under-prediction could lead to understaffing, missed sales opportunities, and decreased customer satisfaction [53].

**Transportation and Logistics:** Over-predicting the demand for transportation might lead to underutilized vehicles or routes, resulting in unnecessary costs. Under-predicting demand might lead to overcrowding and customer dissatisfaction [54].

**Coefficient of Determination  $R^2$**  The Coefficient of Determination ( $R^2$ ), measures how well the model explains the variation in the target variable [55].  $R^2$  is defined as the proportion of the variance in the target variable that the model explains, ranging from 0 to 1. A value of 0 indicates that the model does not explain any variation in the target variable, while a value of 1 means it explains all the variation. It is possible for  $R^2$  to be negative, which indicates that the model performs worse than a simple mean of the target variable.

The formula for R-squared is:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (15)$$

where  $y_i$  is the true value,  $\hat{y}_i$  is the predicted value,  $\bar{y}$  is the mean of the true values, and  $n$  is the number of samples.

### Benefits and Limitations of R-squared

1. Measures the relationship between the model and the response variable: It describes the strength of the relationship on a scale from 0 to 1, which is intuitively understandable.
2. Provides a percentage indicating how much of the variability in the response variable can be explained by the model.
3. Allows comparison of models; higher  $R^2$  values generally indicate a better fit, though this should be used cautiously.

The limitations of  $R^2$  include:

1. It is best suited for linear models, can be misleading for non-linear models or complex relationships.
2. Can increase with the number of predictors, potentially leading to overfitting. Adjusted  $R^2$  is often preferred in multiple regression contexts to account for this.
3. Extreme values can significantly impact  $R^2$ .
4. Can not tell if predictions are systematically over or underestimate actual values.
5. May not accurately reflect the strength of a relationship in small datasets.

**Adjusted  $R^2$**  Adjusted  $R^2$  is a modified version of  $R^2$  that has been adjusted for the number of predictors in the model. It increases only if the new term improves the model more than would be expected by chance, and it decreases when a predictor improves the model by less than expected by chance [56]. The adjusted  $R^2$  is defined as

$$\text{Adjusted } R^2 = 1 - \left( \frac{(1 - R^2)(n - 1)}{n - k - 1} \right), \quad (16)$$

where  $n$  is the number of observations and  $k$  is the number of predictors. This adjustment serves as a penalty for adding unnecessary predictors to the model, with the penalty increasing as the number of predictors increases. Consequently, adjusted  $R^2$  provides a more accurate assessment of model fit, particularly in multiple regression analyses where several predictors are used simultaneously.

Adjusted  $R^2$  is commonly used for model comparison because it does not necessarily increase with the addition of more variables, unlike regular  $R^2$ . It allows for fair comparisons between models of different sizes by penalizing overfitting and rewarding models that achieve a better fit without unnecessarily increasing complexity.

Interpreting adjusted  $R^2$  values is similar to interpreting regular  $R^2$ , with positive values indicating the proportion of variance explained by the model after adjusting for the number of predictors. However, negative values indicate that the model is worse than a model that simply predicts the mean of the dependent variable.

For example, consider a regression model predicting housing prices using various predictors such as square footage, number of bedrooms, and location. The adjusted  $R^2$  value would help determine whether adding additional predictors improves the model's predictive power or simply adds unnecessary complexity.

### 3.2 Classification

Classification is a supervised machine learning task that trains a model to predict the category or class of a given input data point. The primary goal of classification is to help the model establish a mapping between input features and a specific class or category.

There are different types of classification tasks, including binary classification, multi-class classification, and multi-label classification. Binary classification involves training the model to predict one of two classes, such as "spam" or "not spam," for an email. On the other hand, multi-class classification requires the model to predict one of several classes, such as "dog," "cat," or "bird," for an image. In multi-label classification, the model is trained to predict multiple labels for a single data point, such as "dog" and "outdoor," for an image of a dog in the park.

Classification algorithms are based on various techniques such as decision trees, Naive Bayes, k-nearest neighbors, Support Vector Machines, Random Forest, Gradient Boosting, Neural Networks, and others. In the following subsections, we will delve deeper into the most common classification loss functions and performance metrics.

#### 3.2.1 Classification Loss Functions

Several loss functions can be used for classification tasks, depending on the specific problem and algorithm. Table 6 summarizes the most common loss functions used in classification. In the following sections, we describe each of these loss functions with more detail:

**Binary Cross-Entropy Loss (BCE)** Binary Cross-Entropy Loss, also known as log loss, is a prevalent loss function for binary classification problems. It measures the dissimilarity between the predicted probability and the true class label [57]. Cross-entropy originates from information theory, where it measures the dissimilarity between two probability distributions.

In the context of binary classification, where the classes are labeled as 0 or 1, the BCE loss for an individual prediction is defined as:

Table 6: Loss Functions and their applications in classification tasks.

Loss Function	Applications
Binary Cross-Entropy (BCE)	Binary classification, Logistic regression, Neural networks
Weighted Binary Cross-Entropy (WBCE)	Imbalanced binary classification, Rare event prediction
Categorical Cross-Entropy (CCE)	Multiclass classification, Deep learning models
Sparse Categorical Cross-Entropy	Multiclass classification with large number of classes, Natural language processing
Cross-Entropy with Label Smoothing	Multiclass classification with noisy labels, Neural networks to improve generalization
PolyLoss	Imbalanced data, multiple tasks, 2D image classification, 2D and 3D object detection, instance segmentation.
Hinge Loss	Support vector machines, Binary classification with margin maximization

$$L(y, p) = -(y \log(p) + (1 - y) \log(1 - p)) \quad (17)$$

This formula can be intuitively understood by considering it in two parts:

$$\begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1 - p) & \text{if } y = 0, \end{cases} \quad (18)$$

where  $y$  is the true class label and  $p$  is the predicted probability of the class being 1. The loss is minimized when the predicted probability  $p$  matches the actual label  $y$ .

BCE is particularly useful because it is differentiable and provides a probabilistic interpretation of the model's output, essential for gradient-based optimization algorithms. It also naturally penalizes incorrect classifications with high confidence, aligning well with the goals of logistic regression, which uses BCE as its loss function.

However, BCE can be sensitive to class imbalance, where the number of instances in one class significantly outnumbers the other. To address this, a variant called *Weighted Binary Cross-Entropy* can be used, which assigns a higher weight to the minority class, helping to balance the influence of each class on the training process.

$$L_{weighted}(y, p) = -(w_1 \cdot y \log(p) + w_0 \cdot (1 - y) \log(1 - p)), \quad (19)$$

where  $w_1$  and  $w_0$  are weights for the positive and negative classes, respectively. These weights are typically set inversely proportional to the class frequencies.

**Categorical Cross-entropy Loss (CCE)** Categorical Cross-Entropy Loss, also known as multi-class log loss, is used in multi-class classification tasks to measure the dissimilarity between the predicted probabilities and the actual distribution of the outcomes. It extends the concept of binary cross-entropy to multiple classes.

The loss for categorical cross-entropy is defined as the average negative log-likelihood of the correct class:

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C y_{i,j} \log(p_{i,j}), \quad (20)$$

where  $N$  is the number of samples,  $C$  is the number of classes,  $y_{i,j}$  is the true label in a one-hot encoded format, and  $p_{i,j}$  is the predicted probability of class  $j$  for sample  $i$ . This formulation means that the loss calculates the logarithm of the predicted probability for the actual class and penalizes the model based on the distance from the actual label.

One-hot encoding represents the true class of each sample with a vector of length  $C$ , where the index of the true class is marked with a 1, and all other entries are 0. This method ensures that the model's prediction for the correct class is strongly reinforced if accurate and penalized if not.

CCE is critical for models where the accurate prediction of multiple classes is necessary. It is differentiable, making it suitable for optimization algorithms such as gradient descent used in training deep neural networks. Like binary cross-entropy, it severely penalizes incorrect confident predictions, which enhances the learning efficacy of classifiers.

**Sparse Categorical Cross-entropy Loss** Sparse Categorical Cross-Entropy Loss is a variation of the standard categorical cross-entropy loss designed for multi-class classification tasks where the classes are encoded as integers rather than one-hot encoded vectors. This approach is particularly useful in scenarios with a large number of classes, where one-hot encoding would be memory-intensive and less efficient.

The loss for an individual prediction is calculated directly using the integer label as an index:

$$H(y, \hat{y}) = -\log(\hat{y}_{i, y_i}) \quad (21)$$

The overall Sparse Categorical Cross-Entropy Loss for the dataset is the average of these individual losses:

$$H(Y, \hat{Y}) = -\frac{1}{n} \sum_{i=1}^n \log(\hat{y}_{i, y_i}), \quad (22)$$

where  $y_i$  is the integer label representing the true class of the  $i$ -th sample, and  $\hat{y}_{i, y_i}$  is the predicted probability of that class. This formulation allows for a direct reference to the predicted probability corresponding to the actual class label, simplifying the computational process.

Sparse Categorical Cross-Entropy is differentiable, making it suitable for use with gradient descent and other optimization algorithms essential for training deep neural networks. Sparse CCE not only enhances computational efficiency but also aligns well with situations where the categorical labels are naturally ordinal or where the encoding of categories as integers simplifies data preprocessing and handling.

**Cross-Entropy Loss with Label Smoothing** Cross-Entropy loss with label smoothing is an advanced regularization technique used to prevent models from becoming overly confident in their predictions by smoothing the labels. This technique modifies the standard Cross-Entropy loss by adjusting the target labels: a small constant  $\epsilon$  is subtracted from the target class and evenly distributed among the other class labels. This approach is particularly effective in mitigating overconfidence, a common issue when training on large datasets, which can lead to degraded performance on unseen data [58][59].

The formula for Cross-Entropy loss with label smoothing is adjusted as follows:

$$L(y, \hat{y}) = -\sum_{c=1}^C \left[ (1 - \epsilon)y_c \log \hat{y}_c + \frac{\epsilon}{C - 1} (1 - y_c) \log \hat{y}_c \right], \quad (23)$$

where  $y$  represents the true label vector,  $\hat{y}$  the predicted probabilities,  $C$  the number of classes, and  $\epsilon$  a small constant (often set between 0.1 and 0.2). Here, each incorrect class receives a portion of  $\epsilon$  divided by  $C - 1$ , ensuring that the sum of probabilities remains normalized.

Label smoothing adjusts the confidence level of the predictions by penalizing extreme values in the log probabilities, thereby encouraging the model to allocate probability mass more evenly across different classes. This technique has been shown to improve the generalization of models, particularly in scenarios with many categories or when the dataset contains noisy labels. However, the optimal value of  $\epsilon$  may vary depending on specific task requirements and characteristics of the data, often requiring empirical tuning to achieve the best results.

**Negative Log-likelihood** In many machine learning applications, the true distribution  $p$  is represented as a one-hot encoded vector, where only the true class has a probability of 1, and all others have 0. The predicted distribution  $q$  is the output of the model. The cross-entropy loss for a single instance can be expressed as:

$$H(p, q) = -\sum_i p_i \log(q_i), \quad (24)$$

where  $p$  is the true distribution and  $q$  is the predicted distribution. Given that the true distribution  $p$  is one-hot encoded, the summation simplifies to:

$$H(p, q) = -\log(q_{y_t}), \quad (25)$$

where  $y_t$  denotes the true class. This expression is equivalent to the Negative Log Likelihood (NLL), which is defined as:

$$\text{NLL} = -\log(p(y_t|x)) \quad (26)$$

Thus, in classification tasks with one-hot encoded labels, minimizing the cross-entropy loss is equivalent to minimizing the negative log-likelihood. Both metrics serve to encourage the model to assign a higher probability to the true class.

**PolyLoss** PolyLoss [60] generalizes common loss functions like cross-entropy and focal loss. Represents loss functions as a linear combination of polynomial functions, offering a new perspective on how these functions can be adapted for specific tasks and datasets.

The PolyLoss framework is inspired by the Taylor series expansion, where the loss function  $L$  is expressed as:

$$L = \sum_{n=0}^N \alpha_n (1 - p_t)^n, \quad (27)$$

where  $p_t$  is the predicted probability for the target class, and  $\alpha_n$  are the coefficients for the  $n$ -th polynomial term. This framework provides a mechanism for adjusting the polynomial coefficients  $\alpha_n$ , which in turn modulates the sensitivity and emphasis of the loss function on various aspects of the predictions.

**Poly-1 loss** A variant within this framework is the Poly-1 loss, which introduces a single additional hyperparameter  $\epsilon$ . The Poly-1 loss can be formulated as:

$$L_{\text{Poly-1}} = \text{CE} + \epsilon(1 - p_t), \quad (28)$$

where CE is the standard cross-entropy loss, and  $\epsilon$  is a hyperparameter that adjusts the influence of the additional term. By setting  $\epsilon = 0$ , Poly-1 reduces to the standard cross-entropy loss. When  $\epsilon > 0$ , the loss function becomes more sensitive to confident predictions, reducing overfitting in imbalanced datasets or tasks requiring higher precision.

PolyLoss is particularly beneficial in the following scenarios:

1. When dealing with imbalanced datasets, traditional loss functions like cross-entropy and focal loss may not perform optimally. PolyLoss can be tailored to adjust the importance of different classes through its polynomial coefficients, making it more effective for such cases.
2. When working on various tasks or datasets, PolyLoss allows to modify polynomial coefficients for a customized approach, which can lead to better performance across different applications.
3. To simplify the hyperparameter optimization process. The Poly-1 formulation, for example, introduces only one additional hyperparameter, making it easier to tune compared to more complex loss functions.

**Hinge Loss** Hinge Loss is a widely used loss function in maximum-margin classification tasks, particularly with Support Vector Machines (SVMs). It's especially prevalent in scenarios requiring a robust classification margin, such as in one-vs-all classification tasks where an instance is classified into one of several categories [61].

The Hinge Loss for a given instance is formulated as:

$$L(y, f(x)) = \max(0, 1 - y \cdot f(x)), \quad (29)$$

where  $y$  represents the true label of the instance, taking values -1 or 1 in binary classification contexts, and  $f(x)$  is the predicted output from the model for the input  $x$ . The term  $y \cdot f(x)$  reflects the raw margin, which measures how far the predicted value is from the decision boundary in terms of alignment and distance.

The essence of Hinge Loss is that it is zero if the predicted classification is correct and the instance is beyond the margin (the correct side). If the classification is wrong, or if the classification is correct but too close to the decision boundary (the wrong side of the margin), the loss increases linearly with the distance from the margin.

Hinge Loss not only penalizes classification errors but also correct classifications that are not confidently made (i.e., those close to the decision boundary). This characteristic helps in forming a decision boundary that is not only accurate in separating classes but also maximizes the distance (margin) between the nearest points of the classes and the boundary itself, enhancing model generalization.

In addition to its basic form, variants such as the squared hinge loss ( $\max(0, 1 - y \cdot f(x))^2$ ) are used to penalize incorrect classifications more strongly, promoting an even more robust classification boundary. Although primarily associated with binary classification, Hinge Loss can be adapted for multi-class problems through approaches like structured SVMs [62], which handle complex inter-class relationships.

### 3.2.2 Classification Performance Metrics

In this section, we will discuss the various metrics that are commonly used to evaluate the performance of classification models. Table 7 provides a summary of these metrics, and the following subsections describe each one in more detail.



Table 7: Metrics used in classification task.

Common Name	Other Names	Abbr	Definitions	Interpretations
True Positive	Hit	TP	True Sample labeled true	Correctly labeled True Sample
True Negative	Rejection	TN	False Sample labeled false	Correctly labeled False sample
False Positive	False alarm	FP	False sample labeled True	Incorrectly labeled False sample
False Negative	Type I Error Miss, Type II Error	FN	True sample labeled false	Incorrectly label True sample
Recall	True Positive Rate	TPR	$TP/(TP+FN)$	% of True samples correctly labeled
Specificity	True Negative Rate	SPC, TNR	$TN/(TN+FP)$	% of False samples correctly labeled
Precision	Positive Predictive Value	PPV	$TP/(TP+FP)$	% of samples labeled True that really are True
Negative Predictive Value		NPV	$TN/(TN+FN)$	% of samples labeled False that really are False
False Negative Rate		FNR	$FN/(TP+FN)=1-TPR$	% of True samples incorrectly labeled
False Positive Rate	Fall-out	FPR	$FP/(FP+FN)=1-SPC$	% of False samples incorrectly labeled
False Discovery Rate		FDR	$FP/(TP+FP)=1-PPV$	% of samples labeled True that are really False
True Discovery Rate		TDR	$FN/(TN+FN)=1-NPV$	% of samples labeled False that are really True
Accuracy		ACC	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$	Percent of samples correctly labeled
F1 Score		F1	$\frac{(2*TP)}{((2*TP)+FP+FN)}$	Approaches 1 as errors decline

**Confusion Matrix** The confusion matrix is used to define a classification algorithm's performance. It contains the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) that result from the algorithm. The confusion matrix for a binary classification problem is represented in a 2x2 table as shown in Table 8

Table 8: Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

1. True Positives (TP): These are the cases where the model correctly predicts the positive class. In other words, the instances that are actually positive and are correctly identified as positive by the model.
2. True Negatives (TN): These are the cases where the model correctly predicts the negative class. This means the instances are actually negative and are identified as negative by the model.
3. False Positives (FP): Also known as Type I errors, these occur when the model incorrectly predicts the positive class. That is, the instances are actually negative but are wrongly labeled as positive. This is sometimes called "false alarm".
4. False Negatives (FN): Also known as Type II errors, these occur when the model incorrectly predicts the negative class. These instances are actually positive but are mistakenly labeled as negative. This is sometimes referred to as a "miss".

These metrics provide insights into how well the model performs and where it might be making the most mistakes, enabling targeted improvements.

Using the values in the confusion matrix, we can calculate performance metrics such as accuracy, precision, recall, and F1-score.

**Confusion Matrix in Multi-class Classification** In multiclass classification, the confusion matrix is an extension of the binary classification confusion matrix, where there are more than two classes to predict. The matrix helps in visualizing the performance of the classification algorithm across all classes.

For a classification task with  $N$  classes, the confusion matrix will be an  $N \times N$  matrix. Each row of the matrix corresponds to the true class, while each column corresponds to the predicted class. The element at the  $i$ -th row and  $j$ -th column of the matrix, denoted as  $M[i, j]$ , represents the number of instances where the true class is  $i$  but the predicted class is  $j$ .

The diagonal elements ( $M[i, i]$ ) represent the number of points that were correctly classified as class  $i$ . Therefore, for all  $i$  where  $1 \leq i \leq N$ ,  $M[i, i]$  are the true positives for class  $i$ .

The off-diagonal elements ( $M[i, j]$ ) where  $i \neq j$  indicate misclassifications. Specifically,  $M[i, j]$  is the number of observations that belong to class  $i$  (true class) but were predicted as class  $j$  (incorrect class).

For example, if we have three classes, A, B, and C, the confusion matrix might look something like Table 9

Table 9: Confusion Matrix for Multiclass Classification. The diagonal elements  $M[i, i]$  represent the number of points that were correctly classified as class  $i$ . The off-diagonal elements indicate misclassifications. Specifically,  $M[i, j]$  is the number of observations that belong to class  $i$  (true class) but were predicted as class  $j$

	Predicted: A	Predicted: B	Predicted: C
Actual: A	$M[1, 1]$	$M[1, 2]$	$M[1, 3]$
Actual: B	$M[2, 1]$	$M[2, 2]$	$M[2, 3]$
Actual: C	$M[3, 1]$	$M[3, 2]$	$M[3, 3]$

**Multi-class Confusion Matrix Visualization** Multi-class confusion matrices are often presented as heatmaps as shown in Figure 1. This visualization represents the confusion matrix of a multi-class classification model with three classes: Class A, Class B, and Class C. Each cell in the matrix shows the number of samples from the true class (rows) that were predicted as each class (columns). The diagonal cells (from the top left to the bottom right) indicate the number of correct predictions for each class, highlighting the model’s accuracy for those specific categories.

Class A: Out of the total samples belonging to Class A, 95 were correctly predicted as Class A, while 2 were misclassified as Class B, and 3 as Class C.

Class B: For Class B, 90 samples were correctly classified, 4 were incorrectly predicted as Class A, and 6 as Class C.

Class C: In the case of Class C, 85 samples were correctly identified, with misclassifications including 7 as Class A and 8 as Class B. The non-diagonal cells reveal the instances of misclassification. For example, more samples of Class C were misclassified as Class B than as Class A, indicating specific areas where the model’s classification performance could be improved.

**Accuracy** Accuracy is a widely used performance metric in classification tasks, defined as the ratio of correctly classified samples to the total number of samples [63]. For a simple formula representation:

$$Accuracy = \frac{\text{Number of Correctly Classified Samples}}{\text{Total Number of Samples}} \quad (30)$$

In terms of a confusion matrix, which is a table used to describe the performance of a classification model, accuracy is calculated as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (31)$$

where TP, TN, FP, and FN represent the true positives, true negatives, false positives, and false negatives, respectively.

While straightforward and intuitive, accuracy can be misleading in datasets with imbalanced class distributions. For instance, in a dataset where 99 out of 100 samples are negative for a condition (e.g., no cancer), a model predicting ‘No cancer’ for all instances would achieve 99% accuracy, despite not being useful in practical terms.

**Accuracy in Multi-Class Classification** In multi-class scenarios, accuracy is extended by summing the correctly predicted instances across all classes and dividing by the total number of samples:

$$Accuracy = \frac{\sum_{i=1}^C \text{Correct}_i}{\text{Total Samples}} \quad (32)$$

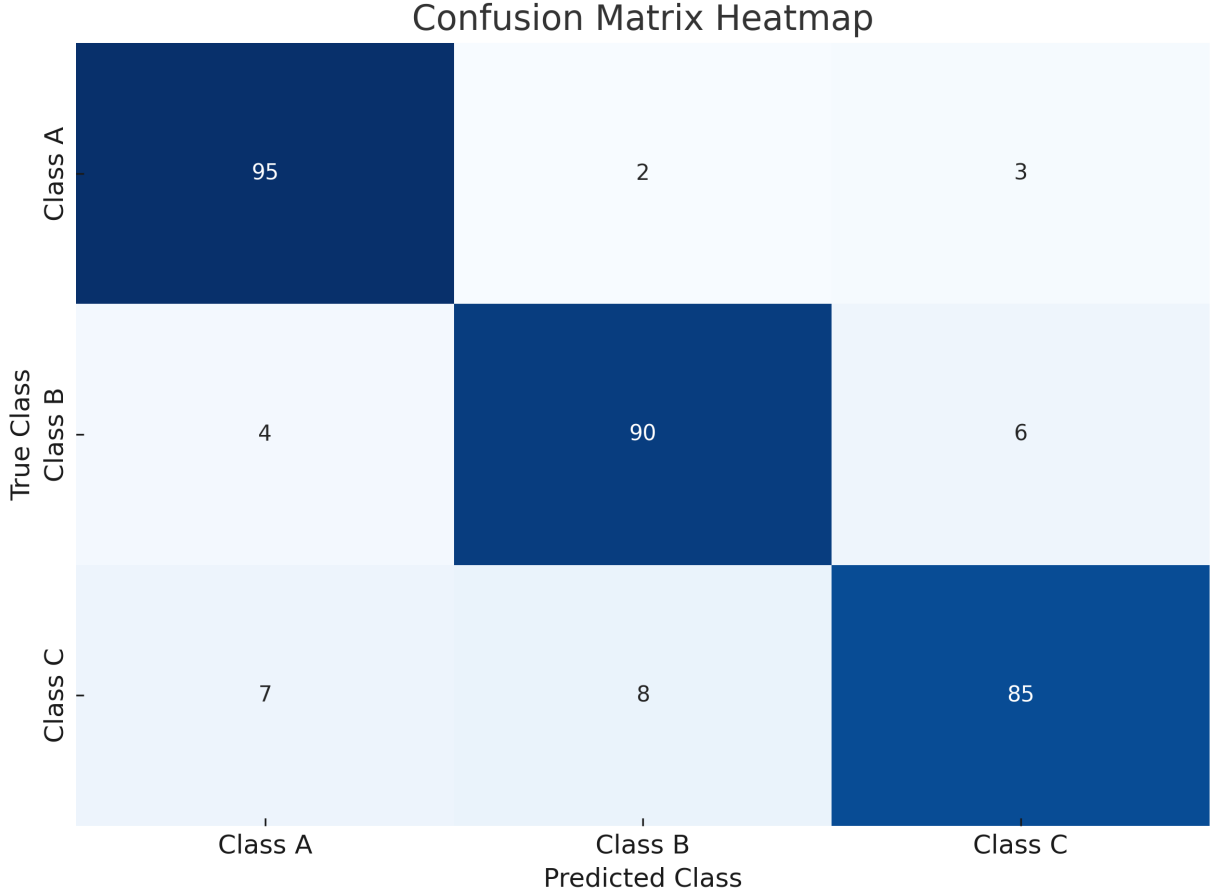


Figure 1: Confusion Matrix Heatmap for Multi-Class Classification. The matrix visualizes the performance of a classification model across three classes: Class A, Class B, and Class C. The values along the diagonal represent the number of correct predictions for each class, whereas off-diagonal values indicate misclassifications. The color intensity in each cell correlates with the number of samples, providing a clear visual distinction between higher and lower frequencies of predictions. This figure aids in identifying the strengths and weaknesses of the model, particularly in distinguishing between the classes..

where  $C$  is the number of classes, and  $Correct_i$  is the number of correctly predicted instances for each class. This measure treats all classes equally, but it does not account for the imbalance among different classes.

**Balanced Accuracy** To address class imbalance, particularly in multi-class settings, balanced accuracy can be considered. This metric calculates the average of recall obtained on each class:

$$Balanced\ Accuracy = \frac{1}{C} \sum_{i=1}^C \frac{TP_i}{TP_i + FN_i} \quad (33)$$

where  $TP_i$  and  $FN_i$  are the true positives and false negatives for each class  $i$ . Balanced accuracy thus provides a more equitable measure of model performance across all classes.

**Precision, PPV, or TDR** Precision, also known as positive predictive value (PPV) or true discovery rate (TDR), measures the accuracy of positive predictions made by a model. It is defined as the number of true positive predictions divided by the total number of predicted positives, which includes both true positives and false positives [64]. The formula for precision can be expressed as:

$$Precision = \frac{TP}{TP + FP}, \quad (34)$$

where  $TP$  represents the number of true positive predictions, and  $FP$  stands for the number of false positive predictions. Precision is particularly critical in domains where the consequences of false positives are significant, such as in medical diagnostics or fraud detection. High precision indicates that the model produces few false positives, leading to more reliable predictions. However, it's crucial to consider precision in conjunction with other performance metrics like recall, as a model could achieve high precision at the cost of recall, thereby missing a substantial number of actual positives.

**Multi-Class Precision** In multi-class classification scenarios, precision is calculated for each class individually and then averaged, which can be done in several ways depending on the context and specific requirements of the evaluation. The two common methods are:

- **Macro-average Precision:** This method calculates precision independently for each class and then takes the average. This approach treats all classes equally, regardless of their support (frequency). Mathematically, it is defined as:

$$\text{Macro-average Precision} = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FP_i}, \quad (35)$$

where  $N$  is the number of classes, and  $TP_i$  and  $FP_i$  are the true positives and false positives for the  $i$ -th class, respectively.

- **Micro-average Precision:** This method aggregates the contributions of all classes to compute the average precision. It calculates the total true positives and false positives across all classes and then applies the precision formula. This approach is influenced more by the class with a larger number of instances. It can be represented as:

$$\text{Micro-average Precision} = \frac{\sum_{i=1}^N TP_i}{\sum_{i=1}^N (TP_i + FP_i)}, \quad (36)$$

where  $N$  is the number of classes.

Both averaging methods are useful but serve different purposes depending on the objective of the analysis. Macro-average is beneficial when you want to treat all classes equally, while micro-average is preferable when you expect a balance between classes to be reflected in the precision metric.

**Recall, Sensitivity, or True Positive Rate (TPR)** Recall, also known as sensitivity or true positive rate (TPR), is a measure of a model's ability to correctly identify positive instances among all actual positives in the dataset. It is calculated as the ratio of true positives to the sum of true positives and false negatives, where a true positive is an instance correctly identified as positive, and a false negative is a positive instance that was incorrectly identified as negative [64]. Mathematically, recall can be expressed as:

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (37)$$

where  $TP$  is the number of true positives, and  $FN$  is the number of false negatives.

A high recall score indicates that the model is successful in capturing a large proportion of positive cases, which is especially important in applications where failing to detect positives (such as diseases or fraudulent transactions) could have severe consequences. However, optimizing for recall may increase the number of false positives, thereby reducing precision. It is common practice to balance recall with precision and other metrics based on the specific needs of the application.

**Multi-Class Recall** In multi-class classification problems, recall must be assessed for each class and then averaged, depending on the specific needs and the nature of the dataset. There are generally two methods to average recall in multi-class settings:

- **Macro-average Recall:** This method calculates recall for each class independently and then takes the average of these measures. This treats all classes equally, regardless of their prevalence. The formula for macro-average recall is:

$$\text{Macro-average Recall} = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FN_i}, \quad (38)$$

where  $N$  is the number of classes,  $TP_i$  is the true positives for the  $i$ -th class, and  $FN_i$  is the false negatives for the  $i$ -th class.

- **Micro-average Recall:** This method aggregates the contributions of all classes to compute the overall recall. It is particularly useful when class imbalance is significant. The total number of true positives and false negatives across all classes are summed up, and the recall formula is applied to these totals:

$$\text{Micro-average Recall} = \frac{\sum_{i=1}^N TP_i}{\sum_{i=1}^N (TP_i + FN_i)}, \quad (39)$$

where  $N$  is the number of classes.

Both averaging techniques are valuable and can be selected based on whether class equality or dataset representation is more important for the analytical goal.

**Precision-Recall Tradeoff** The precision-recall tradeoff describes the inverse relationship between precision and recall. As one metric increases, the other tends to decrease. This inverse relationship is crucial in settings where decisions need to balance the risks of false positives and false negatives.

Consider a machine learning model used for spam detection. If the model is set to be very conservative (high threshold), it is likely to mark only the most obvious spam emails as spam. This approach tends to result in high precision because most emails classified as spam are truly spam, but it may suffer from low recall as some spam emails are missed. Conversely, if the model is set to be more liberal (low threshold), it marks more emails as spam. This method increases recall by identifying most spam emails but at the expense of precision, as more legitimate emails are incorrectly flagged as spam.

This tradeoff between precision and recall is fundamental in many applications. For instance, in medical diagnostics, a high recall might be prioritized to ensure that all potential disease cases are identified, even at the cost of some false positives. On the other hand, in spam detection systems, high precision might be more desirable to prevent legitimate emails from being incorrectly marked as spam, even if some spam emails pass through the filters.

The balance between precision and recall can be illustrated through a Precision-Recall curve (refer to section [3.2.2](#)). This curve shows how these two metrics change based on different thresholds. The F1 score is a commonly used metric for situations where a single metric is required to balance precision and recall without prioritizing either one. We will discuss this metric in more detail in the following section.

**F1-score** The F1 score is a statistical measure that combines precision and recall to provide a single value representing a classification model's overall performance [\[64\]](#). It is defined as the harmonic mean of precision and recall, computed as:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (40)$$

This score considers both the model's ability to correctly identify positive examples (precision) and its ability to identify all positive examples in the dataset (recall). A higher F1 score indicates a better balance between precision and recall, suggesting that the model is not overly skewed towards one metric at the expense of the other. This balance makes the F1 score especially useful in scenarios with imbalanced class distributions, where one class might dominate over others.

The main limitation of the F1 score is that it does not differentiate between the types of errors (false positives and false negatives beyond their proportional impact on precision and recall). Therefore, in certain contexts where the cost of false positives differs significantly from the cost of false negatives, other metrics may be more appropriate.

**F-beta Score** An extension of the F1 score is the F-beta score, which introduces a flexibility in balancing precision and recall according to specific requirements of different applications. It is defined as:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}} \quad (41)$$

where  $\beta$  is a positive real factor, placing more emphasis on recall when  $\beta > 1$  and more on precision when  $\beta < 1$ . This customization allows the metric to be more closely tailored to the specific costs associated with false positives and false negatives in various use cases.

**Specificity or True Negative Rate (TNR)** This metric measures the proportion of negatives that are correctly identified as such by a classification model [65]. It is defined mathematically as the number of true negatives (TN) divided by the sum of true negatives and false positives (FP):

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (42)$$

Specificity is used in fields like medical diagnostics, where a high value indicates that the test is effective at ruling out non-diseased individuals, thus minimizing unnecessary treatments or anxiety. This metric complements recall by focusing on different aspects of the testing outcome. Recall assesses the proportion of actual positives correctly identified, pertinent in scenarios where missing a positive case has severe implications, such as in cancer screening. Specificity, on the other hand, assesses the accuracy of identifying negatives, important in situations where falsely identifying a condition could lead to harmful interventions for example, in screening tests where a positive result could lead to more invasive testing or anxiety. It is also vital in industries like finance, where identifying a transaction as fraudulent when it is not can block legitimate transactions and potentially harm customer relations.

**False Positive Rate (FPR)** The False Positive Rate (FPR) measures the proportion of false positives among all actual negative instances. It is also known as the *Type I Error rate* and is an important complement to the Specificity metric (also called True Negative Rate). The relationship between FPR and specificity is given by  $FPR = 1 - \text{Specificity}$ .

Formally, the FPR is calculated as follows:

$$FPR = \frac{FP}{TN + FP} \quad (43)$$

FPR is influenced by the threshold set for classifying instances as positive or negative. Lowering the threshold generally results in more instances being classified as positive, thus potentially increasing the FPR if many of these are actually negatives. Conversely, raising the threshold tends to decrease the FPR by classifying fewer instances as positive, reducing the risk of false positives.

In practice, FPR is commonly used in the construction of Receiver Operating Characteristic (ROC) curves, where it is plotted on the x-axis against the True Positive Rate (TPR) on the y-axis. This curve is useful for visualizing the trade-off between TPR and FPR at various threshold settings. The ROC curve and its analysis provide insights into the diagnostic ability of a classifier, regardless of the class distribution. See section 3.2.2 for more details on ROC curves and Area Under the Curve (AUC), which provides a single measure of overall model performance across multiple classification thresholds.

**Negative Predictive Value (NPV)** The Negative Predictive Value (NPV) measures the proportion of negative cases that are correctly identified as such. It specifically assesses the accuracy of negative predictions made by a classification model, calculated as follows:

$$NPV = \frac{TN}{TN + FN} \quad (44)$$

This metric is valuable where the consequences of false negatives are severe. For example, in medical diagnostics, a false negative might mean a delay in necessary treatment for a patient, potentially leading to worsened outcomes or even death. Thus, a high NPV is crucial in these situations.

Unlike some metrics such as sensitivity or specificity, NPV is directly affected by the prevalence of the condition being tested for, as it depends on the ratio of negatives (both true and false) in the dataset. This makes NPV especially useful in situations where negative outcomes (non-occurrence of the condition) are much more common than positive ones.

NPV complements the Positive Predictive Value (PPV) or precision, which measures the proportion of positive identifications that were actually correct. NPV and PPV together provide a comprehensive view of a model's predictive accuracy regarding both positive and negative classifications.

In practice, NPV is useful in scenarios with imbalanced classes. For instance, in a disease screening where the disease is rare, the number of negatives (non-disease cases) will vastly outnumber the positives. In such cases, ensuring that the negatives are correctly identified (high NPV) without many false negatives becomes as important as identifying the positives.

The statistical relationship between the prevalence of a condition and NPV suggests that when the prevalence of a disease decreases, it becomes crucial to maintain a high NPV by controlling the number of false negatives. This is especially important in designing and evaluating screening tests for rare diseases where achieving a high NPV is often more meaningful.

**False Discovery Rate (FDR)** The False Discovery Rate (FDR) measures the proportion of false positives among all positive predictions made by a classifier. This metric is crucial in fields where making incorrect positive predictions incurs significant costs or consequences. The formula for FDR is given by:

$$FDR = \frac{FP}{TP + FP} \quad (45)$$

A lower FDR value indicates that a smaller proportion of the positive predictions made by the classifier are false, signifying a higher reliability of positive results.

Unlike the False Positive Rate (FPR), which measures the proportion of false positives out of all actual negatives, the FDR focuses on the proportion of false positives within the group that the model predicts as positive. This distinction makes FDR particularly valuable in scenarios like genomic research or machine learning applications where the primary concern is the precision of the positive predictions rather than the impact on all negatives.

FDR is extensively used in scenarios where the cost of false positives is high, such as in scientific research where false discoveries can lead to incorrect scientific conclusions and waste of resources. In medical testing or fraud detection, controlling FDR ensures that the positive identifications are credible, which is crucial for subsequent decision-making processes.

**Precision-Recall Curve** The precision-recall (PR) curve is a graphical representation that illustrates the trade-off between precision and recall for different threshold values of a classifier. Precision measures the proportion of true positive predictions out of all positive predictions, while recall measures the proportion of true positive predictions out of all actual positive instances. The PR curve plots precision on the y-axis and recall on the x-axis across different thresholds [64].

### Computing the Precision-Recall Curve

1. Begin with a binary classifier capable of predicting a binary outcome and estimating the probability of the positive class, often referred to as *scores*.
2. For every possible threshold (ranging from 0 to 1) applied to these scores, compute both Precision (see Section 3.2.2) and Recall (see Section 3.2.2).
3. Plot a curve with Recall on the X-axis and Precision on the Y-axis. An example is shown in Figure 2 depicting Precision-Recall curves for three different models.

**Interpretation of the Precision-Recall Curve** The effectiveness of a model as visualized by the PR curve can be interpreted as follows:

- *Proximity to the top-right corner*: The closer the curve approaches the top-right corner (1,1) of the plot, the better, indicating high precision and recall across most threshold settings. This ideal point represents a perfect classifier with no false positives or negatives.
- *Area under the curve (AUC-PR)*: The AUC provides a single-number summary of the model's overall ability to achieve high precision and recall simultaneously. Unlike the ROC curve (see section 3.2.2), where a random classifier achieves an AUC of 0.5, the baseline AUC for a PR curve depends on the positive class prevalence.
- *Steepness of the curve*: A steep initial rise of the curve is desirable as it indicates that the model achieves high recall with only a minimal reduction in precision, suggesting an effective balance between the two metrics.



- **Model comparison:** PR curves allow for the direct comparison of different models; a model whose PR curve lies consistently above another indicates superior performance at various threshold levels.

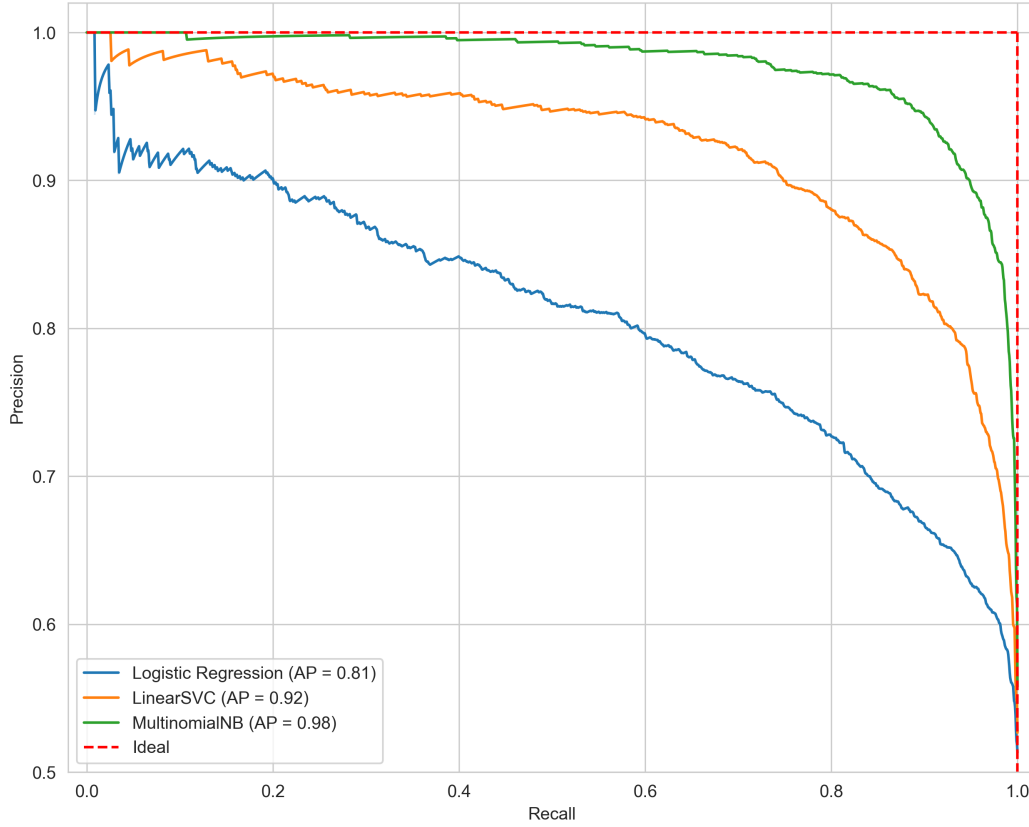


Figure 2: Precision/Recall curves for three models trained on the same data. The dashed line shows the ideal performance. Each model reports its Average Precision metric.

**Precision-Recall Curve in Multi-Class Classification** While the Precision-Recall curve is traditionally used in binary classification settings, it can also be adapted for multi-class classification problems. In multi-class scenarios, a separate PR curve can be computed for each class, treating it as the positive class and all other classes combined as the negative class. This approach allows the evaluation of classifier performance for each specific class in a one-vs-all (OvA) fashion.

**Computing Multi-Class PR Curves** To compute PR curves for multi-class classification, follow these steps:

1. For each class, treat the class as positive and all other classes as negative.
2. Calculate precision and recall for the class across a range of thresholds, just as in binary classification.
3. Plot a PR curve for each class, with recall on the x-axis and precision on the y-axis.

Each curve will provide insights into how well the classifier can identify instances of the respective class, while differentiating them from instances of all other classes.

**Interpretation and Use of Multi-Class PR Curves** The interpretation of each PR curve follows the same principles as in binary classification:

- *Closer to the top-right corner*: A curve that approaches the top-right corner indicates high precision and recall, showing effective class-specific performance.
- *Area under the curve (AUC-PR)*: The AUC for each class's PR curve can be calculated to provide a summary statistic of the model's performance concerning that class. A higher AUC indicates better overall performance in distinguishing the class from others.

**Advantages and Limitations** This approach provides a detailed view of a classifier's performance on a class-by-class basis, which is beneficial in settings where classes are imbalanced or have varying levels of importance. However, interpreting multiple PR curves can be complex, especially when classes are numerous or highly imbalanced. In such cases, additional summary statistics or visualization techniques may be necessary to provide a clearer overall performance picture.

To aggregate performance across multiple classes, macro-averaging and micro-averaging techniques can be used. Macro-averaging computes the average precision and recall across all classes, treating each class equally regardless of its prevalence. Micro-averaging aggregates the contributions of all classes to compute overall precision and recall, which can be particularly relevant in datasets where some classes dominate others.

**Area Under the Receiver Operating Characteristic Curve (AUC-ROC)** The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a widely used metric for evaluating the performance of binary classification models. It measures the model's ability to distinguish between the positive and negative classes across various threshold settings. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) for different thresholds:

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{TN + FP} \quad (46)$$

The AUC-ROC value ranges from 0 to 1, where 1 indicates perfect classification and 0.5 signifies no discriminative ability, equivalent to random guessing.

### Computing the ROC Curve

1. Begin with a binary classifier that predicts outcomes and estimates the probability of the positive class (referred to as *scores*).
2. Calculate the TPR and FPR at each threshold value from 0 to 1 (see Sections [3.2.2](#) and [3.2.2](#)).
3. Plot these values with FPR on the X-axis and TPR on the Y-axis to form the ROC curve. An example of such a plot is shown in Figure [3](#).

**Interpretation of the ROC Curve** The ROC curve provides a graphical representation of a classifier's performance. Key aspects of the ROC curve interpretation include:

- *Diagonal Line*: Represents a random classifier. An ROC curve that lies above the diagonal indicates better than random performance.
- *Area Under the Curve (AUC-ROC)*: Provides a single-value summary of the model's effectiveness, indicating the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. Higher AUC values indicate better performance.
- *Toward the Top-Left Corner*: The ideal position for an ROC curve, indicating both high TPR and low FPR.
- *Comparing Models*: Models can be compared directly by their AUC values; a model with a higher AUC is generally better.

**Limitations in Imbalanced Datasets** While AUC-ROC is a powerful metric, it may not fully represent performance in cases with significant class imbalance. The metric can appear optimistic as it primarily reflects the ranking of positive and negative instances without considering class distribution. In such scenarios, the Precision-Recall curve and its AUC (AUC-PR) might offer a more accurate performance assessment, especially focusing on the minority class.

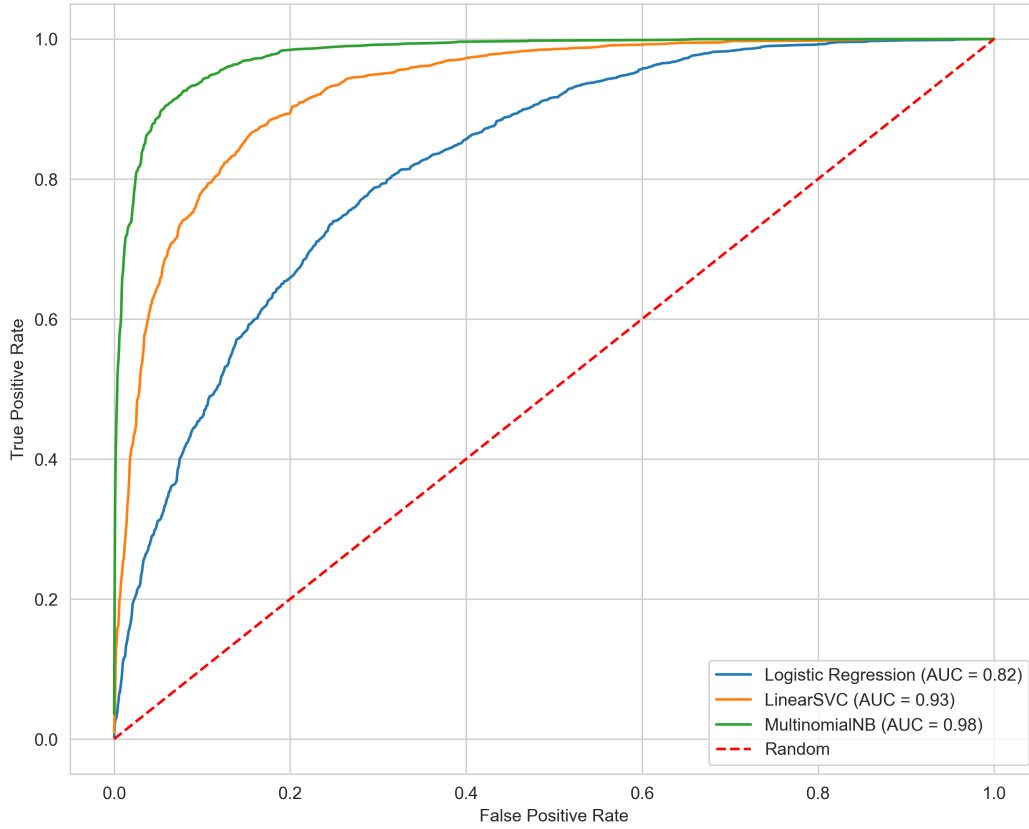


Figure 3: ROC curves for three models trained on the same data. The dashed line shows random performance. Each model reports its Area under the curve (AUC) in the legend.

**Multi-Class Receiver Operating Characteristic Curve (AUC-ROC)** While ROC curves and the corresponding Area Under the Curve (AUC-ROC) are inherently designed for binary classification, they can be extended to multi-class problems through several approaches. These adaptations allow the ROC analysis to maintain its usefulness in evaluating classifiers that need to distinguish among more than two classes.

**One-vs-All (OvA) Approach** The most straightforward method for extending ROC analysis to multi-class settings is the One-vs-All approach. Here, a ROC curve is constructed for each class by treating one class as the positive class and combining all other classes into a single negative class. This results in as many ROC curves as there are classes.

1. For each class, compute the True Positive Rate (TPR) and False Positive Rate (FPR) across various thresholds, treating that class as positive and all others as negative.
2. Plot these rates on the ROC curve, with FPR on the X-axis and TPR on the Y-axis for each class separately.
3. The AUC for each class-specific ROC curve can then be calculated, providing a measure of how well the classifier can distinguish each class from the rest.

**One-vs-One (OvO) Approach** An alternative is the One-vs-One approach, where a ROC curve is generated for every pair of classes. This method is typically used in support vector machines and involves  $(N \times (N - 1))/2$  ROC curves for  $N$  classes, which can become computationally intensive with a large number of classes but provides a very detailed analysis.

**Macro and Micro-Averaging** To synthesize the information across multiple ROC curves in multi-class scenarios, macro-averaging and micro-averaging techniques are employed:

- *Macro-Averaging*: Compute the AUC separately for each class's ROC curve and then average these AUC scores. This method treats each class equally, regardless of class imbalance.
- *Micro-Averaging*: Aggregate the contributions of all classes to compute an overall TPR and FPR across all thresholds, then plot a single ROC curve from these values. This approach gives a performance measure that reflects the classifier's ability to distinguish all classes, weighted by class frequency.

**Visualization and Interpretation** Visualizing multi-class ROC curves can be challenging due to the multiple curves involved. Effective strategies include:

- Plotting each class's ROC curve using different colors or line styles and providing a legend to help identify them.
- Using panels or a grid layout where each panel represents the ROC curve for one class versus all others, or one pair of classes in the case of the OvO approach.
- Providing summary statistics such as the mean AUC or the range of AUCs across classes to quickly gauge overall classifier performance.

*Interpreting these curves requires considering the context of class importance and balance*, as some classes may be more critical to classify correctly than others, or some classes may have more training examples, influencing their representation in the classifier's training process.

## 4 Computer Vision Tasks

Up until now, we have covered loss functions and metrics for two fundamental tasks in Machine Learning, namely regression and classification. In the upcoming sections, we will go deeper into computer vision tasks such as image classification (section 4.1), object detection (section 4.2), image segmentation (section 4.3), face recognition (section 4.4), and generative models (section 4.5).

### 4.1 Image Classification

Image classification involves categorizing an image as a whole into a specific label [66].

Image classification was the first task where deep learning demonstrated significant success, marking a critical moment in the advancement of artificial intelligence. This breakthrough occurred primarily with the development and application of Convolutional Neural Networks (CNNs). In 2012, the AlexNet architecture [67], presented by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, achieved state-of-the-art performance substantially better than the previous best results. This achievement established the potential of deep learning models to outperform traditional machine learning methods that relied heavily on hand-engineered features.

AlexNet's success led to a surge in research and development in deep learning. This resulted in significant improvements in deep learning algorithms and their applications in various domains. Image classification played a key role in this revolution, enabling advances that expanded the scope of deep learning into numerous other applications. This breakthrough has laid a strong foundation for further innovations in the field of AI.

Image classification is used in a diverse range of sectors, demonstrating its broad applicability. In medical imaging, it is used to improve diagnostic accuracy by automatically detecting and classifying abnormalities such as tumors, fractures, and pathological lesions in various types of scans, including X-rays, MRIs, and CT images [68, 69, 70, 71]. In agriculture, it helps to monitor crop health, predict yields, and detect plant diseases through images captured by drones or satellites [72, 73, 74, 75]. The retail industry uses this technology for automatic product categorization, inventory management, and online shopping enhanced with visual search capabilities [76, 77, 78, 79]. Security and surveillance systems use image classification to analyze footage for unauthorized activities, manage crowds, and identify individuals or objects of interest [80, 81, 82]. In environmental monitoring, it is instrumental in tracking deforestation, monitoring changes in the natural environment, and assessing water quality using satellite imagery [82, 83, 84, 85, 86]. In addition, in the manufacturing sector, image classification automates quality control processes by identifying defects and verifying assembly, ensuring that products meet stringent standards with minimal human intervention [87, 88, 89, 90].

### 4.1.1 Image Classification Loss Functions

The loss functions used for image classification are the same as described in section 3.2.1, such as binary cross-entropy, categorical cross-entropy, weighted cross-entropy, focal loss, and hinge loss, as shown in Table 6.

### 4.1.2 Image Classification Metrics

The performance metrics used for image classification are the same described in section 3.2.2 and summarized in Table 7.

## 4.2 Object Detection

Object detection in deep learning is a computer vision technique that involves localizing and recognizing objects in images or videos. It is common in various applications such as autonomous driving [91, 92, 93, 94], surveillance [95, 96, 97], human-computer interaction [98, 99, 100, 101], and robotics [102, 103, 104, 105]. Object detection involves identifying the presence of an object, determining its location in an image, and recognizing the object's class.

### 4.2.1 Object Detection Loss Functions

Object detection models combine various loss functions to address the challenges of localization (regression of bounding box coordinates) and recognition (classification of objects). Table 10 summarizes these loss functions along with their specific applications in the context of object detection.

Loss Function	Type	Application
Multi-Class Log Loss (Cross-Entropy)	Recognition	Used for classifying the categories of the detected objects.
Smooth L1 Loss	Localization	Employed for bounding box regression, less sensitive to outliers than MSE.
Balanced L1 Loss	Localization	Improves upon L1 loss by dynamically scaling the loss based on the magnitude of the errors, enhancing the detection performance under various conditions.
IoU Loss	Localization	Directly maximizes the Intersection over Union metric between predicted and ground truth bounding boxes.
GIoU Loss	Localization	Extends IoU to include aspects of the shape, orientation, and scale of bounding boxes, enhancing localization performance for non-overlapping boxes.
DIoU and CIoU Losses	Localization	Improve the alignment of predicted and ground truth boxes by considering the centroid distance and aspect ratios, respectively.
Focal Loss	Recognition	Addresses class imbalance by focusing on hard-to-classify instances, enhancing the classification of infrequent objects.
YOLO Loss	Both	Specific to the YOLO architecture, combines localization, object presence, and classification losses.
Wing Loss	Localization	Optimizes the localization accuracy for small errors, often used in applications requiring precise localization such as facial landmark detection.

Table 10: Summary of Loss Functions in Object Detection

These loss functions collectively enhance the accuracy and robustness of object detection systems by finely tuning both the recognition of object categories and the precision of bounding box predictions.

In the subsequent sections, we will delve deeper into specific loss functions that have unique characteristics or require additional discussion.

**Smooth L1 Loss** The smooth L1 is a robust loss function commonly used in object detection tasks, introduced in the Fast R-CNN framework [106].

$$L(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{if } |y - \hat{y}| < \beta \\ |y - \hat{y}| - \frac{1}{2}\beta & \text{otherwise,} \end{cases} \quad (47)$$

where  $y$  is the true value,  $\hat{y}$  is the predicted value, and  $\beta$  is a user-specified threshold value.

Smooth L1 loss is closely related to Huber Loss (section 3.1.1), being equivalent to  $\text{huber}(x, y)/\beta$  leading to the following differences [107]:

- As  $\beta \rightarrow 0$ , Smooth L1 loss converges to L1 loss, while Huber loss converges to a constant 0 loss. When  $\beta$  is 0, Smooth L1 loss is equivalent to L1 loss.
- As  $\beta \rightarrow +\infty$ , Smooth L1 loss converges to a constant 0 loss, while Huber loss converges to MSE loss.
- For Smooth L1 loss, as  $\beta$  varies, the L1 segment of the loss has a constant slope of 1. For Huber loss, the slope of the L1 segment is  $\beta$ .

This loss function is used in the Region Proposal Network (RPN) of two-stage object detectors such as Fast R-CNN [106] and Faster R-CNN [108]. It regulates the regression of bounding box coordinates, outperforming traditional MSE in terms of both robustness against outliers and computational efficiency.

**Balanced L1 Loss** Balanced L1 Loss is an enhanced version of the traditional L1 loss, specifically designed to improve the robustness and accuracy of object detection systems in scenarios involving varying error magnitudes. This loss function dynamically adjusts the penalty based on the absolute error, making it particularly effective in handling outliers and improving convergence rates during training [109].

The formula for the Balanced L1 Loss is defined as follows:

$$L_{\text{balanced}}(x) = \begin{cases} \beta \cdot \ln\left(\frac{|x|}{\beta} + 1\right) & \text{if } |x| \geq \beta, \\ \frac{x^2 + \beta^2}{2\beta} & \text{otherwise,} \end{cases} \quad (48)$$

where  $x$  is the error between the predicted value and the ground truth, and  $\beta$  is a threshold parameter that determines the transition point between logarithmic and quadratic behavior.

The Balanced L1 Loss is an effective tool for object detection tasks, especially when the size of objects varies significantly or when precise localization is crucial. It helps to stabilize the training process by reducing the gradient explosion in case of large errors while enabling sensitive detection of small discrepancies. This loss function is recommended when traditional L1 or L2 loss functions do not provide enough robustness against outliers or fail to converge efficiently.

**Intersection over Union (IoU) Loss** IoU is a metric used in object detection that measures the overlap between two bounding boxes. Figure 4 depicts the IoU metric used in object detection. The IoU between two bounding boxes is calculated as

$$IoU = \frac{\text{Area}(B_p \cap B_t)}{\text{Area}(B_p \cup B_t)}, \quad (49)$$

where  $B_p$  and  $B_t$  are the predicted and ground truth bounding boxes, respectively. The IoU loss function is defined as

$$L = 1 - IoU \quad (50)$$

This function encourages the predicted bounding boxes to overlap highly with the ground truth bounding boxes. A high IoU value indicates that the predicted bounding box is close to the ground truth, while a low IoU value indicates that the predicted bounding box is far from the ground truth.

The IoU loss function is commonly used for one-stage detectors [110, 111] as part of a multi-task loss function that includes a classification loss and a localization loss.

Although IoU is one of the most used loss functions and metrics for localization, it has a few limitations:

**Limitations with Non-Overlapping Boxes:** When the predicted box and the ground truth box do not overlap at all, the intersection area is zero. This leads to an IoU of zero, which provides no gradient or signal for the model to adjust the predicted box's position. This lack of overlap may lead to a drastic penalty in IoU, even when the predicted box is close to the ground truth box. This can be problematic for models trained primarily on IoU, as they might struggle to converge or provide reliable feedback for non-overlapping boxes.

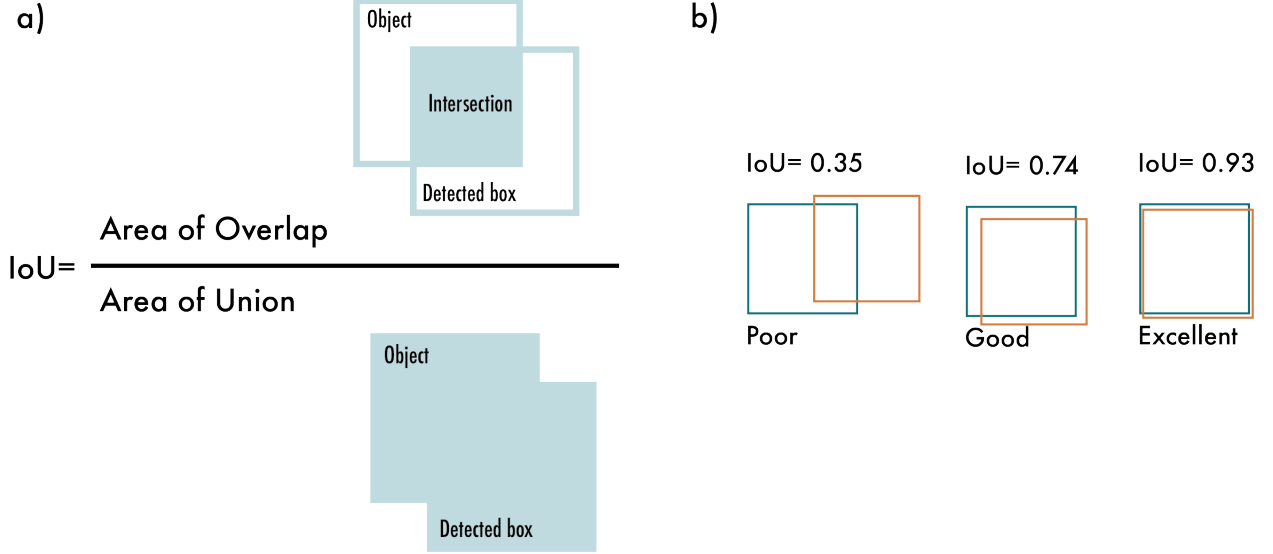


Figure 4: Intersection Over Union (IoU). a) The IoU is calculated by dividing the intersection of the two boxes by the union of the boxes; b) examples of three different IoU values for different box locations.

**Limitations with Partially Overlapping Boxes:** In scenarios where the bounding boxes partially overlap, IoU can exhibit a limited range of values, particularly when the areas of the union and intersection do not vary significantly. This limits the sensitivity of IoU as a measure of bounding box alignment. Also, partially overlapping boxes can lead to similar IoU values even if the positioning or aspect ratios differ greatly, potentially masking significant differences in the alignment.

The Generalized IoU discussed next was introduced to handle these limitations.

**GIoU Loss** GIoU Loss (Generalized Intersection over Union Loss) introduced in 2019 [112] is an extension of the standard IoU Loss, designed to address the limitations of the latter in scenarios involving non-overlapping or partially overlapping bounding boxes. The formula for the GIoU Loss is defined as follows:

$$L_{GIoU}(B_p, B_t) = 1 - IoU + \frac{\text{Area}(C - (B_p \cup B_t))}{\text{Area}(C)}, \quad (51)$$

where  $B_p$  and  $B_t$  are the predicted and ground truth bounding boxes, respectively.  $C$  represents the smallest enclosing box that contains both  $B_p$  and  $B_t$ , while the intersection and union operations are defined over these boxes.

GIoU Loss improves upon traditional IoU Loss in the following ways:

- Penalizes the absence of overlap by incorporating the area outside the union of  $B_p$  and  $B_t$  into the loss calculation.
- It provides a better gradient signal for training, thereby improving the accuracy of bounding box regression tasks.
- Provides consistent feedback throughout the optimization process, which leads to more stable and efficient training.

**Limitations of GIoU:** GIoU improves upon traditional IoU by including the area outside the union of the predicted and ground truth bounding boxes in its loss calculation. However, it has two limitations:

1. **Spatial Misalignment:** GIoU does not directly address the distance between the centroids of bounding boxes, leading to potential misalignment between the predicted and ground truth boxes even when their GIoU value is high.
2. **Aspect Ratio Consistency:** GIoU does not consider the difference in aspect ratios between the bounding boxes, which can lead to suboptimal box regression in situations with varied object shapes where two bounding boxes might intersect and have a relatively high IoU, but their aspect ratios (width-to-height ratios) may differ significantly. This mismatch can lead to inaccurate regression results.



These two limitations are addressed by the DIOU and CIOU discussed next.

**DIOU and CIOU Losses** DIOU (Distance-IoU) and CIOU (Complete-IoU) [113] are loss functions designed to enhance the performance of object detection models, specifically addressing the challenges of localization accuracy in bounding box regression tasks.

DIOU Loss incorporates the distance between the centroids of the predicted and ground truth bounding boxes into the loss calculation, addressing the issue of spatial misalignment. The formula for DIOU Loss is defined as follows:

$$L_{\text{DIOU}}(B_p, B_t) = 1 - \text{IoU} + \frac{d^2(B_p, B_t)}{c^2} \quad (52)$$

where  $B_p$  and  $B_t$  are the predicted and ground truth bounding boxes, respectively.  $d(B_p, B_t)$  represents the Euclidean distance between the centroids of these boxes, and  $c$  denotes the diagonal length of the smallest enclosing box containing both  $B_p$  and  $B_t$ .

CIOU Loss extends DIOU by including an additional term to consider the aspect ratio consistency between the predicted and ground truth bounding boxes. This provides a more comprehensive metric for evaluating the alignment of the boxes. The formula for CIOU Loss is:

$$L_{\text{CIOU}}(B_p, B_t) = L_{\text{DIOU}}(B_p, B_t) + \alpha \cdot v \quad (53)$$

where  $v$  measures the aspect ratio consistency between the predicted and ground truth boxes, and  $\alpha$  is a weighting factor defined as:

$$\alpha = \frac{v}{(1 - \text{IoU}) + v} \quad (54)$$

DIOU and CIOU losses address specific challenges in bounding box regression. DIOU enhances spatial alignment by incorporating the distance between the centroids of the predicted and ground truth bounding boxes, improving localization accuracy. CIOU goes further, including an aspect ratio consistency metric that provides a comprehensive evaluation of box alignment. Both losses offer consistent feedback throughout the training process leading to more accurate and robust object detection models.

**Focal Loss** Introduced by Tsung-Yi Lin et al. [114], Focal Loss is an enhanced version of the standard cross-entropy loss designed to address class imbalance, a common issue where the number of instances in one class significantly outnumbers those in another. This imbalance often results in models that disproportionately favor the majority class, leading to poor identification of minority class instances.

Focal Loss modifies the standard cross-entropy formula to focus training on hard examples that are misclassified and not well represented:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t), \quad (55)$$

where  $p_t$  is the probability of the true class as predicted by the model,  $\alpha_t$  is a weighting factor that adjusts the importance given to each class, and  $\gamma$  is a focusing parameter that modifies the rate at which easy examples contribute to the loss.

Here,  $p_t$  is defined as the model's estimated probability for the actual class label  $y$ . Specifically,  $p_t = p$  if  $y = 1$  and  $p_t = 1 - p$  if  $y = 0$ , effectively focusing the loss calculation on the correct class probability. The parameter  $\alpha_t$  is often set based on the inverse frequency of the classes to mitigate the imbalance effect, while  $\gamma$  helps in reducing the relative loss for well-classified examples, thereby putting more focus on difficult, misclassified cases. Common values for  $\gamma$  range from 2 to 5.

Focal Loss has been widely adopted in fields that require precise identification in imbalanced datasets, such as object detection, semantic segmentation, and more recently, instance segmentation and human pose estimation. Its ability to hone in on harder-to-classify examples makes it particularly valuable in scenarios where the minority classes are crucial yet significantly underrepresented.

Recent applications of Focal Loss include enhancing performance in object detection frameworks like RetinaNet [114] and improving the accuracy of semantic segmentation in medical imaging [115], instance segmentation [116], and human pose estimation [117], demonstrating its effectiveness in tackling class imbalance across different domains.

**YOLO Loss** The YOLO (You Only Look Once) loss [111] is a composite loss function designed specifically for the YOLO family of object detection algorithms. It combines various components to address the dual challenges of localization and classification in object detection tasks.

The YOLO loss consists of three primary components:

1. **Localization Loss:** This loss measures the difference between the predicted and ground truth bounding box coordinates. It is typically a sum of squared errors across the bounding box's center coordinates, width, and height:

$$L_{\text{loc}} = \sum_{i \in \text{box}} \lambda_{\text{coord}} \left[ (x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (w_i - \hat{w}_i)^2 + (h_i - \hat{h}_i)^2 \right], \quad (56)$$

where  $x_i, y_i, w_i, h_i$  and  $\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i$  are the ground truth and predicted bounding box coordinates, respectively, and  $\lambda_{\text{coord}}$  is a weighting factor.

2. **Objectness Loss:** This loss evaluates the confidence score associated with the presence of an object in the predicted box. It is typically a cross-entropy or mean squared error loss:

$$L_{\text{obj}} = \sum_{i \in \text{box}} \left( C_i - \hat{C}_i \right)^2, \quad (57)$$

where  $C_i$  and  $\hat{C}_i$  are the ground truth and predicted confidence scores, respectively.

3. **Classification Loss:** This loss measures the disparity between the predicted class probabilities and the ground truth class labels, typically using a cross-entropy loss:

$$L_{\text{class}} = \sum_{i \in \text{box}} - \sum_{c \in \text{classes}} p_i(c) \cdot \log(\hat{p}_i(c)), \quad (58)$$

where  $p_i(c)$  and  $\hat{p}_i(c)$  are the ground truth and predicted probabilities for class  $c$ , respectively.

YOLO loss is used in the YOLO family of object detection algorithms to balance the competing needs of localization and classification. Its composite nature enables efficient training by offering feedback on three key metrics: box coordinates, object presence, and class labels.

#### 4.2.2 Object Detection Metrics

In this section, we delve into the object detection metrics. In object detection, we also compute the True Positives, False Positives, True Negatives, and False Negatives. The definitions of these metrics are based on the IoU score as follows:

**True Positives in object detection:** The match between the predicted location of an object and its actual location is measured using an Intersection Over Union (IoU) score. The IoU score ranges from 0 to 1, with a score of 1 indicating a perfect match between the predicted and ground-truth locations. Since a perfect match is hard to achieve, we define a threshold value to determine whether a prediction is a true positive. Common values for the threshold are 0.25, 0.5, and 0.75. These thresholds are not fixed and can be adjusted based on the application's requirements. If the IoU score between the predicted and ground-truth boxes is greater than or equal to the defined threshold, the prediction is considered a true positive.

**False Positive in object detection:** Occurs when the model predicts the presence of an object, but the object is not present in the image. This affects the precision metric.

**False Negative in object detection:** Occurs when the model fails to detect an object that is present in an image. This affects the recall metric.

**True Negative in object detection:** Refers to a case where the object detector correctly determines that an object is not present in an image.

**Common IoU thresholds for object detection:**

- 0.5: A threshold of 0.5 is commonly used as a balanced threshold for object detection. A predicted bounding box is considered a true positive if its IoU with the ground truth bounding box is greater than or equal to 0.5.
- 0.75: A threshold of 0.75 is used for applications that require higher precision, such as autonomous driving, where false positive detections can lead to critical consequences.
- 0.25: A threshold of 0.25 is used for applications that require higher recall, such as medical image analysis, where missing detections can lead to an incorrect diagnosis.

The common object detection metrics are:

- Average Precision (AP) and Average Recall (AR)
- Intersection over union (IoU). See details in section 4.2.1
- Precision-Recall Curve. See details in section 3.2.2

**Average Precision (AP)** Object detection models must identify and localize multiple object categories in an image. The AP metric addresses this by calculating each category’s AP separately and then taking the mean of these APs across all categories (that is why it is also called mean average precision or mAP). This approach ensures that the model’s performance is evaluated for each category individually, providing a more comprehensive assessment of the model’s overall performance.

To accurately localize objects in images, AP incorporates the Intersection over Union (IoU) to assess the quality of the predicted bounding boxes. As described previously, IoU is the ratio of the intersection area to the union area of the predicted bounding box and the ground truth bounding box (see Figure 4). It measures the overlap between the ground truth and predicted bounding boxes. The COCO benchmark considers multiple IoU thresholds to evaluate the model’s performance at different levels of localization accuracy.

The two most common object detection datasets are The Pascal VOC [118] and Microsoft COCO [119]. The AP is computed differently in each of these. In the following, we describe how it is computed on each dataset.

**VOC Dataset** This dataset includes 20 object categories. To compute the AP in VOC, we follow the next steps:

1. *Compute Intersection over Union (IoU)*: For each detected object, compute the IoU with each ground truth object in the same image (refer to section 4.2.1 for more details).
2. *Match Detections and Ground Truths*: For each detected object, assign it to the ground truth object with the highest IoU, if the IoU is above the threshold.
3. *Compute Precision and Recall*: For each category, calculate the precision-recall curve by varying the confidence threshold of the model’s predictions (refer to section 3.2.2 for more details). This results in a set of precision-recall pairs.
4. *Sort and interpolate with 11-points*: Sort the precision-recall pairs by recall in ascending order. Then, for each recall level  $r$  in the set  $\{0, 0.1, 0.2, \dots, 1.0\}$ , find the highest precision  $p(r)$  for which the recall is at least  $r$ . This is known as interpolated precision. This process results in a precision-recall curve that is piecewise constant and monotonically decreasing.
5. *Compute Area Under Curve (AUC)*: The Average Precision is then defined as the area under this interpolated precision-recall curve. Since the curve is piecewise constant, this can be computed as a simple sum:  $AP = \text{sum}(p(r)/N)$ , where the sum is over the  $N$  recall levels, and  $p(r)$  is the interpolated precision at recall level  $r$ .

**Microsoft COCO Dataset** This dataset includes 80 object categories and uses a more complex method for calculating AP. Instead of using an 11-point interpolation, it uses a 101-point interpolation, i.e., it computes the precision for 101 recall thresholds from 0 to 1 in increments of 0.01. Also, the AP is obtained by averaging over multiple IoU values instead of just one, except for a common AP metric called  $AP_{50}$ , which is the AP for a single IoU threshold of 0.5. Table 11 shows all the metrics used to evaluate models in the COCO dataset. The steps for computing AP in COCO are the following:

1. *Compute the Intersection over Union (IoU)*: For each detected object, compute the IoU with each ground truth object in the same image.
2. *Match Detections and Ground Truths*: For each detected object, assign it to the ground truth object with the highest IoU, if this IoU is above the threshold.
3. *Compute Precision and Recall*: For each possible decision threshold (confidence score of the detection), compute the precision and recall of the model. This results in a set of precision-recall pairs.
4. *Interpolate Precision*: For each recall level  $r$  in the set  $\{0, 0.01, 0.02, \dots, 1.00\}$  (for the 101-point interpolation used in COCO), find the maximum precision  $p(r)$  for which the recall is at least  $r$ . This is known as interpolated precision.
5. *Compute Area Under Curve (AUC)*: The Average Precision is then defined as the area under this interpolated precision-recall curve. Since the curve is a piecewise constant, this can be computed as a simple sum:

$AP = \text{sum}(p(r))/101$ , where the sum is over the 101 recall levels, and  $p(r)$  is the interpolated precision at recall level  $r$ .

6. *Average over IoU Thresholds*: Repeat steps 2-5 for different IoU thresholds (e.g., 0.5, 0.55, 0.6, ..., 0.95) and average the AP values.
7. *Average over Categories*: Repeat steps 2-6 for each category and average the AP values. This is to prevent categories with more instances from dominating the evaluation.
8. *Average over Object Sizes*: Finally, you can compute AP for different object sizes (small, medium, large) to see how well the model performs on different sizes of objects.

**Average Recall (AR)** Average Recall (AR) is used to evaluate the performance of object detection models. Unlike Precision or Recall, defined at a particular decision threshold, Average Recall is computed by averaging recall values at different levels of Intersection over Union (IoU) thresholds and, if needed, at different maximum numbers of detections per image. This metric is commonly used to report COCO data results [119, 120].

The general steps to compute AR are the following:

1. *Compute the Intersection over Union (IoU)*: For each detected object, compute the IoU with each ground truth object in the same image.
2. *Match Detections and Ground Truths*: For each ground truth object, find the detected object with the highest IoU. If this IoU is above a certain threshold, the detection is considered a true positive, and the ground truth is *matched*. Each ground truth can only be matched once.
3. *Compute Recall*: For each image, recall is the number of matched ground truths divided by the total number of ground truths.
4. *Average over IoU Thresholds*: Repeat steps 2 and 3 for different IoU thresholds (e.g., from 0.5 to 0.95 with step size 0.05), and average the recall values.
5. *Average over Max Detections*: Repeat steps 2-4 for different maximum numbers of detections per image (e.g., 1, 10, 100), and average the recall values. This step is necessary because allowing more detections per image can potentially increase recall but at the cost of potentially more false positives.
6. *Average over Images*: Finally, compute the average recall over all the images in the dataset.

For COCO, the Average Recall measure can also be computed separately for different object sizes (small, medium, and large) to evaluate how well the model works for objects of different sizes.

Table 11: COCO Evaluation Metrics.

Average Precision (AP)	
AP	% AP at IoU=.50:.95 (primary challenge metric)
$AP^{IoU=.50}$	% AP at IoU=.50 (PASCAL VOC metric)
$AP^{IoU=.75}$	% AP at IoU=.75 (strict metric)
AP Across Scales:	
$AP^{small}$	% AP for small objects: area < 32 <sup>2</sup>
$AP^{medium}$	% AP for medium objects: 32 <sup>2</sup> < area < 96 <sup>2</sup>
$AP^{large}$	% AP for large objects: area > 96 <sup>2</sup>
Average Recall (AR):	
$AR^{max=1}$	% AR given 1 detection per image
$AR^{max=10}$	% AR given 10 detections per image
$AR^{IoU=100}$	% AR given 100 detection per image
AR Across Scales:	
$AR^{small}$	% AR for small objects: area < 32 <sup>2</sup>
$AR^{medium}$	% AR for medium objects: 32 <sup>2</sup> < area < 96 <sup>2</sup>
$AR^{large}$	% AR for large objects: area > 96 <sup>2</sup>

### 4.3 Image Segmentation

Image segmentation aims to assign a label or category to each pixel in the image, effectively segmenting the objects at a pixel level. Segmentation is usually performed using deep learning models trained to classify each pixel in the

image based on its features and context. Segmentation methods are mainly classified into three categories: semantic segmentation [121, 122, 123, 124, 125, 126, 127, 128], instance segmentation [129, 130, 131, 132, 133], and panoptic segmentation [134, 135, 136, 137, 138].

*Semantic Segmentation* studies the uncountable stuff in an image. It analyzes each image pixel and assigns a unique class label based on the texture it represents. In a street image, the semantic segmentation’s output will assign the same label to all the cars and the same image to all the pedestrians; it cannot differentiate the objects separately.

*Instance Segmentation* deals with countable things. It can detect each object or instance of a class present in an image and assigns it to a different mask or bounding box with a unique identifier.

*Panoptic Segmentation* presents a unified segmentation approach where each pixel in a scene is assigned a semantic label (due to semantic segmentation) and a unique instance identifier (due to instance segmentation).

Segmentation applications include scene understanding [139, 140, 141, 142, 143], medical image analysis [144, 145, 146, 147, 148, 149, 150], robotic perception [151, 152, 153, 154, 155], autonomous vehicles [156, 157, 158, 159, 160, 161], video surveillance [162, 163, 164, 165, 166], and augmented reality [167, 168, 169, 170, 171].

#### 4.3.1 Segmentation Loss Functions

Common loss functions include cross-entropy loss, Intersection over union (IoU) loss, Focal loss, Dice loss, Tversky loss, and Lovász Loss. The following sections will describe these loss functions and their applications. Table 12 shows a summary of the loss functions used in image segmentation.

Loss Function	Application
Cross-Entropy Loss	Suitable for pixel-wise classification tasks where each pixel belongs to one of several classes.
Binary Cross-Entropy Loss	For binary segmentation tasks where each pixel is classified into one of two classes.
Jaccard Index (IoU) Loss	Measures the Intersection over Union (IoU) between predicted and ground truth masks.
Dice Coefficient Loss	Useful for measuring overlap between predicted segmentation and ground truth, especially in cases of class imbalance.
Tversky Loss	A generalization of Dice loss, balancing false positives and false negatives effectively, making it useful for medical image segmentation.
Focal Loss	Addresses class imbalance by giving more weight to difficult-to-classify pixels, modifying the cross-entropy loss for misclassified samples.
Combined Losses	Combines multiple losses to balance precision and recall. For example, Cross-Entropy + Dice Loss is used to handle both binary and multi-class segmentation.

Table 12: Common loss functions for segmentation tasks.

**Cross Entropy Loss for Segmentation** Cross-entropy loss is commonly used in segmentation tasks; it calculates the dissimilarity between the predicted probabilities and the actual labels across all classes for each pixel. In multi-class segmentation, the cross-entropy loss is defined as the negative log-likelihood of the correct class’s probability at each pixel:

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(p_{i,c}), \quad (59)$$

where  $N$  is the total number of pixels in the image,  $C$  is the number of classes,  $y_{i,c}$  is a binary indicator (0 or 1) if class  $c$  is the correct classification for pixel  $i$ , and  $p_{i,c}$  is the predicted probability of pixel  $i$  being of class  $c$ .

For binary segmentation, where each pixel is categorized into one of two classes (typically representing the object of interest and the background), the Binary Cross Entropy Loss is used:

$$L_{binary} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)], \quad (60)$$

where  $y_i$  is the ground truth label for pixel  $i$  (0 or 1), and  $p_i$  is the predicted probability of the pixel  $i$  belonging to the class with label 1.

Both forms of cross-entropy encourage the model to output probabilities that closely match the true labels, effectively tuning the pixel-level predictions for optimal segmentation accuracy.

**Dice Loss** Dice loss, based on the Dice Similarity Coefficient, also known as Sørensen-Dice index [172], is used in segmentation tasks to quantify the similarity between the predicted segmentation mask and the ground truth mask. This coefficient is particularly useful in data sets with imbalanced classes, such as in medical imaging, where the regions of interest are typically much smaller than the background.

The Dice loss is calculated as:

$$L = 1 - \frac{2 \times |pred \cap gt|}{|pred| + |gt|}, \quad (61)$$

where  $pred$  and  $gt$  represent the sets of pixels belonging to the predicted and ground truth segmentation masks, respectively. Here,  $|pred \cap gt|$  denotes the cardinality of the intersection, or the number of pixels correctly predicted within the region of interest, and  $|pred|$  and  $|gt|$  are the cardinalities of the predicted and ground truth masks, respectively.

The formula effectively measures the overlap between the two masks: a Dice coefficient of 1 indicates perfect agreement, while a coefficient of 0 indicates no overlap.

**Intersection Over Union (IoU) Loss for Segmentation** Intersection Over Union (IoU) loss, also known as Jaccard loss or Jaccard Index (JI), is a widely used metric for evaluating the performance of semantic segmentation models. It assesses the overlap between the predicted segmentation mask and the ground truth on a per-pixel basis, with the objective to maximize the overlap area relative to the total area covered by either the prediction or the ground truth.

The IoU loss is mathematically defined as:

$$IoU = 1 - \frac{1}{N} \sum_{i=1}^N \frac{|y_i \cap \hat{y}_i|}{|y_i \cup \hat{y}_i|}, \quad (62)$$

where  $N$  is the total number of pixels in the image,  $y_i$  is the ground truth mask for pixel  $i$ , and  $\hat{y}_i$  is the predicted mask for pixel  $i$ . Here,  $|y_i \cap \hat{y}_i|$  represents the number of pixels in the intersection of the predicted and ground truth masks for each pixel, while  $|y_i \cup \hat{y}_i|$  represents the number of pixels in the union of the two masks.

The IoU value ranges from 0 to 1, where 0 indicates no overlap and 1 represents perfect agreement between the prediction and the ground truth. This loss function is particularly useful for tasks where the balance between precision and recall is crucial, and it is commonly utilized both as a loss function and as an evaluation metric in various semantic segmentation studies [173, 174, 121, 175].

**Tversky Loss** Tversky loss, introduced by Salehi et al. [176], is an adaptation of the Dice loss specifically designed to address class imbalances in image segmentation tasks. This loss function modifies the traditional Dice coefficient by introducing two hyperparameters,  $\alpha$  and  $\beta$ , which control the penalties for false negatives and false positives, respectively. It is mathematically defined as:

$$Tversky(A, B) = 1 - \frac{|A \cap B|}{|A \cap B| + \alpha|A \setminus B| + \beta|B \setminus A|}, \quad (63)$$

where  $A$  and  $B$  represent the predicted and ground truth segmentation masks, respectively. Here,  $|A \cap B|$  denotes the intersection of the two masks (true positives),  $|A \setminus B|$  indicates the elements in  $A$  but not in  $B$  (false positives), and  $|B \setminus A|$  represents the elements in  $B$  but not in  $A$  (false negatives).

The hyperparameters  $\alpha$  and  $\beta$  allow the tuning of the loss function to focus more on either false positives or false negatives. This capability is useful in medical image segmentation, where the costs of different types of segmentation errors are not always equal. By adjusting  $\alpha$  and  $\beta$ , we can prioritize minimizing the more critical type of error for a specific application.

Tversky loss is advantageous in situations where one class dominates or where misclassification of a particular class has a higher penalty. It provides a way to optimize segmentation models under various conditions.



**Lovász Loss** Introduced by Berman et al. [177], Lovász Loss is designed to directly optimize the Intersection over Union (IoU) measure, providing a smooth, differentiable surrogate for the Jaccard index suitable for gradient-based learning. The Lovász Loss is mathematically defined for a set of predictions and their corresponding ground truth labels, applying a convex Lovász extension of the submodular Jaccard index. The function is computed as:

$$L = \sum_{\text{errors}} \phi(m_i), \quad (64)$$

where  $\phi$  is the Lovász extension of the Jaccard index, and  $m_i$  represents the individual errors between the predicted and ground truth segments. The  $m_i$  values are the margins computed as  $1 - p_i$  if the ground truth label  $y_i$  is 1, or  $p_i$  if  $y_i$  is 0.

The Lovász Loss is unique in its approach as it operates directly on the errors of the IoU between the predicted and actual labels, rather than summing over pixel-wise errors. This method allows the loss to more directly influence the IoU measure, making it highly effective for tasks where the IoU is a more significant indicator of performance than pixel accuracy.

This loss function is especially useful in scenarios where the classes are imbalanced or where the segmentation of small objects is crucial, as it aligns the model’s optimization objectives with these higher-level outcomes.

**Focal Loss for Segmentation** Focal loss by Lin et al. [114], as mentioned before for object detection, addresses the challenge of class imbalance. It has been adapted effectively for use in semantic segmentation. Focal loss modifies the standard cross-entropy loss equation to focus learning better on hard, easily misclassified cases, thus balancing the importance of positive and negative examples.

The Focal Loss is mathematically defined as:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t), \quad (65)$$

where  $p_t$  is the model’s estimated probability for each class being the true class,  $\alpha_t$  is a weighting factor that adjusts the importance of each class, and  $\gamma$  is a focusing parameter that smoothly adjusts the rate at which easy examples are down-weighted.

In the context of segmentation,  $p_t$  is defined as follows:

- If the ground truth label is 1 (positive class),  $p_t = p$ , where  $p$  is the predicted probability of the class.
- If the ground truth label is 0 (negative class),  $p_t = 1 - p$ .

The parameters  $\alpha$  and  $\gamma$  allow for the fine-tuning of the loss function’s sensitivity to class imbalance and misclassified examples. The  $\alpha$  parameter can be used to give more emphasis to rare classes, while  $\gamma$  increases the influence of misclassifications; higher values make the model focus more on hard examples.

Focal Loss has shown significant improvements in performance for segmentation tasks involving highly imbalanced data or where the presence of difficult-to-segment objects requires the model to learn nuanced differences between classes more effectively. It is useful in medical image analysis, aerial imagery, and any scenario where the precision of segmentation is critical, and class imbalance is prevalent.

### 4.3.2 Segmentation Metrics

The common metrics for evaluating segmentation are the Mean Intersection over union (mIoU) (section 4.3.1), pixel accuracy, Average Precision (AP) (section 4.2.2), BF score, and Panoptic Quality. The following sections will explain each, skipping IoU, and AP already discussed.

**Pixel Accuracy** Pixel Accuracy measures the proportion of correctly classified pixels in the entire image. It is calculated by dividing the number of correctly classified pixels by the total number of pixels in the image. The formula for Pixel Accuracy is:

$$\text{Pixel Accuracy} = \frac{\text{Number of correctly classified pixels}}{\text{Total number of pixels in image}} \quad (66)$$

While Pixel Accuracy provides a straightforward assessment of overall model performance, its reliability as a sole metric is compromised in cases of high class imbalance. This metric tends to favor the majority class, potentially overlooking misclassified instances of less frequent classes. Therefore, it is often complemented by other metrics such



Table 13: Common metrics used in different types of segmentation tasks.

Metric Name	Type of Segmentation	Description
Pixel Accuracy	Semantic Segmentation	Measures the percentage of correctly labeled pixels. Best used when class distributions are balanced.
Intersection over Union (IoU)	Semantic Segmentation	Evaluates the overlap between the predicted segmentation and the ground truth over their union. Useful across various classes, particularly when dealing with unbalanced datasets.
Boundary F1 Score	Semantic Segmentation	Measures the harmonic mean of precision and recall for boundary prediction accuracy, particularly useful for assessing the delineation of object boundaries.
Dice Coefficient	Semantic Segmentation	Similar to IoU but considers twice the intersection over the sum of the sizes of both the prediction and the ground truth. Effective for unbalanced datasets.
Precision and Recall	Instance Segmentation	Measure the accuracy and completeness of the detected instances, respectively. Employed to evaluate the detection of instances at various IoU thresholds.
Average Precision (AP)	Instance Segmentation	Computes the average precision across different IoU thresholds; used in competitions and benchmarks like COCO [119].
Mask Mean Average Precision (Mask mAP)	Instance Segmentation	Assesses the quality of the instance masks generated by the model. It averages the AP across all classes and IoU thresholds.
Panoptic Quality (PQ)	Panoptic Segmentation	Combines recognition and segmentation accuracy into a single metric by multiplying segmentation quality (SQ) by recognition quality (RQ).
Segmentation Quality (SQ)	Panoptic Segmentation	Measures the IoU of correctly matched segments to their corresponding ground truth objects, reflecting the segmentation accuracy.
Recognition Quality (RQ)	Panoptic Segmentation	Evaluates how effectively the model identifies and differentiates instances, independent of segmentation performance.

as Intersection over Union (IoU) and the Dice coefficient, which provide insights into the model’s performance across different classes and are less sensitive to class imbalance. Pixel Accuracy might still be useful in scenarios where classes are fairly balanced, and the primary interest is in the general accuracy of pixel classification across the image.

**Boundary F1 Score (BF)** The Boundary F1 Score (BF) [178], commonly known as the BF score, evaluates the precision of image segmentation, specifically focusing on the precision of boundary delineation rather than the segmentation of regions. Unlike metrics such as Intersection over Union (IoU), which assess the overlap between predicted and ground truth regions, the BF score is tailored to applications where accurate boundary detection is crucial, such as in medical image analysis.

The calculation of the BF score involves the following steps:

1. Identify the closest predicted segment for each ground truth segment using a distance measure. This typically involves calculating the mean shortest distance between each point on the predicted boundary and the closest point on the ground truth boundary, as well as vice versa.
2. Calculate the precision as the proportion of predicted segments that are within a predefined distance threshold from any ground truth segment. This is given by  $P = \frac{TP}{TP+FP}$ , where  $TP$  (True Positives) represents the count

of predicted segments adequately close to the ground truth, and  $FP$  (False Positives) denotes the predicted segments that fail to meet this criterion.

3. Determine the recall as the proportion of ground truth segments that are adequately approximated by any predicted segment. This is given by  $R = \frac{TP}{TP+FN}$ , where  $FN$  (False Negatives) refers to ground truth segments that are not adequately approximated by any predicted segment.
4. The BF score is then computed as the harmonic mean of precision and recall:

$$F_{score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (67)$$

The distance threshold, which defines what is considered *close enough*, is typically determined based on a pixel distance or another spatial metric, depending on the application and image resolution. The BF score ranges from 0 to 1, with 1 indicating perfect congruence between the predicted and actual boundaries. It is particularly valuable in applications where boundary precision is critical. However, it is important to note that the choice of threshold can significantly impact the score, and the BF score may not fully reflect the quality of region segmentation.

**Panoptic Quality (PQ)** Panoptic Quality (PQ) is a comprehensive metric proposed for evaluating panoptic segmentation tasks, which involve both semantic and instance segmentation [134]. This metric accounts for the quality of both the segmentation and the recognition of individual instances.

The PQ metric is defined as:

$$PQ = \frac{\sum_{(p,g) \in TP} IoU(p, g)}{|TP|} \times \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}, \quad (68)$$

where:

- $IoU(p, g)$  denotes the intersection-over-union between a predicted segment  $p$  and the corresponding ground truth segment  $g$ .
- $TP$  (True Positives) is the set of matched pairs of predicted and ground truth segments, where a match is determined based on a minimum IoU threshold (commonly 0.5).
- $FP$  (False Positives) is the set of predicted segments that do not match any ground truth segment.
- $FN$  (False Negatives) is the set of ground truth segments that are not matched by any predicted segment.

The PQ metric ranges from 0 to 1, with 1 indicating perfect segmentation and recognition. It is the product of two components:

- **Segmentation Quality (SQ):** This component measures the average IoU of the correctly detected segments (true positives). It reflects how accurately each correctly predicted object has been segmented, emphasizing the quality of the segment boundaries and overlap.
- **Recognition Quality (RQ):** This component measures the ratio of true positives to the sum of true positives, false positives, and false negatives. It captures the accuracy in recognizing and matching the instances of objects, thus balancing precision and recall in detection.

The PQ metric provides a more integrated evaluation compared to the mean IoU, especially in complex scenes with multiple object instances per class. By considering both segmentation accuracy and instance recognition, PQ ensures that models are evaluated not only on the precision of segment boundaries but also on the ability to correctly distinguish and identify all instances within a scene. This makes it particularly valuable in real-world applications where both aspects are critical.

#### 4.4 Face Recognition

Face recognition is the process of accurately matching an individual’s face in an image or video to a corresponding entry in a database of faces. This is often done using deep learning algorithms, such as Convolutional Neural Networks (CNNs) or Transformers [179], that have been trained on large datasets of face images. These algorithms are trained to extract features from images of faces, and then use these features to recognize faces that match those in the database. Face recognition has many applications, including security [180][181], social media [182], and biometric identification systems [183][184].

#### 4.4.1 Face Recognition Loss Functions and Metrics

The loss functions used in face recognition are typically aimed at preserving the structure and similarity of the input faces. They can be divided into two classes: loss functions based on classification and loss functions based on representation learning. Common loss functions based on classification are softmax loss, A-softmax loss, Center loss, Large-Margin cosine loss, and Additive Angular Margin loss. On the other hand, loss functions based on representation learning are Triplet loss, Contrastive loss, Circle loss, and the Barlow twins loss.

The metrics commonly used for face recognition are the same as the ones used for classification, such as accuracy, precision, recall, F1-score, ROC, etc.

In the following subsections, we will describe each of these loss functions.

**Softmax Loss** Softmax Loss computes the cross-entropy between the predicted class probabilities and the true class labels, transforming the negative log-likelihood into a loss value. The final loss is obtained by summing the cross-entropy over all classes and samples.

Let the weight vectors for each class (or each face identity) be denoted as  $W = \{w_1, w_2, \dots, w_n\}$ , where  $n$  is the number of classes (or identities). For a given input image  $x$  belonging to class  $y$ , the linear classifier computes a score  $f(x, W) = Wx + b$ , where  $b$  is the bias term. Here,  $w_i^T x$  represents the dot product between the weight vector  $w_i$  and the feature vector  $x$ .

The softmax function converts these scores into probabilities. The probability of the  $i^{th}$  class is computed as:

$$P(y = i|x; W) = \frac{e^{w_i^T x + b_i}}{\sum_{j=1}^n e^{w_j^T x + b_j}} \quad (69)$$

The softmax loss (also known as cross-entropy loss) for an input-output pair  $(x, y)$  is defined as the negative log-likelihood of the correct class:

$$L_i = -\log(P(y = y_i|x; W)) = -f_{y_i} + \log \sum_{j=1}^n e^{f_j}, \quad (70)$$

where  $f_{y_i}$  is the score corresponding to the true class  $y_i$ , and  $f_j$  are the scores for all classes. The total loss for a batch of data is the mean of  $L_i$  over all examples in the batch.

However, Softmax Loss has limitations in face recognition tasks as it does not explicitly enforce fine-grained control over intra-class and inter-class distances. This can hinder its effectiveness in distinguishing between similar faces and ensuring that different identities are well-separated in the feature space.

**A-Softmax Loss** The A-Softmax loss [185], also known as the SphereFace loss, is designed to address the limitations of traditional softmax loss by incorporating angular information between the feature vectors of face images and their corresponding weight vectors. The A-Softmax loss enhances inter-class separability and reduces intra-class variations by enforcing angular margins, leading to more discriminative features.

Given a weight matrix  $W$ , an input feature vector  $x$ , and a margin parameter  $m$ , the SphereFace loss is calculated as follows:

1. Compute the normalized weight matrix  $W_{norm}$ :

$$W_{norm} = \frac{W}{\|W\|}, \quad (71)$$

where each column vector  $w_i$  in  $W_{norm}$  is a unit vector, i.e.,  $\|w_i\| = 1$ . This normalization ensures that each class's weight vector lies on the unit hypersphere.

2. Compute the cosine of the angle  $\theta$  between the feature vector  $x$  and the normalized weight vector:

$$\cos(\theta) = \frac{W_{norm} \cdot x}{\|x\|}, \quad (72)$$

where  $\|\cdot\|$  denotes the L2 norm.

3. Apply the angular margin  $m$  to the angle  $\theta$ . The angle margin enforces that the decision boundary becomes more stringent, requiring the correct class to have a larger cosine similarity:

$$\cos(m\theta) = \cos(m \cdot \arccos(\cos(\theta))), \quad (73)$$

where  $\arccos(\cos(\theta))$  computes the angle  $\theta$  from the cosine value.

4. Compute the scaled prediction logits:

$$y'_{pred} = \|x\| \cdot \cos(m\theta). \quad (74)$$

5. The final A-Softmax loss  $L$  is calculated as the negative log-likelihood of the correct class:

$$L = -\log \frac{e^{y'_{pred}}}{e^{y'_{pred}} + \sum_{j \neq y} e^{\|x\| \cdot \cos(\theta_j)}}, \quad (75)$$

where the numerator represents the exponentiated scaled prediction for the true class, and the denominator sums the exponentiated scaled predictions for all classes.

The A-Softmax loss introduces an angular margin between different classes, making the features more discriminative. It helps to produce more compact intra-class representations and maximizes inter-class separability, which is particularly beneficial for face recognition tasks, as it enhances the model's ability to differentiate between different identities.

**Center Loss** The center loss [186] aims to reduce intra-class variance while increasing inter-class variance in feature space by minimizing the distances between feature representations of samples and their respective class centers. This loss function is designed to encourage samples from the same class to cluster closely around a central point, known as the class center, thus improving the discriminative power of the learned features.

The center loss is defined as the Euclidean distance between the feature representation of a sample and the corresponding class center in the feature space. It is typically combined with the primary loss function, such as softmax loss, to ensure both discriminative feature learning and correct classification.

The mathematical formulation of the center loss is:

$$L_{center} = \frac{1}{2} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{c}_{y_i}\|_2^2, \quad (76)$$

where:

- $\mathbf{x}_i$  is the feature representation of the  $i^{th}$  sample.
- $y_i$  is the corresponding class label for the  $i^{th}$  sample.
- $\mathbf{c}_{y_i}$  is the center of class  $y_i$  in the feature space.
- $n$  is the total number of samples in the batch.

The class centers  $\mathbf{c}_{y_i}$  are updated during the training process. They are calculated by taking the mean of the feature representations of all samples belonging to the same class in the current mini-batch. To prevent abrupt updates and ensure stability, a learning rate parameter  $\alpha$  is often used to update the centers incrementally:

$$\mathbf{c}_{y_i}^{(t+1)} = \mathbf{c}_{y_i}^{(t)} + \alpha \left( \mathbf{x}_i - \mathbf{c}_{y_i}^{(t)} \right), \quad (77)$$

where  $\mathbf{c}_{y_i}^{(t)}$  is the class center at the  $t^{th}$  iteration, and  $\alpha$  is the learning rate for the center updates.

The total loss function for the training process is typically a combination of the primary loss (e.g., softmax loss) and the center loss:

$$L = L_{softmax} + \lambda L_{center}, \quad (78)$$

where  $\lambda$  is a weight parameter that balances the influence of the center loss against the primary loss.

The inclusion of center loss helps the model produce features that are not only well-separated across different classes but also tightly clustered within the same class. This is particularly beneficial for tasks like face recognition, where distinguishing between individuals is critical.

**CosFace: Large-Margin Cosine Loss** CosFace loss [187], also known as the Large Margin Cosine Loss, is designed to enhance the discriminative power of deeply learned features by maximizing the decision margin in the cosine space. The cosine of the angle between the feature vector  $\mathbf{x}_i$  and the weight vector  $\mathbf{w}_j$  of the  $j^{th}$  class is computed as:

$$\cos \theta_j = \frac{\mathbf{w}_j^T \mathbf{x}_i}{\|\mathbf{w}_j\| \|\mathbf{x}_i\|}, \quad (79)$$

where  $\|\cdot\|$  denotes the L2 norm.

To introduce a margin in the cosine space, CosFace modifies the cosine similarity corresponding to the ground truth class  $y_i$  by subtracting a cosine margin  $m$ . The adjusted cosine similarity for the ground truth class is given by:

$$\cos \theta_{y_i} - m. \quad (80)$$

The CosFace loss for an input-output pair  $(\mathbf{x}_i, y_i)$  is then defined as:

$$L_i = -\log \frac{e^{s(\cos \theta_{y_i} - m)}}{e^{s(\cos \theta_{y_i} - m)} + \sum_{j \neq y_i}^n e^{s \cos \theta_j}}, \quad (81)$$

where  $s$  is a scaling factor that helps in optimizing the convergence of the training process by adjusting the distribution of logits.

The key advantage of CosFace is its ability to enforce a large margin between classes in the normalized cosine space, which leads to better separation of features across different classes. Unlike SphereFace, which uses angular margins, CosFace directly manipulates the cosine similarity, avoiding potential issues related to the non-monotonicity of the cosine function. Additionally, since both the feature vectors and the weight vectors are normalized, the model cannot reduce the loss by simply scaling the norms, thereby ensuring that discriminative learning focuses on the angular separation of features.

**ArcFace: Additive Angular Margin Loss** The ArcFace loss [188], also known as Additive Angular Margin Loss, enhances the discriminative power of the softmax loss by incorporating an angular margin penalty into the target logit. This technique aims to achieve better class separation by adding an additional margin in the angular space.

The ArcFace method introduces an additive angular margin  $m$  to the angle  $\theta_{y_i}$  corresponding to the ground truth class  $y_i$ . The modified cosine similarity for the ground truth class is defined as:

$$\cos(\theta_{y_i} + m), \quad (82)$$

where  $\theta_{y_i}$  is the angle between the feature vector  $\mathbf{x}_i$  and the weight vector  $\mathbf{w}_{y_i}$  corresponding to the true class  $y_i$ .

The ArcFace loss for an input-output pair  $(\mathbf{x}_i, y_i)$  is then defined as:

$$L_i = -\log \frac{e^{s \cos(\theta_{y_i} + m)}}{e^{s \cos(\theta_{y_i} + m)} + \sum_{j \neq y_i}^n e^{s \cos \theta_j}}, \quad (83)$$

where  $s$  is a scaling factor that controls the smoothness of the decision boundary.

The introduction of the margin  $m$  effectively increases the decision margin in the angular space, pushing the features of different classes further apart while keeping features of the same class closely grouped. This margin can be interpreted as an additional arc length on the hypersphere, which the feature vectors and weight vectors are projected onto, with a radius determined by the scaling factor  $s$ . The angular nature of the margin ensures that the decision boundaries are determined purely by the angles, making the feature separation more robust to variations in feature magnitude.

ArcFace has been shown to provide superior performance in face recognition tasks compared to other loss functions like Triplet Loss, achieving better inter-class discrepancy (separation between different classes) while maintaining high intra-class similarity (compactness within the same class) [189]. This characteristic makes ArcFace particularly effective for applications that require high precision to distinguish between closely related classes, such as identity verification and authentication systems.

**Triplet Loss** Triplet Loss is one of the most well-known loss functions for face recognition, designed to encourage the model to distinguish between images of the same person and images of different persons. The core idea behind Triplet Loss [190, 191] is to ensure that the distance between an anchor image and a positive image (of the same person) is smaller than the distance between the anchor image and a negative image (of a different person) by at least a predefined margin.

Given an anchor image,  $A$ , a positive image,  $P$ , and a negative image,  $N$ , the loss function is defined as

$$L_{triplet} = \max\{0, \|f_A - f_P\|_2^2 - \|f_A - f_N\|_2^2 + \alpha\}, \quad (84)$$

where  $f_A$ ,  $f_P$ , and  $f_N$  are the feature embeddings of the anchor, positive, and negative images, respectively, and  $\alpha$  is the margin parameter. This margin ensures that the model maintains a certain minimum distance between the embeddings of different classes.

The term  $\|f_A - f_P\|_2^2$  represents the squared L2 distance between the anchor embeddings and the positive image, while  $\|f_A - f_N\|_2^2$  represents the squared L2 distance between the embeddings of the anchor and the negative image.

The goal of the triplet loss is to minimize this loss function, which effectively pulls the positive pairs closer together in the embedding space and pushes the negative pairs farther apart. By doing so, the model learns a feature space where images of the same person have similar embeddings, and images of different persons have distinctly separate embeddings.

The choice of the margin  $\alpha$  is crucial; if it is too small, the model might not learn enough discriminative features, whereas if it is too large, it can lead to slow convergence or even instability during training. Careful selection and tuning of  $\alpha$  are necessary to achieve optimal performance.

**Contrastive Loss** The contrastive loss [192] is designed to learn a feature representation or embedding that projects similar images close to each other and dissimilar images far apart in the feature space. This loss is commonly used with a Siamese neural network architecture, which processes pairs of images to learn these relationships.

The contrastive loss function is defined as

$$L = \frac{1}{N} \sum_{i=1}^N \left[ y_i \|f(x_i^a) - f(x_i^p)\|_2^2 + (1 - y_i) \max\left(0, m - \|f(x_i^a) - f(x_i^n)\|_2^2\right) \right], \quad (85)$$

where:

- $f(x_i^a)$  is the feature representation of the anchor image  $x_i^a$ .
- $f(x_i^p)$  is the feature representation of the positive image  $x_i^p$ .
- $f(x_i^n)$  is the feature representation of the negative image  $x_i^n$ .
- $y_i$  is a binary label indicating whether the pair (anchor and positive) is similar (1) or dissimilar (0).
- $m$  is the margin hyperparameter that determines the minimum distance required between dissimilar pairs.
- $N$  is the number of image pairs in the batch.

The term  $\|f(x_i^a) - f(x_i^p)\|_2^2$  measures the squared Euclidean distance between the feature representations of the anchor and positive images. For similar pairs, the goal is to minimize this distance. In contrast, the term  $\max\left(0, m - \|f(x_i^a) - f(x_i^n)\|_2^2\right)$  ensures that the distance between two pairs is at least the margin  $m$ . If the distance is already greater than  $m$ , the contribution to the loss is zero.

The margin  $m$  acts as a threshold and controls how aggressively the model should separate dissimilar embeddings. The choice of  $m$  affects the model's ability to differentiate between similar and dissimilar pairs. One advantage of contrastive loss is that it facilitates extending a trained model to new or unseen classes, as it focuses on learning a meaningful feature space rather than classifying it into predefined categories.

One limitation of this loss function is that the margin  $m$  is a constant for all dissimilar pairs, which may not account for varying degrees of dissimilarity; the model might implicitly assume equal dissimilarity between all negative pairs. Additionally, the absolute notion of similarity and dissimilarity defined by the label  $y_i$  may not generalize well across different contexts, as what is considered similar in one context may not be perceived the same in another [189, 193, 194].

**Circle Loss** Circle loss [195] enhances the discriminative power of learned features by simultaneously maximizing the similarity of positive pairs and minimizing the similarity of negative pairs, while maintaining a *circle-like* decision boundary. This approach addresses challenges such as imbalanced data and complex distributions, which are prevalent in tasks like face recognition.

The Circle loss is defined as:

$$\begin{aligned}
 \alpha_{pos_i} &= \max(O_{pos} - s_{pos_i}, 0) \\
 \alpha_{neg_j} &= \max(s_{neg_j} - O_{neg}, 0) \\
 \text{sum}_{pos} &= \sum_i e^{-\gamma \cdot \alpha_{pos_i} \cdot (s_{pos_i} - O_{pos})} \\
 \text{sum}_{neg} &= \sum_j e^{\gamma \cdot \alpha_{neg_j} \cdot (s_{neg_j} - O_{neg})} \\
 L &= \log(1 + \text{sum}_{pos} \cdot \text{sum}_{neg})
 \end{aligned} \tag{86}$$

Where:

- $s_{pos_i}$  and  $s_{neg_j}$  represent the pairwise similarity scores for positive pairs (same class) and negative pairs (different classes), respectively.
- $O_{pos}$  and  $O_{neg}$  are user-defined margins for positive and negative pairs. Typically,  $O_{pos}$  is set smaller than the expected similarity for positive pairs, and  $O_{neg}$  is set larger than the expected similarity for negative pairs, ensuring  $O_{pos} < O_{neg}$ .
- $\alpha_{pos_i}$  and  $\alpha_{neg_j}$  are slack variables ensuring that the similarity scores for positive pairs are greater than  $O_{pos}$ , and for negative pairs are less than  $O_{neg}$ . These are computed using  $\max(0, \cdot)$  to ignore pairs that already satisfy the margin requirements.
- $\text{sum}_{pos}$  and  $\text{sum}_{neg}$  are the scaled and exponentiated sums of positive and negative similarities, respectively. The exponential scaling, modulated by  $\gamma$ , emphasizes differences, particularly for pairs that do not meet the margin requirement.  $\gamma$  acts as a scaling factor controlling this emphasis.
- The overall loss  $L$  is calculated as the logarithm of  $1 + \text{sum}_{pos} \cdot \text{sum}_{neg}$ . This formulation drives  $\text{sum}_{pos}$  to be low (favoring high similarity for positive pairs) and  $\text{sum}_{neg}$  to be low (favoring low similarity for negative pairs). The logarithmic transformation stabilizes the loss, mitigating the impact of outliers and ensuring the argument remains positive.

Circle loss offers a clear convergence target in the similarity space defined by  $(s_{neg}, s_{pos})$ , directing optimization toward the point  $(O_{neg}, O_{pos})$ . However, selecting the appropriate margins  $(O_{neg}, O_{pos})$  can be challenging and typically requires hyperparameter tuning, often through cross-validation. A common practice is to begin with smaller margins and gradually increase them, avoiding overly large margins that can hinder learning or excessively small margins that fail to enforce sufficient discrimination.

In some implementations,  $O_{pos}$  and  $O_{neg}$  are not independent but are related by a constant margin  $m$ , simplifying the tuning process. By setting a fixed gap between positive and negative pairs, this approach ensures a consistent separation, making it necessary to only tune one of the margins or the margin gap  $m$ .

**Barlow Twins Loss** The Barlow Twins loss [196] is a self-supervised learning approach designed to learn highly informative and non-redundant representations. The main idea is to make the outputs of a twin network, processing two different augmentations of the same image, as similar as possible while reducing redundancy between the dimensions of these representations.

Given a batch size  $N$  and the dimensionality  $D$  of the embeddings, the network processes two different augmentations of the same input data, producing the embeddings  $z_a$  and  $z_b$ . These embeddings are then normalized to have zero mean and unit variance along the batch dimension, resulting in  $z_{a_{norm}}$  and  $z_{b_{norm}}$ .

The computation of the Barlow Twins loss involves the following steps:

1. Compute the Cross-Correlation Matrix  $C$ :

$$C = \frac{1}{N} z_{a_{norm}}^T \cdot z_{b_{norm}} \tag{87}$$



2. Compute the Difference Matrix  $C_{diff}$ : Subtract the identity matrix  $I$  from  $C$  and square the elements:

$$C_{diff} = (C - I)^2 \quad (88)$$

3. Scale Off-Diagonal Elements: Apply a scaling factor  $\lambda$  to the off-diagonal elements of  $C_{diff}$ :

$$C_{diff_{ij}} = \begin{cases} C_{diff_{ij}}, & \text{if } i = j \\ \lambda C_{diff_{ij}}, & \text{if } i \neq j \end{cases} \quad (89)$$

4. Compute the Barlow Twins Loss  $L$ : Sum all the elements in the adjusted  $C_{diff}$ :

$$L = \sum_{i,j} C_{diff_{ij}} \quad (90)$$

The loss function is designed to achieve two main goals: (1) making the diagonal elements of the cross-correlation matrix close to 1, which ensures that the representations of the two views are highly correlated; and (2) making the off-diagonal elements close to 0, which reduces redundancy among different components of the learned representations.

This method does not rely on a fixed number of classes and avoids data expansion issues as it does not require explicit negative examples. However, achieving good performance often requires a large dimensionality of the final representation. Additionally, the model's robustness can be affected by certain distortions to the inputs, as noted in [189].

Barlow Twins loss is effective in scenarios where it is important to learn representations that are both invariant to different augmentations of the same input and decorrelated across different dimensions, making it suitable for various self-supervised learning tasks.

**SimSiam Loss** SimSiam [197] is a self-supervised learning method designed to learn meaningful representations by encouraging the similarity between two augmented views of the same image. Although not explicitly designed for face recognition, SimSiam can be employed to learn robust face representations.

Given two different augmentations,  $x_1$  and  $x_2$ , of the same image, these are passed through a neural network and a prediction head to obtain feature embeddings  $z_1, z_2$  and predictions  $p_1, p_2$ . The loss function is defined as the negative cosine similarity between the predictions and the corresponding features from the alternate view:

$$L = -\frac{1}{2} \left[ \frac{p_1^T z_2}{\|p_1\| \cdot \|z_2\|} + \frac{p_2^T z_1}{\|p_2\| \cdot \|z_1\|} \right], \quad (91)$$

where  $T$  denotes the transpose, and  $\|\cdot\|$  denotes the L2 norm.

This loss encourages the model to align the predictions  $p_1$  and  $p_2$  with the features  $z_2$  and  $z_1$ , respectively, from different augmented views. An essential aspect of SimSiam is the stop-gradient operation applied to  $z_2$  in the first term and  $z_1$  in the second term of the loss function. This operation prevents gradients from backpropagating through these feature vectors, which is crucial for avoiding trivial solutions where the model could simply collapse by producing identical outputs for all inputs.

The advantages of SimSiam Loss are the following:

1. **No Need for Negative Pairs:** Unlike contrastive learning methods that require negative examples, SimSiam does not need such pairs, simplifying the training process and potentially improving efficiency.
2. **Stop-Gradient Operation:** The stop-gradient operation is a key feature that helps prevent the collapse into trivial solutions, a common problem in self-supervised learning.
3. **Simplicity:** SimSiam employs a straightforward loss function and a symmetric network architecture, making it simpler compared to other self-supervised methods. It focuses on making representations from different views of the same image similar.

The disadvantages include:

1. **Hyperparameter Sensitivity:** The method's performance is sensitive to hyperparameters like learning rate and weight decay, necessitating careful tuning to achieve optimal results. Incorrect settings can significantly impact performance.

2. **Dependence on Data Augmentation:** The effectiveness of SimSiam heavily relies on the quality and choice of data augmentations, which requires domain expertise and considerable effort to optimize.
3. **Non-semantic Features:** A limitation common to many self-supervised learning methods, including SimSiam, is that the learned features may not always capture high-level semantic information. They might excel in identifying low-level patterns but may lack in semantic depth.

SimSiam offers an approach to self-supervised learning by leveraging the similarity between augmented views without the need for negative samples, but it also requires careful consideration of augmentations and hyperparameters to be most effective.

## 4.5 Image Generation

Image generation in deep learning involves using artificial neural networks to generate new images. This task has progressed significantly with models such as Variational Autoencoders (VAEs) [198, 199, 200, 201], Generative Adversarial Networks (GANs) [202, 203, 204, 205, 206], Normalized Flow models (NFs) [207, 208, 209, 210, 211], Energy-Based Models (EBMs) [212, 213, 214, 215, 216, 217], and Diffusion Models [218, 219, 220, 221, 222]. These models can generate high-quality images that can be used in various applications such as image super-resolution [223, 224, 225, 226], denoising [227, 228, 229], inpainting [230, 231, 232, 233], and style transfer [234, 235, 236, 237].

*Variational Autoencoders (VAEs)* are generative models that use deep learning techniques to create new data and learn latent representations of the input data. They consist of an encoder and a decoder. The encoder compresses input data into a lower-dimensional latent space, while the decoder reconstructs the original data from points in the latent space. The process is trained to minimize the difference between original and reconstructed data and ensure the latent space approximates a standard Gaussian distribution. VAEs can generate new data by feeding the decoder points sampled from the latent space.

*Generate Adversarial Networks (GANs)* involve two neural networks, a Generator, and a Discriminator, playing a game against each other. The Generator tries to create data similar to the training data, while the Discriminator tries to distinguish between the real and generated data. Through this process, both networks improve: the Generator learns to produce increasingly realistic data, while the Discriminator becomes better at distinguishing between real and artificial data. This adversarial process continues until an equilibrium is reached, at which point the Generator is producing realistic data and the Discriminator is, at best, randomly guessing whether the data is real or generated. This equilibrium is conceptually referred to as a Nash equilibrium in game theory [238].

*Normalizing Flows* use invertible transformations to generate diverse outputs and provide an exact and tractable likelihood for a given sample, enabling efficient sampling and density estimation. However, they can be computationally intensive to train.

*Energy-Based Models (EBMs)* learn a scalar energy function to distinguish real data points from unlikely ones. A neural network often parameterizes this function and learns from the data. Sampling in EBMs is typically done via Markov Chain Monte Carlo (MCMC) methods. EBMs can represent a wide variety of data distributions but can be challenging to train due to the intractability of the partition function and the computational expense of MCMC sampling.

*Diffusion models* use a random process to transform simple data distribution, like Gaussian noise, into the desired target data distribution. This is controlled by a trained neural network, allowing the generation of high-quality data, like images, through a smooth transition from noise to the desired data.

The following sections will review the common lost functions and performance metrics used for image generation.

### 4.5.1 Image Generation Loss functions

The loss function in a VAE consists of the reconstruction loss and the Kullback-Leibler Divergence Loss (KL) (refer to section 4.5.1). The reconstruction loss measures how well the decoded data matches the original input data. The KL divergence measures how much the learned distribution in the latent space deviates from a target distribution, usually a standard normal distribution. KL-divergence is used as a regularization term to ensure that the distributions produced by the encoder remain close to a unit Gaussian, penalizing the model if the learned distributions depart from it.

The most common loss function used in GANs is the adversarial loss, which is the sum of the cross-entropy loss between the generator's predictions and the real or fake labels. Later, WGAN [239] applied the Wasserstein distance as an alternative to training GANs to improve stability and avoid mode collapse that occurs when the generator network stops learning the underlying data distribution and begins to produce a limited variety of outputs, rather than a diverse range of outputs that accurately represent the true data distribution.

Normalizing Flows are typically trained using maximum likelihood estimation. Given a dataset, the aim is to maximize the log-likelihood of the data under the model by minimizing the negative log-likelihood of the data.

During training, Energy-based models (EBMs) minimize a loss function that encourages the energy function to assign lower energy values to data points from the training data and higher energy values to other points. Different types of EBMs use different loss functions, such as Contrastive Divergence (CD) [240], Maximum Likelihood Estimation (MLE) [241], and Noise-Contrastive Estimation (NCE) [242].

Diffusion models use a denoising loss function based on the Mean Absolute Error (MAE) or the Mean Squared Error (MSE) between the original and reconstructed data.

In the next sections, we describe in detail each of these losses.

**Reconstruction Loss** Reconstruction loss is a essential in tasks involving image generation and reconstruction, such as in Variational Autoencoders (VAEs), image restoration, and image synthesis. The primary purpose of the reconstruction loss is to ensure that the output from the decoder network is as close as possible to the original input image. This loss penalizes the differences between the original and reconstructed images, encouraging the model to learn a faithful representation of the input data.

One common form of reconstruction loss is the Mean Squared Error (MSE), which measures the average squared differences between the original and reconstructed images. The MSE loss is defined as:

$$L_{recon} = \frac{1}{N} \sum_{i=1}^N \|x_i - \hat{x}_i\|^2, \quad (92)$$

where  $x_i$  is the original image,  $\hat{x}_i$  is the reconstructed image, and  $N$  is the number of images in the batch. The MSE loss is particularly suitable for continuous-valued image data, as it directly quantifies the pixel-wise differences.

For binary images, Binary Cross-Entropy (BCE) loss is often used, as it is more appropriate for modeling binary outputs. The BCE loss can be defined as:

$$L_{BCE} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)], \quad (93)$$

where  $y_i$  is the binary value of the original image, and  $\hat{y}_i$  is the binary value of the reconstructed image. The BCE loss measures the divergence between the true binary values and the predicted probabilities, making it suitable for tasks involving binary or probabilistic outputs.

Reconstruction loss is widely used in various deep learning tasks beyond VAEs, including:

1. **Image Restoration:** In tasks such as denoising, super-resolution, and inpainting, reconstruction loss helps the model generate cleaner, high-resolution, or complete versions of images.
2. **Image Generation:** Generative models like GANs and VAEs use reconstruction loss to ensure that generated images resemble real samples.
3. **Anomaly Detection:** Reconstruction loss is used in models designed to detect anomalies by measuring how well the model reconstructs normal versus anomalous data.

**Kullback-Leibler Divergence Loss** The KL divergence loss, also known as the KL divergence loss, measures the difference between the predicted probability distribution and the true probability distribution of the classes [243]. The KL divergence loss ensures that the predicted probabilities are close to the true probabilities, which can be useful in cases where the true probabilities are known or can be approximated.

The KL divergence loss is defined as

$$KL(p||q) = \sum_{i=1}^n p(x_i) \log\left(\frac{p(x_i)}{q(x_i)}\right), \quad (94)$$

where  $p(x_i)$  is the true probability of class  $x_i$  and  $q(x_i)$  is the predicted probability of class  $x_i$ .

The KL divergence loss is often used in generative models such as variational autoencoders, which aim to approximate the true data distribution. It is also used in reinforcement learning to ensure that the agent's learned policy is close to the optimal policy.

One disadvantage of using the KL divergence loss is that it is sensitive to zero probabilities in the true probability distribution. This can lead to numerical instability and can make the loss function difficult to optimize. To overcome this issue, a common practice is to add a small value (e.g.,  $1e-7$ ) to the true probability distribution to avoid zero probabilities.

**Adversarial Loss** Adversarial loss is the fundamental loss function used in Generative Adversarial Networks (GANs). It is derived from a minimax game framework involving two neural networks: the generator ( $G$ ) and the discriminator ( $D$ ). The adversarial loss is essential for training the generator, as it aims to generate data indistinguishable from real data, while the discriminator learns to differentiate between real and generated data.

The concept of adversarial loss was first introduced by Goodfellow et al. in [202]. The authors demonstrated that this loss function facilitates the convergence of the generator towards a Nash equilibrium, where the generator's output is so realistic that the discriminator cannot reliably distinguish between real and generated data.

The adversarial loss for a GAN can be expressed as:

$$L_{adv}(G, D) = -\mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] - \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))], \quad (95)$$

where:

- $G$  represents the generator network that maps a noise vector  $z$  to a data space, generating an image  $G(z)$ .
- $D$  represents the discriminator network that outputs the probability that an input image is real.
- $x$  denotes real images drawn from the real data distribution  $p_{data}(x)$ .
- $z$  denotes latent vectors sampled from a noise distribution  $p_z(z)$ .

In this setup, the generator aims to minimize the adversarial loss by producing realistic images  $G(z)$  that fool the discriminator, while the discriminator tries to maximize the loss by correctly identifying real images from generated ones. This adversarial objective can be framed as a two-player game with the following optimization goals:

For the Discriminator  $D$ , maximize the probability of correctly classifying real and generated samples:

$$\max_D \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \quad (96)$$

For the Generator  $G$ , minimize the likelihood of the discriminator identifying the generated samples as fake:

$$\min_G \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \quad (97)$$

In practice, the generator is often trained to maximize  $\mathbb{E}_{z \sim p_z(z)}[\log D(G(z))]$ , which can help mitigate issues such as vanishing gradients when the discriminator becomes too strong.

The training process involves alternately updating the generator and the discriminator. The generator improves its output to produce more realistic images, while the discriminator becomes better at distinguishing real from generated data. This dynamic continues until the generator's images are indistinguishable from real ones, ideally reaching an equilibrium where the discriminator can no longer differentiate between real and fake samples with certainty.

Adversarial loss is an effective tool for training GANs, enabling the generation of highly realistic data. However, it can also lead to challenges such as mode collapse and instability during training, requiring careful balancing and regularization techniques.

**Wasserstein Loss** The Wasserstein loss, introduced by Arjovsky et al. in [239], is used in generative models like Generative Adversarial Networks (GANs) to measure the divergence between the real data distribution and the generated data distribution. Unlike the traditional adversarial loss, Wasserstein loss is based on the Wasserstein distance (also known as Earth Mover's Distance), which quantifies the cost of transporting mass from one distribution to match another.

Mathematically, the Wasserstein loss is defined as:

$$W(p_{data}, p_{gen}) = \inf_{\gamma \in \Gamma(p_{data}, p_{gen})} \int_{x, y} \|x - y\| d\gamma(x, y), \quad (98)$$

where:

- $p_{data}$  and  $p_{gen}$  represent the real and generated data distributions, respectively.
- $\Gamma(p_{data}, p_{gen})$  is the set of all joint distributions  $\gamma(x, y)$  whose marginals are  $p_{data}$  and  $p_{gen}$ .
- $\gamma(x, y)$  represents a coupling between  $p_{data}$  and  $p_{gen}$ , effectively describing how mass should be transported to transform one distribution into the other.
- $\|x - y\|$  is a chosen metric (typically Euclidean distance) between samples  $x$  and  $y$  from the real and generated distributions, respectively.

The Wasserstein distance measures the minimum work required to transform the generated distribution into the real distribution, where work is defined as the integral of the distance between points  $x$  and  $y$ , weighted by the joint distribution  $\gamma(x, y)$ . This distance provides a meaningful measure even when the real and generated distributions have disjoint supports, which is a common issue with traditional GAN losses.

The Wasserstein loss has a couple of advantages over the adversarial loss. On the one hand, the Wasserstein loss tends to produce more stable training dynamics compared to traditional adversarial losses. This stability arises because the Wasserstein distance provides continuous and differentiable gradients, even when the support of the distributions does not overlap. On the other hand, unlike traditional GANs, which can suffer from mode collapse (where the generator produces a limited variety of outputs), Wasserstein GANs (WGANs) mitigate this problem. The Wasserstein distance ensures that the generator is penalized proportionally to the distance between distributions, encouraging the generation of diverse outputs.

The implementation of Wasserstein loss in GANs typically requires some modifications, such as using weight clipping or gradient penalty (as seen in WGAN-GP [244]) to enforce the Lipschitz constraint on the discriminator's function. These modifications are necessary to ensure the validity of the distance measure and the stability of the training process.

**Negative Log-Likelihood in Normalizing Flows** Normalizing Flows are a class of generative models that aim to learn a bijective transformation (or a series of transformations) that maps a complex data distribution to a simpler, typically standard, base distribution, such as a Gaussian. The key feature of Normalizing Flows is the ability to compute the exact likelihood of the data by leveraging the change of variables formula, which requires the transformation to be invertible and differentiable with a tractable Jacobian determinant.

Given a data point  $x$ , let  $z = f_\theta(x)$  denote the transformation of  $x$  to the latent space under the flow model parameterized by  $\theta$ , and  $p_z(z)$  denote the density of the base distribution. The log-likelihood of  $x$  under the model is computed as:

$$\log p_\theta(x) = \log p_z(f_\theta(x)) + \log \left| \det \frac{\partial f_\theta(x)}{\partial x} \right|, \quad (99)$$

where the second term,  $\log \left| \det \frac{\partial f_\theta(x)}{\partial x} \right|$ , is the log absolute determinant of the Jacobian matrix of the transformation  $f_\theta$ . This term accounts for the change in volume induced by the transformation, ensuring the correct density transformation from the base distribution to the data distribution.

The training objective for Normalizing Flows involves maximizing the likelihood of the data, which is equivalent to minimizing the negative log-likelihood. Given a dataset with  $N$  data points  $x_1, x_2, \dots, x_N$ , the loss function  $L(\theta)$  is expressed as:

$$L(\theta) = -\frac{1}{N} \sum_{i=1}^N \log p_\theta(x_i) = -\frac{1}{N} \sum_{i=1}^N \left[ \log p_z(f_\theta(x_i)) + \log \left| \det \frac{\partial f_\theta(x_i)}{\partial x_i} \right| \right]. \quad (100)$$

During training, this loss function is minimized using stochastic gradient descent or its variants, allowing the model to learn the parameters  $\theta$  of the flow transformations.

Normalizing Flows can be computationally intensive to train due to the necessity of computing and backpropagating through the Jacobian of the transformations. Consequently, specialized architectures such as RealNVP [207] and Glow [208] are designed to facilitate efficient computation of the Jacobian determinant, often using techniques like affine coupling layers and invertible  $1 \times 1$  convolutions, which simplify the computation of the determinant and its gradient.

**Contrastive Divergence** Contrastive Divergence (CD) is a method used in training Energy-Based Models (EBMs) to approximate the gradient of the log-likelihood, which involves an expectation over all possible data configurations. Calculating this expectation exactly is typically intractable. CD approximates it by running a Markov chain for a limited number of steps.

Given a dataset consisting of  $N$  data points  $x_1, x_2, \dots, x_N$ , the log-likelihood of the data under the model is:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^N \log p(x_i; \theta), \quad (101)$$

where  $\theta$  represents the parameters of the model, and  $p(x; \theta)$  is the probability of a data point  $x$ , defined as  $p(x; \theta) = \frac{e^{-E(x; \theta)}}{Z(\theta)}$ , with  $E(x; \theta)$  being the energy function and  $Z(\theta)$  the partition function.

The gradient of the log-likelihood with respect to the parameters is:

$$\frac{\partial}{\partial \theta} L(\theta) = \frac{1}{N} \sum_{i=1}^N \frac{\partial}{\partial \theta} \log p(x_i; \theta). \quad (102)$$

This gradient can be decomposed into two terms:

1. Positive Phase: This term is computed from the data distribution and is straightforward to compute:

$$\frac{1}{N} \sum_{i=1}^N \frac{\partial E(x_i; \theta)}{\partial \theta}. \quad (103)$$

2. Negative Phase: This term involves an expectation over the model's distribution and is generally intractable:

$$-\langle \frac{\partial E(x; \theta)}{\partial \theta} \rangle_p, \quad (104)$$

where  $\langle \cdot \rangle_p$  denotes an expectation with respect to the model distribution.

In CD, the negative phase is approximated by running a Markov chain starting from a data point for a few steps, obtaining a sample to estimate the expectation. This leads to the CD- $k$  algorithm, where  $k$  represents the number of Gibbs sampling steps. The update to the parameters after observing a data point  $x$  is then:

$$\Delta \theta = \eta \left( \frac{1}{N} \sum_{i=1}^N \frac{\partial E(x_i; \theta)}{\partial \theta} - \langle \frac{\partial E(x'; \theta)}{\partial \theta} \rangle_{CD} \right), \quad (105)$$

where  $\eta$  is the learning rate,  $\langle \cdot \rangle_{CD}$  denotes the expectation with respect to the distribution defined by the Markov chain after  $k$  steps, and  $x'$  is the sample obtained after  $k$  steps of Gibbs sampling starting from the data point  $x$ .

Contrastive Divergence is often more computationally efficient than other methods, such as Persistent Contrastive Divergence (PCD) [245] or Mean-Field methods [246], because it requires running the Markov chain for only a few steps instead of until reaching equilibrium. However, this introduces a bias in the gradient estimate of the log-likelihood. CD is suitable when the bias is acceptable or can be mitigated by increasing the number of steps in the Markov chain.

#### 4.5.2 Image Generation Metrics

Some of the common metrics used for Image generation are:

- Peak Signal-to-Noise Ratio (PSNR)
- Structural Similarity Index (SSIM)
- Inception Score (IS) [247]
- Frechet Inception Distance (FID) [248]

In the following sections, we will dive into each of these metrics.

**Peak Signal-to-Noise Ratio (PSNR)** Peak Signal-to-Noise Ratio (PSNR) is a traditional metric used to quantify the quality of image and video codecs by comparing the original and reconstructed (compressed and then decompressed) images or videos. It measures the fidelity of the reconstruction, indicating how closely the reconstructed image resembles the original.



In image generation tasks, PSNR is often used to evaluate the quality of the generated images, particularly in applications such as image super-resolution, denoising, and inpainting. PSNR compares the generated image to a ground truth high-quality image, providing an objective assessment of reconstruction quality.

The PSNR is defined in terms of the mean squared error (MSE), which measures the average squared differences between the original and the reconstructed images. For two  $m \times n$  monochrome images  $I$  and  $K$ , where one image is a noisy approximation of the other, MSE is calculated as:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2 \quad (106)$$

PSNR is then defined as:

$$PSNR = 10 \cdot \log_{10} \left( \frac{(MAX_I)^2}{MSE} \right), \quad (107)$$

where  $MAX_I$  is the maximum possible pixel value of the image. For example, in an 8-bit grayscale image, this value is 255.

PSNR is expressed in decibels (dB) and provides an easily interpretable score: higher PSNR values generally indicate better quality reconstructions. However, it is important to note that PSNR may not always correlate with perceived visual quality. For example, two images with similar perceptual quality can have significantly different PSNR values if they differ in pixel-wise detail.

While PSNR is useful for evaluating tasks with a clear ground truth, such as super-resolution or denoising, it is less effective for generative tasks where the goal is to produce new, high-quality images rather than exact reconstructions. In such cases, metrics like the Inception Score (IS) or Fréchet Inception Distance (FID) are often preferred, as they better capture perceptual quality and diversity [249].

**Structural Similarity Index (SSIM)** The Structural Similarity Index (SSIM) [250] is a metric used to assess the similarity between two images. It is particularly useful in image generation tasks for comparing generated images with target (real) images, providing a quantitative measure of how closely synthetic images produced by a generative model resemble actual images.

SSIM considers three main components: luminance, contrast, and structure. These components reflect perceptual characteristics that are important in human vision, making SSIM a more perceptually relevant measure compared to simpler metrics like Mean Squared Error (MSE) or Peak Signal-to-Noise Ratio (PSNR). The SSIM value ranges from -1 to 1, where 1 indicates perfect similarity, 0 indicates no correlation, and -1 indicates complete dissimilarity.

The calculation of the SSIM index involves the following steps:

1. **Luminance Comparison:** Measures the similarity in brightness between two images.
2. **Contrast Comparison:** Evaluates the similarity in contrast or intensity variation.
3. **Structural Comparison:** Assesses the similarity in structural information, focusing on the pattern of pixel intensities.

The formula for SSIM between two images  $x$  and  $y$  is given by:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (108)$$

where:

- $\mu_x$  and  $\mu_y$  are the means of the pixel intensities in images  $x$  and  $y$ , representing luminance.
- $\sigma_x$  and  $\sigma_y$  are the standard deviations of the pixel intensities in  $x$  and  $y$ , representing contrast.
- $\sigma_{xy}$  is the covariance of the pixel intensities between  $x$  and  $y$ , representing structural similarity.
- $C_1$  and  $C_2$  are small constants added to avoid instability when the denominator is close to zero, typically defined as  $C_1 = (K_1L)^2$  and  $C_2 = (K_2L)^2$ , where  $L$  is the dynamic range of the pixel values (e.g., 255 for 8-bit grayscale images), and  $K_1$  and  $K_2$  are small constants.



SSIM is more robust to variations in brightness and contrast compared to MSE and PSNR because it considers local patterns in pixel intensities that are consistent with human visual perception. However, while SSIM is valuable for evaluating the perceptual quality of images, it does not fully capture the diversity or the distributional properties of the generated data.

For a comprehensive evaluation of generative models, SSIM is often used alongside other metrics such as the Inception Score (IS) and Fréchet Inception Distance (FID). These additional metrics help assess the model’s ability to capture the underlying data distribution and generate diverse and high-quality images.

**Inception Score (IS)** The Inception Score (IS) is a metric used to evaluate the quality and diversity of images generated by generative models, such as GANs. The score was introduced by Salimans et al. [247] and is computed using an Inception-v3 classifier pre-trained on the ImageNet dataset.

The IS combines two key aspects: the quality of the generated images and the diversity of the objects within them. The score is based on the idea that a good generative model should produce images that are both recognizable (high quality) and varied across different classes (diversity).

To calculate the Inception Score, the generated images are passed through the Inception-v3 network, and the resulting softmax probabilities are used. The score is calculated using the following formula:

$$IS = \exp(\mathbb{E}_{x \sim p_g(x)}[KL(p(y|x)||p(y))]), \quad (109)$$

where:

- $p_g(x)$  represents the distribution of the generated images.
- $p(y|x)$  is the conditional probability of class  $y$  given image  $x$ , as predicted by the Inception model.
- $p(y) = \mathbb{E}_{x \sim p_g(x)}[p(y|x)]$  is the marginal class distribution over all generated images.

The Kullback-Leibler (KL) divergence term measures how different the conditional distribution  $p(y|x)$  is from the marginal distribution  $p(y)$ . A high KL divergence indicates that the model generates images with distinct class labels (diversity), while a low entropy in  $p(y|x)$  suggests high confidence in the predictions (quality).

The Inception Score provides a single scalar value that reflects both the quality and diversity of the generated images. It is easy to calculate and does not require access to the true data distribution, making it a convenient evaluation metric.

The limitations of the Inception Score include the following:

1. *Dependence on the Inception Model:* IS relies heavily on the Inception-v3 model, which is trained on ImageNet. This dependence may limit the score’s applicability in domains significantly different from ImageNet, as the model’s learned features might not generalize well.
2. *Mismatch with Human Perception:* The IS does not always correlate with human judgments of image quality. It can give high scores to images that are diverse and recognizable, even if they lack photorealism.
3. *Inability to Detect Mode Collapse:* The score may not effectively detect mode collapse (where the model generates a limited variety of images) if the few modes present are sufficiently diverse.
4. *Insensitivity to Class Balance:* The IS does not account for the frequency with which different classes are generated, potentially overlooking scenarios where the model produces certain classes disproportionately.
5. *Unreliable with Few Samples:* When calculated with a small number of samples, the IS can be unreliable due to high variance, making it less robust in such cases.

**Fréchet Inception Distance (FID)** The Fréchet Inception Distance (FID) [248] is a metric used to evaluate the quality of images generated by a model by comparing their statistical properties to those of real images. Unlike the Inception Score, which primarily focuses on the class distributions of generated images, FID measures the distance between the distributions of real and generated images in a feature space, typically using the activations of a specific layer in an Inception network.

To compute FID, generated images (denoted as  $X$ ) and real images (denoted as  $Y$ ) are passed through an Inception network, and activations from a chosen layer are extracted. These activations are represented as  $X'$  and  $Y'$ , respectively. The FID assumes that these activations follow a multivariate Gaussian distribution, characterized by the mean and covariance matrix. Let  $\mu_x$  and  $\mu_y$  denote the means of the activations for  $X'$  and  $Y'$ , and  $\Sigma_x$  and  $\Sigma_y$  their respective covariance matrices. The Fréchet distance between these two Gaussian distributions is defined as:

$$FID(X', Y') = \|\mu_x - \mu_y\|^2 + Tr(\Sigma_x + \Sigma_y - 2(\Sigma_x \Sigma_y)^{1/2}), \quad (110)$$

where:

- $\|\mu_x - \mu_y\|^2$  is the squared Euclidean distance between the means of the two distributions, measuring differences in their central tendency.
- $Tr$  denotes the trace of a matrix, which is the sum of its diagonal elements.
- $(\Sigma_x \Sigma_y)^{1/2}$  is the matrix square root of the product of the covariance matrices, representing the similarity in the spread and orientation of the distributions. This term is well-defined since covariance matrices are positive semi-definite.

The FID has two key advantages. It considers both the average features and the diversity of the generated images by taking into account the mean and covariance of the feature representations. Additionally, FID is less affected by noise compared to the Inception Score, making it a more dependable measure of generative model quality.

However, the FID has limitations. Similar to the Inception Score, FID relies on the Inception network, which is pre-trained on the ImageNet dataset. This reliance can limit its effectiveness when evaluating images from domains significantly different from ImageNet, as the extracted features may not be as relevant. Additionally, FID assumes that the feature distributions are Gaussian, which may not always hold true, potentially affecting the accuracy of the metric.

FID has become a standard metric for evaluating generative models because it effectively captures differences in both quality and diversity between the real and generated images.

## 5 Natural Language Processing

Natural Language Processing (NLP) is a subfield of artificial intelligence that focuses on the interaction between computers and human languages. It involves the development of algorithms and models that allow computers to understand, interpret, and generate human language in a way that is both meaningful and useful. NLP encompasses a wide range of tasks, from basic text processing and understanding to complex language generation and translation.

The main goal of NLP is to bridge the gap between human communication and machine understanding, allowing more natural and efficient interactions with technology. This includes enabling machines to comprehend the subtleties of human language, such as syntax, semantics, context, and sentiment.

The main tasks in NLP are the following:

1. **Text Classification:** Involves categorizing text into predefined classes or categories. Examples include sentiment analysis, spam detection, and topic classification.
2. **Language Modeling:** Predicts the probability of a sequence of words. It is a foundational task in NLP, crucial for applications like text generation and autocomplete.
3. **Machine Translation:** Converts text from one language to another. Popular applications include translating documents, websites, and real-time speech translation.
4. **Named Entity Recognition (NER):** Identifies and classifies proper names, such as people, organizations, and locations, within text.
5. **Part-of-Speech Tagging:** Involves labeling each word in a sentence with its corresponding part of speech, such as noun, verb, adjective, etc.
6. **Sentiment Analysis:** Determines the sentiment expressed in a piece of text, which can be positive, negative, or neutral.
7. **Text Summarization:** Automatically generates a concise summary of a larger body of text. It can be extractive (selecting key sentences) or abstractive (generating new sentences).
8. **Question Answering:** Involves building systems that can automatically answer questions posed in natural language.
9. **Speech Recognition and Synthesis:** Converts spoken language into text (speech recognition) and vice versa (text-to-speech synthesis).

Each of these tasks requires specialized approaches and techniques, including the use of various loss functions and performance metrics to optimize and evaluate models. NLP continues to evolve rapidly, driven by advances in machine learning and the increasing availability of large datasets. The following sections will explore the key loss functions and performance metrics used in these tasks, providing information on their applications and relevance.

## 5.1 Loss Functions used in NLP

In the context of Natural Language Processing (NLP), loss functions play a vital role in guiding models to learn meaningful representations of language, improve predictions, and perform various language-related tasks effectively.

While some loss functions, such as Cross-Entropy Loss and Hinge Loss, are broadly applicable across different domains like classification and computer vision, they are also extensively used in NLP due to the nature of language tasks that often involve classification and sequence prediction. In addition to these, there are specific loss functions tailored to handle the unique challenges presented by NLP tasks, such as sequence generation and language modeling.

In this section, we will explore the various loss functions commonly employed in NLP. We will begin with general-purpose loss functions like Cross-Entropy Loss and Hinge Loss, which are frequently used in text classification and other discriminative tasks. Following that, we will discuss the Negative Log-Likelihood (NLL), a fundamental loss for probabilistic models in language modeling. Finally, we will cover specialized losses designed for sequence generation tasks, which are crucial for applications such as machine translation and text summarization.

Each loss function has distinct characteristics and is chosen based on the specific requirements of the NLP task at hand. By understanding these loss functions and their applications, we can better appreciate how they contribute to the development of robust and accurate NLP models.

### 5.1.1 Cross-Entropy Loss (Token-Level)

Cross-entropy loss is an essential loss function in NLP, especially prominent in tasks where the model predicts a sequence of tokens. This loss function measures the dissimilarity between the predicted probability distribution over the vocabulary and the actual distribution, which is typically represented as a one-hot encoded vector indicating the correct token.

The cross-entropy loss is used in the following NLP tasks:

1. In language modeling [251], the objective is to predict the next word in a sequence given the preceding words. The model outputs a probability distribution over the vocabulary for the next word. The Cross-Entropy Loss penalizes the model when the predicted distribution assigns low probability to the actual next word. This training approach helps the model learn to generate coherent and contextually appropriate text.
2. In machine translation [179], the model converts a sentence from a source language to a target language. During training, the Cross-Entropy Loss is computed for each token in the target sentence, guiding the model to produce translations that closely match the reference translations. The loss function helps align the generated sequence with the correct sequence in the target language.
3. For summarization tasks [252], the model generates a concise summary of a longer text. The token-level Cross-Entropy Loss is used to train the model to generate summaries that are accurate and informative, by minimizing the difference between the predicted summary tokens and the reference summary tokens.
4. Tasks such as part-of-speech [253] tagging and named entity recognition [254] involve labeling each token in a sentence. The Cross-Entropy Loss is applied at the token level, encouraging the model to assign high probabilities to the correct labels for each word in the sequence.

The token-level Cross-Entropy loss for a predicted sequence  $\hat{y} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_T)$  and a target sequence  $y = (y_1, y_2, \dots, y_T)$  is defined as:

$$L_{CE} = -\frac{1}{T} \sum_{t=1}^T \log p(\hat{y}_t = y_t | y_{<t}), \quad (111)$$

where  $T$  is the length of the sequence,  $p(\hat{y}_t = y_t | y_{<t})$  represents the predicted probability of the correct token  $y_t$  at position  $t$ , given the preceding tokens in the target sequence  $y_{<t}$ .

Token-level Cross-Entropy is well suited for tasks that require precise sequence generation, as it directly penalizes deviations from the correct sequence at the token level. However, while it ensures that each token is predicted accurately, it may not fully capture the quality of the overall sequence, especially in cases where the final output is evaluated using metrics that consider global sequence properties.

### 5.1.2 Hinge Loss

In the context of NLP, Hinge loss is employed when the task involves separating data points into two distinct classes, such as in sentiment analysis or binary text classification tasks.

Hinge loss is defined as:

$$L_{\text{hinge}} = \max(0, 1 - yf(x)), \quad (112)$$

where  $y$  represents the true label,  $f(x)$  is the model's predicted score for the input  $x$ , and  $y \in \{-1, +1\}$ . The function  $f(x)$  typically corresponds to the decision function of an SVM, which outputs a score indicating the distance of the sample from the decision boundary. The labels are encoded as  $-1$  or  $+1$  to represent the two classes.

The Hinge Loss penalizes the model when the prediction is incorrect or when the prediction is correct but not confident enough. Specifically, if  $yf(x) \geq 1$ , the loss is zero, indicating that the prediction is both correct and sufficiently confident. However, if  $yf(x) < 1$ , the model incurs a loss proportional to the distance from the correct margin, pushing the model to adjust its parameters to improve classification confidence.

In NLP, Hinge Loss is used in tasks such as:

1. Tasks like spam detection, where the goal is to classify emails or messages as spam or not spam, and sentiment analysis, where the goal is to classify text as having positive or negative sentiment. Hinge Loss helps in defining a clear margin between the two classes, enhancing the robustness of the classification model.
2. Hinge Loss can also be extended to sequence labeling [255, 62] tasks using Structured SVMs, where it helps in learning the correct sequence of labels for inputs such as part-of-speech tagging or chunking. In these cases, the loss is structured to consider the sequence dependencies among predictions.
3. Hinge loss is used in sparse models for NLP that aim to use a relatively small number of features to map inputs to outputs like label sequences or parse trees. These models minimize a regularized empirical risk functional composed of a hinge loss term for goodness of fit and a regularizer term to promote sparsity [256].
4. In discriminatively-trained multiclass linear models in NLP. Compared to log loss, hinge loss with L1 or L2 regularization tends to produce sparser models, although not as sparse as L1-regularized log loss. For models with only indicator features, there is a critical regularizer weight threshold below which L1 and L2 hinge loss produce similar sparsity [257].
5. In adaptive loss functions for document-level relation extraction models. The adaptive hinge balance loss helps these models achieve high performance on both precision and recall [258].
6. Hinge loss is one of the multivariate performance measures used in adversarial prediction frameworks for NLP tasks. These frameworks aim to more closely align model predictions with application performance by treating prediction as an adversarial game between a loss-minimizing model and a loss-maximizing evaluation [259].

Hinge Loss is useful in scenarios where a clear decision boundary is required, and the goal is to maximize the margin between different classes. It is also used in sparse, structured, and adversarial NLP models to promote desirable properties like sparsity, precision-recall balance, and alignment with application-specific metrics. However, it is mainly applied in binary classification settings and is less common for multi-class problems unless adapted through methods like one-vs-all classification.

### 5.1.3 Cosine Similarity Loss

Cosine similarity loss [260] is a metric-based loss function commonly used in NLP for tasks involving the comparison of text embeddings or other vector representations. The objective of Cosine Similarity Loss is to maximize the cosine similarity between vectors representing similar items (e.g., sentences, words, or documents) while minimizing the similarity between vectors representing dissimilar items. This loss is useful in tasks such as semantic textual similarity, paraphrase identification, and information retrieval.

Cosine similarity measures the cosine of the angle between two non-zero vectors in a multi-dimensional space. It ranges from  $-1$  to  $1$ , where  $1$  indicates that the vectors are identical,  $0$  indicates orthogonality (no similarity), and  $-1$  indicates that the vectors are diametrically opposed.

Given two vectors  $\mathbf{u}$  and  $\mathbf{v}$ , the cosine similarity is defined as:

$$\text{cosine\_similarity}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}, \quad (113)$$

where  $\mathbf{u} \cdot \mathbf{v}$  is the dot product of the vectors, and  $\|\mathbf{u}\|$  and  $\|\mathbf{v}\|$  are the magnitudes (norms) of the vectors.

To define a loss function based on cosine similarity, we aim to minimize the cosine similarity for dissimilar pairs and maximize it for similar pairs. One common formulation of Cosine Similarity Loss is:

$$L = 1 - \text{cosine\_similarity}(\mathbf{u}, \mathbf{v}), \quad (114)$$

where  $L$  is the loss, and  $\mathbf{u}$  and  $\mathbf{v}$  are embeddings of similar items. For dissimilar pairs, the loss can be defined to penalize high cosine similarity scores.

Cosine similarity loss is used in the following tasks:

1. Semantic Textual Similarity (STS), where the goal is to determine how similar two pieces of text are in terms of meaning. Cosine Similarity Loss can be used to train models to produce embeddings for sentences or phrases such that semantically similar texts have high cosine similarity scores.
2. Paraphrase Identification, where two sentences convey the same meaning in different words, Cosine Similarity Loss helps in learning representations that bring paraphrases closer together in the embedding space. This loss encourages the model to capture the semantic equivalence between sentences.
3. Information Retrieval, where documents and queries are often represented as vectors. Cosine Similarity Loss is used to ensure that relevant documents have high cosine similarity with the query vector, improving the retrieval of relevant information.
4. Text Clustering, to ensure that similar texts are grouped together. By maximizing the cosine similarity within clusters and minimizing it between clusters, the model can learn to identify meaningful groupings of text data.

The cosine similarity loss has the following limitations.

- A cosine similarity score of zero indicates orthogonality, suggesting no similarity. However, in some cases, it may not be clear whether this is due to actual dissimilarity or a lack of information in the embeddings.
- Cosine Similarity Loss is typically used with dense vector representations, such as word embeddings or sentence embeddings. It may not be directly applicable to sparse representations, which require additional considerations for normalization and similarity computation.

## 5.2 Losses for Sequence Generation

Sequence generation tasks in NLP involve predicting a sequence of words or tokens given an input, such as translating a sentence from one language to another, summarizing a document, or generating a response in a conversation. The challenges associated with sequence generation include maintaining fluency and coherence and ensuring that the sequence generated accurately reflects the intended meaning or information. This section explores the loss functions commonly used to optimize models for these tasks.

### 5.2.1 Minimum Risk Training (MRT)

While token-level losses like Cross-Entropy are effective, they may not always align perfectly with evaluation metrics like BLEU [261] or ROUGE [262], which consider the quality of the entire sequence. Minimum Risk Training (MRT) offers an alternative approach that directly optimizes the evaluation metric of interest, providing a more holistic measure of sequence quality by focusing on the entire sequence rather than individual tokens.

Minimum Risk Training [263] seeks to minimize the expected risk associated with generating sequences. The risk is defined using a loss function  $\Delta(\hat{y}, y)$ , which measures the difference between a predicted sequence  $\hat{y}$  and a target sequence  $y$ , according to a specific evaluation metric. This approach is particularly useful for tasks such as machine translation and text summarization, where the quality of the output sequence is best assessed in its entirety.

The MRT objective is formulated as:

$$L_{MRT} = \sum_{\hat{y} \in \mathcal{Y}} p(\hat{y}|x; \theta) \Delta(\hat{y}, y), \quad (115)$$

where  $\mathcal{Y}$  is the set of all possible output sequences,  $p(\hat{y}|x; \theta)$  represents the model's predicted probability of generating sequence  $\hat{y}$  given the input  $x$ , parameterized by  $\theta$ , and  $\Delta(\hat{y}, y)$  is the loss function evaluating the discrepancy between the predicted sequence  $\hat{y}$  and the target sequence  $y$ .

However, computing this expectation exactly is intractable due to the vast number of possible sequences  $\mathcal{Y}$ . Instead, MRT approximates this by sampling a subset of sequences  $\mathcal{Y}_s$  from the model's distribution. The loss function is then given by:

$$L_{MRT} \approx \sum_{\hat{y} \in \mathcal{Y}_s} \frac{\exp(\alpha \cdot \log p(\hat{y}|x; \theta))}{\sum_{\hat{y}' \in \mathcal{Y}_s} \exp(\alpha \cdot \log p(\hat{y}'|x; \theta))} \Delta(\hat{y}, y), \quad (116)$$

where  $\alpha$  is a scaling factor that controls the sharpness of the distribution over the sampled sequences. The numerator represents the scaled probability of the sequence  $\hat{y}$ , while the denominator normalizes this over the set of sampled sequences  $\mathcal{Y}_s$ . The use of a softmax function ensures that the computed expectation respects the relative probabilities of the sequences.

MRT is effective in tasks where the ultimate goal is to optimize a specific evaluation metric that reflects the quality of the entire generated sequence. Some common applications include machine translation, text summarization, and dialogue systems.

In machine translation, MRT can directly optimize for metrics like BLEU, which evaluates the overlap between the generated translation and a reference translation. By minimizing the expected risk based on BLEU scores, the model learns to produce translations that better match human judgments. In summarization tasks, metrics such as ROUGE are used to measure the similarity between the generated summary and a reference summary. MRT can be utilized to optimize for these metrics, ensuring that generated summaries capture the most critical information from the source text. For generating conversational responses, MRT can be tailored to optimize metrics that capture the relevance and informativeness of the responses, leading to more natural and engaging dialogues.

The primary advantage of MRT is its ability to directly optimize for the metric used to evaluate the model's outputs. This leads to a closer alignment between training objectives and evaluation criteria, which can result in significant improvements in performance, particularly in tasks where sequence-level coherence and relevance are crucial.

However, MRT involves approximating expectations over potentially large sets of sequences, which can be computationally demanding. The choice of the scaling factor  $\alpha$  and the specific loss function  $\Delta(\hat{y}, y)$  also plays a critical role in the effectiveness of MRT, requiring careful tuning to ensure that the model effectively learns from the training data.

### 5.2.2 REINFORCE Algorithm

The REINFORCE [264] algorithm is a fundamental approach in reinforcement learning (RL) and is used in NLP for training models in sequence generation tasks where the evaluation metrics are non-differentiable. It is particularly useful in scenarios where the desired output sequence's quality cannot be directly optimized through standard backpropagation techniques, such as optimizing BLEU scores in machine translation or ROUGE scores in summarization.

REINFORCE is a policy gradient method that optimizes the expected reward by adjusting the model parameters to increase the likelihood of generating high-reward sequences. The core idea is to treat the sequence generation process as a series of actions taken by an agent, where each action corresponds to selecting the next token in the sequence. The agent's policy, which defines the probability distribution over actions, is updated to maximize the expected cumulative reward.

The expected reward  $R$  for a policy  $\pi$  parameterized by  $\theta$  is defined as:

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} [R(\tau)], \quad (117)$$

where  $\tau = (x_1, x_2, \dots, x_T)$  represents a generated sequence (trajectory), and  $R(\tau)$  is the reward associated with that sequence. The goal is to find the parameters  $\theta$  that maximize  $J(\theta)$ .

To optimize  $J(\theta)$ , the gradient with respect to the parameters  $\theta$  is computed using the policy gradient theorem [265]:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} [R(\tau) \nabla_\theta \log \pi_\theta(\tau)], \quad (118)$$

where  $\pi_\theta(\tau)$  is the probability of generating the sequence  $\tau$  under the policy  $\pi_\theta$ . The term  $R(\tau) \nabla_\theta \log \pi_\theta(\tau)$  represents the product of the reward and the gradient of the log-probability of the sequence, which serves as the direction to adjust the policy parameters  $\theta$ .

The REINFORCE algorithm updates the policy parameters  $\theta$  using the following rule:

$$\Delta \theta = \eta R(\tau) \nabla_\theta \log \pi_\theta(\tau), \quad (119)$$

where  $\eta$  is the learning rate. This update rule increases the probability of generating sequences with higher rewards and decreases the probability of generating sequences with lower rewards.



Similar to Minimum Risk Training, REINFORCE is useful in NLP tasks involving sequence generation, where the quality of the output sequence is measured by non-differentiable metrics such as machine translation, text summarization and dialogue systems.

In machine translation, REINFORCE can be used to optimize translation quality directly based on metrics like BLEU, which evaluate the overlap between generated and reference translations. By treating the generation of each word as an action, the algorithm updates the model to produce more accurate and fluent translations. In summarization tasks, metrics such as ROUGE assess the relevance and completeness of the generated summary. REINFORCE optimizes the model to maximize these metrics, ensuring that the generated summaries are informative and concise. For generating conversational responses, REINFORCE can optimize custom metrics that capture the quality of responses, such as relevance, coherence, and engagement. This leads to more natural and contextually appropriate dialogues.

While REINFORCE provides a powerful method for optimizing non-differentiable evaluation metrics, it also poses the following challenges:

- The gradient estimates in REINFORCE can have high variance, making the training process unstable and slow to converge. Techniques such as reward normalization or using a baseline function can help reduce variance and stabilize learning.
- REINFORCE requires a large number of samples to accurately estimate the expected reward and its gradient, which can be computationally expensive, especially for complex tasks with large action spaces.
- In sequence generation tasks, the reward is typically received only after generating the entire sequence, which can make it difficult to assign credit to individual actions (tokens) within the sequence. Strategies like reward shaping or using intermediate rewards can help address this issue.

### 5.3 Performance Metrics used in NLP

Performance metrics provide quantifiable measures to assess how well a model performs a given task. In NLP, these metrics vary widely depending on the nature of the task and the type of output the model produces. For tasks involving classification, metrics like Accuracy, Precision, Recall, F1 Score, and AUC-ROC are crucial for determining the model's ability to correctly classify text into categories. While these metrics are commonly used in general classification problems, their application in NLP involves specific considerations, such as handling imbalanced datasets and the multi-label nature of some tasks.

Metrics like BLEU [261] and ROUGE [262] are essential for sequence generation and text comparison tasks. These metrics evaluate the quality of generated text, such as translations or summaries, by comparing it to reference texts. They provide information on the fluency, adequacy, and informativeness of the generated content.

Other metrics, such as Perplexity [266], are specific to language modeling and measure how well a model predicts a sequence of words. Exact Match is often used in tasks like question answering, where the precision of the model's output is critical.

In this section, we will explore these performance metrics in detail, focusing on their application in NLP tasks. We will discuss their definitions, how they are calculated, and the specific challenges and considerations when applying them to NLP models. This analysis will help in understanding the strengths and limitations of each metric and guide the selection of appropriate metrics for evaluating different NLP applications.

#### 5.3.1 Accuracy

Accuracy is used in classification tasks where the goal is to correctly predict the category or label of a given input. It is defined as the ratio of correctly predicted instances to the total number of instances, providing a straightforward measure of the model's overall correctness.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (120)$$

Accuracy is used in the following NLP Tasks:

1. Text Classification tasks such as sentiment analysis, spam detection, and topic classification. In these tasks, each text input is assigned a discrete label, and accuracy measures the proportion of texts that are correctly classified. For instance, in sentiment analysis, accuracy indicates how often the model correctly identifies a piece of text as positive, negative, or neutral.



2. Sequence labeling tasks like part-of-speech tagging, accuracy measures the percentage of words that are assigned the correct part-of-speech label. This provides an overall sense of how well the model understands the syntactic structure of sentences.
3. In Named Entity Recognition (NER), accuracy can be used to evaluate how correctly the model identifies and categorizes named entities in a text. This includes recognizing entities such as names of people, organizations, locations, etc., and classifying them into predefined categories.
4. Language detection, where the model determines the language in which a given text is written. The metric reflects the proportion of texts that are correctly identified in terms of their language.

While accuracy is a useful and easy-to-understand metric, it has well-known limitations, such as imbalanced data and context sensitivity, which do not account for the quality or meaning of the output.

### 5.3.2 Precision, Recall, and F1 Score

Since we have introduced these metrics before in sections [3.2.2](#), [3.2.2](#) and [3.2.2](#) we will focus on their use for NLP tasks.

In NLP, precision, recall, and F1 Score are crucial for tasks where the identification of specific categories or entities is critical. These metrics are beneficial in scenarios where the dataset is imbalanced or where the costs of false positives and false negatives differ significantly.

Precision is commonly used in NLP applications such as:

1. Named Entity Recognition (NER) where it measures the proportion of correctly identified entities out of all entities predicted by the model. High precision means that when the model predicts an entity, it is likely correct, which is crucial in applications like information extraction.
2. For text classification tasks such as spam detection, precision indicates the proportion of correctly identified spam messages among all messages labeled as spam. High precision is essential in applications where the cost of false positives is high, such as filtering important emails as spam.

Recall is commonly used in NLP applications such as:

- Named Entity Recognition (NER) [\[254\]](#), where it measures the proportion of actual entities that the model correctly identifies. High recall is essential when missing an entity has significant consequences, such as missing important mentions in legal documents.
- In sentiment analysis [\[267\]](#), recall can indicate how well the model identifies all instances of a particular sentiment, such as all positive reviews. High recall ensures that most relevant instances are captured, even if it means tolerating more false positives.

The **F1 score** is the harmonic mean of precision and recall, providing a single metric that balances both concerns. It is handy when there is an uneven class distribution or when false positives and false negatives carry different costs.

F1 score in NLP is used in applications such as:

- Text classification and NER, where it balances the need for both high precision and high recall. For instance, in legal or medical document analysis, both false positives and false negatives can be costly, making the F1 Score a critical metric.
- Information retrieval and question answering systems, where the F1 Score can help evaluate the system's ability to retrieve relevant documents or answer questions correctly, balancing between retrieving relevant results and avoiding irrelevant ones.

### 5.3.3 AUC-ROC

We previously introduced this metric in Section [3.2.2](#). Here, we will focus on its applications for NLP.

In NLP, AUC-ROC is used in tasks where the output is probabilistic, and the focus is on distinguishing between two classes, such as in binary text classification tasks. The AUC-ROC metric measures the trade-off between the true positive rate (TPR) and false positive rate (FPR) across different threshold settings, providing a single scalar value that summarizes the model's performance.

AUC-ROC is widely used in several NLP tasks where binary classification is involved, such as:

1. Spam detection [268], where the goal is to classify emails or messages as spam or not spam. The AUC-ROC metric provides a comprehensive evaluation of the model’s ability to correctly classify spam and non-spam messages across various thresholds. A higher AUC-ROC value indicates better overall performance, especially in differentiating between the two classes.
2. Binary sentiment analysis [267], where the objective is to classify text as expressing positive or negative sentiment, AUC-ROC helps in assessing how well the model distinguishes between positive and negative sentiments. This is particularly useful when the model outputs probabilities for each class, as the AUC-ROC can evaluate the quality of these probabilistic predictions.
3. Binary topic classification [269, 270, 271], where the task is to determine whether a document belongs to a specific topic or not, AUC-ROC measures the model’s accuracy in identifying relevant documents. This metric is essential for applications like content filtering, where false positives can lead to incorrect content recommendations.

### 5.3.4 BLEU Score

The BLEU (Bilingual Evaluation Understudy) [261] score is a widely used metric for evaluating the quality of text generated by NLP models, especially in machine translation. Measures how closely machine-generated text matches a set of reference translations provided by humans. The BLEU score calculates the precision of n-grams—continuous sequences of  $n$  items (typically words)—in the candidate (generated) text concerning the reference text. The metric ranges from 0 to 1, with 1 indicating a perfect match with the reference and 0 indicating no match.

The BLEU score is computed using the following steps:

1. Calculate the precision for n-grams of various lengths (e.g., unigrams, bigrams, trigrams, etc.). Precision is the ratio of the number of n-grams in the candidate text that are also in the reference text to the total number of n-grams in the candidate text.
2. To prevent the candidate text from getting a high score by simply repeating n-grams from the reference, BLEU applies a clipping technique. The count of each n-gram in the candidate is clipped to the maximum number that occurs in any reference translation.
3. Compute the geometric mean of the precision scores for different n-gram lengths.
4. Apply a brevity penalty (BP) to penalize excessively short candidate translations that may have high precision. This penalty is based on the ratio of the length of the candidate translation to the length of the reference translation.

The BLEU score is defined as:

$$BLEU = BP \cdot \exp \left( \sum_{n=1}^N w_n \log p_n \right), \quad (121)$$

where  $BP$  is the brevity penalty,  $w_n$  is the weight for n-gram precision (often uniformly set to  $\frac{1}{N}$ ),  $p_n$  is the precision for n-grams of length  $n$ .

The brevity penalty is calculated as:

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{(1 - \frac{r}{c})}, & \text{if } c \leq r \end{cases} \quad (122)$$

where  $c$  is the length of the candidate translation, and  $r$  is the length of the reference translation.

The BLEU score is predominantly used in the following NLP tasks:

1. Machine Translation [179], where it provides an automatic measure of translation quality by comparing machine-generated translations against human reference translations. A high BLEU score indicates that the generated translation closely matches the reference in terms of word choice and order.
2. Text Summarization [252], to assess the quality of automated summaries by comparing them to human-generated summaries. However, its use in summarization is less common due to its focus on precision over recall and its inability to account for paraphrasing.
3. Text Generation [272, 273], where it helps evaluate how well the generated text aligns with expected responses or content, though it is less commonly used due to the open-ended nature of these tasks.

While the BLEU score is a valuable metric for evaluating text generation quality, it has some limitations, including the following:

- BLEU focuses purely on n-gram overlap and does not account for the semantic meaning of the text. As a result, a sentence that conveys the correct meaning using different words may receive a low BLEU score.
- BLEU score is highly dependent on the quality and number of reference translations. A more diverse set of references can help capture acceptable variations in translation.
- The brevity penalty helps discourage short translations, but the overall emphasis on precision can undervalue longer, accurate translations that capture the complete meaning but use fewer n-grams present in the references.

### 5.3.5 ROUGE Score

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [262] score is a set of metrics used for evaluating automatic summarization and machine translation. It compares the overlap between the candidate text (the machine-generated summary or translation) and reference text(s) (human-generated), focusing on recall-oriented measures. ROUGE is widely used in NLP for tasks where capturing the relevant content and coverage of the reference text is essential.

ROUGE scores are based on the comparison of n-grams, word sequences, and word pairs between the candidate and reference texts. The key variants of ROUGE include ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S.

**ROUGE-N** measures the overlap of n-grams between the candidate and reference text. For example, ROUGE-1 considers unigrams (individual words), while ROUGE-2 considers bigrams (pairs of consecutive words).

$$\text{ROUGE-N} = \frac{\sum_{\text{gram}_n \in \text{Ref}} \min(\text{Count}_{\text{match}}(\text{gram}_n), \text{Count}_{\text{cand}}(\text{gram}_n))}{\sum_{\text{gram}_n \in \text{Ref}} \text{Count}_{\text{Ref}}(\text{gram}_n)} \quad (123)$$

where  $\text{Count}_{\text{match}}(\text{gram}_n)$  is the number of n-grams in both the candidate and reference,  $\text{Count}_{\text{cand}}(\text{gram}_n)$  is the number of n-grams in the candidate, and  $\text{Count}_{\text{Ref}}(\text{gram}_n)$  is the number of n-grams in the reference.

**ROUGE-L** measures the longest common subsequence (LCS) between the candidate and reference. Unlike n-gram matching, ROUGE-L captures sentence-level structure similarity and can identify long-distance dependencies.

$$\text{ROUGE-L} = \frac{\text{LCS}(R, C)}{\text{Length}(R)} \quad (124)$$

where  $\text{LCS}(R, C)$  is the length of the longest common subsequence of the reference (R) and candidate (C) sentences.

**ROUGE-W** is a weighted version of ROUGE-L, which accounts for the importance of consecutive matches by giving higher weight to longer consecutive matches.

**ROUGE-S** measures the overlap of skip-bigrams, which are pairs of words in their sentence order that allow for gaps. This metric captures the relationship between non-consecutive words.

$$\text{ROUGE-S} = \frac{\sum_{\text{skip bigram} \in \text{Ref}} \min(\text{Count}_{\text{match}}(\text{skip bigram}), \text{Count}_{\text{cand}}(\text{skip bigram}))}{\sum_{\text{skip bigram} \in \text{Ref}} \text{Count}_{\text{Ref}}(\text{skip bigram})} \quad (125)$$

ROUGE scores are particularly valuable in evaluating the following NLP tasks:

1. Text summarization, to assess the quality of automatically generated summaries by comparing them to human-written summaries. It measures how well the summary captures the important information from the original text.
2. In machine translation [274], ROUGE can provide insights, particularly in assessing the recall of content from the reference translations.
3. In broader text generation tasks [275], ROUGE can help evaluate the informativeness and completeness of the generated text compared to reference texts, such as in question answering or dialogue systems.

ROUGE has the following considerations and limitations:

- Like BLEU, ROUGE is sensitive to exact wording and may not fully capture the quality of paraphrased or synonymously expressed content. This limitation means that high-quality variations in language might be undervalued.

- The quality of ROUGE scores can depend significantly on the number and diversity of reference texts. More diverse references can better capture the range of valid outputs, providing a fairer assessment of the candidate text.
- ROUGE, being n-gram-based, primarily measures surface-level similarity and may not adequately assess the deeper semantic meaning or factual correctness of the text.

In recap, ROUGE scores are an essential metric for evaluating summarization and other NLP tasks where capturing the breadth of reference content is critical. While they provide a useful measure of content overlap, they should often be complemented with other metrics or human judgment to obtain a comprehensive evaluation of text quality.

### 5.3.6 Perplexity

Perplexity is a commonly used metric for evaluating language models in NLP. It measures how well a probabilistic model predicts a sample and is particularly useful for models that generate or understand natural language text, such as those used in language modeling and text generation tasks. Perplexity assesses a language model’s uncertainty by evaluating how surprised it is by the actual outcome. Perplexity was initially introduced in 1977 in the context of speech recognition by Frederick Jelinek et al. [266] but nowadays is widely used to evaluate large language models [276] [277] [278].

Perplexity is defined as the exponentiation of the entropy of the model’s predictions. For a given sequence of words, a language model assigns probabilities to each word based on the preceding words. Perplexity essentially measures the geometric mean of the inverse probabilities of the words in the sequence, which corresponds to the average branching factor when predicting the next word.

Mathematically, for a language model predicting a sequence of words  $w_1, w_2, \dots, w_T$ , the perplexity is defined as:

$$\text{Perplexity}(W) = \exp \left( -\frac{1}{T} \sum_{t=1}^T \log p(w_t | w_{1:t-1}) \right), \quad (126)$$

where:

- $W = w_1, w_2, \dots, w_T$  is the sequence of words,
- $p(w_t | w_{1:t-1})$  is the conditional probability of the word  $w_t$  given the preceding words  $w_{1:t-1}$ ,
- $T$  is the total number of words in the sequence.

Lower perplexity indicates that the language model is better at predicting the sequence of words, implying greater certainty and accuracy in its predictions.

Perplexity is widely used in evaluating the performance of language models across various NLP tasks such as:

1. Language Modeling as a standard metric for evaluating language models, such as n-gram models [279] [280], recurrent neural networks (RNNs) [281], and transformers [276] [277]. It provides a direct measure of how well the model predicts a held-out test set, with lower perplexity indicating better predictive performance.
2. In tasks like text generation, chatbots, and dialogue systems, perplexity helps assess the fluency and coherence of generated text. Lower perplexity suggests that the model can generate more grammatically and contextually appropriate text.
3. Speech Recognition and Machine Translation, where perplexity is also used in assessing the language models embedded within speech recognition and machine translation systems. It helps evaluate the system’s ability to generate plausible sequences of words or phrases based on the input.

Although perplexity is a valuable metric, it has some limitations:

- It is sensitive to the choice of the test corpus. Different corpora can have varying levels of complexity, affecting the perplexity score. Therefore, comparisons of perplexity should be made on the same dataset or corpora with similar characteristics.
- It can be influenced by the size of the model’s vocabulary. Larger vocabularies can increase the range of possible predictions, potentially leading to higher perplexity scores. Normalization techniques or adjustments may be needed to account for differences in vocabulary size.

- It assumes that the probabilities output by the model are well-calibrated. If the model is overconfident in its predictions, perplexity may not accurately reflect the model’s performance.

Although it has its limitations, perplexity is still an important metric for evaluating language models. It helps us to understand the model’s predictive abilities and the quality of the text it generates. Perplexity is a crucial measure for determining how well a model grasps the underlying structure of a language, and it is widely used for developing and evaluating NLP systems.

### 5.3.7 Exact Match (EM)

The Exact Match (EM) [282] is a straightforward and widely used evaluation metric in NLP, particularly in tasks where the correctness of the entire predicted output is crucial. Measures the proportion of predictions that match the reference outputs exactly, without any differences in words, order, or formatting. EM is a stringent metric that provides a clear indication of how often a model’s predictions are entirely correct.

Exact Match evaluates a model’s accuracy by determining whether the predicted output matches the reference output exactly, without any deviations. It is a binary metric that assigns a score of 1 if the prediction matches the reference and 0 otherwise. The EM score is then calculated as the percentage of predictions that are exactly correct.

Mathematically, the EM score can be defined as:

$$\text{Exact Match} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(y_i = \hat{y}_i), \quad (127)$$

where:

- $N$  is the total number of examples,
- $y_i$  is the reference output for the  $i$ -th example,
- $\hat{y}_i$  is the predicted output for the  $i$ -th example,
- $\mathbb{I}(\cdot)$  is an indicator function that returns 1 if the condition inside is true and 0 otherwise.

The Exact Match metric is used in the following NLP tasks:

1. In extractive question answering (QA), the task is to extract the exact answer span from a given context. EM measures the percentage of questions for which the model’s predicted answer exactly matches the correct answer. It is a critical metric in QA, as even a slight deviation from the reference answer (such as missing a word or a different word order) results in a score of 0.
2. In Natural Language Inference (NLI) tasks, the goal is to determine the relationship between a premise and a hypothesis (e.g., entailment, contradiction, or neutral). EM can be used to evaluate whether the model’s predicted label matches the correct label exactly.
3. While less commonly used in machine translation and summarization than BLEU or ROUGE, EM can still be relevant in contexts where exact wording is critical, such as translating legal documents or generating summaries that must adhere strictly to a specific format.
4. In applications like code generation or formula generation, where the output must follow strict syntactical and semantic rules, EM provides a measure of the model’s ability to produce correct and precise outputs.

Exact match has the following limitations:

- It is a very strict metric, as it requires complete agreement between the predicted and reference outputs. This stringency makes it a strong indicator of perfect correctness but can also lead to low scores if the task allows for multiple valid answers or paraphrasing.
- It does not provide partial credit for predictions that are mostly correct but differ slightly from the reference. This limitation can be significant in tasks like open-domain QA or summarization, where multiple correct answers or phrasings might exist.
- EM focuses purely on the surface form of the text and does not account for semantic similarity. As a result, it may not be suitable for tasks where capturing the meaning or intent is more important than the exact wording.

## 6 Discussion

Throughout this paper, we have reviewed a variety of loss functions and metrics utilized in deep learning, from general tasks, including regression, classification, to computer vision tasks such as image classification, object detection, image segmentation, and face recognition. We end our review with an exploration of loss functions and performance metrics in the field of Natural Language Processing. Our review highlights the importance of selecting an appropriate loss function and evaluation metric, depending on the specific task at hand and the characteristics of the data.

For regression tasks, for instance, Mean Squared Error (MSE) and Mean Absolute Error (MAE) are widely used due to their simplicity and interpretability. However, there are more robust alternatives such as Huber loss or application-specific such as quantile or Poisson loss. Additionally, while RMSE and MAE are commonly used for evaluation, they may need to adequately capture the performance of models on all types of data, leading to the use of additional metrics such as  $R^2$  and Adjusted  $R^2$ .

In classification, it is noted that while binary cross-entropy loss is standard for binary classification tasks, options such as focal loss and weighted binary cross-entropy can provide robustness in situations where classes are imbalanced. Also, more than accuracy is required to provide a complete picture of a model's performance, especially in imbalanced datasets. This requires the use of additional metrics such as precision, recall, F1 score, and AUC-ROC.

The complexity increases when we consider challenging tasks in computer vision and natural language processing. Here, loss functions are about more than just calculating the difference between the predicted and true values. For instance, the YOLO loss in object detection, Kullback-Leibler, or the Triplet loss, consider other aspects such as location, probability distributions, and multiple instances or entities.

Metrics such as Average Precision (AP) and Average Recall (AR) for object detection, panoptic Quality (PQ) for Panoptic segmentation, and BLEU score for machine translation go beyond typical accuracy-based metrics to evaluate the quality of instance identification, segmentation, and accurate translations.

As discussed, each loss function and metric has pros and cons, and their appropriateness may depend on the specific application or the characteristics of the data set. Understanding these trade-offs is critical for the design of effective deep learning systems.

Looking ahead, there are opportunities for developing new loss functions and metrics that are more robust to data anomalies, consider specific practical constraints, or are tailored to new and emerging deep learning tasks. We also see potential in automated methods that intelligently select or combine different loss functions and metrics based on the given task and data.

## 7 Conclusion

The choice of the loss function and metric can profoundly influence the performance of a model; hence understanding their nature and implications is helpful for anyone working in deep learning.

From regression to complex tasks such as object detection, face recognition, generative models, and natural language processing, we have highlighted the importance of using appropriate loss functions and evaluation metrics. Furthermore, considering the characteristics of the dataset, specifically class imbalances and outliers, is critical when designing and evaluating the models. There is no one-size-fits-all loss function or metric for every task. This stresses the continued necessity for researchers to develop task-specific or adaptable loss functions and metrics and further refine the performance and applicability of deep learning models.

A promising direction for future work is the exploration of automated methods to intelligently select or combine loss functions and metrics, thereby reducing the manual effort and potential bias involved in their selection. In addition, developing robust loss functions using artificial intelligence that can effectively handle data anomalies and practical constraints is a promising area for exploration.

Accurate modeling and prediction of complex phenomena are becoming increasingly critical in our rapidly evolving digital world. By improving our understanding and usage of loss functions and metrics in deep learning, we can make significant contributions to the advancement of this technology. This sets the ground for the development of more sophisticated, effective and reliable models in the future.

## 8 Acknowledgments

We thank the National Council for Science and Technology (CONACYT) for its support through the National Research System (SNI).

## Declaration of generative AI and AI-assisted technologies in the writing process

We acknowledge the use of three AI tools: Grammarly Assistant to improve the grammar, clarity, and overall readability of the manuscript, Perplexity for finding relevant academic works, and GPT-4o to help with the wording and proofreading of the manuscript.

## References

- [1] J. Smith, “Deep learning for image recognition,” *Journal of Artificial Intelligence*, vol. 15, no. 2, pp. 78–95, 2018.
- [2] E. Jones, “Object detection using neural networks,” *Computer Vision Review*, vol. 8, no. 4, pp. 112–128, 2020.
- [3] M. Gonzalez, “Image segmentation using neural networks,” *Pattern Recognition*, vol. 22, no. 1, pp. 36–52, 2017.
- [4] L. Wang, “Facial recognition using neural networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1100–1112, 2019.
- [5] W. Chen, “Image generation using adversarial neural networks,” *ACM Transactions on Graphics*.
- [6] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.
- [7] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising,” *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.
- [8] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” *International Conference on Learning Representations (ICLR)*, 2018.
- [9] L. Chen, “Automatic speech recognition with neural networks,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1200–1212, 2019.
- [10] W. Liu, “Speech emotion recognition with neural networks,” *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 500–512, 2020.
- [11] M. Kim, “Multilingual speech recognition with neural networks,” *Speech Communication*, vol. 101, pp. 50–63, 2018.
- [12] P. Zhang, “Robust speech recognition with neural networks,” *Computer Speech and Language*, vol. 65, pp. 101–120, 2021.
- [13] H. Wu, “End-to-end speech recognition with neural networks,” *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1631–1635, 2017.
- [14] J. Smith, “Sentiment analysis using natural language processing,” *Journal of Artificial Intelligence*, vol. 15, no. 2, pp. 78–95, 2018.
- [15] E. Jones, “Named entity recognition using natural language processing,” *Computational Linguistics*, vol. 46, no. 4, pp. 112–128, 2020.
- [16] M. Gonzalez, “Text summarization using natural language processing,” *ACM Transactions on Information Systems*, vol. 35, no. 2, pp. 36–52, 2017.
- [17] L. Wang, “Part-of-speech tagging using natural language processing,” *Journal of Machine Learning Research*, vol. 20, no. 7, pp. 1100–1112, 2019.
- [18] W. Chen, “Question answering using natural language processing,” *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [19] D. Harrison Jr and D. L. Rubinfeld, “Hedonic housing prices and the demand for clean air,” *Journal of environmental economics and management*, vol. 5, no. 1, pp. 81–102, 1978.
- [20] H.-x. Zhao and F. Magoulès, “A review on the prediction of building energy consumption,” *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6, pp. 3586–3592, 2012.



- [21] F. E. Harrell Jr, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati, "Regression modelling strategies for improved prognostic prediction," *Statistics in medicine*, vol. 3, no. 2, pp. 143–152, 1984.
- [22] S. Shen, H. Jiang, and T. Zhang, "Stock market forecasting using machine learning algorithms," *Department of Electrical Engineering, Stanford University, Stanford, CA*, pp. 1–5, 2012.
- [23] E. C. Malthouse and R. C. Blattberg, "Can we predict customer lifetime value?," *Journal of interactive marketing*, vol. 19, no. 1, pp. 2–16, 2005.
- [24] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 2006.
- [25] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.
- [26] A. Auslender and M. Teboulle, "Interior gradient and epsilon-subgradient descent methods for constrained convex minimization," *Mathematics of Operations Research*, vol. 29, no. 1, pp. 1–26, 2004.
- [27] Y. Bai, Q. Jiang, and J. Sun, "Subgradient descent learns orthogonal dictionaries," *arXiv preprint arXiv:1810.10702*, 2018.
- [28] P. Bianchi, W. Hachem, and S. Schechtman, "Stochastic subgradient descent escapes active strict saddles on weakly convex functions," *arXiv preprint arXiv:2108.02072*, 2021.
- [29] L. Xiao, "Dual averaging method for regularized stochastic learning and online optimization," *Advances in Neural Information Processing Systems*, vol. 22, 2009.
- [30] P. J. Huber, "Robust estimation of a location parameter," *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518, 1992.
- [31] R. A. Saleh, A. Saleh, *et al.*, "Statistical properties of the log-cosh loss function used in machine learning," *arXiv preprint arXiv:2208.04564*, 2022.
- [32] R. Koenker and K. F. Hallock, "Quantile regression," *Journal of economic perspectives*, vol. 15, no. 4, pp. 143–156, 2001.
- [33] M.-Y. Chen and J.-E. Chen, "Application of quantile regression to estimation of value at risk," *Review of Financial Risk Management*, vol. 1, no. 2, p. 15, 2002.
- [34] J. Bruzda, "Multistep quantile forecasts for supply chain and logistics operations: bootstrapping, the garch model and quantile regression based approaches," *Central European Journal of Operations Research*, vol. 28, no. 1, pp. 309–336, 2020.
- [35] J. B. Bremnes, "Probabilistic wind power forecasts using local quantile regression," *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, vol. 7, no. 1, pp. 47–54, 2004.
- [36] W. P. Gaglianone and L. R. Lima, "Constructing density forecasts from quantile regressions," *Journal of Money, Credit and Banking*, vol. 44, no. 8, pp. 1589–1607, 2012.
- [37] L. Massidda and M. Marrocu, "Quantile regression post-processing of weather forecast for short-term solar power probabilistic forecasting," *Energies*, vol. 11, no. 7, p. 1763, 2018.
- [38] A. Zarnani, S. Karimi, and P. Musilek, "Quantile regression and clustering models of prediction intervals for weather forecasts: A comparative study," *Forecasting*, vol. 1, no. 1, pp. 169–188, 2019.
- [39] J. Zietz, E. N. Zietz, and G. S. Sirmans, "Determinants of house prices: a quantile regression approach," *The Journal of Real Estate Finance and Economics*, vol. 37, pp. 317–333, 2008.
- [40] R. Winkelmann, "Reforming health care: Evidence from quantile regressions for counts," *Journal of Health Economics*, vol. 25, no. 1, pp. 131–145, 2006.
- [41] A. C. Cameron and P. K. Trivedi, *Regression analysis of count data*, vol. 53. Cambridge university press, 2013.
- [42] H. Yang, K. Ozbay, and B. Bartin, "Effects of open road tolling on safety performance of freeway mainline toll plazas," *Transportation research record*, vol. 2324, no. 1, pp. 101–109, 2012.
- [43] J. Viel, "Poisson regression in epidemiology," *Revue D'epidemiologie et de Sante Publique*, vol. 42, no. 1, pp. 79–87, 1994.
- [44] Y. Mouatassim and E. H. Ezzahid, "Poisson regression and zero-inflated poisson regression: application to private health insurance data," *European actuarial journal*, vol. 2, no. 2, pp. 187–204, 2012.
- [45] H. Shen and J. Z. Huang, "Forecasting time series of inhomogeneous poisson processes with application to call center workforce management," *The Annals of Applied Statistics*, vol. 2, no. 2, pp. 601–623, 2008.

- [46] P. Avila Clemenshia and M. Vijaya, "Click through rate prediction for display advertisement," *International Journal of Computer Applications*, vol. 975, p. 8887, 2016.
- [47] D. Lambert, "Zero-inflated poisson regression, with an application to defects in manufacturing," *Technometrics*, vol. 34, no. 1, pp. 1–14, 1992.
- [48] D. W. Osgood, "Poisson-based regression analysis of aggregate crime rates," *Journal of quantitative criminology*, vol. 16, pp. 21–43, 2000.
- [49] P. Goodwin and R. Lawton, "On the asymmetry of the symmetric mape," *International journal of forecasting*, vol. 15, no. 4, pp. 405–408, 1999.
- [50] X. Tian, H. Wang, and E. Erjiang, "Forecasting intermittent demand for inventory management by retailers: A new approach," *Journal of Retailing and Consumer Services*, vol. 62, p. 102662, 2021.
- [51] D. Ramos, P. Faria, Z. Vale, and R. Correia, "Short time electricity consumption forecast in an industry facility," *IEEE Transactions on Industry Applications*, vol. 58, no. 1, pp. 123–130, 2021.
- [52] S. Kumar Chandar, "Grey wolf optimization-elman neural network model for stock price prediction," *Soft Computing*, vol. 25, pp. 649–658, 2021.
- [53] M. Lawrence, M. O'Connor, and B. Edmundson, "A field study of sales forecasting accuracy and processes," *European Journal of Operational Research*, vol. 122, no. 1, pp. 151–160, 2000.
- [54] K.-F. Chu, A. Y. Lam, and V. O. Li, "Deep multi-scale convolutional lstm network for travel demand and origin-destination predictions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3219–3232, 2019.
- [55] N. J. Nagelkerke *et al.*, "A note on a general definition of the coefficient of determination," *Biometrika*, vol. 78, no. 3, pp. 691–692, 1991.
- [56] J. Miles, "R-squared, adjusted r-squared," *Encyclopedia of statistics in behavioral science*, 2005.
- [57] R. Goodman, J. Miller, and P. Smyth, "Objective functions for neural network classifier design," in *Proceedings. 1991 IEEE International Symposium on Information Theory*, pp. 87–87, 1991.
- [58] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.
- [59] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?," *Advances in neural information processing systems*, vol. 32, 2019.
- [60] Z. Leng, M. Tan, C. Liu, E. D. Cubuk, X. Shi, S. Cheng, and D. Anguelov, "Polyloss: A polynomial expansion perspective of classification loss functions," *arXiv preprint arXiv:2204.12511*, 2022.
- [61] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri, "Are loss functions all the same?," *Neural computation*, vol. 16, no. 5, pp. 1063–1076, 2004.
- [62] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, and Y. Singer, "Large margin methods for structured and interdependent output variables.," *Journal of machine learning research*, vol. 6, no. 9, 2005.
- [63] J. R. Taylor, *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books, 2 sub ed., 1996.
- [64] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.
- [65] R. Parikh, A. Mathai, S. Parikh, G. C. Sekhar, and R. Thomas, "Understanding and using sensitivity, specificity and predictive values," *Indian journal of ophthalmology*, vol. 56, no. 1, p. 45, 2008.
- [66] "Image classification." <https://paperswithcode.com/task/image-classification>. Accessed: 2024-04-25.
- [67] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
- [68] N. Varshney, N. Shalini, M. Sharma, V. K. Yadav, D. V. Saravanan, and N. Kumar, "Resnet transfer learning for enhanced medical image classification in healthcare," *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*, vol. 1, pp. 1–7, 2023.
- [69] V. S. A. Kagolanu, L. Thimmareddy, K. L. Kanala, and B. Sirisha, "Multi-class medical image classification based on feature ensembling using deepnets," *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 540–544, 2022.

- [70] L. Saleh and L. Zhang, "Medical image classification using transfer learning and network pruning algorithms," *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1932–1938, 2023.
- [71] G. Harika, K. Keerthi, D. H. Kommineni, and K. Soumya, "Classification of cervical cancer using resnet-50," *2023 Global Conference on Information Technologies and Communications (GCITC)*, pp. 1–8, 2023.
- [72] S. A. Shah, G. M. Lakho, H. A. Keerio, M. N. Sattar, G. Hussain, M. Mehdi, R. B. Vistro, E. A. Mahmoud, and H. O. Elansary, "Application of drone surveillance for advance agriculture monitoring by android application using convolution neural network," *Agronomy*, 2023.
- [73] Y. Yuan, "Computer vision and deep learning for precise agriculture: A case study of lemon leaf image classification," *Journal of Physics: Conference Series*, vol. 2547, 2023.
- [74] A. D. Nidhis, C. N. V. Pardhu, K. C. Reddy, and K. Deepa, "Cluster based paddy leaf disease detection, classification and diagnosis in crop health monitoring unit," *Computer Aided Intervention and Diagnostics in Clinical and Medical Images*, 2019.
- [75] A. Tendolkar, A. Choraria, M. M. M. Pai, S. Girisha, G. Dsouza, and K. S. Adithya, "Modified crop health monitoring and pesticide spraying system using ndvi and semantic segmentation: An agrocopter based approach," *2021 IEEE International Conference on Autonomous Systems (ICAS)*, pp. 1–5, 2021.
- [76] J. Peng, C. Xiao, and Y. Li, "Rp2k: A large-scale retail product dataset for fine-grained image classification," 2021.
- [77] M. I. H. Shihab, N. Tasnim, H. Zunair, L. K. Rupy, and N. Mohammed, "Vista: Vision transformer enhanced by u-net and image colorfulness frame filtration for automatic retail checkout," 2022.
- [78] A. Boriya, S. S. Malla, R. Manjunath, V. Velicheti, and M. Eirinaki, "Viser: A visual search engine for e-retail," *2019 First International Conference on Transdisciplinary AI (TransAI)*, pp. 76–83, 2019.
- [79] M. Alghamdi, H. A. Mengash, M. Aljebreen, M. Maray, A. A. Darem, and A. S. Salama, "Empowering retail through advanced consumer product recognition using aquila optimization algorithm with deep learning," *IEEE Access*, vol. 12, pp. 71055–71065, 2024.
- [80] C. S. Pillai, "Real time object detection using convolution neural network based classification in video surveillance systems," 2019.
- [81] T. M and J. Singh, "Unusual crowd activity detection in video using cnn, lstm and opencv," *International Journal for Research in Applied Science and Engineering Technology*, 2023.
- [82] R. Ayad and F. Q. Al-Khalidi, "Convolutional neural network (cnn) model to mobile remote surveillance system for home security," *International Journal of Computing and Digital Systems*, 2023.
- [83] S. Ramadasan, K. Vijayakumar, S. Prabha, and E. Karthickeien, "Forest region extraction and evaluation from satellite images using cnn segmentation," *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, pp. 1–5, 2024.
- [84] M. Ahmed, R. Mumtaz, Z. Anwar, A. Shaukat, O. Arif, and F. Shafait, "A multi-step approach for optically active and inactive water quality parameter estimation using deep learning and remote sensing," *Water*, 2022.
- [85] F. Farahnakian, L. Zelioli, M. Middleton, I. Seppä, T. P. Pitkänen, and J. Heikkonen, "Cnn-based boreal peatland fertility classification from sentinel-1 and sentinel-2 imagery," *2023 IEEE International Symposium on Robotic and Sensors Environments (ROSE)*, pp. 1–7, 2023.
- [86] M. Ilteralp, S. Arıman, and E. Aptoula, "A deep multitask semisupervised learning approach for chlorophyll-a retrieval from remote sensing images," *Remote. Sens.*, vol. 14, p. 18, 2021.
- [87] A. Abdullah, M. Jawahar, N. Manogaran, G. Subbiah, K. Seeranagan, B. Balusamy, and A. C. Saravanan, "Leather image quality classification and defect detection system using mask region-based convolution neural network model," *International Journal of Advanced Computer Science and Applications*, 2024.
- [88] A.-M. A. Mamun, M. R. Hossain, and M. M. Sharmin, "Detection and classification of metal surface defects using lite convolutional neural network (lcnn)," *Material Science & Engineering International Journal*, 2024.
- [89] K. A. Pranoto, W. Caesarendra, I. Petra, G. M. Królczyk, M. D. Surindra, and P. W. Yoyo, "Burrs and sharp edge detection of metal workpiece using cnn image classification method for intelligent manufacturing application," *2023 IEEE 21st International Conference on Industrial Informatics (INDIN)*, pp. 1–7, 2023.
- [90] Y. Yang, L. Pan, J. Ma, R. Yang, Y. Zhu, Y. Yang, and L. Zhang, "A high-performance deep learning algorithm for the automated optical inspection of laser welding," *Applied Sciences*, 2020.

- [91] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and M. Zieba, “End to end learning for self-driving cars,” *arXiv preprint arXiv:1604.07316*, 2016.
- [92] F. Codevilla, M. Müller, A. López, and V. Koltun, “End-to-end driving via conditional imitation learning,” *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–9, 2018.
- [93] A. Sadeghian, V. Kosaraju, P. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, “Sophie: An attentive gnn for predicting driving paths from sensor data,” *IEEE International Conference on Computer Vision (ICCV)*, pp. 9267–9276, 2019.
- [94] Z. Chen, Y. Liu, X. Li, X. Yang, and Q. Zhang, “Reinforcement learning-based motion planning for autonomous driving in dynamic environments,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 2, pp. 736–747, 2020.
- [95] R. Chalapathy and A. Menon, “Anomaly detection using deep one-class classification,” *arXiv preprint arXiv:1802.06360*, 2017.
- [96] Y. Li, M. Zhang, Z. Chen, and Q. Yang, “Human activity recognition using recurrent neural networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 766–779, 2018.
- [97] C. Zhang, H. Zhang, L. Zhang, and Q. Li, “Crowd analysis using deep learning: A survey,” *IEEE Access*, vol. 4, pp. 212–228, 2016.
- [98] E. Y. Kim, E. Park, and J. H. Kim, “Emotion recognition based on physiological signals for human-computer interaction,” *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 457–470, 2018.
- [99] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, “Deep learning-based electroencephalography analysis: A comprehensive review,” *Journal of Neural Engineering*, vol. 15, no. 1, p. 011001, 2018.
- [100] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” *Advances in Neural Information Processing Systems (NIPS)*, pp. 3104–3112, 2014.
- [101] S. Li and W. Deng, “Deep facial expression recognition: A survey,” *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 321–341, 2017.
- [102] S. Levine, P. Pastor, A. Krizhevsky, and J. Ibarz, “Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection,” *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 421–436, 2018.
- [103] T. Zhang, Z. McCarthy, Y. Yang, E. Schmerling, and C. J. Tomlin, “Deep reinforcement learning for robot motion planning in dynamic environments,” *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3713–3720, 2019.
- [104] M. Gualtieri and A. Singh, “Pick-and-place using deep reinforcement learning,” *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8, 2018.
- [105] A. Hussein, H. Rad, and C. R. Smith, “Imitation learning for robot manipulation using deep neural networks,” *arXiv preprint arXiv:1703.08612*, 2017.
- [106] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.
- [107] “torch.nn.smoothl1loss.” <https://pytorch.org/docs/stable/generated/torch.nn.SmoothL1Loss.html>. Accessed: 2024-05-01.
- [108] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *Advances in neural information processing systems*, vol. 28, 2015.
- [109] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, “Libra r-cnn: Towards balanced learning for object detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 821–830, 2019.
- [110] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 21–37, Springer, 2016.
- [111] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.
- [112] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658–666, 2019.

- [113] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, “Distance-iou loss: Faster and better learning for bounding box regression,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12993–13000, Apr. 2020.
- [114] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.
- [115] M. S. Hossain, J. M. Betts, and A. P. Paplinski, “Dual focal loss to address class imbalance in semantic segmentation,” *Neurocomputing*, vol. 462, pp. 69–87, 2021.
- [116] T. He, Y. Liu, C. Shen, X. Wang, and C. Sun, “Instance-aware embedding for point cloud instance segmentation,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 255–270, Springer, 2020.
- [117] Z. Luo, Z. Wang, Y. Huang, L. Wang, T. Tan, and E. Zhou, “Rethinking the heatmap regression for bottom-up human pose estimation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13264–13273, 2021.
- [118] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [119] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*, pp. 740–755, Springer, 2014.
- [120] R. Padilla, W. L. Passos, T. L. Dias, S. L. Netto, and E. A. Da Silva, “A comparative analysis of object detection metrics with a companion open-source toolkit,” *Electronics*, vol. 10, no. 3, p. 279, 2021.
- [121] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.
- [122] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
- [123] L.-C. Chen, Y. Zhu, G. Papandreou, and et al., “Encoder-decoder with atrous separable convolution for semantic image segmentation,” *European Conference on Computer Vision (ECCV)*, pp. 801–818, 2018.
- [124] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 4, pp. 834–848, 2017.
- [125] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.
- [126] O. Oktay, J. Schlemper, L. L. Folgoc, and et al., “Attention u-net: Learning where to look for the pancreas,” *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 138–146, 2018.
- [127] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *European Conference on Computer Vision (ECCV)*, pp. 346–361, 2014.
- [128] S. Zheng, S. Jayasumana, B. Romera-Paredes, and et al., “Conditional random fields as recurrent neural networks,” *International Conference on Computer Vision (ICCV)*, pp. 1529–1537, 2015.
- [129] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.
- [130] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network for instance segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8759–8768, 2018.
- [131] A. Kirillov, K. He, R. Girshick, and P. Dollár, “Panoptic feature pyramid networks,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6399–6408, 2019.
- [132] A. Kirillov, Y. Wu, K. He, and R. Girshick, “Pointrend: Image segmentation as rendering,” *European Conference on Computer Vision (ECCV)*, pp. 405–421, 2020.
- [133] X. Wang, T. Kong, C. Shen, X. You, and L. Li, “Solo: Segmenting objects by locations,” *IEEE International Conference on Computer Vision (ICCV)*, pp. 6319–6328, 2019.
- [134] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, “Panoptic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9404–9413, 2019.
- [135] Y. Xiong, H. Zhu, D. Lin, and et al., “Upsnet: A unified panoptic segmentation network,” *European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

- [136] L. Liu, X. Wang, C. Shen, X. You, and L. Li, “Panoptic feature pyramid networks,” *IEEE International Conference on Computer Vision (ICCV)*, pp. 2704–2713, 2017.
- [137] B. Cheng, M. D. Zhang, Z. Zhang, and et al., “Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation,” *European Conference on Computer Vision (ECCV)*, pp. 53–69, 2020.
- [138] P. Wang, Y. Chen, I. Stanton, and et al., “Panoptic segmentation transformer,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4077–4086, 2021.
- [139] L.-J. Li, R. Socher, and L. Fei-Fei, “Towards total scene understanding: Classification, annotation and segmentation in an automatic framework,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2036–2043, IEEE, 2009.
- [140] B.-S. Hua, Q.-H. Pham, D. T. Nguyen, and et al., “Scenenet: Understanding real world indoor scenes with synthetic data,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3419–3428, 2016.
- [141] B. Zhou, H. Zhao, X. Puig, and et al., “Scene parsing through ade20k dataset,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641, 2017.
- [142] S. Song, S. Lichtenberg, and J. Xiao, “Sunrgb-d: A rgb-d scene understanding benchmark suite,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 567–576, 2015.
- [143] F. Yu, Y. Zhang, S. Song, and et al., “Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop,” *arXiv preprint arXiv:1506.03365*, 2015.
- [144] D. L. Pham, C. Xu, and J. L. Prince, “Current methods in medical image segmentation,” *Annual review of biomedical engineering*, vol. 2, no. 1, pp. 315–337, 2000.
- [145] K. Yan, X. Wang, L. Lu, and et al., “Deeplesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8262–8271, 2018.
- [146] Y. Liu, K. Gadepalli, M. Norouzi, and et al., “Densely connected convolutional networks for medical image segmentation,” *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 232–239, 2017.
- [147] M. Havaei, A. Davy, D. Warde-Farley, and et al., “Brain tumor segmentation with deep neural networks,” *Medical Image Analysis*, vol. 35, pp. 18–31, 2017.
- [148] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” *3D Vision (3DV)*, pp. 565–571, 2016.
- [149] K. Kamnitsas, C. Ledig, V. F. Newcombe, and et al., “Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation,” *Medical Image Analysis*, vol. 36, pp. 61–78, 2017.
- [150] H. Chen, X. Qi, Q. Dou, and et al., “Med3d: Transfer learning for 3d medical image analysis,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 12, pp. 2804–2813, 2019.
- [151] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, “Cross-view semantic segmentation for sensing surroundings,” *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4867–4873, 2020.
- [152] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 652–660, 2017.
- [153] A. Toshev and C. Szegedy, “DeepPose: Human pose estimation via deep neural networks,” *European Conference on Computer Vision (ECCV)*, pp. 492–505, 2014.
- [154] S. Jégou, M. Drozdal, D. Vazquez, and et al., “The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation,” *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1175–1183, 2017.
- [155] F. N. Iandola, S. Han, M. W. Moskewicz, and et al., “Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 685–694, 2016.
- [156] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, “Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5108–5115, IEEE, 2017.
- [157] V. Mnih, K. Kavukcuoglu, D. Silver, and et al., “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [158] X. Chen, A. Liu, and M. Liu, “Deep sensor fusion for autonomous vehicles: A comprehensive review,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 1, pp. 258–268, 2020.

- [159] Y. Gal and Z. Ghahramani, “Uncertainty in deep learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, no. 11, pp. 1892–1908, 2016.
- [160] A. Alahi, V. Goel, V. Ramanathan, and et al., “Social lstm: Human trajectory prediction in crowded spaces,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–971, 2016.
- [161] H. Chen, Q. Li, Y. Zhao, and et al., “A survey of deep learning techniques for autonomous driving,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 829–846, 2021.
- [162] S. Ammar, T. Bouwmans, N. Zaghden, and M. Neji, “Deep detector classifier (deepdc) for moving objects segmentation and classification in video surveillance,” *IET Image processing*, vol. 14, no. 8, pp. 1490–1501, 2020.
- [163] J. Redmon and A. Farhadi, “Yolo9000: Better, faster, stronger,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–6525, 2017.
- [164] S. Yeung, O. Russakovsky, and G. Mori, “End-to-end learning of action detection from frame glimpses in videos,” *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 2678–2686, 2016.
- [165] D. Tran, L. Bourdev, and R. Fergus, “Learning spatiotemporal features with 3d convolutional networks,” *IEEE International Conference on Computer Vision (ICCV)*, pp. 4489–4497, 2015.
- [166] M. Hasan, J. Choi, and J. Neumann, “Learning deep representations for anomaly detection in video surveillance,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1889–1898, 2016.
- [167] T.-y. Ko and S.-h. Lee, “Novel method of semantic segmentation applicable to augmented reality,” *Sensors*, vol. 20, no. 6, p. 1737, 2020.
- [168] Q. Cao, L. Liu, L. Zhou, and et al., “Real-time object recognition for augmented reality applications,” *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 4933–4942, 2018.
- [169] Y. Kim and S. Kim, “Augmented reality visual tracking using convolutional neural networks,” *IEEE Access*, vol. 6, pp. 3079–3087, 2018.
- [170] B. Tekin, I. Katircioglu, and M. Salzmann, “Real-time hand pose estimation for augmented reality applications using convolutional neural networks,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 5, pp. 1781–1790, 2018.
- [171] J. Bessenes and A. Kovashka, “Augmented reality with generative neural networks for synthetic object placement,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 11, pp. 2984–2992, 2018.
- [172] T. A. Sorensen, “A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons,” *Biol. Skar.*, vol. 5, pp. 1–34, 1948.
- [173] M. A. Rahman and Y. Wang, “Optimizing intersection-over-union in deep neural networks for image segmentation,” in *International symposium on visual computing*, pp. 234–244, Springer, 2016.
- [174] F. van Beers, A. Lindström, E. Okafor, and M. A. Wiering, “Deep neural networks with intersection over union loss for binary image segmentation,” in *ICPRAM*, pp. 438–445, 2019.
- [175] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, “Semantic segmentation using adversarial networks,” *arXiv preprint arXiv:1611.08408*, 2016.
- [176] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, “Tversky loss function for image segmentation using 3d fully convolutional deep networks,” in *Machine Learning in Medical Imaging: 8th International Workshop, MLMI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Proceedings 8*, pp. 379–387, Springer, 2017.
- [177] M. Berman, A. R. Triki, and M. B. Blaschko, “The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4413–4421, 2018.
- [178] G. Csurka, D. Larlus, and F. Perronnin, “What is a good evaluation measure for semantic segmentation?,” in *British Machine Vision Conference*, pp. 10–5244, 2013.
- [179] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [180] M. Owayjan, A. Dergham, G. Haber, N. Fakih, A. Hamoush, and E. Abdo, “Face recognition security system,” in *New trends in networking, computing, E-learning, systems sciences, and engineering*, pp. 343–348, Springer, 2015.



- [181] I. M. Sayem and M. S. Chowdhury, “Integrating face recognition security system with the internet of things,” in *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, pp. 14–18, IEEE, 2018.
- [182] P. Indrawan, S. Budiyo, N. M. Ridho, and R. F. Sari, “Face recognition for social media with mobile cloud computing,” *International Journal on Cloud Computing: Services and Architecture*, vol. 3, no. 1, pp. 23–35, 2013.
- [183] K. I. Chang, K. W. Bowyer, and P. J. Flynn, “Multimodal 2d and 3d biometrics for face recognition,” in *2003 IEEE International SOI Conference. Proceedings (Cat. No. 03CH37443)*, pp. 187–194, IEEE, 2003.
- [184] A. W. Senior and R. M. Bolle, “Face recognition and its application,” *Biometric Solutions: For Authentication in an E-World*, pp. 83–97, 2002.
- [185] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.
- [186] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pp. 499–515, Springer, 2016.
- [187] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, “Cosface: Large margin cosine loss for deep face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5265–5274, 2018.
- [188] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.
- [189] C. Manwatkar, “How to choose a loss function for face recognition,” 2023. Accessed: 2023-06-28.
- [190] M. Schultz and T. Joachims, “Learning a distance metric from relative comparisons,” *Advances in neural information processing systems*, vol. 16, 2003.
- [191] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.
- [192] S. Chopra, R. Hadsell, and Y. LeCun, “Learning a similarity metric discriminatively, with application to face verification,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, pp. 539–546, IEEE, 2005.
- [193] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, “Sampling matters in deep embedding learning,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2840–2848, 2017.
- [194] E. Hoffer and N. Ailon, “Deep metric learning using triplet network,” in *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12–14, 2015. Proceedings 3*, pp. 84–92, Springer, 2015.
- [195] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, “Circle loss: A unified perspective of pair similarity optimization,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6398–6407, 2020.
- [196] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow twins: Self-supervised learning via redundancy reduction,” in *International Conference on Machine Learning*, pp. 12310–12320, PMLR, 2021.
- [197] X. Chen and K. He, “Exploring simple siamese representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.
- [198] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [199] S. R. Bowman, L. Vilnis, O. Vinyals, and et al., “Generating sentences from a continuous space,” *CoRR*, vol. abs/1511.06349, 2016.
- [200] J. An and S. Cho, “Variational autoencoder based anomaly detection using reconstruction probability,” *Special Lecture on IE*, vol. 2, no. 1, 2015.
- [201] S. Zhao, J. Song, and S. Ermon, “Infovae: Balancing learning and inference in variational autoencoders,” *CoRR*, vol. abs/1706.02262, 2017.
- [202] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [203] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, pp. 2672–2680, Curran Associates, Inc., 2014.

- [204] L. Yu, W. Zhang, J. Wang, and Y. Yu, “Seqgan: Sequence generative adversarial nets with policy gradient,” 2017.
- [205] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017.
- [206] L.-C. Yang, S.-Y. Chou, and Y.-H. Yang, “Midinet: A convolutional generative adversarial network for symbolic-domain music generation,” 2017.
- [207] L. Dinh, J. Sohl-Dickstein, and S. Bengio, “Density estimation using real nvp,” *arXiv preprint arXiv:1605.08803*, 2016.
- [208] D. P. Kingma and P. Dhariwal, “Glow: Generative flow with invertible 1x1 convolutions,” *Advances in neural information processing systems*, vol. 31, 2018.
- [209] W. Grathwohl, R. T. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud, “Ffjord: Free-form continuous dynamics for scalable reversible generative models,” *arXiv preprint arXiv:1810.01367*, 2018.
- [210] D. J. Rezende and S. Mohamed, “Variational inference with normalizing flows,” 2016.
- [211] D. J. Rezende and S. Mohamed, “Variational inference with normalizing flows,” *Journal of Machine Learning Research (JMLR)*, vol. 18, no. 1, pp. 201–244, 2017.
- [212] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, “A tutorial on energy-based learning,” *Predicting structured data*, vol. 1, no. 0, 2006.
- [213] Y. Du and I. Mordatch, “Implicit generation and modeling with energy based models,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [214] T. Tieleman, “Training restricted boltzmann machines using approximations to the likelihood gradient,” in *Neural Computation*, pp. 1064–1071, 01 2008.
- [215] T. Tieleman, “Training restricted boltzmann machines using approximations to the likelihood gradient,” in *Proceedings of the 25th International Conference on Machine Learning, ICML ’08*, (New York, NY, USA), p. 1064–1071, Association for Computing Machinery, 2008.
- [216] D. Belanger and A. McCallum, “Structured prediction energy networks,” in *International Conference on Machine Learning (ICML)*, p. 983–992, 2016.
- [217] R. E. Tillman, T. Balch, and M. Veloso, “Privacy-preserving energy-based generative models for marginal distribution protection,” *Transactions on Machine Learning Research*, 2023.
- [218] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *International Conference on Machine Learning*, pp. 2256–2265, PMLR, 2015.
- [219] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [220] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, “Your classifier is secretly an energy based model and you should treat it like one,” 2020.
- [221] J. R. Hershey, J. L. Roux, and F. Weninger, “Deep unfolding: Model-based inspiration of novel deep architectures,” 2014.
- [222] Y. Song, Y. Zhang, and S. Ermon, “Deep variational diffusion models,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.
- [223] S. C. Park, M. K. Park, and M. G. Kang, “Super-resolution image reconstruction: a technical overview,” *IEEE signal processing magazine*, vol. 20, no. 3, pp. 21–36, 2003.
- [224] C. Dong, C. C. Loy, K. He, and et al., “Learning a deep convolutional network for image super-resolution,” in *European Conference on Computer Vision (ECCV)*, pp. 184–199, 2014.
- [225] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” *CoRR*, vol. abs/1609.04802, 2016.
- [226] Y. Zhang, K. Li, K. Li, and et al., “Image super-resolution using very deep residual channel attention networks,” in *European Conference on Computer Vision (ECCV)*, p. 294–310, 2018.
- [227] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin, “Deep learning on image denoising: An overview,” *Neural Networks*, vol. 131, pp. 251–275, 2020.
- [228] J. Xie, L. Xu, and E. Chen, “Image denoising and inpainting with deep neural networks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, p. 341–349, 2012.

- [229] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, “Noise2Noise: Learning Image Restoration without Clean Data,” in *Proceedings of the 35th International Conference on Machine Learning*, pp. 2971–2980, PMLR, 2018.
- [230] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, “Image inpainting,” in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 417–424, 2000.
- [231] J. Yu, Z. L. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, “Free-form image inpainting with gated convolution,” *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4470–4479, 2018.
- [232] D. Pathak, P. Krahenbuhl, J. Donahue, and et al., “Context encoders: Feature learning by inpainting,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2536–2544, 2016.
- [233] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, “Image inpainting for irregular holes using partial convolutions,” 2018.
- [234] L. A. Gatys, A. S. Ecker, and M. Bethge, “A neural algorithm of artistic style,” *arXiv preprint arXiv:1508.06576*, 2015.
- [235] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” 2016. cite arxiv:1603.08155.
- [236] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1510–1519, 2017.
- [237] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, “Learning texture transformer network for image super-resolution,” 2020.
- [238] J. Nash Jr, “Non-cooperative games,” in *Essays on Game Theory*, pp. 22–33, Edward Elgar Publishing, 1996.
- [239] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *International conference on machine learning*, pp. 214–223, PMLR, 2017.
- [240] M. A. Carreira-Perpinan and G. Hinton, “On contrastive divergence learning,” in *International workshop on artificial intelligence and statistics*, pp. 33–40, PMLR, 2005.
- [241] I. J. Myung, “Tutorial on maximum likelihood estimation,” *Journal of mathematical Psychology*, vol. 47, no. 1, pp. 90–100, 2003.
- [242] M. Gutmann and A. Hyvärinen, “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models,” in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304, JMLR Workshop and Conference Proceedings, 2010.
- [243] I. Csiszar, “ $I$ -Divergence Geometry of Probability Distributions and Minimization Problems,” *The Annals of Probability*, vol. 3, no. 1, pp. 146 – 158, 1975.
- [244] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein gans,” *Advances in neural information processing systems*, vol. 30, 2017.
- [245] G. Desjardins, A. Courville, Y. Bengio, P. Vincent, and O. Delalleau, “Tempered markov chain monte carlo for training of restricted boltzmann machines,” in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 145–152, JMLR Workshop and Conference Proceedings, 2010.
- [246] M. Opper and D. Saad, *Advanced mean field methods: Theory and practice*. MIT press, 2001.
- [247] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” *Advances in neural information processing systems*, vol. 29, 2016.
- [248] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in neural information processing systems*, vol. 30, 2017.
- [249] D. Mukherjee, P. Saha, D. Kaplun, A. Sinitca, and R. Sarkar, “Brain tumor image generation using an aggregation of gan models with style transfer,” *Scientific Reports*, vol. 12, no. 1, p. 9141, 2022.
- [250] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [251] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., “Improving language understanding by generative pre-training,” 2018.
- [252] J.-M. Torres-Moreno, *Automatic text summarization*. John Wiley & Sons, 2014.
- [253] H. Schmid, “Part-of-speech tagging with neural networks,” *arXiv preprint cmp-lg/9410018*, 1994.

- [254] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, “Unsupervised named-entity extraction from the web: An experimental study,” *Artificial intelligence*, vol. 165, no. 1, pp. 91–134, 2005.
- [255] B. Taskar, C. Guestrin, and D. Koller, “Max-margin markov networks,” *Advances in neural information processing systems*, vol. 16, 2003.
- [256] A. F. Martins, M. A. Figueiredo, and N. A. Smith, “Structured sparsity in natural language processing: models, algorithms and applications,” in *Tutorial Abstracts at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012.
- [257] R. Moore and J. DeNero, “L1 and l2 regularization for multiclass hinge loss models,” in *Symposium on machine learning in speech and language processing*, 2011.
- [258] J. Wang, X. Le, X. Peng, and C. Chen, “Adaptive hinge balance loss for document-level relation extraction,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3872–3878, 2023.
- [259] H. Wang, *Adversarial Prediction Framework for Information Retrieval and Natural Language Processing Metrics*. PhD thesis, University of Illinois at Chicago, 2017.
- [260] J. T. Hoe, K. W. Ng, T. Zhang, C. S. Chan, Y.-Z. Song, and T. Xiang, “One loss for all: Deep hashing with a single cosine similarity based learning objective,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 24286–24298, 2021.
- [261] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.
- [262] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, pp. 74–81, 2004.
- [263] S. Shen, Y. Cheng, Z. He, W. He, H. Wu, M. Sun, and Y. Liu, “Minimum risk training for neural machine translation,” *arXiv preprint arXiv:1512.02433*, 2015.
- [264] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Machine learning*, vol. 8, pp. 229–256, 1992.
- [265] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” *Advances in neural information processing systems*, vol. 12, 1999.
- [266] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker, “Perplexity—a measure of the difficulty of speech recognition tasks,” *The Journal of the Acoustical Society of America*, vol. 62, no. S1, pp. S63–S63, 1977.
- [267] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? sentiment classification using machine learning techniques,” *arXiv preprint cs/0205070*, 2002.
- [268] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, “A bayesian approach to filtering junk e-mail,” in *Learning for Text Categorization: Papers from the 1998 workshop*, vol. 62, pp. 98–105, Citeseer, 1998.
- [269] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” in *European conference on machine learning*, pp. 137–142, Springer, 1998.
- [270] D. D. Lewis, “Naive (bayes) at forty: The independence assumption in information retrieval,” in *European conference on machine learning*, pp. 4–15, Springer, 1998.
- [271] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [272] Y. Heryanto and A. Triayudi, “Evaluating text quality of gpt engine davinci-003 and gpt engine davinci generation using bleu score,” *SAGA: Journal of Technology and Information System*, vol. 1, no. 4, pp. 121–129, 2023.
- [273] M. S. Amin, A. Mazzei, L. Anselma, *et al.*, “Towards data augmentation for drs-to-text generation,” in *CEUR WORKSHOP PROCEEDINGS*, vol. 3287, pp. 141–152, CEUR-WS, 2022.
- [274] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” *arXiv preprint arXiv:1910.13461*, 2019.
- [275] Y. Wang, J. Jiang, M. Zhang, C. Li, Y. Liang, Q. Mei, and M. Bendersky, “Automated evaluation of personalized text generation using large language models,” *arXiv preprint arXiv:2310.11593*, 2023.
- [276] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

- [277] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [278] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [279] H. Masataki, Y. Sagisaka, K. Hisaki, and T. Kawahara, “Task adaptation using map estimation in n-gram language modeling,” in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 783–786, IEEE, 1997.
- [280] H. Li, D. Cai, J. Xu, and T. Watanabe, “*n*-gram is back: Residual learning of neural text generation with *n*-gram language model,” *arXiv preprint arXiv:2210.14431*, 2022.
- [281] M. Sundermeyer, H. Ney, and R. Schlüter, “From feedforward to recurrent lstm neural networks for language modeling,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 517–529, 2015.
- [282] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “Squad: 100,000+ questions for machine comprehension of text,” *arXiv preprint arXiv:1606.05250*, 2016.