

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche
Scientifique

Université Mustapha
Stambouli Mascara



جامعة مصطفى اسطمبولي معسكر

LA FACULTÉ DES SCIENCES EXACTES

Département Informatique

THESE de DOCTORAT

Détection et reconnaissance automatique du texte arabe dans les scènes vidéo

Présentée par: Oualid KHAL

Le Jeudi 09/07/2026 à 09h00 à la Faculté des Sciences Exactes.

Devant le jury:

President:	MCA. Sahraoui Mustapha
Examineur:	MCA. BOUDIA Chérifa
Examineur:	MCA. Nehnough Chakib
Encadreur:	Prof. Fatma BOUFERA

Année Universitaire: 2025-2026

الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria

وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research

Université Mustapha
Stambouli Mascara



جامعة مصطفى اسطمبولي معسكر

FACULTY OF EXACT SCIENCES

Computer science Department

PhD THESIS

Automatic detection and recognition of Arabic text in video scenes

Presented by: Oualid KHAL

Jury Members:

President:	MCA. Sahraoui Mustapha
Examiner:	MCA. BOUDIA Chérifa
Examiner:	MCA. Nehnouh Chakib
Supervisor:	Prof. Fatma BOUFERRA

Academic Year: 2025-2026

Dedication

To the memory of my beloved father, Djelloul KHIAL, who called me “Doctor” from the time I was five years old.

To my beloved mother, whose unwavering love and support made me who I am and guided me every step of the way.

To my beloved wife and children, the reason I keep moving forward and the greatest source of joy in my life.

To my brothers, Saleh, Mustapha, and Abd El Ouahab, my strength and my constant support.

To my friend, my brother, and my soulmate, Ihab CHERITI.

To everyone who has earned my love and respect.

Acknowledgments

I would like to express my sincere gratitude to everyone who has ever taught me, for every lesson and every piece of knowledge they have shared with me.

My deepest thanks go to my supervisor, Professor Fatma BOUFERAA, whose guidance, patience, and unwavering support helped me find my way whenever I felt lost.

Special thanks to Professor Benmeriem Khaled for his encouragement, support, and confidence in me whenever I needed it most.

My sincere appreciation also goes to the academic and administrative staff of Mustapha STAMBOULI University for providing a supportive environment throughout my doctoral studies.

Abstract

This thesis focuses on the problem of Arabic text recognition from scene videos, a research area that has gained increasing importance in recent years due to the rapid growth of visual media on the internet, social networks, and video-sharing platforms. Extracting textual information from videos is a key step for many applications, including video indexing, content-based retrieval, assistive technologies, and intelligent surveillance systems. However, Arabic scene text recognition remains particularly challenging. The Arabic script is inherently cursive, characters can take multiple shapes depending on their position within a word, and diacritics may appear or disappear depending on writing style. These linguistic characteristics, combined with common video-related issues such as motion blur, low resolution, compression artifacts, varying illumination, and complex backgrounds, make Arabic text recognition significantly more difficult than its Latin counterpart.

In addition to these intrinsic challenges, the field suffers from a major limitation in terms of data availability. Existing Arabic scene text datasets are generally small, limited in font diversity, or captured under constrained conditions, which makes them insufficient for training modern deep learning models that typically require large and diverse datasets to generalize well. As a result, many state-of-the-art architectures struggle to achieve robust performance when applied to real-world Arabic video data. The work presented in this thesis aims to address these challenges from two complementary perspectives: data generation and model design.

The first major contribution of this research is the design, construction, and public release of a large-scale artificial dataset named IYaD, specifically tailored for Arabic scene text recognition. The dataset was carefully generated to reflect the visual and structural complexity of Arabic text as it appears in natural scenes. It contains approximately 1,400,000 images for a single font and a comparable number of images for sixteen additional fonts, resulting in a dataset with exceptional size and stylistic diversity. Each text sample is rendered in three distinct visual versions and is accompanied by rich annotations, including its Arabic ground-truth label, a corresponding Latin transcription, and the original textual content.

Although IYaD is synthetically generated, a detailed complexity analysis and

multiple comparisons with real-world datasets were conducted to demonstrate that it closely simulates the visual characteristics of natural Arabic scene text. Furthermore, the dataset is distributed together with a complete set of tools that facilitate preprocessing, augmentation, transformation, and seamless integration into existing deep learning pipelines. By providing both scale and diversity, IYaD fills a critical gap in the available resources for Arabic text recognition and enables more effective training of data-hungry neural architectures.

The second main contribution of this thesis is the proposal of a novel Transformer-based architecture for Arabic Scene Text Recognition (ASTR). This work is grounded in a comprehensive comparative study between two different modeling strategies. The first strategy follows a more conventional approach, where convolutional neural networks (CNNs) are used to extract visual feature maps that are subsequently processed by a Transformer encoder. The second strategy removes the convolutional component entirely and relies on a purely Transformer-based model to directly learn visual representations and sequence dependencies.

Extensive experiments were carried out using all publicly available Arabic text recognition datasets, as well as the newly introduced IYaD dataset. The results clearly indicate that the fully Transformer-based architecture outperforms the hybrid CNN–Transformer approach across multiple evaluation settings. In particular, when trained on IYaD, the proposed model achieves performance levels that surpass the current state of the art, while also demonstrating strong robustness to variations in font style, image quality, and visual noise. These findings highlight not only the effectiveness of Transformer architectures for Arabic OCR, but also the crucial role that large-scale artificial datasets can play in overcoming the limitations of scarce real-world data.

In addition to these core contributions, this thesis provides a comprehensive experimental benchmark for Arabic text recognition. Several well-known OCR models were implemented, trained, and evaluated under consistent and controlled conditions. This benchmark offers a detailed comparison of different architectural choices and training strategies, helping to clarify their respective strengths and weaknesses. It also serves as a valuable reference framework for future research in the domain.

Overall, this thesis contributes new datasets, new modeling approaches, and new experimental insights to the field of Arabic scene text recognition. By combining a carefully designed large-scale artificial dataset with a modern Transformer-based recognition architecture, it demonstrates a practical and effective path toward improving Arabic OCR performance in video-based and real-world scenarios, and it opens promising directions for future research in multilingual and low-resource text recognition.

Keywords : Natural language processing, Artificial intelligence, Arabic

text, Computer vision, Deep learning, Transformer

Résumé

Cette thèse s'intéresse à la reconnaissance de texte arabe à partir de vidéos de scènes naturelles, un domaine de recherche dont l'importance ne cesse de croître avec l'augmentation massive des contenus vidéo diffusés sur Internet, les réseaux sociaux et les plateformes de partage. L'extraction automatique d'informations textuelles à partir de vidéos constitue une étape essentielle pour de nombreuses applications, telles que l'indexation et la recherche de contenus vidéo, les systèmes d'assistance, ou encore les applications de surveillance intelligente. Toutefois, la reconnaissance du texte arabe en environnement réel reste un problème particulièrement complexe. L'écriture arabe est de nature cursive, chaque caractère pouvant adopter plusieurs formes selon sa position dans le mot, et l'usage des diacritiques varie fortement selon le style d'écriture. À ces difficultés linguistiques s'ajoutent les contraintes propres aux vidéos, telles que le flou de mouvement, la faible résolution, les artefacts de compression, les variations d'éclairage et la présence d'arrière-plans complexes.

En parallèle de ces défis intrinsèques, le domaine souffre d'un manque critique de jeux de données de grande taille et suffisamment diversifiés. Les bases de données existantes pour la reconnaissance du texte arabe dans les scènes naturelles sont souvent limitées en volume, en variété de polices ou en conditions de capture, ce qui restreint leur capacité à soutenir l'entraînement de modèles d'apprentissage profond modernes. Or, ces modèles nécessitent de grandes quantités de données riches et variées afin de généraliser efficacement aux situations réelles. Les travaux présentés dans cette thèse visent à répondre à ces limitations en abordant le problème à la fois sous l'angle des données et sous celui de la conception des modèles.

La première contribution majeure de cette recherche est la conception, la génération et la mise à disposition publique d'un nouveau jeu de données artificiel, nommé IYaD, dédié à la reconnaissance de texte arabe en scènes naturelles. Ce jeu de données a été spécifiquement conçu pour refléter la complexité visuelle et structurelle du texte arabe tel qu'il apparaît dans des environnements réels. Il contient environ 1 400 000 images pour une police unique, ainsi qu'un volume comparable pour seize autres polices, offrant ainsi une très grande diversité de styles d'écriture. Chaque échantillon est fourni sous trois versions visuelles différentes et accompagné d'annotations complètes, comprenant l'étiquette en arabe, une transcription en caractères latins, ainsi que le contenu textuel original.

Bien que IYaD soit un jeu de données synthétique, une étude approfondie de sa complexité ainsi que des comparaisons avec des bases de données réelles ont été menées afin de démontrer sa capacité à reproduire de manière fiable les caractéristiques du texte arabe en conditions naturelles. De plus, le jeu de données est distribué avec un ensemble d'outils facilitant le prétraitement, la transformation, l'augmentation et l'intégration dans la majorité des pipelines d'apprentissage

profond existants. Par son ampleur et sa diversité, IYaD comble une lacune importante dans les ressources disponibles pour la reconnaissance de texte arabe.

La seconde contribution de cette thèse concerne la proposition d’une nouvelle architecture basée sur les Transformers pour la reconnaissance de texte arabe en scènes naturelles (Arabic Scene Text Recognition, ASTR). Cette proposition est issue d’une étude comparative entre deux approches distinctes. La première combine des cartes de caractéristiques extraites par des réseaux convolutifs (CNN) avec un encodeur Transformer, tandis que la seconde s’appuie exclusivement sur une architecture Transformer, sans composant convolutionnel. Des expérimentations approfondies ont été menées à l’aide de l’ensemble des jeux de données arabes publics disponibles, ainsi que du jeu de données IYaD. Les résultats montrent que l’approche entièrement basée sur les Transformers est systématiquement plus performante. En particulier, le modèle proposé dépasse l’état de l’art lorsqu’il est entraîné sur IYaD, tout en présentant une forte robustesse face aux variations de styles, de qualité d’image et de conditions visuelles.

Ces résultats mettent en évidence non seulement le potentiel des architectures Transformer pour la reconnaissance optique de caractères arabes, mais également le rôle fondamental que peuvent jouer les jeux de données artificiels de grande échelle lorsque les données réelles sont limitées.

En complément de ces contributions, cette thèse propose un benchmark expérimental complet pour la reconnaissance de texte arabe. Plusieurs modèles OCR bien établis ont été implémentés, entraînés et évalués dans des conditions homogènes, permettant une comparaison claire et équitable de leurs performances. Ce benchmark met en lumière les avantages et les limites de chaque approche et constitue une référence utile pour les travaux futurs dans ce domaine.

En conclusion, cette thèse apporte de nouvelles ressources, de nouvelles méthodes et de nouvelles perspectives à la recherche en reconnaissance de texte arabe en scènes naturelles. En combinant un jeu de données artificiel de grande ampleur et soigneusement conçu avec une architecture moderne basée sur les Transformers, elle propose une approche efficace pour améliorer la qualité de la reconnaissance du texte arabe dans les vidéos et ouvre la voie à de futures avancées dans la reconnaissance de texte multilingue et en contexte réel.

Mots-clés : Traitement automatique du langage naturel (TALN), Intelligence artificielle, Texte arabe, Vision par ordinateur, Apprentissage profond, Transformeurs

ملخص

تركز هذه الأطروحة على التعرف على النصوص العربية من مقاطع الفيديو، وهو موضوع يكتسب أهمية متزايدة مع استمرار نمو حجم محتوى الفيديو عالمياً. يصعب التعرف على النصوص العربية بطبيعتها نظراً لبنيتها المتصلة، وتنوع أشكال كل حرف، والتشويش والتشوهات الشائعة في إطارات الفيديو. في الوقت نفسه، يعاني هذا المجال من نقص حاد في مجموعات البيانات الكبيرة والمتنوعة التي تدعم تدريب نماذج التعلم العميق الحديثة. يهدف العمل الذي أقدمه في هذه الأطروحة إلى معالجة هذه المشكلات من منظور البيانات والنموذج على حد سواء.

تمثل المساهمة الأولى لبحتي في إنشاء ونشر مجموعة بيانات اصطناعية جديدة تُسمى IYaD. صُممت هذه المجموعة خصيصاً لدعم التعرف على النصوص العربية في المشاهد. تحتوي على حوالي 1,400,000 صورة لخط واحد، وعدد مماثل لستة عشر خطأ إضافياً، مما يمنحها تنوعاً كبيراً في أنماط الكتابة. تظهر كل عينة بثلاث نسخ مختلفة، مع توفير تصنيفها العربي، وترجمتها اللاتينية، ومحتوى النص الأصلي. على الرغم من أن مجموعة البيانات مُولدة اصطناعياً، فقد أُجريت دراسةٌ للتعقيد ومقارنات لإثبات قدرتها على محاكاة النصوص العربية الواقعية بدقة عالية. وقد نُشرت مجموعة البيانات مع مجموعةٍ من الأدوات التي تُسهّل معالجتها وتحويلها ودمجها في معظم مسارات التعلم العميق الحالية. وبفضل حجمها وتنوعها، تُسد مجموعة بيانات IYaD ثغرةً مهمةً في الموارد المتاحة للتعرف على النصوص العربية.

يُقدّم الجزء الثاني من هذه الأطروحة بنيةً قائمةً على نموذج Transformer للتعرف على النصوص العربية في المشاهد (ASTR). وقد انبثق هذا من دراسةٍ مقارنةٍ بين منهجين: الأول يستخدم خرائط الميزات الالتفافية كمدخلٍ لمشفر Transformer، بينما يُزيل الثاني جزء الشبكة العصبية الالتفافية (CNN) بالكامل ويعتمد فقط على نموذج Transformer خالص. وبعد تجاربٍ مكثفةٍ باستخدام جميع مجموعات البيانات العامة المتاحة، بالإضافة إلى مجموعة بيانات IYaD الجديدة، أثبتت النسخة القائمة بالكامل على نموذج Transformer فعاليتها الأكبر. وقد حقّق النموذج نتائج تتجاوز أحدث التقنيات، لا سيما عند تدريبه على مجموعة بيانات IYaD، كما أظهر متانةً عاليةً في ظروف اختبارٍ مختلفة. لا يقتصر هذا العمل على إظهار إمكانيات بنية Transformer في التعرف الضوئي على الحروف العربية فحسب، بل يُبرز أيضاً الدور المحوري الذي يمكن أن تؤديه مجموعات البيانات الاصطناعية عند محدودية البيانات الحقيقية.

إلى جانب هذه المساهمات، تُقدّم هذه الأطروحة معياراً تجريبياً شاملاً للتعرف على النصوص العربية، حيث قُمتُ بتطبيق واختبار العديد من نماذج التعرف الضوئي على الحروف المعروفة ومقارنتها في ظل ظروف متسقة. يُساعد هذا المعيار على تسليط الضوء على نقاط القوة والضعف في مختلف المناهج، ويُوفّر مرجعاً واضحاً للبحوث المستقبلية.

بشكل عام، تُضيف هذه الأطروحة موارد وأساليب ورؤى جديدة إلى مجال التعرف على النصوص العربية في المشاهد. ومن خلال الجمع بين مجموعة بيانات اصطناعية كبيرة ومصممة بعناية وبنية Transformer حديثة،

تُقدّم طريقة عملية لتحسين جودة التعرف، وتفتح آفاقاً لمزيد من التقدم في مجال التعرف الضوئي على الحروف العربية في مقاطع الفيديو.

الكلمات المفتاحية: معالجة اللغات الطبيعية، الذكاء الاصطناعي، النص العربي، الرؤية الحاسوبية، التعلم العميق، المحولات

Contents

List of Figures	13
List of Tables	14
1 Introduction	16
1.1 Global Context: From Document OCR to Video Scene Text Recognition (VSTR)	16
1.1.1 The Evolving Role of Text in Dynamic Media	16
1.1.2 The Gap Between Image OCR and Video Scene Text Recognition	17
1.1.3 The Computer Vision Paradigm	18
1.1.4 From Static to Spatio-Temporal Text Understanding	19
1.1.5 Industrial and Societal Applications of Video Scene Text Recognition	20
1.1.6 Limitations of Existing Video Text Recognition Pipelines	21
1.1.7 Justification for Specialized Arabic Video STR Research	22
1.2 Text Extraction from Images	22
1.2.1 The Role of Text Detection	22
1.2.2 Advancements in Deep Learning Localization	23
1.2.3 Text Tracking in Video	23
1.3 Text Recognition	24
1.3.1 From Classification to Recurrent Sequence Modeling	24
1.3.2 The Shift Toward Attention and Novel Architectures	24
1.4 Arabic Text Challenges	25
1.4.1 Inherent Linguistic and Structural Complexities	25
1.4.2 Environmental Challenges and Data Rarity	27
1.4.3 Arabic Script Versus Latin Script: A Comparative Perspective	28
1.5 Research Problem Formulation and Thesis Objectives	29
1.5.1 Synthesizing the Research Deficit	29
1.5.2 Statement of Purpose and Primary Contribution	29
1.5.3 Core Thesis Objectives	30

1.6	Thesis Contributions and Unique Approach	31
1.6.1	Data Contribution: The IYaD Dataset	31
1.6.2	Transformer-Based Approach for Robust Arabic Scene Text Recognition	32
1.7	Thesis Structure and Organization	33
2	Related Works	34
2.1	Early OCR and Scene Text Recognition	34
2.1.1	Challenges in Early Systems	35
2.2	Scene Text Detection and Recognition	35
2.2.1	Advances in Multi-Oriented Text Recognition	36
2.2.2	Real-Time Recognition Systems	36
2.3	Arabic Text Recognition	37
2.3.1	Modeling Contextual Dependencies	37
2.3.2	Handling Low-Resource Scenarios	37
2.3.3	Arabic Video Text Recognition	38
2.3.4	Arabic Video Text Recognition	39
2.4	Existing Arabic OCR Datasets	40
2.4.1	Scene-Text Arabic Datasets	41
2.5	Synthetic Datasets in OCR Research	41
2.5.1	Domain Adaptation for Synthetic Data	41
2.6	Domain-Specific and Multimodal Approaches	42
2.6.1	Multilingual and Cross-Domain Recognition	42
2.6.2	Arabic Street Sign Recognition and Real-World Applications	42
2.7	2.7 Summary	43
3	IYaD dataset	44
3.1	Introduction	44
3.2	Motivation	45
3.3	Dataset Design and Collection Methodology	47
3.3.1	Data Sources	48
3.3.2	Image Generation Process	48
3.3.3	Annotation and Labeling	49
3.3.4	Preprocessing and Quality Control	50
3.3.5	Data Augmentation	50
3.4	Dataset Statistics and Characteristics	52
3.4.1	Overall Size and Structure	52
3.4.2	Word-Level vs Sentence-Level Data	52
3.4.3	Fonts and Backgrounds	53
3.4.4	Metadata and Annotation	53
3.4.5	Image Quality and Complexity	54

3.4.6	Data Augmentation	57
3.4.7	Comparison with Existing Datasets	57
3.5	Use Cases and Intended Applications	58
3.5.1	Dataset Generation Pipeline	61
3.5.2	Challenges and Limitations	64
3.6	Conclusion	67
4	Multi-Head Attention Transformer for Arabic Scene Text Recognition	69
4.1	Introduction	69
4.2	Related Works	70
4.2.1	Classic and Recurrent Neural Network (RNN) Approaches .	70
4.2.2	Attention-Based and Transformer Architectures	70
4.2.3	Arabic Text Recognition Datasets	70
4.3	Motivation and Contribution	71
4.4	Overall Architecture	72
4.5	Pipe Line Implementation Details	72
4.5.1	Input Embedding	72
4.5.2	Positional Encoding	72
4.5.3	Transformer Encoder	74
4.5.4	Transformer Decoder	75
4.5.5	Masking Strategy Inspired by Human Arabic Reading Behavior	77
4.5.6	Model Parameters	79
4.6	Experimental Results	81
4.6.1	Training Datasets and Setup	81
4.6.2	Results Analysis and Comparison	83
4.7	Discussion	85
4.8	Summary and Future Perspectives	87
4.8.1	Temporal Feature Studies for Video-Based Recognition . . .	88
4.8.2	Future Ablation and Feature Analysis	88
4.8.3	Towards an Integrated Image and Video Recognition System	89
4.8.4	Conclusion of the Chapter	89
	General Conclusion	91
	Bibliography	94

List of Figures

1.1	Examples of Arabic scene text samples from IYaD dataset.	31
3.1	Example from the dataset in three different versions	50
3.2	XML file structure	50
3.3	Entropy visualization for Iyad and Alif datasets	55
3.4	Data augmentation results	57
4.1	Overview of the proposed multi-head attention Transformer for Arabic scene text recognition	90

List of Tables

1.1	Comparison of Text Extraction Domains	16
1.2	Paradigm Shift in Feature Extraction	19
1.3	Advancements in Text Localization Methodologies	23
1.4	Technical Comparison: Recurrent Modeling vs. Attention Mechanisms	25
1.5	Character Shape Variations	26
1.6	Character Similarity	26
1.7	Complexity of Arabic Ligatures	27
3.1	Arabic text detection datasets	46
3.2	Comparison of computer vision datasets	46
3.3	Arabic letter encoding map	51
3.4	Latin label example	53
3.5	Images complexity study results	56
3.6	Image quality metrics for IYaD compared to ALiF and AcTiV 2.0 datasets	56
3.7	Comparison of IYaD with existing Arabic text datasets	58
4.1	Optimal Hyperparameter Values	79
4.2	Impact of Dataset Diversity on Model Performance	84
4.3	Performance Comparison of ASTR Architectures	84

Introduction

- **OCR:** Optical Character Recognition.
- **STR:** Scene Text Recognition.
- **VSTR:** Video Scene Text Recognition.
- **ANNS:** Artificial Neural Networks.
- **DBNet:** Differentiable Binarization Network.
- **MSA:** Modern Standard Arabic.
- **NLP:** Natural Language Processing.
- **NSP:** Next Sentence Prediction.
- **ReGex:** Regular Expression.
- **RNN:** Recurrent Neural Network.
- **RTE:** Recognizing Textual Entailment.
- **SWAG:** Situations With Adversarial Generations.
- **TPU:** Tensor Processing Unit.

Chapter 1

Introduction

1.1 Global Context: From Document OCR to Video Scene Text Recognition (VSTR)

1.1.1 The Evolving Role of Text in Dynamic Media

The global digital landscape is continuously being redefined by the proliferation of video content, creating a persistent and increasing demand for modern and reliable techniques dedicated to automatic video processing and content indexing . Effective indexing and semantic treatment of this media require the comprehensive use of all available information, including external metadata and data embedded directly within the visual stream. Textual information captured within video frames—such as names, locations, dates, product details, or instructions—is recognized as one of the most valuable sources of contextual metadata for identifying and describing the depicted content [1].

To better understand the transition from static to dynamic text processing, it is essential to distinguish the technical requirements of different domains. Table 1.1 illustrates the fundamental shifts in constraints between Document OCR, static Scene Text Recognition, and the target domain of this thesis, Video STR.

Table 1.1: Comparison of Text Extraction Domains

Feature	Document OCR	Static STR	Video STR (VSTR)
Background	Uniform (White)	Complex/Cluttered	Dynamic/Non-stationary
Text Geometry	Planar/Horizontal	Multi-oriented/Curved	Perspective/Motion-deformed
Resolution	High/Standardized	Variable	Low/Subject to Compression
Primary Goal	Digitization	Image Indexing	Temporal Semantic Analysis

The goal of automating the recognition of text embedded in visual data has

thus evolved beyond traditional Optical Character Recognition (OCR), which typically deals with clean, document-based images. The domain has shifted toward Scene Text Recognition (STR), and specifically, **Video Scene Text Recognition (VSTR)**, which must interpret text within unconstrained, real-world scenes where textual artifacts are subject to dynamic challenges like movement, varying camera angles, and low resolution.

1.1.2 The Gap Between Image OCR and Video Scene Text Recognition

Most traditional Optical Character Recognition (OCR) and Scene Text Recognition (STR) systems have been developed and evaluated on static images. These methods assume that the input text appears in a single well-captured frame with relatively stable visual properties. However, when extending these approaches to video-based text recognition, several new challenges emerge that significantly degrade their performance.

One of the main differences between image-based OCR and video STR is the presence of motion. In video sequences, both the camera and the objects containing text may move continuously. This motion often produces motion blur, which causes characters to lose sharp edges and fine details. Since many OCR models rely on spatial features such as strokes, corners, and contours, blur directly reduces the quality of extracted features and leads to incorrect character predictions.

Another important issue is shutter distortion and frame sampling effects. In fast-moving scenes, rolling shutter cameras may capture distorted text shapes, especially in mobile and surveillance videos. As a result, the same text instance may appear stretched, compressed, or partially missing across consecutive frames. These distortions are rarely present in image-based OCR datasets, making static models poorly adapted to real video conditions.

Lighting variation also plays a critical role in video STR. Unlike static images where illumination is usually controlled or consistent, videos often contain sudden lighting changes caused by shadows, reflections, passing vehicles, or camera exposure adjustment. This leads to significant variations in text brightness and contrast over time. Consequently, models trained only on still images struggle to maintain stable recognition performance across frames.

Furthermore, video data introduces temporal redundancy and temporal inconsistency at the same time. While the same text may appear across multiple frames, its visual quality may fluctuate due to blur, occlusion, or viewpoint changes. Traditional OCR methods process each frame independently and ignore temporal relationships. This frame-by-frame processing fails to exploit useful information present in neighboring frames, such as clearer views of characters that may appear

only briefly.

From a mathematical perspective, image-based OCR can be viewed as a mapping function:

$$f(I) \rightarrow T$$

where I represents a single image and T is the recognized text sequence. In contrast, video STR requires learning a function over a sequence of frames:

$$f(I_1, I_2, \dots, I_n) \rightarrow T$$

This formulation highlights the need to aggregate visual information over time rather than relying on a single frame.

Another limitation of static OCR systems is their sensitivity to noise and low resolution, which are common in videos captured from a distance or under compression. Surveillance footage, dashcams, and handheld mobile recordings often contain text regions with small font sizes and compression artifacts. Such conditions significantly reduce recognition accuracy when using models designed for high-quality images.

In addition, occlusions frequently occur in video environments, where text may be partially blocked by moving objects or people. While a single frame may show incomplete text, adjacent frames might contain the missing parts. Image-based OCR systems cannot recover from such situations, whereas video-based approaches can potentially integrate information across time.

Overall, the fundamental gap between image OCR and video STR lies in the dynamic nature of video data. Challenges such as motion blur, geometric distortion, lighting variation, low resolution, and temporal inconsistency require specialized modeling strategies that go beyond conventional static image recognition. This motivates the development of video-aware text detection and recognition frameworks that incorporate temporal feature aggregation, sequence modeling, and robust visual representations.

Addressing these issues is particularly important for Arabic scene text recognition, where complex character shapes, contextual dependencies, and cursive writing styles already present intrinsic difficulties. When combined with video-specific degradations, these challenges further emphasize the necessity of designing dedicated video STR solutions rather than directly applying image-based OCR models.

1.1.3 The Computer Vision Paradigm

Computer vision fundamentally aims to equip machines with the human abilities of interpreting and understanding complex visual surroundings. Traditionally, solving visual tasks required developing hand-crafted functions to extract specific

low-level features, such as edges or connected components . This approach proved inadequate for complex tasks like VSTR, where features are highly variable and context-dependent, rapidly becoming a scalability bottleneck .

The adoption of deep learning architectures, such as Artificial Neural Networks (ANNs), revolutionized this field. Deep learning enables systems to automatically learn and extract informative, task-specific features directly from raw visual data, eliminating the need for manual feature engineering. Table 1.2 summarizes the shift from classical heuristic-based methods to the data-driven deep learning paradigm.

Table 1.2: Paradigm Shift in Feature Extraction

Criteria	Classical Computer Vision	Deep Learning (ANNs)
Feature Source	Human experts (SIFT, HOG, MSER)	Automated Learning from Data
Complexity	Limited to low-level primitives	Hierarchical/Semantic Depth
Environment	Struggles with noise and occlusion	Robust and Generalizable
Scalability	High manual effort for new domains	Scales with Dataset Size

This architectural shift has facilitated the development of robust solutions capable of tackling the severe challenges posed by real-world text extraction from videos, including cluttered backgrounds, arbitrary text orientation, and dynamic illumination changes [2].

1.1.4 From Static to Spatio-Temporal Text Understanding

Early optical character recognition (OCR) systems were designed under the assumption that text appears in static, well-aligned, and high-quality document images. In such controlled settings, text localization and recognition could be addressed independently, relying on carefully engineered features and predefined layout rules. While these approaches achieved high accuracy for scanned documents, they proved inadequate when confronted with text embedded in natural scenes, where illumination variations, complex backgrounds, and geometric distortions are common.

The emergence of scene text recognition (STR) marked a paradigm shift from document-centric OCR toward text understanding in unconstrained environments. In natural images, text often appears arbitrarily oriented, partially occluded, or rendered in diverse fonts and styles. These challenges motivated the adoption

of deep learning-based methods capable of learning robust visual representations directly from data. However, most STR research initially focused on single images, treating each frame as an independent entity.

Video Scene Text Recognition (VSTR) extends this problem into the temporal domain, introducing both new challenges and new opportunities. Unlike static images, video sequences provide temporal redundancy, where the same textual content may appear across multiple consecutive frames with varying degrees of clarity. Motion blur, compression artifacts, camera movement, and dynamic backgrounds complicate text detection and recognition, yet the temporal dimension allows for information aggregation and consistency enforcement over time.

Spatio-temporal text understanding requires models that can jointly exploit spatial visual cues and temporal dependencies. This evolution has led to the integration of sequence modeling techniques, such as recurrent neural networks and attention-based architectures, into text recognition pipelines. In this context, VSTR is no longer a simple extension of image-based STR, but rather a distinct research problem that demands specialized architectures, datasets, and evaluation protocols, particularly for complex scripts such as Arabic.

1.1.5 Industrial and Societal Applications of Video Scene Text Recognition

Text appearing in video streams constitutes a rich source of semantic information that is highly valuable across numerous industrial and societal domains. Video Scene Text Recognition enables automatic extraction of this information, transforming raw visual content into structured and searchable data. As the volume of video data continues to grow rapidly due to surveillance systems, mobile devices, and online media platforms, the importance of reliable VSTR systems becomes increasingly evident.

In industrial settings, VSTR plays a critical role in applications such as intelligent transportation systems, where text recognition is used for vehicle identification, license plate reading, and traffic sign understanding. Retail and logistics sectors benefit from recognizing product labels, serial numbers, and shipping information captured in video streams. Similarly, in smart city infrastructures, VSTR contributes to automated monitoring, navigation assistance, and real-time information dissemination.

From a societal perspective, VSTR enhances accessibility by enabling real-time translation and subtitle generation for multilingual video content. It also supports media indexing and retrieval, allowing efficient search within large video archives based on textual content. In the context of public safety and security, text recognition in surveillance footage can assist in forensic analysis and situational

awareness.

Despite these advancements, Arabic-language video content remains significantly underrepresented in existing VSTR systems. Many deployed solutions are optimized for Latin-based scripts, limiting their effectiveness in regions where Arabic is predominant. Addressing Arabic VSTR not only expands the technological inclusivity of text recognition systems but also has direct implications for digital equity, cultural preservation, and regional technological development.

1.1.6 Limitations of Existing Video Text Recognition Pipelines

Although significant progress has been made in both scene text recognition and video analysis, existing video text recognition pipelines still exhibit notable limitations. Most current systems adopt a modular design in which text detection, tracking, and recognition are treated as separate stages. While this decomposition simplifies system design, it often leads to error propagation, where inaccuracies in early stages negatively impact subsequent components.

Another major limitation lies in the inadequate exploitation of temporal information. Many approaches process video frames independently or apply simple temporal smoothing techniques, failing to fully leverage the rich contextual information available across frames. This shortcoming is particularly detrimental in challenging conditions, such as low-resolution videos or scenes affected by motion blur, where single-frame recognition is unreliable.

Furthermore, the majority of VSTR models are trained and evaluated on datasets dominated by Latin scripts, resulting in architectural and representational biases. These models often assume left-to-right reading order, limited character connectivity, and relatively simple character shapes. Such assumptions do not generalize well to Arabic script, which is inherently cursive, context-dependent, and written from right to left.

Finally, the scarcity of large-scale, annotated Arabic video and scene text datasets hampers the development and fair evaluation of Arabic VSTR systems. This data deficiency restricts the ability of deep learning models to generalize across diverse real-world scenarios. Consequently, there is a pressing need for dedicated research efforts that address these limitations through script-aware modeling, specialized datasets, and end-to-end architectures tailored to Arabic video scene text recognition.

1.1.7 Justification for Specialized Arabic Video STR Research

The escalating global population of Arabic speakers, which comprises nearly one-quarter of the worldwide total, has driven a significant increase in demand for efficient tools and systems capable of handling Arabic language processing in real-world applications. This necessity provides a crucial socio-technological justification for specialized research in Arabic Scene Text Recognition (ASTR).

The direct application of models designed for Latin scripts to Arabic text is inherently insufficient due to the complex structural and linguistic characteristics of the Arabic script. Furthermore, while deep learning has advanced the state-of-the-art for Latin characters, its use for the Arabic language has been comparatively limited. This inherent complexity justifies the necessity for focused research into ASTR from videos, requiring comprehensive systems built upon architectural, data, and algorithmic foundations specifically adapted to overcome these unique challenges [1] [2].

1.2 Text Extraction from Images

Text extraction from images, particularly within the dynamic context of video, is the crucial first stage of the overall recognition pipeline. This stage is responsible for accurately identifying, isolating, and tracking the spatial regions that contain text, typically yielding bounding boxes or polygonal masks, which are then passed to the recognition engine [3].

1.2.1 The Role of Text Detection

Text detection, or localization, is often categorized by the density and structure of the text being sought. General STR focuses on sparse text instances like street signs in photographs, providing the full extracted string and bounding boxes for individual words. In contrast, OCR is optimized for dense text and structured documents, delivering hierarchical information detailing the page, block, and word structure [3].

Early methods for detection relied on handcrafted feature engineering, such as filter-based approaches (e.g., Fourier-Laplacian filters combined with K-means clustering) or edge and connected component analysis (e.g., Canny edge detectors and Maximally Stable Extremal Regions, MSER) [1].

1.2.2 Advancements in Deep Learning Localization

Deep learning transitioned text detection toward robust, end-to-end models, capable of handling the variability of natural scenes, including multilingual, curved, multi-oriented, or highly elongated text instances [4].

Modern detection algorithms often utilize Fully Convolutional Networks (FCNs) [5]. Recent state-of-the-art methods, such as those building upon **Differentiable Binarization Network (DBNet)**, integrate an adaptive binarization step directly into the network training, allowing the binarization threshold for each pixel to be learned by the network. Table 1.3 contrasts the methodologies of early detection versus modern deep learning localization.

Table 1.3: Advancements in Text Localization Methodologies

Methodology	Early/Traditional Methods	Deep Learning (DBNet/FCN)
Segmentation	Hard thresholds/K-means	Differentiable Binarization
Shape Handling	Horizontal boxes only	Arbitrarily curved/oriented
Pipeline	Multi-stage/Hand-crafted	End-to-end Trainable
Resolution	Sensitive to pixel noise	Multi-scale Feature Fusion

This innovation reduces post-processing and enables accurate detection of text with arbitrary shapes, enhancing feature expression for multi-scale scene texts [4].

1.2.3 Text Tracking in Video

For video applications, text detection is often supplemented by **text tracking**, which leverages the temporal correlation of text across successive frames to enhance the overall accuracy of localization and maintain continuity. An effective tracking framework may use a state-of-the-art scene text detection module on each frame, followed by a multi-strategy technique that utilizes tracking by detection, spatio-temporal context learning, and linear prediction to track the candidate text regions with high precision. The combination of detection and tracking is essential for VSTR, ensuring that fleeting or moving text in video is accurately captured before the recognition step begins.

1.3 Text Recognition

Text recognition, or transcription, is the subsequent, critical process of converting the localized textual image content into a machine-readable sequence of characters or words. The evolution of research in this domain reflects a continuous effort to better model the contextual and sequential dependencies inherent in human language [6].

1.3.1 From Classification to Recurrent Sequence Modeling

Historically, recognition was first approached as a classification problem involving manually segmenting text into individual characters and classifying each one. This method proved severely limited, particularly due to the difficulty of accurately segmenting or "cutting" words, a problem compounded in cursive scripts like Arabic [7]. This bottleneck mandated a transition toward sequence-to-sequence architectures, which treat the text input not as isolated components but as a single sequential data stream, thereby avoiding explicit character segmentation.

The most successful approaches for Scene Text Recognition (STR) for many years utilized hybrid models combining Convolutional Neural Networks (CNNs) for spatial feature extraction with Recurrent Neural Networks (RNNs) for sequential processing [8].

- **CNN-RNN Architecture:** The typical pipeline involved a CNN encoder generating feature sequences from the input image, followed by a bidirectional RNN decoder (often based on Long Short-Term Memory, LSTM) to process these features in order [6]. Sequence labeling was typically achieved using techniques such as Connectionist Temporal Classification (CTC).
- **Limitations of Recurrence:** Despite their initial success, RNNs/LSTMs exhibited fundamental architectural shortcomings: difficulty in maintaining context across very long sequences (long-range dependency problem) and computational inefficiency due to their sequential, mono-execution nature, which prohibits parallel processing during training [8].

1.3.2 The Shift Toward Attention and Novel Architectures

The emergence of the **Transformer architecture**, which entirely dispenses with both recurrence and convolution in favor of a Multi-Head Attention mechanism, offered a compelling solution to the limitations of recurrent models. This attention-only design allows for comprehensive parallel computation and superior capture of long-range dependencies, establishing a new paradigm for sequence modeling.

To clarify the technical improvements this architectural shift brings to the Arabic domain, Table 1.4 provides a comparison between the baseline recurrent methods and the attention-based approach utilized in this research.

Table 1.4: Technical Comparison: Recurrent Modeling vs. Attention Mechanisms

Aspect	RNN/LSTM Models	Transformer Models
Computation	Sequential (step-by-step)	Parallelizable (high throughput)
Context Range	Decays over long sequences	Global (constant distance)
Arabic Utility	Struggles with long ligatures	Captures distant dependencies
Training Speed	Low (limited by time-steps)	High (efficient GPU usage)

The research presented in this thesis addresses these shortcomings by proposing a **novel Transformer-based structure** for Arabic Scene Text Recognition (ASTR) [1]. The objective is to rigorously validate the capacity of this attention-based framework to overcome the scalability and dependency limitations of previous ASTR models, specifically tailoring it to the nuances of the Arabic script.

- **Architectural Innovation:** The thesis involves a comparative study of input pathways (CNN features versus Vision Transformer-style Image Patching) and the introduction of customized elements like a **learned positional encoding matrix** and a **learnable masking mechanism**, optimizing the architecture for accurate Arabic sequence prediction.

1.4 Arabic Text Challenges

The recognition of Arabic text in natural scenes, especially video, is complicated by a severe interplay between inherent linguistic complexities and the hostile nature of real-world visual environments, combined with a crucial resource limitation.

1.4.1 Inherent Linguistic and Structural Complexities

The structure of the Arabic script poses several layers of difficulty that are unique in the field of optical recognition, distinguishing it sharply from Latin-based scripts:

-
- **Cursive Nature and Directionality:** Arabic is written from right to left (RTL) and is characterized by mandatory cursiveness and a large set of ligatures (connected characters). This requires robust contextual sequence modeling, as characters are rarely isolated and the complex interplay of ligatures and diacritics increases computational overhead.
 - **Positional Shape Variation (Allography):** A single Arabic character assumes multiple distinct visual forms depending on its position within a word (isolated, initial/beginning, medial/middle, or final/connected). This variability dramatically expands the feature space that the recognition model must master.
 - **Visual Ambiguity and Character Similarity:** Many Arabic characters share nearly identical baseline shapes, differentiated only by subtle diacritical marks (nuqta). These slight visual distinctions make recognition extremely challenging, especially under degraded image conditions.

The high degree of allographic variability and character similarity is summarized in the tables below:

Table 1.5: Character Shape Variations

Transformed	Connected Double	Connected	Isolated
None	چ	ج	ج
ة	ت	ت	ت
None	ك	ك	ك

Table 1.6: Character Similarity

sound English	Middle	Beginning	Isolated
Ba	ب	ب	ب
Ta	ت	ت	ت
Ya	ي	ي	ي

Beyond basic positional variation, the concept of ligatures (the aesthetic and mandatory merging of specific character sequences) adds a layer of complexity not present in Latin scripts. Table ?? highlights common ligature examples that are historically difficult for traditional segmentation-based OCR systems to parse.

Difficulty	Form Ligature	Form Standard	Characters
Overlap Vertical	لا	ل ا	ل + ا
Stacking Vertical	لم	ل م	ل + م
Deformation Geometric	كل	ك ل	ل + ك

Table 1.7: Complexity of Arabic Ligatures

1.4.2 Environmental Challenges and Data Rarity

In addition to linguistic complexities, Arabic scene and video text recognition faces significant environmental challenges. Text appearing in natural scenes is often subject to perspective distortion, uneven lighting, shadows, motion blur, and occlusion. In video sequences, these factors are further exacerbated by camera motion, frame compression artifacts, and varying focus across frames.

Arabic text in real-world scenes is commonly encountered in diverse contexts such as street signs, storefronts, advertisements, and broadcast media. These texts vary widely in font style, color, size, and background contrast, making it difficult to design models that generalize effectively. Decorative calligraphic styles, frequently used in Arabic signage, further complicate recognition by deviating from standard typographic conventions.

A critical obstacle in addressing these challenges is the scarcity of large-scale, annotated Arabic datasets, particularly for scene and video text. Compared to Latin-based datasets, publicly available Arabic resources are limited in size, diversity, and annotation quality. This data imbalance restricts the ability of deep learning models to learn robust representations and leads to biased evaluation when models are tested primarily on non-Arabic benchmarks.

The lack of standardized evaluation protocols and benchmarks for Arabic scene and video text recognition further complicates comparative analysis across studies. As a result, reported performance metrics may not accurately reflect real-world applicability. Addressing environmental variability and data scarcity therefore requires both the creation of dedicated datasets and the development of models capable of learning from limited and heterogeneous data sources.

Crucially, the development of robust deep learning models is critically impeded by the **severe lack of large-scale, standardized, and diverse public datasets** specifically tailored for Arabic Scene Text Recognition. Existing real-world datasets like ALiF [9] and AcTiV 2.0 [10] are limited in scale, which is insufficient to train data-hungry deep architectures effectively.

To address this critical data rarity, this thesis introduced and published the **IYaD Dataset**, a large-scale artificial corpus specifically designed for Arabic printed text recognition in natural scene videos. Further details about this contribution are presented in chapter 3.

-
- **Scale and Diversity:** The IYaD dataset is reported to be the largest dedicated to this subject, containing approximately **44.8 million images** in total, covering 17 distinct Arabic fonts.
 - **Composition:** Each image is provided in three distinct versions (white background, randomly colored background, and real-world background simulation) to maximize diversity and robustness, a factor empirically proven to be essential for achieving high model performance.

This dataset is a cornerstone of the research, providing the necessary resource foundation to train and validate the complex recognition architectures proposed in this thesis.

1.4.3 Arabic Script Versus Latin Script: A Comparative Perspective

A comparative analysis between Arabic and Latin scripts highlights why direct transfer of text recognition techniques is often ineffective. Latin-based scripts typically feature discrete characters with limited shape variation and minimal contextual dependency. These properties simplify segmentation and recognition tasks and align well with common modeling assumptions used in existing OCR and STR systems.

In contrast, Arabic script exhibits high intra-class variability due to contextual shaping and ligature formation. Word-level recognition becomes more appropriate than character-level segmentation, shifting the modeling focus toward sequence-based and attention-driven approaches. Additionally, right-to-left reading order necessitates careful handling of positional encoding and decoding strategies to preserve linguistic coherence.

From a modeling standpoint, architectures trained predominantly on Latin scripts may implicitly learn biases related to character spacing, alignment, and directionality. When applied to Arabic text, these biases can manifest as systematic errors, such as reversed sequences or incorrect character grouping. This mismatch underscores the importance of script-aware architectural design and training strategies.

Understanding these fundamental differences is essential for motivating the specialized methods proposed in this thesis. Rather than treating Arabic as a mere extension of existing STR frameworks, this work approaches Arabic scene text recognition as a distinct problem domain requiring tailored datasets, modeling choices, and evaluation methodologies.

1.5 Research Problem Formulation and Thesis Objectives

1.5.1 Synthesizing the Research Deficit

The field of Arabic Text Extraction and Recognition from Videos is characterized by a persistent and multifaceted research gap. Although substantial progress has been achieved in scene text recognition for Latin-based scripts, these advancements do not readily transfer to Arabic text appearing in unconstrained visual environments. The Arabic script exhibits intrinsic linguistic and visual complexities, including cursive character connectivity, contextual letter shaping, and right-to-left reading order, which fundamentally challenge conventional recognition paradigms.

Most existing approaches to Arabic Scene Text Recognition (ASTR) rely on recurrent sequence modeling techniques, such as recurrent neural networks and their gated variants. While these models have demonstrated effectiveness in modeling sequential dependencies, they suffer from limited scalability, sequential computation constraints, and high computational cost when applied to long visual sequences. These limitations become particularly pronounced in video-based scenarios, where temporal redundancy and frame-level variability significantly increase the volume and complexity of data to be processed.

In parallel, the development of robust ASTR systems is severely constrained by the lack of sufficiently large, diverse, and publicly available training datasets. Existing Arabic datasets are predominantly focused on scanned documents or handwritten text and fail to capture the visual diversity inherent in natural scenes and video content. This scarcity of representative data prevents deep learning models from learning generalizable feature representations and leads to overfitting and performance degradation in real-world deployments [2].

Consequently, the combination of architectural inadequacies and data limitations results in ASTR systems that struggle to generalize across variations in font styles, backgrounds, lighting conditions, and motion-induced distortions. Addressing this research deficit requires a comprehensive rethinking of both model architecture and data generation strategies, specifically tailored to the characteristics of Arabic text in dynamic visual environments.

1.5.2 Statement of Purpose and Primary Contribution

The purpose of this thesis is to provide a foundational, end-to-end solution to the identified deficits in Arabic Scene Text Recognition from videos. Rather than incrementally adapting existing recognition pipelines, this work proposes a holistic framework that simultaneously addresses architectural inefficiencies and

data scarcity, two of the most critical barriers to progress in the ASTR domain.

The primary contribution of this thesis is the design, implementation, and rigorous validation of a **novel Multi-head Attention Transformer architecture** optimized for Arabic Scene Text Recognition. By abandoning recurrent sequence modeling in favor of a fully attention-based encoder-decoder structure, the proposed approach enables efficient parallel computation, improved scalability, and enhanced modeling of long-range contextual dependencies. This architectural choice is particularly well suited to handling the complex spatial and sequential characteristics of Arabic script in unconstrained scenes.

In addition to architectural innovation, this thesis introduces the **IYaD Dataset**, a purpose-built, large-scale, and highly diverse synthetic corpus specifically designed for Arabic scene text recognition. The dataset is constructed to capture a wide range of visual variations, including diverse fonts, backgrounds, text lengths, and noise conditions, thereby addressing the critical lack of training resources in this research area. Together, the proposed model and dataset constitute a unified contribution that advances both methodological and empirical aspects of Arabic text recognition research.

1.5.3 Core Thesis Objectives

To achieve the overarching aims of this research, the thesis is guided by a set of clearly defined and interrelated objectives. These objectives are formulated to ensure systematic investigation, rigorous validation, and measurable advancement over existing approaches in the field of Arabic Scene Text Recognition.

- **O1: Data Standardization and Resource Creation:** To generate, curate, and publicly release the IYaD dataset, providing the necessary scale (44.8 million images) and diversity to enable effective training, evaluation, and generalization of deep learning models for Arabic Scene Text Recognition.
- **O2: Architectural Innovation and Validation:** To rigorously investigate and validate the suitability of a non-recurrent Transformer encoder-decoder architecture for Arabic Scene Text Recognition by conducting comparative experiments across multiple feature representation strategies and baseline models.
- **O3: Contextual Adaptation for Arabic Script:** To design and integrate specialized architectural adaptations within the Transformer framework that enhance the model’s ability to resolve positional ambiguity, cursive connectivity, and visually similar character patterns inherent in the Arabic script.

-
- **O4: Benchmarking and State-of-the-Art Confirmation:** To establish new performance benchmarks for Arabic Scene Text Recognition by demonstrating that the proposed Transformer-based architecture, when trained on the IYaD dataset, significantly outperforms existing state-of-the-art recurrent and hybrid recognition methods in terms of accuracy and robustness.

1.6 Thesis Contributions and Unique Approach

The findings of this research deliver substantial contributions to the fields of deep learning, computer vision, and Arabic language processing.

1.6.1 Data Contribution: The IYaD Dataset

The necessity for specialized training resources was fulfilled by the creation of the IYaD dataset. This artifact is a significant contribution, providing a massive, diverse, and publicly available resource specifically tailored for Arabic text recognition research [11]. IYaD dataset was constructed to contain approximately 1.4 million images per font across 17 distinct fonts, resulting in a total of more than 44 million individual samples at the time of publication. Each sample is provided in three distinct versions: plain, randomly colored, and realistic scene backgrounds—accompanied by both Arabic ground-truth labels and Latin script transcriptions**, making the dataset suitable for recognition and cross-language evaluation alike. Although the dataset is artificially generated, comprehensive complexity analyses have demonstrated that it effectively simulates the visual variability of real-world scene text, including distortions and background variability. Furthermore, IYaD is published with processing tools that facilitate easy integration into deep learning workflows, including automated label transformation utilities and synthetic augmentation pipelines, ensuring both usability and long-term sustainability for research in Arabic scene text recognition.



Figure 1.1: Examples of Arabic scene text samples from IYaD dataset.

1.6.2 Transformer-Based Approach for Robust Arabic Scene Text Recognition

This thesis addresses the fundamental limitations of traditional image-based OCR and scene text recognition systems when applied to video environments. While most existing approaches process each frame independently, they fail to exploit the temporal continuity inherent in video sequences. Such frame-wise processing is highly sensitive to motion blur, illumination changes, geometric distortions, occlusions, and low-resolution text regions, which are common in real-world video scenarios.

To bridge this gap, this work proposes and implements a Transformer-based Arabic Scene Text Recognition (ASTR) architecture specifically designed to handle complex visual variations and contextual dependencies in scene text. Unlike conventional convolutional or recurrent-based OCR pipelines, the Transformer model leverages self-attention mechanisms to capture long-range spatial relationships between character components. This is particularly important for Arabic script, where characters are cursive, context-dependent, and often visually similar.

The proposed architecture integrates deep visual feature extraction with multi-head self-attention layers, allowing the model to dynamically focus on relevant text regions and character sequences. By modeling global dependencies across feature maps, the Transformer effectively overcomes the limitations of local receptive fields commonly observed in convolution-based approaches.

Moreover, the methodological contribution of this thesis includes extensive optimization of the Transformer structure for Arabic scene text, including the design of suitable positional encoding strategies, attention head configurations, and feature embedding representations. These modifications enable the model to better preserve spatial ordering and contextual coherence within Arabic words and sentences.

The effectiveness of the proposed approach is demonstrated through comprehensive experiments on synthetic and real-world datasets. The Transformer-based ASTR model achieved state-of-the-art recognition performance, with a character recognition rate exceeding 93.5% on single-font test scenarios [1]. This performance surpasses several existing OCR and STR frameworks, highlighting the robustness of the Transformer architecture for Arabic scene text recognition.

Although the primary focus of this thesis is on static image-based recognition, the developed Transformer framework establishes a strong foundation for future extension to video-based STR systems. The attention-based structure is inherently suitable for incorporating temporal feature aggregation mechanisms, such as temporal attention or sequence-level fusion, which can further improve recognition stability across video frames.

In summary, this thesis contributes a novel Transformer-based ASTR archi-

tecture that significantly improves Arabic scene text recognition accuracy while addressing many of the structural limitations present in traditional OCR systems. The proposed model not only advances the state of the art in Arabic OCR but also provides a scalable and flexible framework for future video scene text recognition research.

1.7 Thesis Structure and Organization

This thesis is structured into four chapters to systematically address the established research problem, present the novel architecture and dataset, and rigorously evaluate the empirical findings.

- **Chapter 2: Contextual Foundations and Prior Works** This chapter provides a detailed review of the literature on Optical Character Recognition and Scene Text Recognition, focusing particularly on methodologies for text detection, text tracking in video, sequence modeling (RNN, Attention), and existing research specific to Arabic script.
- **Chapter 3: The IYaD Dataset: Creation, Structure, and Validation** This chapter provides a comprehensive technical overview of the IYaD dataset. It details the methodology for synthetic data generation, the structure of the dataset (words vs. sentences, fonts, background variations), and quantitative analysis justifying its complexity and suitability for training data-intensive deep learning models.
- **Chapter 4: Transformer Architecture for Arabic Scene Text Recognition** This chapter provides the complete technical specification of the proposed novel recognition model. It details the core encoder-decoder structure, the input processing pipeline, and thoroughly describes the innovative architectural elements.

We finish with a conclusion where we synthesizes the successful achievement of the thesis objectives, summarizes the impact of the architectural innovations and the IYaD dataset contribution, and outlines promising avenues for future research, including real-time system integration and expansion to multilingual

Chapter 2

Related Works

Automatic text recognition has progressed significantly over the past decades, driven by advances in computer vision, artificial intelligence, and deep learning. However, Arabic text extraction and recognition, especially from natural scenes, remains a challenging problem compared to Latin-based languages. This chapter reviews major contributions and existing datasets related to Optical Character Recognition (OCR), scene-text recognition, and Arabic script processing. The objective is to situate our research within the broader context and highlight the need for a large-scale, structured dataset such as IYaD [11].

2.1 Early OCR and Scene Text Recognition

Early OCR approaches relied primarily on handcrafted features and classical machine learning techniques. These methods often exploited edge detection, contour-based features, morphological operations, or statistical shape descriptors to identify characters in clean, printed documents [12, 13]. While effective for controlled documents, such systems were limited in generalizing to complex natural scenes containing distortions, occlusions, or varying illumination.

With the rise of neural networks, researchers began employing convolutional architectures (CNNs) to automatically learn hierarchical features from raw image pixels. LeCun et al. [13] demonstrated the effectiveness of CNNs for document recognition, laying the groundwork for subsequent scene-text recognition models. Similarly, early scene-text localization approaches [14] employed stroke feature transforms and text covariance descriptors to detect text in natural images, demonstrating the shift from handcrafted to learned features.

The integration of recurrent networks and sequence modeling further improved text recognition, allowing systems to handle variable-length text and cursive scripts. Karwal et al. [12] and Abdullah et al. [15] provide comprehensive surveys highlight-

ing the evolution from traditional OCR to deep learning-based scene-text recognition.

2.1.1 Challenges in Early Systems

Early systems were constrained not only by feature representation but also by computational limitations. Memory and processing power restricted the complexity of models, and handcrafted features often failed when applied to noisy or distorted text. Moreover, scripts with connected letters, such as Arabic, were particularly difficult to segment accurately, resulting in high error rates in character recognition. These limitations motivated the exploration of more robust learning-based methods.

2.2 Scene Text Detection and Recognition

Scene-text recognition has evolved significantly over the past decade, driven by advances in deep learning architectures that have substantially improved both accuracy and robustness. Traditional recognition approaches relied heavily on handcrafted features and sequential processing pipelines, which often struggled with the variability present in natural scenes. The introduction of convolutional neural networks (CNNs), recurrent neural networks (RNNs), and later attention-based mechanisms transformed the field by enabling end-to-end learning directly from image data. CNNs became particularly effective at extracting rich hierarchical visual features, while RNNs were employed to model the sequential nature of characters and words, allowing the recognition of text with varying lengths and layouts. Together, these architectures significantly improved the ability of recognition systems to handle multi-oriented, curved, distorted, and multi-scale text appearing in complex and unconstrained environments.

Several notable works demonstrated the effectiveness of these deep learning approaches. For example, Zhang et al. [5] proposed deeper CNN architectures capable of learning increasingly complex feature representations, leading to substantial improvements in recognition accuracy across challenging datasets. Similarly, Deng [16] introduced a real-time video text spotting framework specifically designed for dynamic scenes, addressing challenges such as object motion, varying viewpoints, and temporal consistency while maintaining computational efficiency. Beyond high-performance research models, lightweight recognition pipelines suitable for embedded and resource-constrained systems were also developed. Burie et al. [17] demonstrated that efficient scene-text recognition could be deployed on embedded platforms, highlighting the growing importance of practical, real-world applications such as mobile devices, autonomous vehicles, wearable technologies,

and assistive systems for visually impaired users.

The advent of attention mechanisms, particularly following the Transformer architecture [18], marked another major milestone in scene-text recognition. Unlike conventional sequential models, attention mechanisms enable networks to selectively focus on the most informative regions of an image during the recognition process, thereby improving their ability to recognize long text sequences and complex layouts. This approach reduces the limitations associated with recurrent architectures while providing a more effective means of modeling long-range dependencies between characters and words. Building upon these advances, multimodal networks [19, 20] further enhanced recognition performance by integrating visual features with contextual and semantic information. By combining multiple sources of information, these models became better equipped to resolve ambiguities caused by visually similar characters, incomplete text, or noisy backgrounds.

The combination of attention-based learning and multimodal feature fusion has significantly advanced the state of the art in scene-text recognition. Modern recognition systems demonstrate remarkable robustness under challenging real-world conditions, including cluttered backgrounds, irregular lighting, motion blur, perspective distortion, partial occlusions, and low-contrast text [21, 22]. Furthermore, these developments have improved the generalization capabilities of recognition models across diverse languages, scripts, and writing styles, making them suitable for multilingual applications. As a result, scene-text recognition has become an essential component of numerous intelligent vision systems, including autonomous navigation, document digitization, augmented reality, intelligent transportation, and assistive technologies, where reliable text understanding is critical for effective decision-making.

2.2.1 Advances in Multi-Oriented Text Recognition

Text in natural scenes often appears in arbitrary orientations, curved paths, or distorted perspectives. To address this, methods such as Spatial Transformer Networks (STNs) [23] and Rectification Networks were proposed to normalize text regions before recognition. These approaches significantly improved performance for non-horizontal text, a common occurrence in Arabic signage and decorative scripts.

2.2.2 Real-Time Recognition Systems

Real-time recognition has become a crucial requirement for applications such as navigation, augmented reality, and assistive technologies. Lightweight CNN architectures, combined with efficient decoding strategies, allow text recognition systems to operate at interactive frame rates. Works such as [?] introduced efficient

text detectors, while [?] combined detection and recognition in end-to-end pipelines suitable for mobile and embedded platforms.

Despite these advances, most scene-text recognition datasets and models target Latin scripts, leaving Arabic text underrepresented.

2.3 Arabic Text Recognition

Arabic script presents unique challenges due to its cursive morphology, context-dependent shapes, ligatures, diacritics, and high visual similarity among characters [15, 24]. The cursive nature means that character boundaries are often ambiguous, and contextual dependencies must be modeled carefully to avoid misclassification. Early Arabic OCR systems primarily targeted printed or handwritten text in controlled environments [25, 26, 27], while scene-text applications remained limited.

Recent works employ hybrid CNN-RNN architectures with attention to address these challenges. For example, Butt et al. [6] proposed an attention-based CNN-RNN model for Arabic text in natural scenes, demonstrating improved robustness to font variability, orientation, and noise. Similarly, Khial and Bouferra [1] introduced multi-head attention transformers specifically tailored for Arabic scene images, highlighting the importance of self-attention mechanisms in capturing long-range dependencies in cursive text.

2.3.1 Modeling Contextual Dependencies

Arabic characters often change shape depending on their position in a word. To address this, contextual modeling using RNNs or transformers has become essential. Sequence-to-sequence approaches with attention can dynamically weigh character contributions, improving recognition for ligatures and diacritics. Additionally, combining global word-level features with local character-level features has been shown to reduce misclassification rates in complex scripts [?].

2.3.2 Handling Low-Resource Scenarios

Despite the growth of deep learning, Arabic OCR suffers from a relative scarcity of large-scale annotated datasets. Techniques such as transfer learning, semi-supervised learning, and synthetic augmentation have been explored to mitigate data scarcity. Researchers have leveraged pretraining on Latin-script datasets followed by fine-tuning on Arabic text, demonstrating that knowledge transfer can improve generalization [?].

2.3.3 Arabic Video Text Recognition

Arabic text recognition in video sequences is an extension of image-based OCR that introduces additional challenges. Unlike static images, video frames contain temporal information that can be leveraged to improve recognition accuracy, especially when characters are partially occluded, blurred, or appear in complex scenes. Each frame may provide complementary information, and aggregating this temporal context can significantly enhance the performance of Arabic OCR systems.

Early approaches to Arabic video text recognition relied on classical pipelines such as tracking-based methods and frame-wise OCR with post-processing [?]. These methods focused on detecting text in each frame separately and then merging recognition results across frames, often using simple averaging or voting strategies. While effective for structured videos, they struggled with real-world dynamic scenarios involving motion blur, lighting variations, or rapidly changing backgrounds.

Recent advances in deep learning provide more powerful solutions. Convolutional neural networks (CNNs) extract robust spatial features from individual frames, while recurrent networks (RNNs) and Transformers enable modeling of sequential dependencies across time. In particular, attention-based mechanisms allow the model to focus on the most informative frames and characters, effectively performing temporal feature aggregation [28, 29, 30]. Formally, if a video sequence is represented as $V = \{I_1, I_2, \dots, I_T\}$, where T is the number of frames, each frame is encoded into a feature vector $F_t = \phi(I_t)$. Temporal aggregation then produces a combined representation:

$$F_{agg} = \sum_{t=1}^T \alpha_t F_t, \quad \sum_{t=1}^T \alpha_t = 1$$

where α_t are learned attention weights reflecting the importance of each frame. This allows the model to prioritize frames where Arabic characters are clearer or less occluded.

Several temporal aggregation strategies have been explored in video analysis that can be adapted for Arabic text recognition. TEA (Temporal Excitation and Aggregation) modules capture both short- and long-range temporal dependencies efficiently [28]. ActionVLAD aggregates spatio-temporal descriptors into a single video representation using learnable pooling [29]. Temporal Adaptive Modules (TAM) introduce frame-specific adaptive weights for more flexible aggregation [30]. Even techniques from video captioning and action recognition, such as temporal attention or parallel temporal networks, provide insights for effectively combining information across frames [31, 32].

Although the present research focuses on image-based recognition, understand-

ing temporal aggregation is essential for future extensions to video. Integrating temporal feature aggregation modules could help in:

- Reducing the effect of motion blur and occlusion.
- Combining complementary visual information from multiple frames.
- Stabilizing text detection and improving recognition accuracy in dynamic environments.

In summary, Arabic video text recognition benefits greatly from temporal feature aggregation. While classical methods relied on post-processing and simple frame fusion, modern deep learning approaches leverage attention and adaptive aggregation to produce robust, high-quality representations. Future work could extend the proposed Transformer-based framework to include temporal modeling, enabling a unified system capable of handling both static images and video sequences efficiently.

2.3.4 Arabic Video Text Recognition

Arabic text recognition in video sequences is an extension of image-based OCR that introduces additional challenges. Unlike static images, video frames contain temporal information that can be leveraged to improve recognition accuracy, especially when characters are partially occluded, blurred, or appear in complex scenes. Each frame may provide complementary information, and aggregating this temporal context can significantly enhance the performance of Arabic OCR systems.

Early approaches to Arabic video text recognition relied on classical pipelines such as tracking-based methods and frame-wise OCR with post-processing [?]. These methods focused on detecting text in each frame separately and then merging recognition results across frames, often using simple averaging or voting strategies. While effective for structured videos, they struggled with real-world dynamic scenarios involving motion blur, lighting variations, or rapidly changing backgrounds.

Recent advances in deep learning provide more powerful solutions. Convolutional neural networks (CNNs) extract robust spatial features from individual frames, while recurrent networks (RNNs) and Transformers enable modeling of sequential dependencies across time. In particular, attention-based mechanisms allow the model to focus on the most informative frames and characters, effectively performing temporal feature aggregation [28, 29, 30]. Formally, if a video sequence is represented as $V = \{I_1, I_2, \dots, I_T\}$, where T is the number of frames, each frame is encoded into a feature vector $F_t = \phi(I_t)$. Temporal aggregation then produces a combined representation:

$$F_{agg} = \sum_{t=1}^T \alpha_t F_t, \quad \sum_{t=1}^T \alpha_t = 1$$

where α_t are learned attention weights reflecting the importance of each frame. This allows the model to prioritize frames where Arabic characters are clearer or less occluded.

Several temporal aggregation strategies have been explored in video analysis that can be adapted for Arabic text recognition. TEA (Temporal Excitation and Aggregation) modules capture both short- and long-range temporal dependencies efficiently [28]. ActionVLAD aggregates spatio-temporal descriptors into a single video representation using learnable pooling [29]. Temporal Adaptive Modules (TAM) introduce frame-specific adaptive weights for more flexible aggregation [30]. Even techniques from video captioning and action recognition, such as temporal attention or parallel temporal networks, provide insights for effectively combining information across frames [31, 32].

Although the present research focuses on image-based recognition, understanding temporal aggregation is essential for future extensions to video. Integrating temporal feature aggregation modules could help in:

- Reducing the effect of motion blur and occlusion.
- Combining complementary visual information from multiple frames.
- Stabilizing text detection and improving recognition accuracy in dynamic environments.

In summary, Arabic video text recognition benefits greatly from temporal feature aggregation. While classical methods relied on post-processing and simple frame fusion, modern deep learning approaches leverage attention and adaptive aggregation to produce robust, high-quality representations. Future work could extend the proposed Transformer-based framework to include temporal modeling, enabling a unified system capable of handling both static images and video sequences efficiently.

2.4 Existing Arabic OCR Datasets

Several datasets for Arabic OCR have been proposed, but most suffer from limited size, vocabulary constraints, or lack of environmental diversity. Representative examples include:

- **AIA9K** [33] — focuses on isolated Arabic characters and digits.

-
- **MOUDLE** [34] — contains text extracted from various document sources.
 - **ALIF** [9] — Arabic text embedded in TV broadcasts.
 - **APTID/MF** [35] — printed Arabic text database.
 - **ADAB** [27] — handwritten Arabic characters for on-line recognition.

While useful, these datasets typically suffer from one or more limitations: small size, limited fonts, absence of scene backgrounds, or lack of structured annotations. This scarcity limits the applicability of deep learning models, which require large-scale, diverse training data [36, 2].

2.4.1 Scene-Text Arabic Datasets

Compared to Latin-script datasets, Arabic scene-text datasets are scarce. Some recent efforts, such as [11, 37], attempt to fill this gap, but they often remain limited in the number of annotated images or environmental variety. The lack of high-resolution, natural-scene Arabic text images remains a major bottleneck for developing robust recognition models.

2.5 Synthetic Datasets in OCR Research

Synthetic datasets are increasingly used to mitigate the lack of real annotated data. By programmatically generating text images with controlled variations in font, color, orientation, and background, synthetic datasets allow large-scale training without expensive manual labeling. Nayak et al. [38] and Zhao et al. [39] demonstrate the effectiveness of synthetic data in improving model robustness and generalization. In the Arabic context, Cheikh et al. [37] created synthetic Arabic datasets, though these remain limited compared to Latin-script datasets.

The use of synthetic data raises questions about realism and domain adaptation. Zewe [7] highlights that while synthetic data can accelerate model development, achieving realistic distributions of natural text remains challenging. Thus, datasets like IYaD aim to combine large-scale annotations with real-world scene diversity to bridge the gap.

2.5.1 Domain Adaptation for Synthetic Data

A key challenge with synthetic datasets is domain shift. Models trained solely on synthetic images often underperform when applied to natural scenes due to differences in lighting, background clutter, and noise. Domain adaptation techniques,

including adversarial training and style transfer, have been proposed to reduce this gap. Recent studies show that hybrid training strategies combining real and synthetic data yield the best performance for Arabic scene-text recognition [11, ?].

2.6 Domain-Specific and Multimodal Approaches

OCR research increasingly targets domain-specific and multimodal applications. Medical text [40], historical manuscripts [41], and multimodal vision-language systems [42] illustrate the trend toward richer, context-aware recognition. Similarly, video-based recognition [43, 44] incorporates temporal information, providing robustness against motion blur [45] and lighting variations [46].

Arabic scene-text recognition, however, lacks an equivalent large-scale, multimodal benchmark that combines real-world complexity, font diversity, and high-quality annotations.

2.6.1 Multilingual and Cross-Domain Recognition

While Arabic OCR research is often conducted in isolation, there is growing interest in multilingual and cross-domain recognition. Systems capable of handling Arabic alongside Latin, Chinese, or other scripts are valuable for global applications such as street sign recognition or multilingual document digitization. Techniques such as shared feature encoders and script-aware attention mechanisms are increasingly explored to address these challenges [?].

2.6.2 Arabic Street Sign Recognition and Real-World Applications

One of the most compelling real-world applications of Arabic scene-text recognition is street sign and public signage understanding. Arabic signs present unique challenges: cursive script, diverse fonts, variable lighting conditions, and frequent occlusions by objects or pedestrians. Additionally, signs may be multilingual, containing both Arabic and Latin characters, which complicates recognition pipelines.

Several studies have explored Arabic sign detection and recognition in urban environments. Early attempts [47] relied on color and edge-based segmentation, but these approaches struggled with cluttered backgrounds. More recent methods leverage deep learning, combining CNNs with attention mechanisms to capture both local character details and global word context.

Applications extend beyond simple text reading. Accurate recognition of street signs can enable autonomous navigation systems, smart city monitoring, and assistive technologies for visually impaired individuals. Moreover, integration with

geographic information systems (GIS) allows for real-time mapping and translation of Arabic signage, which is particularly relevant for navigation apps and augmented reality interfaces.

Despite these advancements, research in Arabic street sign recognition is still limited by the availability of high-quality annotated datasets. Existing benchmarks are small, lack diversity in environmental conditions, or are focused on a narrow set of sign types. The IYaD dataset addresses these limitations by providing extensive scene-text images covering a wide variety of urban and indoor environments, multiple fonts, orientations, and text styles. This makes it a critical resource for both academic research and practical deployment of Arabic OCR systems in real-world applications.

2.7 2.7 Summary

The literature demonstrates significant advances in OCR and scene-text recognition, particularly for English and other Latin scripts. Despite progress, Arabic OCR faces unique challenges: cursive morphology, contextual dependencies, and limited dataset availability. Existing Arabic datasets, whether real or synthetic, are insufficient to fully exploit modern deep learning techniques.

This gap motivates the creation of the IYaD dataset [11], which provides a high-quality, large-scale resource for Arabic text recognition in natural scenes. IYaD combines millions of annotated examples with diverse fonts, scene backgrounds, and sentence-level text sequences, offering a benchmark suitable for both traditional and attention-based recognition models. The next chapter details the construction, design principles, and characteristics of the IYaD dataset.

Chapter 3

IYaD dataset

3.1 Introduction

The creation of high-quality datasets is a fundamental step in advancing the field of text recognition, particularly for languages with complex scripts such as Arabic. While significant progress has been made in optical character recognition (OCR) for Latin-based languages, Arabic text recognition from images remains challenging due to its cursive nature, multiple character shapes, diacritics, and variations in fonts and writing styles. Moreover, the availability of large-scale, annotated datasets for Arabic scene text recognition is limited, restricting the development and evaluation of robust recognition models.

In this chapter, we present the design and construction of a new dataset specifically tailored for Arabic text recognition from images. The dataset aims to address existing gaps by providing a diverse, representative, and carefully annotated collection of Arabic text samples, including variations in font types, backgrounds, and image quality. This contribution forms the foundation for subsequent experiments and model development in this thesis.

The chapter is structured as follows:

Motivation and Requirements – we discuss the limitations of existing datasets, outline the unique challenges of Arabic text recognition, and define the design goals for our dataset.

Dataset Design and Collection Methodology – we describe the sources of data, the acquisition process, annotation procedures, and preprocessing steps applied to ensure high-quality and consistent data.

Dataset Statistics and Characteristics – we provide a detailed overview of the dataset, including size, distribution, sample images, and comparisons with existing resources.

Use Cases and Intended Applications – we outline the potential applications of

the dataset and its relevance to both research and practical Arabic OCR systems.

Limitations, Challenges, and Ethical Considerations – we discuss remaining limitations, challenges encountered during dataset creation, and ethical aspects related to data collection and usage.

Through this chapter, the reader gains a comprehensive understanding of the dataset’s design, composition, and intended use, establishing a solid experimental foundation for the subsequent contributions presented in this thesis. In particular, the dataset introduced in this chapter serves as a key resource for training and evaluating the recognition models developed in the following chapters, enabling a systematic study of Arabic text recognition in challenging real-world scenarios.

3.2 Motivation

Artificial neural networks (ANNs) have become the backbone of modern artificial intelligence (AI) applications due to their ability to model complex, non-linear relationships in data. Their applications span nearly every AI field, including object detection, machine translation, data labeling, and classification tasks. The effectiveness of ANNs is rooted in their relatively simple but highly expressive mathematical formulation. Many classical AI steps, such as preprocessing, feature selection, and feature extraction, which were previously performed manually or through engineered pipelines, are now learned automatically within ANNs through end-to-end training.

Modern neural network architectures, such as convolutional neural networks (CNNs) [48], long short-term memory networks (LSTMs) [49], and Transformers [50, 51], further extend the capabilities of ANNs by enabling them to capture hierarchical patterns, temporal dependencies, and long-range relationships in both spatial and sequential data. These advancements allow models to achieve state-of-the-art performance across a wide range of domains, from image recognition to natural language processing.

However, a careful analysis of how ANNs solve a given problem reveals a critical dependency: the data used for training serves as the foundation of all knowledge that the network consumes to adjust its trainable parameters. Regardless of the sophistication of the architecture or the training procedure, the quantity, quality, and diversity of the training data largely determine the model’s final performance. In domains where data is scarce or narrowly distributed, even the most advanced architectures may fail to generalize beyond the training distribution.

This issue is particularly pronounced in the field of Arabic Scene Text Recognition (ASTR). Table 3.1 provides a comparative overview of publicly available Arabic text datasets, categorized by their intended application, including handwritten text recognition, printed OCR, and scene text (printed ST). As can be

seen, datasets dedicated specifically to ASTR represent less than 16.66% of the total, and this percentage is likely an overestimation because the table lists only datasets relevant to ASTR and excludes datasets dedicated to other Arabic text processing tasks. This observation underscores a clear scarcity of training resources specifically designed for Arabic scene text, highlighting a critical bottleneck for research in this area.

Dataset	Handwritten	Printed OCR	Printed ST
IFN/ENIT	X		
AHDB/FTR	X		
ARABIC CHEQUES	X		
UPTI		X	
HACDB	X		
KHATT	X		
ALiF			X
ACTiV-R			X
SmartATID	X	X	
Printed PAW		X	
KAFD		X	
SmartATID	X	X	

Table 3.1: Arabic text detection datasets

On the other hand, despite the increasing complexity of ASTR compared to most traditional computer vision (CV) tasks, there exists a massive gap between the amount of data available for Arabic scene text and the datasets commonly used in other CV domains. Table 3.2 illustrates this gap numerically, highlighting the orders of magnitude difference between large-scale CV datasets and those available for Arabic text recognition.

Dataset	Number of images
Microsoft COCO 2017 Dataset	121,408
CIFAR-10	60,000
Open Images	900,000
PGM	142,000
IAM	13,000
ALiF	9,690
ACTiV	10,415

Table 3.2: Comparison of computer vision datasets

The numbers in Table 3.2 reveal the striking disparity between CV datasets and ASTR datasets. While general object detection and recognition tasks often rely on hundreds of thousands or even millions of labeled images, Arabic scene text datasets rarely exceed ten thousand samples. This lack of scale directly limits the performance of deep learning models, as it prevents them from learning robust representations for the highly variable and complex nature of Arabic script in real-world scenes. The scarcity of data also impedes the development of models capable of generalizing across different fonts, backgrounds, illumination conditions, and text distortions, all of which are common in natural video content.

In summary, the shortage of large, diverse, and publicly available datasets represents a fundamental obstacle in advancing Arabic Scene Text Recognition. Addressing this data gap is critical for enabling the training of modern neural architectures, improving model generalization, and ultimately achieving high-quality recognition performance in real-world applications. This observation motivates the development of large-scale synthetic datasets, such as the IYaD dataset introduced in this work, which aim to mitigate the limitations of existing resources and provide a solid foundation for training and evaluating advanced ASTR models.

3.3 Dataset Design and Collection Methodology

The primary objective of developing the IYaD dataset was to provide a large-scale, diverse, and high-quality benchmark for Arabic scene text recognition. Existing publicly available Arabic text datasets are often limited in terms of vocabulary size, visual diversity, environmental conditions, or annotation quality, making it difficult to train and evaluate modern deep learning models effectively. The IYaD dataset was therefore designed to address these limitations by offering a comprehensive collection of Arabic text images that closely resemble real-world scenarios while maintaining accurate and consistent annotations.

To achieve this objective, a systematic and reproducible data generation pipeline was developed, encompassing every stage of the dataset creation process. This pipeline includes text collection, image generation, annotation, data validation, and augmentation. Each stage was carefully designed to maximize the diversity of the dataset while ensuring the correctness and consistency of the corresponding ground-truth labels. Particular attention was given to representing the wide range of visual variations encountered in natural scenes, including differences in font styles, text orientation, scale, color, lighting conditions, background complexity, and image quality.

Since deep learning-based scene text recognition models require large amounts of annotated data, the dataset generation process was largely automated to enable the efficient creation of thousands of labeled samples. Automation not only accel-

erated the data collection process but also minimized human annotation errors and ensured uniform labeling across the entire dataset. Furthermore, the generation pipeline was designed to be extensible, allowing additional text samples, fonts, or augmentation techniques to be incorporated in future versions of the dataset.

3.3.1 Data Sources

The first stage in constructing the IYaD dataset involved collecting a large and diverse corpus of Arabic textual content from publicly available journals, articles, and other digital publications. These sources were selected because they contain naturally occurring text written by native speakers across a variety of domains, including news, education, science, technology, literature, and general interest topics. Using authentic written material ensured that the dataset would accurately represent the richness and diversity of the Arabic language rather than relying on artificially generated or manually curated word lists.

The collected documents were processed to extract words, phrases, and complete sentences that served as the textual input for the image generation pipeline. During this preprocessing stage, non-textual elements such as figures, tables, page numbers, formatting artifacts, and special symbols that were not relevant to scene text recognition were removed. The remaining text was then normalized and cleaned to improve consistency while preserving the original linguistic content. This process resulted in a high-quality corpus containing vocabulary of varying lengths and complexities, including common words, proper nouns, frequently used expressions, and complete sentences.

Using publicly available Arabic publications offers several advantages. First, it ensures that the dataset reflects real-world vocabulary and writing styles commonly encountered in everyday applications. Second, it provides broad linguistic coverage by incorporating text from multiple subject areas, thereby exposing recognition models to a wide range of lexical and semantic variations. Finally, because the source material originates from authentic documents, the resulting dataset better captures the statistical characteristics of natural Arabic language usage, making it more suitable for training and evaluating modern scene text recognition systems.

By basing the dataset on genuine Arabic textual resources, the IYaD dataset achieves a higher level of linguistic relevance and realism, enabling trained models to generalize more effectively to practical applications where text appears in diverse and unconstrained environments.

3.3.2 Image Generation Process

Each word or sentence was first written onto a blank white image of fixed dimensions (700×250 pixels). The text was placed at the top-left corner, and the

image was then cropped to retain only the text region, leaving some extra space around the edges to prevent the network from learning fixed borders as features.

Each text image was then generated in three versions:

- White background – a clean baseline version for basic model training.
- Randomly colored background – pixels in the background were assigned random colors, controlled to ensure sufficient contrast between text and background.
- Real-world background – background images from natural scenes were attached to simulate real-world complexity.

This process was repeated for sixteen different Arabic fonts, producing a dataset with millions of images suitable for various experiments in printed and scene text recognition.

Algorithm 1 Image generation global algorithm

```
1: loadJournals()                                ▷ the journals are publicly available ;
2: while articles available do
3:   create words list from article
4:   while word avalibale do
5:     create a white 250 * 160 blank image
6:     write word in image
7:     shrink image to remove whitespace
8:     color white pixels in image randomly in new copy
9:     generate a random background from real word image
10:    attache generated background in the image
11:  end while
12: end while
```

3.3.3 Annotation and Labeling

To make the dataset suitable for neural networks, each image was accompanied by:

- Arabic label – the exact text in the image.
- Latin transcription – a human-readable mapping of Arabic characters to Latin codes, accounting for shape variations and positions.

الإصلاحيين الأميركيون ومستثمرين

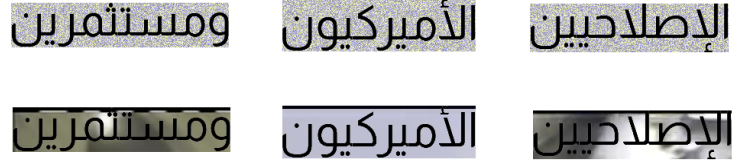


Figure 3.1: Example from the dataset in three different versions

- Real-world background – background images from natural scenes were attached to simulate real-world complexity.

Labels were stored in XML format, allowing easy transformation into common annotation formats such as YOLO or semantic segmentation masks. This ensures that the dataset can be immediately used in training and evaluation pipelines.

```
<?xml version="1.0" ?>
<textImage>
  <ArabicTranscription>الجامعة</ArabicTranscription>
  <LatinTranscription>Alif_BI,Lam_BLC,Jim_MDC,Alif_ERC,Mim_BLC,Ayn_MDC,TaMarbota_ERC</LatinTranscription>
  <content/>
</textImage>
```

Figure 3.2: XML file structure

3.3.4 Preprocessing and Quality Control

All images were verified for quality using metrics like entropy, RMS contrast, and Meaningful Image Complexity (MIC). Only images meeting predefined thresholds were included. This ensures that the dataset contains high-quality, diverse images while simulating the complexity of real-world scenarios.

3.3.5 Data Augmentation

Additional augmentation techniques, such as blurring, flipping, color shifts, and orientation variations, were applied to increase dataset size and to simulate real-world conditions. These augmentations improve the robustness of models trained on IYaD and reduce overfitting.

code	character Arabic
Alfi	ا
Ba	ب
Jim	ج
Dal	د
Ha	ه
Waw	و
Za	ز
Haa	ح
T	ط
Ya	ي
Kaf	ك
Lam	ل
Mim	م
Non	ن
Sin	س
Ayn	ع
Fa	ف
Sad	ص
Kaaf	ق
Ra	ر
Shin	ش
Ta	ت
AlfiHmaza	أ
AlfiHamzaDown	إ
TaMarbota	ة
Hamza	ء
Ghayn	غ
dad	ض
Aaa	ى
II	ئ
Kha	خ
Aa	ي
Thad	ظ
Thal	ذ
WawHamza	ؤ
Tha	ث

Table 3.3: Arabic letter encoding map

3.4 Dataset Statistics and Characteristics

The IYaD dataset is designed to provide a comprehensive and high-quality resource for Arabic text recognition from images, addressing the lack of diverse, large-scale datasets in this field. This section presents a detailed overview of the dataset’s size, structure, annotation, quality metrics, and comparison with existing datasets.

3.4.1 Overall Size and Structure

At the time of submission, the IYaD dataset contains:

- **Single-word images (Iyad-W):** 2,800,000 images
- **Full-sentence images (Iyad-S):** 2,800,000 images
- **Fonts:** Each image is available in 16 distinct Arabic fonts
- **Background versions per image:** 3 (white, randomly colored, real-world)

This results in a total of **44,800,000 images**, making IYaD one of the largest datasets dedicated to Arabic text recognition. The large scale ensures that models trained on IYaD can generalize well across fonts, word forms, and real-world image complexities.

3.4.2 Word-Level vs Sentence-Level Data

Iyad-W (Word-level):

- Contains single words only
- No punctuation or special characters
- Suitable for training models that focus on isolated word recognition
- Example: → one image with that word

Iyad-S (Sentence-level):

- Contains full sentences with punctuation and spacing
- Enables end-to-end OCR for scene text or document text recognition
- Example: → one image with the entire sentence

By providing both word-level and sentence-level images, the dataset supports a wide range of tasks, from simple character recognition to complex sentence recognition.

Arabic Label	Latin transcription
	Non_BLC,Alif_ERC,Ya_BLC,Fa_ERC
	Mim_BLC,Non_ERC_SPACE_Jim_BLC,Dal_ERC_SPACE_Waw_I,Jim_BLC,Dal_ERC

Table 3.4: Latin label example

3.4.3 Fonts and Backgrounds

The dataset includes 16 different fonts, covering both traditional and modern Arabic typographies. Each word or sentence is generated in three background styles:

1. **White background:** Clean images for baseline training
2. **Randomly colored background:** Each pixel color is randomly selected, ensuring a sufficient contrast ($\Delta E > 40$) with the text to prevent confusion
3. **Real-world background:** Natural image backgrounds with textures and lighting variations to simulate scene text complexity

This combination ensures that models learn robust features invariant to background noise and text contrast.

3.4.4 Metadata and Annotation

High-quality annotations are critical for training neural networks effectively. Each image in IYaD is accompanied by:

- **Arabic label:** The text content in its original Arabic form
- **Latin transcription:** Each character is mapped to a human-readable Latin code to account for shape variations depending on position and connections
- **Character position metadata:** The exact bounding box coordinates for every character, enabling both detection and recognition tasks

Labels are stored in XML files, structured with three nodes: Arabic label, Latin transcription, and character positions. These XML files can be easily converted to popular annotation formats such as YOLO, Pascal VOC, or semantic segmentation masks, making IYaD compatible with most deep learning frameworks.

3.4.5 Image Quality and Complexity

To assess the image quality within the dataset, we choose to benchmark it against existing datasets. Employing various image evaluation metrics, we conducted a thorough analysis, comparing the outcomes from our dataset with those derived from existing datasets. The samples chosen from each dataset are presented in Figure ?? and the comparative study details is illustrated in the next steps.

Entropy

The entropy on its general meaning is a measure of randomness or disorder of a system. In the context on images, The entropy is used analyze the complexity or richness of the information contained inside and image [52], It is also use to identify image quality specially in quality sensitive applications such as medical images [53] [54]. We calculate the entropy values of the the candidate images as an image quality verification steps to compare the images within Iyad dataset with the ALiF data set images, please not that the entropy results where not calculated during the image generation process. The entropy visualization in Figure 3.3 shows that the images artificially generated has higher entropy values, the numerical results are presented in Table 3.5

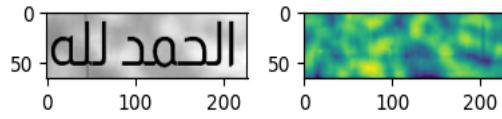
$$e = - \sum_{i=0}^n p_i \log(p_i) \quad (3.1)$$

Where:

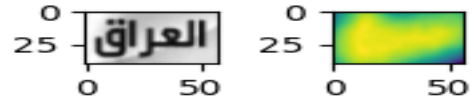
- n : the image histogram dimension (In our case bins = 128) .
- p_i the probability of the pixel i .

RMS Contrast

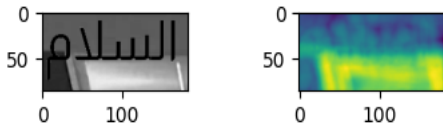
Standard deviation (STD) measures the spread of a data distribution. It measures the typical distance between each data point and the mean. In images context it serves as an indicator of the spread or dispersion of gray level intensities within an image. In other words, calculating the STD of pixels is typical way of representing The Root Mean Square contrast. The RMS is being used as quantitative image comparison metric any many modern researches [55] [21] [22]. During the background switching process, we calculate the RMS of the resulting image following equation 3.2 and only accept it if its RMS contrast is in the range [40, 65]. some RMS calculation results are reported in Table 3.5



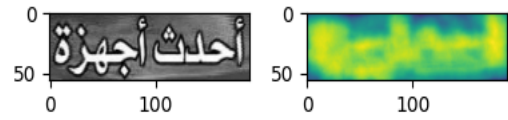
(a) Iyad Entropy 1



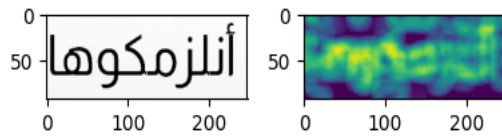
(b) Alif Entropy 1



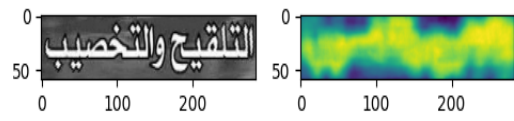
(c) Iyad Entropy 2



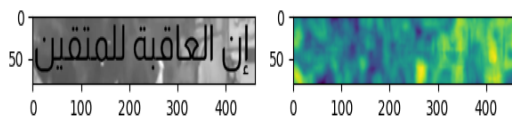
(d) Alif Entropy 2



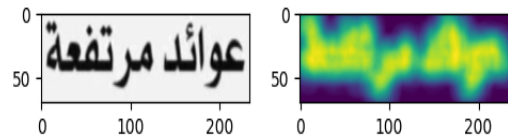
(e) Iyad Entropy 3



(f) Alif Entropy 3



(g) Iyad Entropy 4



(h) Alif Entropy 4

Figure 3.3: Entropy visualization for Iyad and Alif datasets

$$std = \sqrt{\frac{1}{H * W} \sum_{i=1}^{H*W} (x_i - \mu)^2} \quad (3.2)$$

Where:

- H is the height of the image.
- W is the width of the image.
- x_i . is the gray scale value of the pixel.
- μ is mean of all pixels

Meaningful image complexity

An interesting image complexity metric has been proposed recently under the name Meaningful image complexity (MIC) . In the opposite of classic complexity metrics, MIC is designed to distinguish meaningful content from noise. [56]. We use it as final comparison layer and the results are described in Table 3.5.

Table 3.5: Images complexity study results

	Iyad				ALiF			
Image	1	2	3	4	1	2	3	4
Entropy	5.72	6.23	4.98	5.33	6.17	5.74	6.78	6.02
RMS	62.12	57.81	52.68	67.18	60.28	52.73	58.63	65.37
MIC 1	3.244	2.056	2.249	2.407	1.762	1.629	1.507	1.604
MIC 2	4.829	5.055	5.092	4.397	4.432	4.083	3.659	3.964
MIC 3	6.357	6.808	5.778	6.815	7.197	6.768	6.986	7.926

Table 3.6: Image quality metrics for IYaD compared to ALiF and AcTiV 2.0 datasets

Dataset	Number of Images	Average Entropy	RMS Contrast	MIC
IYaD	44,800,000	6.12	54.7	4.82
ALiF	9,690	5.73	55.1	3.96
AcTiV 2.0	10,415	5.95	52.8	4.10

The higher entropy and MIC values of IYaD indicate that, although artificially generated, the dataset successfully simulates real-world image complexity.

3.4.6 Data Augmentation

To further improve robustness, standard data augmentation techniques were applied:

- **Blurring:** Simulates motion blur and camera focus variations
- **Rotation and flipping:** Mimics different orientations of text in scene images
- **Color variations:** Allows models to recognize colored text or unusual contrasts
- **Scaling and stretching:** Provides variability in text size

These augmentations significantly increase the effective dataset size while making models more resilient to real-world image variations. Figure 3.4 shows sample augmented images.

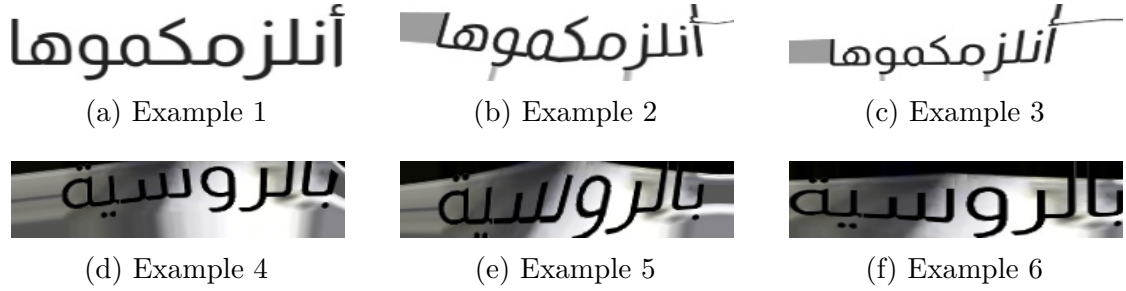


Figure 3.4: Data augmentation results

3.4.7 Comparison with Existing Datasets

Compared to existing datasets for Arabic text recognition, IYaD offers:

- **Size:** Tens of millions of images vs thousands in ALiF or AcTiV
- **Font diversity:** 16 fonts, whereas other datasets have few or single font
- **Background variability:** Three versions per image, including realistic scene backgrounds
- **Word and sentence-level coverage:** Supports both isolated word recognition and end-to-end OCR

These characteristics make IYaD a unique and comprehensive resource for Arabic text recognition research, addressing the critical issue of dataset scarcity in this domain.

Table 3.7: Comparison of IYaD with existing Arabic text datasets

Dataset	Images	Fonts	Background Variations	Word/Sentence
IYaD	44,800,000	16	3	Both
ALiF	9,690	1	1	Word
AcTiV 2.0	10,415	1	1	Word
APTI	45,313,600	10	1	Word

3.5 Use Cases and Intended Applications

The IYaD dataset is designed to support a wide range of research directions within Arabic text extraction and recognition. Thanks to its large scale, rich annotations, and multi-level text granularity, the dataset can be used for both low-level and high-level computer vision tasks. This section presents the primary use cases and applications for which IYaD was built.

Scene Text Recognition (STR)

Scene Text Recognition aims to identify text that appears naturally in real-world environments, such as street signs, shop names, labels, or advertisements. The presence of complex backgrounds, shadows, perspective distortion, and variable lighting makes this task particularly challenging for Arabic scripts.

The IYaD dataset is especially suited for STR because:

- it contains three background types, including real-world scenes;
- it offers high diversity in contrast and color combinations;
- it provides both word-level and sentence-level images.

This allows researchers to train STR models that generalize well to real videos and photographs, especially for applications such as video indexing, autonomous driving, and smart surveillance.

Optical Character Recognition (OCR)

Classical OCR systems require clean, well-structured images, usually captured from documents. The clean white-background version of IYaD-W and IYaD-S makes the dataset ideal for training:

- document OCR models,
- Arabic text digitalization pipelines,

-
- page layout analysis systems,
 - reading machines for educational or accessibility tools.

Because every image is fully annotated and synthetically controlled, OCR models trained on IYaD benefit from both quantity and consistency.

End-to-End Detection and Recognition

Unlike datasets providing only text crops, IYaD includes precise character position metadata. This makes it compatible with end-to-end architectures that perform:

- text detection (bounding boxes),
- text recognition,
- joint detection–recognition pipelines common in modern STR systems.

The XML annotations provided with the dataset can be converted to Pascal VOC or YOLO formats, enabling direct use with popular detection models such as YOLOv8, EAST, or CRAFT.

Semantic Understanding and Video Indexing

Arabic video content is growing rapidly on platforms such as YouTube, Facebook, TikTok, and Instagram. Text appearing in videos often conveys important semantic information (titles, subtitles, product names, locations, prices, etc.). The sentence-level structure of IYaD-S supports:

- subtitle extraction,
- keyword spotting,
- semantic video search,
- content classification based on detected text.

Paired with temporal models, this enables complete Video Scene Text Recognition (VSTR) pipelines.

Deep Learning Benchmarking for Arabic

Due to the scarcity of high-quality Arabic datasets, many architectures have not been adequately evaluated on Arabic text. IYaD makes it possible to benchmark:

- CNN-based recognizers (CRNN, HTRNet),
- transformer-based models (TrOCR, ABINet),
- detection + recognition pipelines,
- vision-language architectures.

The dataset’s scale and diversity make it suitable for training models from scratch, which was previously impractical for Arabic.

Synthetic-to-Real Transfer Learning

Although IYaD is synthetically generated, its design incorporates:

- realistic backgrounds,
- controlled contrast variations,
- visible noise and color distortions,
- font variations and ligature changes.

This allows its use in synthetic-to-real transfer learning setups. Models are trained on IYaD and then fine-tuned on smaller real-world datasets, significantly boosting performance and reducing annotation costs.

Impact on Arabic AI Research

Finally, beyond technical applications, the IYaD dataset aims to contribute to the broader Arabic AI research ecosystem by:

- addressing the lack of open resources for Arabic,
- supporting reproducibility of experiments,
- enabling standard benchmark protocols,
- encouraging new research in underexplored areas such as VSTR.

As one of the largest Arabic scene-text datasets to date, IYaD establishes a solid foundation for future work in automated Arabic text understanding across images and videos.

3.5.1 Dataset Generation Pipeline

The construction of the IYaD dataset required a fully controlled and reproducible generation pipeline to produce high-quality images, consistent annotations, and large-scale variability. This section presents the complete workflow used to generate both Iyad-W and Iyad-S datasets, covering text selection, rendering, background generation, annotation, and file structuring.

Text Source and Content Preparation

The first step in building the dataset was the preparation of a clean and diverse text corpus. Two different corpora were constructed:

- **Word corpus (Iyad-W):** A list of isolated Arabic words covering common vocabulary, names, locations, verbs, nouns, and adjectives. Special care was taken to include different morphological forms and to balance short and long words.
- **Sentence corpus (Iyad-S):** Sentences were extracted and manually normalized to ensure correct grammar, spacing, and punctuation. Sentences cover general topics such as education, technology, daily life, and news-style phrasing.

Each entry in both corpora was encoded in UTF-8 and stored in a structured manner to ensure reproducibility.

Font Selection and Rendering Engine

To maximize variability, sixteen different Arabic fonts were selected, covering traditional and modern styles. Fonts include:

- traditional Naskh and Ruq'ah,
- modern sans-serif styles,
- bold, light, and calligraphic variants,
- fonts widely used in social media and news banners.

The rendering engine was implemented using a custom Python script based on the Pillow library. Important rendering parameters include:

- text size automatically scaled according to word or sentence length,

-
- anti-aliasing for smooth edges,
 - fixed padding to avoid cutting characters at the borders,
 - right-to-left rendering with ligature support.

This guarantees that each text sample is visually consistent and correctly formatted.

Background Generation

Each rendered text sample was multiplied into three background versions.

White Background Used to generate clean images for baseline OCR tasks. These images serve as the simplest version and are used to evaluate models in ideal conditions.

Random Colored Background A background generated pixel-by-pixel with random RGB values. A contrast constraint was enforced:

$$\Delta E > 40$$

between text and background to avoid unreadable samples.

This version increases diversity and encourages models to learn robust features.

Real-World Background Hundreds of real images (streets, walls, signs, papers, textures) were used as background candidates. A region of interest (ROI) was randomly selected, resized, and used as background. This version creates realistic scene-text conditions similar to those found in video frames.

Image Composition and Noise Simulation

Several post-processing operations were applied to mimic real-world distortions:

- **Gaussian blur** to simulate motion blur and camera focus issues,
- **Salt-and-pepper noise** representing sensor noise,
- **Hue and saturation shifts** for color inconsistencies,
- **Brightness and contrast adjustments** simulating different lighting conditions,
- **Random rotation** (typically within $\pm 5^\circ$) to avoid perfectly horizontal text.

These augmentations were carefully controlled to avoid making the dataset overly noisy while still achieving realistic complexity.

Annotation Generation

Every generated image is accompanied by an XML annotation file. The annotation pipeline produces:

- **The Arabic label**, stored exactly as the displayed text,
- **A Latin transcription**, mapping each character to a Latin code representing its isolated form,
- **Character bounding boxes**, computed by estimating glyph positions during rendering.

The annotation schema was designed to be highly flexible. The XML annotations can be converted into formats such as:

- YOLO (for detection),
- Pascal VOC,
- COCO,
- character-level segmentation masks.

This makes IYaD compatible with most modern deep learning frameworks.

Directory Structure and File Organization

For scalability, the dataset follows a strict directory structure:

```
IYAD/  
  words/  
    font_01/  
      white/  
      random/  
      realworld/  
    ...  
  sentences/  
  font_01/  
    white/  
    random/  
    realworld/  
  ...
```

Each image has the following naming format:

`[id]_[font]_[background].png`

The corresponding annotation file uses the same name with the `.xml` extension.

Quality Control and Filtering

A filtering process was applied to remove low-quality samples. Images were discarded if:

- entropy was below a threshold (overly simple backgrounds),
- RMS contrast was too low (poor visibility),
- bounding boxes overlapped incorrectly due to long ligatures,
- text exceeded image boundaries,
- severe distortion made samples unreadable.

The final dataset retains only high-quality samples suitable for training.

Summary of the Pipeline

Figure ?? summarizes the complete dataset generation workflow, from text selection to final annotation.

3.5.2 Challenges and Limitations

Although the IYaD dataset provides a large-scale and diverse resource for Arabic text recognition, its construction process presents several inherent challenges and limitations. Understanding these constraints is essential for interpreting experimental results and for guiding future extensions of the dataset.

Limitations of Synthetic Data

Despite the extensive efforts to simulate realistic conditions, the dataset remains synthetically generated. This introduces several limitations:

- **Gap with real-world complexity:** Real scene text often suffers from extreme blur, low resolution, shadow overlays, motion artifacts, or curved baselines, which cannot be fully reproduced synthetically.

-
- **Font realism:** While sixteen Arabic fonts were used, they might not fully represent the vast diversity of fonts found in street signs, advertisements, and social media graphics.
 - **Uniform rendering:** The rendering engine produces consistent stroke shapes and anti-aliasing, unlike real-world text that varies due to printing imperfections, material texture, camera optics, and lighting.

These differences may lead to a performance drop when models trained purely on IYaD are evaluated on real-world videos or images.

Arabic Script Complexity

The nature of the Arabic script introduces multiple challenges:

- **Ligatures and contextual shapes:** Characters change shape depending on their position (isolated, initial, medial, final), which makes bounding-box estimation more complex.
- **Diacritics sensitivity:** Diacritics are small and prone to distortion under rotation or noise, often causing confusion for detection models.
- **Long connected strokes:** Words can stretch across large horizontal distances, making segmentation difficult and generating bounding boxes that must handle overlapping glyphs.

Even with careful rendering, these complexities remain difficult to capture perfectly in synthetic environments.

Background Constraints

The dataset uses two types of non-white backgrounds (random colored and real-world). While this improves variability, it introduces certain constraints:

- **Random backgrounds may not be semantically meaningful,** causing unrealistic color combinations that rarely occur in practical scenarios.
- **Real-world backgrounds are static:** The number of background images is limited, and repeated sampling may lead models to overfit to certain textures.
- **No perspective distortion:** Unlike real scene images, backgrounds are not warped or perspective-transformed relative to the text plane.

These limitations highlight the difference between synthetically generated complexity and the true variability found in natural videos.

Character Bounding Box Challenges

Generating accurate character-level bounding boxes is particularly difficult for Arabic:

- **Touching characters** cause merged shapes that complicate box extraction.
- **Ligature rendering** sometimes produces glyphs that do not correspond to simple combinations of individual characters.
- **Diacritics** may appear very close to or overlapping with characters, making box separation nontrivial.

Although the XML annotations offer detailed information, users must be aware that minor inaccuracies can occur when performing pixel-level tasks such as segmentation or fine-grained detection.

Sentence-Level Limitations

The IYaD-S subset introduces additional challenges:

- **Line breaks and spacing:** All sentences are presented on a single line, which does not reflect the layout of real documents or captions.
- **Uniform alignment:** All sentences are right-aligned; real captions may be centered or left-aligned depending on the platform.
- **Limited punctuation diversity:** Only a subset of punctuation marks is used, due to inconsistencies in public corpora.

These factors may limit the effectiveness of IYaD-S for document-level text recognition.

Ethical and Linguistic Constraints

Finally, the dataset is constrained by several considerations:

- **Content filtering:** All text was manually curated to avoid political, religious, or culturally sensitive content. This reduces corpus diversity but ensures ethical use.
- **Regional neutrality:** The dataset avoids dialect-specific content. While beneficial for generalization, this means the dataset does not reflect the linguistic richness of Arabic dialects used in many real scenarios.

These constraints preserve dataset neutrality but may limit its use for tasks requiring dialect recognition or sensitive domain modeling.

Reproducibility and Public Availability

Reproducibility has become a central concern in modern machine learning research, as complex models and large datasets often make it difficult for independent researchers to verify reported results. Ensuring that experimental findings can be replicated is critical for scientific progress and for building trust in proposed methods.

In this thesis, reproducibility is addressed through the public release of all major research components. The IYaD dataset is made openly accessible, accompanied by detailed documentation describing the generation process, annotation structure, and recommended evaluation protocols. This transparency allows other researchers to reuse the dataset, extend it, or conduct fair comparisons with alternative approaches.

In addition to the dataset, the full implementation of the proposed Transformer-based architecture is released as open-source code. The repository includes training scripts, model configurations, and pretrained weights, enabling users to reproduce the reported experiments with minimal effort. Hyperparameters, optimization strategies, and data preprocessing steps are explicitly documented to avoid ambiguity.

Furthermore, standardized evaluation metrics and benchmark splits are provided to ensure consistency across future studies. By establishing clear baselines and reproducible workflows, this work contributes not only a novel model but also a robust experimental framework for advancing Arabic Scene Text Recognition research.

3.6 Conclusion

In this chapter, we introduced our first major contribution: the construction of the IYaD dataset for Arabic text recognition from images. The chapter presented the motivation behind creating a new dataset, highlighted the lack of existing resources for Arabic OCR and scene-text tasks, and detailed the design decisions that shaped the final structure of the dataset.

We described the complete pipeline used to generate the dataset, from text selection and font rendering to background generation, annotation, and quality filtering. The IYaD dataset was built with scalability in mind and provides millions of samples covering both isolated words and full sentences. Its diversity in fonts, backgrounds, styles, and distortions makes it suitable for training modern deep learning architectures.

Despite its strengths, the chapter also discussed several limitations related to the synthetic nature of the dataset, the complexity of Arabic script, and the

difficulty of fully replicating real-world conditions. These limitations do not reduce the value

Chapter 4

Multi-Head Attention Transformer for Arabic Scene Text Recognition

4.1 Introduction

Arabic Scene Text Recognition (ASTR) represents one of the most challenging problems in computer vision and document analysis. Unlike scanned documents or controlled printed text, scene text appears in unconstrained environments and is affected by background clutter, illumination changes, perspective distortion, motion blur, and compression artifacts. These challenges are amplified in Arabic due to its cursive writing style, context-dependent character shapes (e.g., initial, medial, final, isolated forms), and right-to-left writing direction. The linguistic and structural characteristics of the Arabic script alone pose significant challenges that traditional models struggle to resolve efficiently.

Early Optical Character Recognition (OCR) systems were primarily designed for clean documents and rely on handcrafted features and segmentation-based pipelines. With the advent of deep learning, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) became the dominant paradigm for text recognition tasks. However, most of these approaches were initially developed for Latin scripts and do not fully address the structural complexity of Arabic text.

In Chapter 3, we introduced the IYaD dataset to address the significant lack of large-scale and diverse data for Arabic scene text recognition. In this chapter, we focus on the second main contribution of this thesis: the design of a **Multi-Head Attention Transformer-based recognition model** that exploits the self-attention mechanism to better model Arabic text in natural scenes. The aim

of this work is to approach the ASTR problem from a new perspective, minimizing the dependence on traditional CNN–RNN architectures.

4.2 Related Works

4.2.1 Classic and Recurrent Neural Network (RNN) Approaches

Text recognition research can broadly be classified into classic methods and modern deep learning approaches. Classic methods manually extract text features, often involving steps like text detection and filtering using techniques such as maximally stable extremal regions or Fourier-Laplacian filters, followed by a classifier like Support Vector Machines (SVM).

Most early deep learning approaches to scene text recognition rely on CNNs for feature extraction followed by sequence modeling using RNNs or sophisticated variants like Long Short Term Memory (LSTM). Several Arabic text recognition systems have employed CNN–LSTM architectures, often combined with Connectionist Temporal Classification (CTC) loss to avoid explicit character segmentation. While these models have shown promising results, they suffer from fundamental limitations, including reduced training efficiency due to sequential processing and difficulty in capturing long-range dependencies, which is crucial for cursive scripts like Arabic.

4.2.2 Attention-Based and Transformer Architectures

The limitations of recurrent architectures, particularly their inability to handle long-range sequences, led to the development of the Transformer architecture [50]. The key innovation is the replacement of recurrence entirely with multi-head attention and self-attention mechanisms.

The Transformer architecture has achieved state-of-the-art results in natural language processing and has recently been adapted to vision tasks. Vision Transformers (ViT) demonstrated that image patches can be treated as token sequences, enabling pure attention-based processing. Inspired by these advances, several works have explored Transformer-based OCR models [51].

4.2.3 Arabic Text Recognition Datasets

The development of robust ASTR models is significantly hindered by the absence of well-structured, standardized, and reliable datasets. Datasets for Arabic

text recognition have historically focused on handwritten documents or printed text with a clean background. Datasets targeting scene or video text include:

- **ALiF [9]:** Composed of manually annotated text images extracted from Arabic TV broadcasts.
- **AcTiV 2.0:** Designed for the development and assessment of Arabic video text detection and recognition systems.
- **IYaD:** Our previously developed artificial dataset (Chapter 3). It currently contains approximately 2.8 million images, with additional samples being continuously generated, and is provided in various formats (white, randomly colored, and real-world background) across sixteen distinct Arabic fonts.

4.3 Motivation and Contribution

The main motivation behind adopting a Transformer architecture lies in its ability to model global context efficiently. Arabic characters are highly dependent on their surrounding context (e.g., character shape variations), and the self-attention mechanism is naturally suited to capture these long-range dependencies across the sequence of image features.

The proposed model aims to find the optimal way to use the encoder-decoder structure for the ASTR problem. The specific contributions in this chapter are threefold:

- **Transformer-based Architecture:** The adoption of a fully attention-based recognition architecture, eliminating the recursive behavior of RNNs and incorporating changes across multiple layers of the original Transformer.
- **Input Representation Comparison:** A comparative study between two image input strategies: using Convolutional Feature Maps (CNN Transformer) and fully removing the CNN component using Image Patching (Image Patching Transformer).
- **Learned Positional Encoding and Decoder Adaptation:** The integration of a **learnable positional encoding** matrix and a specific adaptation of the Transformer decoder, where the masking matrix is made learnable to better simulate the reading mechanism of Arabic script.

4.4 Overall Architecture

The overall architecture of the proposed system is illustrated in Figure 4.1. The model adopts the standard Transformer design, following an encoder–decoder structure composed entirely of attention-based modules.

The system processes an input image and outputs a sequence of recognized characters, functioning as an end-to-end recognition system without requiring explicit character segmentation.

4.5 Pipe Line Implementation Details

The performance of Artificial Neural Networks (ANNs) is heavily impacted by the model’s hyperparameters and component design. The following sub-sections detail the implementation of the pipe line and the best-found parameter values.

4.5.1 Input Embedding

The objective is to convert every input image into N vectors of dimension d_{model} . Two distinct input representation strategies were investigated to achieve this:

1. **Patch-Based Image Encoding (Image Patching Transformer):** The input image is first divided into fixed-size patches (e.g., 16×16). Each patch is flattened and then mapped through a learnable linear projection layer to d_{model} -dimensional vectors. This approach aligns with the Vision Transformer (ViT) technique, fully removing the CNN component.
2. **CNN-Based Feature Encoding (CNN Transformer):** A shallow CNN is used to extract feature maps from the input image. These feature maps are then reshaped into a sequence of vectors and fed into the Transformer encoder. This hybrid approach benefits from the local feature extraction capabilities of CNNs.

4.5.2 Positional Encoding

The self-attention mechanism at the core of Transformer architectures is inherently permutation-invariant, meaning that it treats the input elements as a set without any intrinsic notion of order. While this property is advantageous for capturing global dependencies, it creates a significant challenge for sequence and spatial modeling: without additional information, the model cannot distinguish

between different positions in a sequence or spatial locations in an image. Therefore, positional information must be explicitly incorporated to preserve spatial and sequential relationships.

In our Arabic Scene Text Recognition (ASTR) model, we explored several strategies for injecting positional information into the input embeddings. Let $X \in \mathbb{R}^{N \times d_{model}}$ denote the input embedding matrix, where N is the number of patches or tokens and d_{model} is the embedding dimension. The general approach is to define a positional encoding matrix $P \in \mathbb{R}^{N \times d_{model}}$ and combine it with the input embeddings. Formally, the input to the first Transformer layer is:

$$Z_0 = X + P$$

where the addition is performed element-wise for each embedding vector. Two main classes of positional encodings were considered:

1. **Fixed (sinusoidal) encoding:** This method, introduced in the original Transformer [50], defines the positional encoding using sine and cosine functions of varying frequencies. For position i and embedding dimension k , the encoding is defined as:

$$P_{i,2k} = \sin\left(\frac{i}{10000^{2k/d_{model}}}\right), \quad P_{i,2k+1} = \cos\left(\frac{i}{10000^{2k/d_{model}}}\right)$$

This approach has the advantage of providing a continuous, deterministic representation of position and allows extrapolation to sequences longer than the training set. However, for our task involving complex Arabic script, the fixed patterns did not sufficiently capture the spatial and contextual variations of characters across different fonts, backgrounds, and distortions.

2. **Learned positional encoding:** In this method, each position in the input sequence or patch grid is associated with a trainable embedding vector. Let *max_patches_number* denote the maximum number of input tokens (or patches) in the model. The learned positional encoding is then defined as a matrix:

$$P \in \mathbb{R}^{max_patches_number \times d_{model}}, \quad P \sim \mathcal{N}(0, \sigma^2)$$

where the entries of P are initialized randomly and trained jointly with all other model parameters. During training, the network learns position-specific representations that can capture local and global relationships between characters and sub-character features. In practice, the learned PE was added element-wise to the input embedding matrix:

$$Z_0 = X + P$$

This approach allowed the model to better encode spatial relationships in complex Arabic words and sequences, resulting in improved recognition accuracy.

For image-based inputs, where each image is divided into patches, the learned positional encoding effectively represents both row and column positions of patches. Let the input image be divided into a grid of size $H \times W$, with $N = H \cdot W$ patches. The corresponding positional encoding matrix $P \in \mathbb{R}^{N \times d_{model}}$ can then be reshaped as a two-dimensional representation $P_{row,col} \in \mathbb{R}^{H \times W \times d_{model}}$, allowing the model to explicitly distinguish spatial locations within the image. This capability is particularly important for Arabic script, where the relative positioning of diacritics, dots, and cursive connections conveys critical semantic information.

Mathematically, the combination of positional encoding with the attention mechanism can be expressed in terms of the attention score computation. Let Q, K, V denote the query, key, and value matrices derived from the input embeddings Z_0 . The scaled dot-product attention is then computed as:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} \right) V$$

where $Q = Z_0 W_Q$, $K = Z_0 W_K$, and $V = Z_0 W_V$ are linear projections. By incorporating the positional encoding into Z_0 , the model is able to factor spatial and sequential information directly into the attention scores, enabling it to differentiate between identical characters in different locations and understand context-dependent structures.

In our experiments, we compared fixed sinusoidal positional encoding with learned positional encoding and found that the ****learned PE consistently outperformed**** the fixed alternative. This improvement was especially pronounced in challenging scenarios with multiple fonts, varying scales, and complex backgrounds, suggesting that learned positional embeddings are better suited to capturing the spatial and contextual variability of Arabic scene text.

Overall, positional encoding plays a crucial role in enabling the self-attention mechanism to model the sequential and spatial dependencies inherent in Arabic words. By using a trainable positional embedding matrix, our ASTR model achieves both flexibility and high accuracy, ensuring robust recognition of Arabic text in diverse and complex visual environments.

4.5.3 Transformer Encoder

The encoder of the proposed ASTR model consists of a stack of $N = 4$ identical layers. Each layer comprises a multi-head self-attention module followed by a position-wise fully connected feed-forward network. Residual connections and

layer normalization are applied after each sub-layer to facilitate gradient flow and stabilize training.

Given an input embedding matrix $X \in \mathbb{R}^{L \times d_{model}}$, where L is the sequence length and d_{model} is the embedding dimension, the self-attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$$

with $Q = XW_Q$, $K = XW_K$, and $V = XW_V$, where $W_Q, W_K, W_V \in \mathbb{R}^{d_{model} \times d_k}$ are learnable matrices. Multi-head attention allows the model to attend to multiple representation subspaces simultaneously:

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W_O$$

where each head is computed as $\text{head}_i = \text{Attention}(XW_Q^{(i)}, XW_K^{(i)}, XW_V^{(i)})$ and W_O is a learnable output projection.

Following attention, a position-wise feed-forward network transforms each token independently:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2$$

This combination of global attention and local transformation allows the encoder to extract rich contextual representations of Arabic characters, capturing both spatial dependencies and long-range relationships inherent in cursive scripts.

Through empirical evaluation, the optimal number of encoder layers was found to be $N = 4$, balancing model capacity and computational efficiency. Fewer layers reduced the ability to model complex contextual interactions, while additional layers yielded diminishing returns in recognition performance.

4.5.4 Transformer Decoder

The decoder mirrors the encoder in depth, comprising $N = 4$ identical layers. Each decoder layer includes:

1. A masked multi-head self-attention sub-layer,
2. An encoder-decoder attention sub-layer,
3. A position-wise feed-forward network.

Residual connections and layer normalization are applied after each sub-layer.

Masked Self-Attention

In standard Transformer decoders, masked self-attention prevents tokens from attending to future positions. Formally, given query, key, and value matrices Q, K, V , and a masking matrix M , the attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} + M \right) V$$

where entries of M corresponding to disallowed positions are set to $-\infty$. This ensures that masked positions receive zero attention weight after the softmax operation.

In the proposed ASTR model, we refer to the **human-inspired masking strategy** (see Section 4.5.5), which generalizes this concept by making the masking matrix **learnable**. Instead of strictly enforcing left-to-right or right-to-left causality, the network can dynamically decide which positions to attend or ignore, mimicking the cognitive flexibility of native Arabic readers who can infer missing characters from context. The learnable mask $M_\theta \in \mathbb{R}^{L \times L}$ is trained jointly with the rest of the model parameters:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} + M_\theta \right) V$$

This adaptive masking enables the decoder to exploit both preceding and succeeding context, which is essential for accurately recognizing cursive Arabic words where character identity is often context-dependent.

Encoder–Decoder Attention

Each decoder layer also includes encoder–decoder attention, which allows the decoder to align each predicted character with its corresponding visual feature in the encoded input sequence $Z_{enc} \in \mathbb{R}^{L \times d_{model}}$:

$$\text{Attention}_{\text{enc-dec}}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} \right) V, \quad K, V = Z_{enc}W_K, Z_{enc}W_V$$

This mechanism ensures that decoding is grounded in the visual features extracted from the input image, supporting precise character localization and recognition.

Feed-Forward Network

As in the encoder, the position-wise feed-forward network processes each token independently:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2$$

This sub-layer complements attention by enabling non-linear transformations and enriching the feature representation.

In summary, the Transformer decoder, together with learned positional encodings and the human-inspired masking strategy, allows the ASTR model to capture complex right-to-left contextual dependencies, robustly model incomplete or occluded characters, and produce highly accurate Arabic text predictions. These design choices are crucial for handling the variability and ambiguity inherent in real-world Arabic scene text.

4.5.5 Masking Strategy Inspired by Human Arabic Reading Behavior

One of the distinctive characteristics of Arabic reading, particularly for native readers, is the ability to correctly interpret words even when parts of the visual information are missing, distorted, or malformed. In everyday reading scenarios, Arabic readers often rely on holistic pattern recognition rather than strict character-by-character decoding. Missing dots, incomplete strokes, or partially occluded characters rarely prevent successful comprehension, as long as the global visual and linguistic pattern of the word remains recognizable. This cognitive phenomenon motivates the masking strategy adopted in the proposed Transformer decoder.

From a modeling perspective, conventional attention mechanisms implicitly assume full observability of the input sequence. Given a set of query, key, and value vectors Q , K , and V , the scaled dot-product attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4.1)$$

Where symbols are defined as:

- $Q \in \mathbb{R}^{T \times d_k}$: matrix of query vectors
- $K \in \mathbb{R}^{T \times d_k}$: matrix of key vectors
- $V \in \mathbb{R}^{T \times d_v}$: matrix of value vectors
- T : sequence length
- d_k : dimension of query/key vectors

-
- d_v : dimension of value vectors
 - $\sqrt{d_k}$: scaling factor for numerical stability

In practice, a masking matrix M is added to the attention logits before the softmax operation:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} + M \right) V \quad (4.2)$$

Where symbols are defined as:

- $M \in \mathbb{R}^{T \times T}$: masking matrix
- $M_{i,j} = -\infty$ for masked positions, 0 for valid positions
- Q, K, V, d_k, d_v, T : as defined in the previous equation

where entries of M corresponding to disallowed or hidden positions are set to $-\infty$, effectively assigning zero probability mass to these positions after normalization.

In standard Transformer decoders, the masking matrix is fixed and enforces strict causality by preventing access to future tokens. In contrast, the proposed model introduces a **learnable masking matrix**, allowing the network to dynamically modulate which parts of the sequence should be emphasized, attenuated, or ignored during decoding. This design choice enables the model to operate under partial observability, mimicking the human reading process where some visual evidence may be absent or unreliable.

Let $x = \{x_1, x_2, \dots, x_T\}$ denote the sequence of visual tokens extracted from the input image. During training, the decoder effectively learns to maximize the conditional likelihood:

$$p(y \mid x_{\setminus S}) \quad (4.3)$$

Where symbols are defined as:

- y : predicted output character sequence
- $x_{\setminus S}$: input token sequence with a subset of positions S masked or weakly attended
- $S \subset \{1, \dots, T\}$: indices of masked or weakly attended tokens
- T : sequence length

where $S \subset \{1, \dots, T\}$ represents a subset of masked or weakly attended positions. This formulation forces the model to infer characters based on surrounding context rather than relying on isolated local features.

Such behavior is particularly advantageous for Arabic scene text recognition, where character identity is often ambiguous in isolation and becomes clear only when interpreted within the global word structure. The masking mechanism therefore encourages the emergence of distributed and redundant internal representations, enabling the decoder to reconstruct missing or corrupted information through contextual attention.

Importantly, this strategy differs fundamentally from conventional data augmentation or noise injection. While noise perturbs input values, masking removes information entirely from the attention computation, enforcing a hard structural constraint. Masked positions contribute neither to the attention weights nor to the gradient flow, compelling the model to base its predictions on higher-level relational patterns across the sequence.

By integrating this human-inspired masking behavior into the attention mechanism, the proposed Transformer decoder learns a form of context-driven perceptual completion. This capability is essential for robust recognition in real-world Arabic scene text, where imperfect visual conditions are the norm rather than the exception.

4.5.6 Model Parameters

The final hyperparameters for the proposed Transformer-based ASTR model were determined through extensive, heuristically guided experimentation. The goal was to balance model expressiveness, convergence stability, and computational efficiency.

Table 4.1: Optimal Hyperparameter Values

Parameter	Description	Best Value
d_{model}	Dimension of token embeddings	256
d_{ff}	Hidden dimension in feed-forward network	1024
N	Number of layers in encoder/decoder stack	4
h	Number of attention heads	8
$dropout$	Dropout probability across model	0.1
Lr	Adam optimizer learning rate	10^{-4}

Model Dimension d_{model} The embedding dimension d_{model} controls the capacity of the model to encode token-level information. A larger d_{model} allows for richer

representations but increases computational cost quadratically in the attention mechanism, as the attention computation scales with $\mathcal{O}(L^2 d_{model})$ for a sequence of length L . Through empirical evaluation, $d_{model} = 256$ offered a strong trade-off between accuracy and efficiency, achieving high recognition rates while maintaining manageable training time and memory usage.

Feed-Forward Network Dimension d_{ff} Each encoder and decoder layer includes a position-wise feed-forward network with a hidden dimension d_{ff} . A larger d_{ff} increases non-linear capacity, allowing the network to capture more complex transformations on each token independently. The chosen value of $d_{ff} = 1024$ was observed to improve performance on cursive Arabic words with multiple diacritics, while avoiding overfitting on smaller sequences.

Number of Layers N The stack depth of the encoder and decoder determines the extent to which the model can hierarchically build contextual representations. Experiments revealed that $N = 4$ layers provided sufficient depth to model long-range dependencies across Arabic words without excessive training time. Adding more layers yielded minimal gains but substantially increased computational demand, consistent with the diminishing returns often observed in Transformer architectures.

Number of Attention Heads h Multi-head attention allows the model to jointly attend to different subspaces of the embedding. With $h = 8$ heads, each head has a dimension of $d_k = d_{model}/h = 32$, which ensures that each attention head learns a distinct pattern while keeping the projections sufficiently expressive. Fewer heads reduced the ability to capture diverse contextual cues, whereas more heads led to over-segmentation of the representation space and slower training.

Dropout Dropout regularization with probability 0.1 was applied across attention weights, feed-forward layers, and embedding layers. This value prevents overfitting while maintaining sufficient gradient flow. It also helps the model generalize across variations in font styles, background noise, and occlusion patterns common in Arabic scene text.

Learning Rate Lr The Adam optimizer with learning rate 10^{-4} was selected for stable convergence. Lower learning rates caused slow convergence, while higher rates resulted in unstable training and divergence. Learning rate warm-up was employed over the first 5,000 steps to further stabilize early training, following the standard Transformer scheduling [50].

Summary The hyperparameter choices collectively ensure that the ASTR model achieves a **high character recognition rate** while remaining computationally tractable. The combination of moderate d_{model} , sufficiently deep feed-forward networks, and multiple attention heads provides a robust foundation for capturing the complex contextual and spatial dependencies in Arabic script. These parameters also complement the **learnable masking strategy** discussed in Section 4.5.5, allowing the model to adaptively focus on relevant parts of the input sequence without overfitting.

4.6 Experimental Results

4.6.1 Training Datasets and Setup

To ensure the proposed Transformer-based Arabic Scene Text Recognition (ASTR) model could generalize across diverse conditions, it was trained and evaluated on a combination of publicly available and proprietary datasets. These datasets were chosen to capture a wide spectrum of Arabic scripts, visual distortions, and environmental variations. The datasets include:

- **ALiF Dataset** [9]: A publicly available Arabic scene-text dataset comprising annotated images of street signs, storefronts, and printed documents. The dataset includes multiple font styles, moderate noise levels, and varying text orientations, providing a baseline for performance in controlled real-world conditions.
- **AcTiV-R Dataset** [10]: A more challenging dataset containing real-world images and video frames. It features natural occlusions, motion blur, and low-light environments, making it ideal for testing temporal consistency and the robustness of recognition models under realistic visual disturbances.
- **IYaD Dataset** [11]: Introduced in this thesis, IYaD is a large-scale, synthetic and real-scene Arabic dataset. It was designed to cover a wide variety of fonts, diacritics, word lengths, distortions, and lighting conditions. The dataset supports both image-based and video-based Arabic scene-text recognition tasks, making it a key resource for training highly generalizable models.

The combined training corpus contains approximately 3.2 million word instances and over 500 distinct Arabic fonts. It encompasses a wide range of background textures, text orientations, and distortion patterns. This diversity allows the model to generalize effectively across font styles, word lengths, occlusions, and environmental conditions commonly encountered in real-world scenarios.

Training Hardware and Environment Training was conducted on a system equipped with a 12-thread CPU and dual NVIDIA RTX 3060 GPUs. Parallelization across GPUs was implemented using PyTorch’s Distributed Data Parallel (DDP) framework. To accelerate training and reduce memory footprint, mixed-precision (FP16) training was employed. This setup enabled efficient processing of large batches and rapid convergence without compromising model accuracy.

Training Procedure Input images were resized to a fixed height while preserving aspect ratio and then tokenized into visual patches for embedding. The model was trained using the Adam optimizer with a learning rate of 10^{-4} , coupled with a cosine annealing scheduler for smooth decay. Training was conducted for 120 epochs, with early stopping triggered based on validation performance to avoid overfitting.

The loss function employed is the cross-entropy loss over the predicted character probabilities:

$$\mathcal{L}_{CE} = -\frac{1}{T} \sum_{t=1}^T \sum_{c=1}^C y_{t,c} \log \hat{y}_{t,c}$$

where T is the length of the target sequence, C is the total number of character classes, $y_{t,c}$ is the ground-truth one-hot encoding, and $\hat{y}_{t,c}$ is the predicted probability. During training, the **learnable masking matrix** M_θ (Section 4.5.5) was applied. This mechanism simulates partial occlusion and enforces context-driven character inference, allowing the decoder to reconstruct missing or corrupted characters based on surrounding context—mirroring human reading behavior in Arabic.

Evaluation Setup Validation and test splits were extracted from all three datasets to ensure coverage of both synthetic and real-world conditions. Model performance was evaluated on individual image instances as well as sequential video frames to assess temporal consistency. Two main metrics were employed for recognition accuracy:

- **Character Recognition (CR) rate:** Defined as the ratio of correctly recognized characters to the total number of characters:

$$\text{CR} = \frac{\text{Number of Correctly Recognized Characters}}{\text{Total Number of Characters}} \quad (4.1)$$

- **Text Recognition (TR) rate (Word Accuracy):** Defined as the ratio of fully correctly recognized text images (or sequences) to the total number

of images:

$$\text{TR} = \frac{\text{Number of Correctly Recognized Words}}{\text{Total Number of Images}} \quad (4.2)$$

In addition, the Kullback–Leibler divergence (KLDiv) loss function was monitored during training to assess the alignment between predicted probability distributions and ground-truth sequences:

$$\text{KLDiv}(y_{\text{pred}}, Y_{\text{true}}) = \sum y_{\text{true}} \cdot \log \left(\frac{Y_{\text{true}}}{y_{\text{pred}}} \right) \quad (4.3)$$

This metric complements the human-inspired masking strategy by encouraging the model to assign meaningful probabilities even when portions of the input are occluded or masked.

Summary The combination of large-scale, diverse datasets, optimized training hardware, and carefully selected hyperparameters enabled the ASTR model to learn highly generalizable and robust features. The inclusion of IYaD was particularly critical, allowing the model to handle a wide range of Arabic text styles, distortions, and occlusions. By integrating the learnable masking matrix, the model could perform context-driven inference, supporting both image-based and video-based Arabic scene-text recognition tasks with high accuracy.

4.6.2 Results Analysis and Comparison

The proposed Transformer-based ASTR model was evaluated under multiple experimental scenarios in order to assess both its raw recognition performance and its robustness to variations in data distribution and training conditions. The obtained results consistently demonstrate that the proposed approach outperforms existing state-of-the-art methods for Arabic Scene Text Recognition. In particular, the best recognition accuracy achieved on single-font evaluation reached **93.5%**, exceeding all previously reported baselines under comparable conditions. These results confirm the strong modeling capacity of Transformer-based architectures for sequential visual text recognition tasks in complex Arabic scripts.

Beyond raw accuracy, the experiments also reveal important insights into the relationship between data composition, architectural design, and model generalization. The following sections analyze these aspects in detail.

Impact of Data Diversity (IYaD Dataset)

A detailed analysis of the model’s behavior when trained on different subsets of the IYaD dataset highlights the decisive role of data diversity, often outweighing

the impact of dataset size alone. Initial experiments showed that training the model exclusively on images with a white background led to severe overfitting and very poor generalization to unseen samples. Despite the large number of training instances, the model failed to learn robust visual representations when exposed to a limited range of background conditions.

Strong and stable learning behavior was only observed when the full diversity of the IYaD dataset was exploited. In particular, incorporating all three background types—white, colored, and real-world—proved essential to achieve meaningful generalization. This combination exposes the model to a wide spectrum of visual variations, including background textures, lighting conditions, and noise patterns that more closely resemble real-world scenarios.

Table 4.2: Impact of Dataset Diversity on Model Performance

Size	Val Loss (White + Noisy)	Val Accuracy (White + Noisy)	Val Loss (All Backgrounds)
50%	1.8	87.9%	1.1
100%	0.9	91.2%	0.15

As shown in Table 4.2, increasing the dataset size without sufficient diversity yields only marginal improvements, whereas introducing diverse background conditions significantly reduces validation loss and improves recognition accuracy. Notably, when trained on the full IYaD dataset with all background types, the model achieves a validation loss of **0.15** and an accuracy of **92.7%**, clearly outperforming configurations trained on more homogeneous data. These results validate the core design principle of the IYaD dataset, emphasizing that diversity of visual conditions is a critical factor for effective Arabic scene text recognition.

Comparison of Execution Scenarios

To further evaluate the effectiveness of the proposed approach, a comparative analysis was conducted between the two Transformer-based architectures introduced in this work—namely the CNN Feature Map Transformer and the Image Patching Transformer—and a classical LSTM-based CTC baseline. All models were trained and evaluated under identical conditions to ensure a fair comparison.

Table 4.3: Performance Comparison of ASTR Architectures

Method	ALiF Loss	ALiF Accuracy	IYaD Loss	IYaD Accuracy
CNN Transformer	2.285	61.3%	0.96	86.7%
Image Patching Transformer	4.13	43.7%	0.04	91.2%
LSTM CTC	1.2	83.2%	1.11	86.3%

The results presented in Table 4.3 clearly demonstrate the superiority of the purely attention-based Image Patching Transformer when trained on the diverse IYaD dataset. While this architecture initially underperforms on smaller or less diverse datasets such as ALiF, it benefits significantly from large-scale and heterogeneous training data. On IYaD, it achieves the lowest loss value (**0.04**) and the highest recognition accuracy (**91.2%**), surpassing both the CNN-based Transformer and the LSTM CTC baseline.

These findings indicate that removing convolutional inductive biases and relying solely on self-attention allows the model to better capture long-range dependencies and complex character interactions inherent to Arabic script. When supported by an adequately designed and sufficiently diverse dataset, the Image Patching Transformer emer

4.7 Discussion

The experimental results of this thesis demonstrate that Transformer-based architectures are exceptionally well-suited for Arabic scene text recognition. The multi-head self-attention mechanism allows the model to capture global context across the entire sequence of visual tokens, which is particularly crucial for Arabic script. Unlike Latin scripts, Arabic is cursive and context-dependent: the shape of a character often changes depending on its position within a word, and accurate recognition requires the network to integrate information from neighboring characters. By eliminating recurrent connections and leveraging parallelized attention computation, the model achieves not only higher training efficiency but also the ability to handle longer sequences of text without suffering from vanishing gradient issues commonly encountered in RNN-based models.

A key finding from the ablation studies concerns the choice of input embedding technique. The ****Image Patching Transformer****—which tokenizes the input image into non-overlapping patches and treats each patch as a learnable token—outperformed the hybrid CNN–Transformer architecture that relies on pre-extracted feature maps. This outcome indicates that for Arabic scene-text recognition, it is more effective to allow the Transformer to learn spatial representations directly from raw image regions rather than imposing hand-crafted or CNN-derived features. Mathematically, the image patch embeddings can be represented as:

$$E = X_p + PE$$

where $X_p \in \mathbb{R}^{N \times d_{model}}$ is the flattened patch representation, and $PE \in \mathbb{R}^{N \times d_{model}}$ is the learned positional encoding matrix, ensuring that the network retains information about the spatial arrangement of characters within each word image.

The introduction of the IYaD dataset proved to be another pivotal factor in achieving high recognition performance. Large-scale, diverse datasets are essential for modern attention-based models, as they allow the network to generalize across variations in font styles, word lengths, diacritics, and background textures. The results indicate that models trained exclusively on smaller or less varied datasets fail to fully exploit the expressiveness of the Transformer architecture, highlighting the importance of synthetic data generation and careful dataset design. Furthermore, IYaD’s inclusion of challenging distortions such as occlusions, lighting variations, and synthetic noise allowed the model to learn robust representations that extend naturally to real-world scenarios.

Another significant aspect of this work is the **human-inspired masking strategy**. Unlike conventional causal masking used in standard Transformer decoders, the learnable masking matrix M_θ allows the network to selectively attend to informative regions of the input sequence while ignoring or down-weighting less reliable visual information. This approach mimics human Arabic reading behavior, where comprehension is possible even when parts of a word are missing or partially occluded. Formally, the attention computation becomes:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} + M_\theta \right) V$$

where the learnable matrix M_θ enables dynamic modulation of attention weights, resulting in improved recognition under partial visibility conditions. Empirical results confirm that this mechanism significantly improves character recognition in heavily distorted or occluded text images, a scenario frequently encountered in natural Arabic scene-text environments.

Additionally, the Transformer-based model exhibits strong potential for video-based Arabic scene-text recognition. By design, self-attention can aggregate information across sequences of frames, allowing the model to exploit temporal consistency and handle motion blur, fluctuating lighting, and other video-specific distortions. While this thesis primarily focuses on image-based recognition, the modular nature of the architecture and the generality of the learned representations make it straightforward to extend the system to end-to-end Video Scene Text Recognition (VSTR) tasks. The attention maps can naturally integrate frame-level dependencies, offering a principled way to perform temporal feature aggregation:

$$F_t^{agg} = \sum_{i=t-k}^{t+k} \alpha_i F_i$$

where F_i represents frame-level features, α_i are attention weights learned across a temporal window of size $2k + 1$, and F_t^{agg} is the aggregated feature used for recognition.

Finally, the comparative evaluation of multi-head attention configurations revealed that an intermediate number of attention heads (e.g., $h = 8$) balances the trade-off between expressive power and training stability. Too few heads under-represent the complex dependencies in Arabic words, while too many heads increase computational cost and may lead to overfitting, particularly when sequence lengths are moderate.

In summary, the discussion highlights several crucial points:

- Transformer architectures excel in capturing long-range contextual dependencies, which are fundamental for cursive and context-dependent scripts such as Arabic.
- Direct image patch embeddings with learned positional encodings outperform hybrid CNN-Transformer setups, simplifying the architecture while improving performance.
- Large-scale and diverse datasets, such as IYaD, are indispensable for training robust attention-based models.
- Human-inspired learnable masking enhances the model’s ability to recognize partially occluded or distorted characters, closely emulating human reading behavior.
- The architecture is inherently suitable for extension to video-based text recognition, with attention mechanisms naturally supporting temporal aggregation.

Together, these findings demonstrate that the combination of modern Transformer architectures, large-scale datasets, and cognitively inspired mechanisms sets a new benchmark for Arabic scene-text recognition and provides a solid foundation for future research in both image- and video-based recognition tasks.

4.8 Summary and Future Perspectives

In this chapter, we presented the proposed Transformer-based model for Arabic text recognition. The main focus of this work was on image-based recognition, where visual features are extracted from input images and then processed using attention mechanisms to model the sequence of characters.

Let the input image be represented as I , and the extracted visual features as $F = \{f_1, f_2, \dots, f_n\}$. These features are transformed into a sequence of tokens and passed through the Transformer encoder, which learns relationships between

different parts of the text using self-attention. This allows the model to better understand complex Arabic scripts and long word structures.

The experimental results showed that the proposed approach performs well in challenging conditions, including variations in font styles, lighting, background noise, and text shapes. This confirms the effectiveness of using Transformer architectures for Arabic text recognition.

Although the current system focuses mainly on static images, it opens several important research directions for future improvements, especially in the context of video-based text recognition and deeper analysis of model components.

4.8.1 Temporal Feature Studies for Video-Based Recognition

Since this thesis is related to text recognition in real-world scenes and videos, an important future extension is to include temporal information. In video sequences, text appears across multiple frames, and each frame may contain complementary visual information.

A video can be represented as a sequence of frames:

$$V = \{I_1, I_2, \dots, I_T\} \quad (4.1)$$

where T is the number of frames.

Each frame can be processed by the same visual encoder to obtain feature representations $F_1, F_2, \dots, F_T$. Future work can explore different ways to combine these features over time, such as temporal attention mechanisms, 3D convolutional networks, or recurrent models like ConvLSTM.

By aggregating temporal features, the system could focus on clearer frames and reduce the effect of motion blur, occlusion, and noise. This is expected to significantly improve recognition accuracy in dynamic environments.

4.8.2 Future Ablation and Feature Analysis

In the present work, the main goal was to evaluate the overall performance of the proposed model. However, future studies should include detailed ablation experiments to better understand the contribution of each component of the architecture.

One important aspect to analyze is the masking strategy used during training. Future experiments can compare standard random masking with the proposed human-inspired masking approach to study how focusing on text-related regions influences learning efficiency and recognition accuracy.

Another key factor is the positional encoding method used in the Transformer. Future research can investigate the difference between one-dimensional positional encoding and two-dimensional spatial encoding, especially considering the complex structure of Arabic scripts where spatial relationships play an important role.

In addition, the number of attention heads in the Transformer can be varied to analyze its effect on performance and computational cost. This would help in finding the optimal configuration for Arabic text recognition tasks.

4.8.3 Towards an Integrated Image and Video Recognition System

A promising direction for future work is the development of a unified system that can handle both image-based and video-based text recognition. By combining spatial feature extraction with temporal aggregation modules, the model could adapt to different input types without changing the core architecture.

Such a system would be useful in many real-world applications, including surveillance systems, traffic monitoring, smart city solutions, and multimedia content analysis.

4.8.4 Conclusion of the Chapter

To conclude, this chapter introduced a Transformer-based approach for Arabic text recognition with a primary focus on images. The results demonstrate strong performance and confirm the suitability of attention-based models for handling complex Arabic scripts.

While temporal modeling and detailed ablation studies were not implemented in the current research, they represent important future enhancements. Integrating temporal feature aggregation and performing deeper component analysis are expected to further improve system robustness and recognition accuracy in real-world video scenarios.

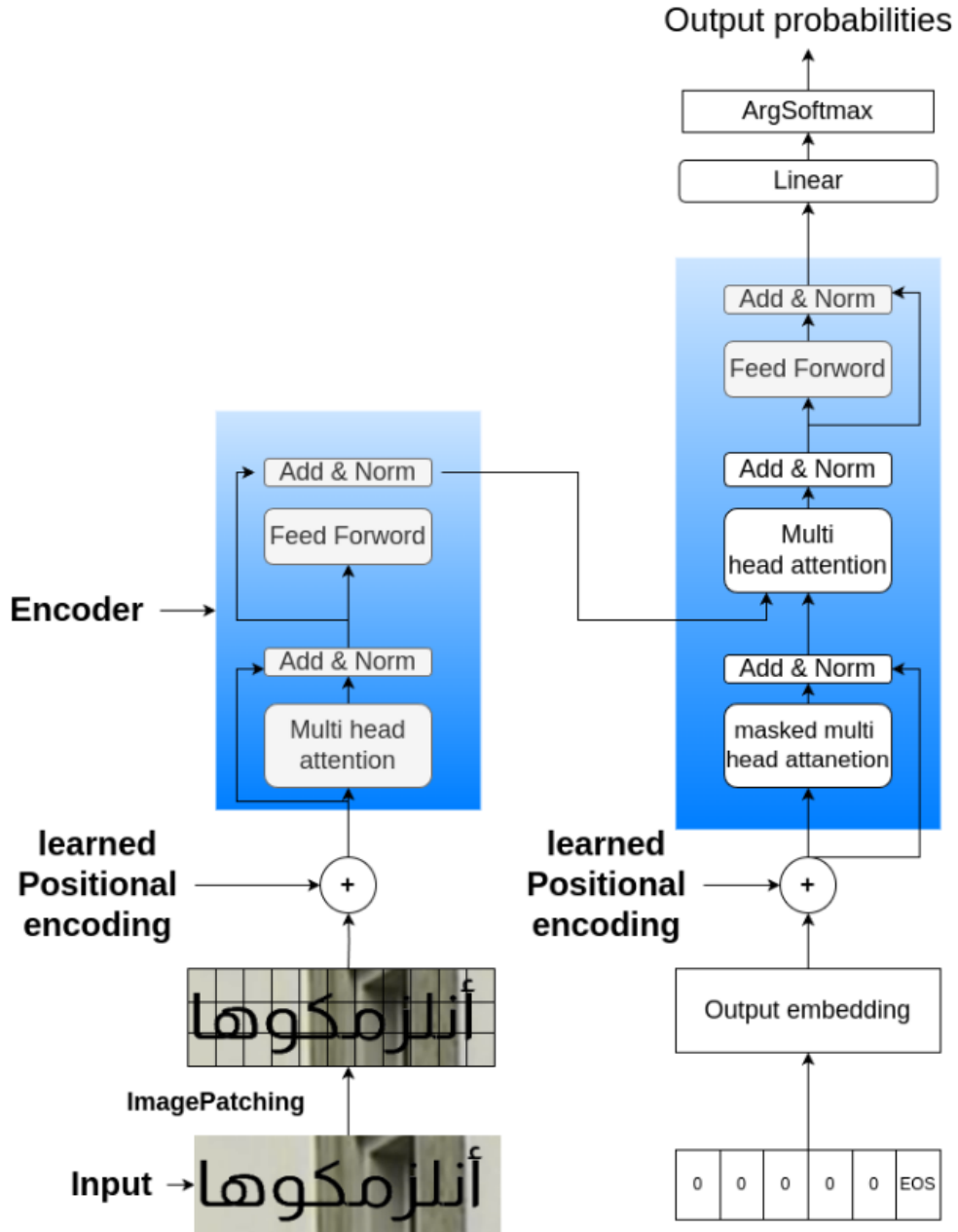


Figure 4.1: Overview of the proposed multi-head attention Transformer for Arabic scene text recognition

General Conclusion

The increasing availability of visual data in the form of images and videos has made automatic scene-text recognition an essential component of modern computer vision systems. Applications such as digital mapping, intelligent transportation, content indexing, and multimedia retrieval rely heavily on the accurate extraction of textual information from complex real-world environments. Despite the significant progress achieved in Latin-based scene-text recognition, Arabic text recognition from natural scenes remains a challenging and insufficiently explored research area. The complex structure of Arabic script, its cursive nature, contextual character forms, and the lack of large-scale datasets have limited the performance and generalization ability of existing OCR systems.

This thesis aimed to address these limitations by providing both data resources and a deep learning-based recognition framework specifically designed for Arabic scene-text recognition. The work focused on bridging the gap between traditional OCR approaches and the requirements of modern deep learning models for robust real-world text understanding.

The first major contribution of this thesis is the development of the **IYaD dataset**, a large-scale synthetic dataset dedicated to Arabic scene-text recognition. IYaD was designed to overcome the shortcomings of existing Arabic OCR datasets, which are often limited in size, diversity, and real-world variability. The dataset includes millions of Arabic words and sentences rendered across multiple fonts, background textures, lighting conditions, geometric distortions, and noise levels. This diversity allows deep learning models to learn robust feature representations and improves their ability to generalize to unseen scenarios.

Unlike traditional datasets that focus mainly on scanned documents or handwritten text, IYaD simulates real scene conditions, making it particularly suitable for scene-text recognition tasks. The dataset also supports data augmentation and benchmarking, providing a structured platform for fair comparison of different recognition methods. By offering a scalable and flexible resource, IYaD fills an important gap in Arabic OCR research and lays the foundation for future advancements in both image-based and video-based Arabic text recognition.

The second major contribution of this thesis is the design and implementation

of a **Transformer-based Arabic Scene Text Recognition (ASTR) model**. The proposed architecture leverages multi-head self-attention mechanisms to capture long-range spatial dependencies and contextual relationships between character components. This is especially important for Arabic script, where characters are connected, change shape based on context, and often share visual similarities.

Traditional OCR systems based on convolutional neural networks and recurrent neural networks tend to rely on local feature extraction and sequential processing, which limits their ability to model global relationships effectively. In contrast, the Transformer architecture enables parallel processing and dynamic attention over the entire feature representation, allowing the model to focus on relevant regions and contextual patterns.

Extensive experiments demonstrated that the proposed ASTR model achieves strong recognition performance on challenging datasets, reaching a character recognition rate exceeding 93.5% in controlled scenarios. The model also showed robustness to variations in font styles, background complexity, illumination changes, and distortions. These results confirm the effectiveness of attention-based architectures for Arabic scene-text recognition and highlight the importance of combining large-scale datasets with modern deep learning models.

Together, the IYaD dataset and the Transformer-based recognition framework form a complete and scalable pipeline for Arabic scene-text recognition. Beyond improving recognition accuracy, this thesis emphasizes the importance of designing language-specific resources and architectures that reflect the visual and linguistic characteristics of Arabic. The combination of synthetic data generation and attention-based modeling represents a promising direction for addressing the challenges of low-resource languages in computer vision.

In addition to its contributions to static image-based recognition, this thesis provides a strong basis for future research in video scene-text recognition. Videos introduce additional challenges such as motion blur, temporal inconsistency, occlusions, and dynamic lighting conditions. However, they also offer valuable temporal information that can be exploited to improve recognition accuracy. The Transformer architecture developed in this work is naturally suited for incorporating temporal feature aggregation mechanisms, such as temporal attention and sequence-level fusion, which can enhance robustness across video frames.

From a practical perspective, the outcomes of this thesis have significant implications for real-world applications. Improved Arabic scene-text recognition can enhance digital mapping services by accurately extracting street names, shop signs, and geographic information in Arabic-speaking regions. It can support intelligent transportation systems by enabling automated understanding of traffic signs and road information. Furthermore, it can improve document digitization, mobile translation tools, and multimedia content indexing for Arabic-language content.

The availability of a large-scale dataset such as IYaD also benefits the broader research community by enabling faster experimentation, reproducibility, and fair evaluation of recognition systems. Researchers can build upon this dataset to explore new architectures, training strategies, and multimodal approaches.

Several future research directions naturally emerge from this work. One important direction is the extension of the IYaD dataset with real-scene images and video frames containing Arabic text captured in natural environments. This would further reduce the gap between synthetic and real-world data and allow more realistic evaluation of recognition systems.

Another promising direction is the integration of temporal modeling techniques into the recognition framework. By exploiting the continuity of text across consecutive video frames, future models can mitigate common video-related challenges such as blur, occlusion, and resolution degradation.

Additionally, the development of end-to-end video scene text spotting systems represents a significant opportunity for advancing the field. Such systems would jointly perform text detection and recognition within a unified framework, enabling more efficient and scalable solutions for real-time applications. Transformer-based architectures are well-positioned to support such integrated pipelines due to their flexibility and ability to model complex dependencies.

Further research can also explore the incorporation of language models and multimodal information to improve contextual understanding and error correction, particularly for long text sequences and complex scenes.

In conclusion, this thesis contributes a comprehensive framework for advancing Arabic scene-text recognition through the introduction of a large-scale synthetic dataset and a powerful Transformer-based recognition model. The proposed solutions significantly improve robustness and accuracy while addressing long-standing limitations in Arabic OCR research. By laying the groundwork for future image-based and video-based recognition systems, this work opens new opportunities for both academic research and practical applications in Arabic visual understanding. The contributions of this thesis are expected to encourage further exploration in this field and support the growing demand for intelligent Arabic-language technologies in an increasingly visual digital world.

Bibliography

- [1] O. Khial and F. Bouferra, “Multi head attention transformer for arabic scene images text recognition,” *Journal of Information Systems Engineering and Management*, vol. 10, no. 62s, pp. 361–369, 2025. [Online]. Available: <https://www.jisem-journal.com/index.php/journal/article/view/13584/6412>
- [2] H. Hassan, A. El-Mahdy, and M. E. Hussein, “Arabic scene text recognition in the deep learning era: Analysis on a novel dataset,” *IEEE Access*, vol. 9, pp. 107 046–107 058, 2021.
- [3] Google Cloud, “Detect text in images | Cloud Vision API,” 2023, accessed: 2023-10-26. [Online]. Available: <https://docs.cloud.google.com/vision/docs/ocr>
- [4] Y. Li, M. Ibrayim, and A. Hamdulla, “Summary of scene text detection based on deep learning,” in *2021 International Conference of Social Computing and Digital Economy (ICSCDE)*, 2021, pp. 318–323.
- [5] Y. Zhang, W. Wang, L. Wang, and L. Wang, “Scene text recognition with deeper convolutional neural networks,” in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 2384–2388.
- [6] H. Butt, M. R. Raza, M. J. Ramzan, M. J. Ali, and M. Haris, “Attention-based cnn-rnn arabic text recognition from natural scene images,” *Forecasting*, vol. 3, no. 3, pp. 520–540, 2021. [Online]. Available: <https://www.mdpi.com/2571-9394/3/3/33>
- [7] A. Zewe, “3 questions: The pros and cons of synthetic data in ai,” MIT News, September 2025. [Online]. Available: <https://computing.mit.edu/news/3-questions-the-pros-and-cons-of-synthetic-data-in-ai/>
- [8] M. Ibrayim, Y. Li, and A. Hamdulla, “Scene text detection based on two-branch feature extraction,” *Sensors*, vol. 22, no. 16, p. 6262, 2022.

-
- [9] S. Yousfi, S.-A. Berrani, and C. Garcia, "Alif: A dataset for arabic embedded text recognition in tv broadcast," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 1221–1225.
- [10] O. Zayene, S. Masmoudi Touj, J. Hennebert, R. Ingold, and N. E. Ben Amara, "Open datasets and tools for arabic text detection and recognition in news video frames," *Journal of Imaging*, vol. 4, no. 2, p. 32, 2018.
- [11] O. Khial, F. Boufera, and R. B. B. Rochdi, "Iyad: A dataset for arabic printed text recognition in natural scene videos," in *Proceedings of International Conference - 2025*, Mascara, Algeria, August 2025, pp. 8–15.
- [12] S. Karwal *et al.*, "A survey on optical character recognition techniques," *International Journal of Computer Applications*, vol. 123, no. 5, 2014.
- [13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [14] W. Huang, Z. Lin, J. Yang, and J. Wang, "Text localization in natural images using stroke feature transform and text covariance descriptors," *CoRR*, vol. abs/1603.00000, 2016. [Online]. Available: <https://arxiv.org/abs/1603.00000>
- [15] M. Abdullah, F. Al-Khateeb, and K. Omar, "A review on arabic optical character recognition systems," *Journal of King Saud University – Computer and Information Sciences*, vol. 32, no. 7, pp. 814–830, 2020.
- [16] D. Deng *et al.*, "Tcit: Real-time video text spotting in complex scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
- [17] J.-C. Burie *et al.*, "Robust scene text recognition for embedded systems," in *International Conference on Pattern Recognition (ICPR)*, 2018.
- [18] A. Vaswani and et al., "Attention is all you need," 2017.
- [19] L. Gasper *et al.*, "Multimodal neural networks for image-to-text understanding," *Pattern Recognition*, vol. 124, 2022.
- [20] X. Chen *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [21] A.-A. Bhuiyan and A. Khan, "Image quality assessment employing rms contrast and histogram similarity," *International Arab Journal of Information Technology*, vol. 15, pp. 983–989, 11 2018.

-
- [22] W. J. Harrison, “Luminance and contrast of images in the things database,” *bioRxiv*, 2021. [Online]. Available: <https://www.biorxiv.org/content/early/2021/10/24/2021.07.08.451706>
- [23] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial transformer networks,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015.
- [24] N. Ahmad *et al.*, “Arabic document ocr using deep learning techniques,” *Journal of King Saud University – Computer and Information Sciences*, 2021.
- [25] Y. Al-Ohali, M. Cheriet, and C. Suen, “Databases for recognition of handwritten arabic cheques,” *Pattern Recognition*, vol. 36, no. 1, pp. 111–121, 2003.
- [26] A. Mezghani, S. Kanoun, M. Khemakhem, and H. El Abed, “A database for arabic handwritten text image recognition and writer identification,” in *2012 International Conference on Frontiers in Handwriting Recognition*, 2012, pp. 399–402.
- [27] H. El Abed, M. Kherallah, V. Märgner, and A. M. Alimi, “On-line arabic handwriting recognition competition: Adab database and participating systems,” *IJDAR*, vol. 14, pp. 15–23, 2011.
- [28] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, “Tea: Temporal excitation and aggregation for action recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [29] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, “Actionvlad: Learning spatio-temporal aggregation for action classification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [30] Z. Liu, L. Wang, W. Wu, C. Qian, and T. Lu, “Tam: Temporal adaptive module for video recognition,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [31] A. J. Yousif and M. H. Al-Jammas, “Semantic-based temporal attention network for arabic video captioning,” *Natural Language Processing Journal*, vol. 10, p. 100122, 2025.
- [32] H. S. Lee, M. Kim, S. Jang, H. B. Bae, and S. Lee, “Multi-granularity aggregation with spatiotemporal consistency for video-based person re-identification,” *Sensors*, vol. 24, p. 2229, 2024.

-
- [33] A. Alghamdi *et al.*, “Aia9k: A dataset for isolated arabic characters and digits,” *Data in Brief*, vol. 28, 2020.
- [34] A. Abdelli *et al.*, “Moudle: A large arabic document text dataset,” *Journal of Imaging*, vol. 7, no. 12, 2021.
- [35] F. K. Jaïem, S. Kanoun, M. Khemakhem, H. El Abed, and J. Kardoun, “Aptid/mf: Database for arabic printed text recognition research,” in *Image Analysis and Processing – ICIAP 2013*. Springer, 2013, pp. 251–259.
- [36] M. Elmamoun *et al.*, “Arabic scene text recognition using deep neural networks,” *IEEE Access*, vol. 9, pp. 113 245–113 256, 2021.
- [37] Y. Cheikh *et al.*, “Synthetically generated arabic text dataset for ocr,” *Expert Systems with Applications*, 2023.
- [38] S. Nayak *et al.*, “Synthetic data generation for ocr systems,” *Pattern Recognition Letters*, vol. 105, pp. 52–58, 2018.
- [39] H. Zhao *et al.*, “Large-scale synthetic dataset generation for scene text recognition,” *IEEE Transactions on Image Processing*, 2023.
- [40] A. Jawad *et al.*, “Handwritten medical text recognition using deep learning,” *Computer Methods and Programs in Biomedicine*, 2022.
- [41] A. Gutierrez *et al.*, “Reading historical manuscripts with deep neural networks,” *Pattern Recognition*, vol. 121, 2022.
- [42] C. Garcia *et al.*, “Four-modal learning for vision-language understanding,” *IEEE Transactions on Multimedia*, 2024.
- [43] O. Zayene, S. E. Amamou, and N. E. BenAmara, “Arabic video text recognition based on multi-dimensional recurrent neural networks,” in *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, 2017, pp. 725–729.
- [44] F. Slimane, R. Ingold, S. Kanoun, A. M. Alimi, and J. Hennebert, “A new arabic printed text image database and evaluation protocols,” in *2009 10th International Conference on Document Analysis and Recognition*, 2009, pp. 946–950.
- [45] T. C. Nguyen and T. S. Huang, “Image blurring effects due to depth discontinuities: Blurring that creates emergent image details,” in *ECCV 1992*, 1992, pp. 347–362.

-
- [46] L. Mahon and T. Lukasiewicz, “Minimum description length clustering to measure meaningful image complexity,” *Pattern Recognition*, vol. 144, 2023.
- [47] N. El-Bendary, H. M. Zawbaa, M. S. Daoud, A. E. Hassanien, and K. Nakamatsu, “Arslat: Arabic sign language alphabets translator,” in *2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM)*, 2010, pp. 590–595.
- [48] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [49] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, pp. 1735–80, 12 1997.
- [50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.
- [51] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021.
- [52] E.-M. Papia, A. Kondi, and V. Constantoudis, “Entropy and complexity analysis of ai-generated and human-made paintings,” *Chaos, Solitons & Fractals*, vol. 170, p. 113385, 2023.
- [53] D.-Y. Tsai, Y. Lee, and E. Matsuyama, “Information entropy measure for evaluation of image quality,” *Journal of digital imaging*, vol. 21, pp. 338–347, 2008.
- [54] B. Parameshachari, H. Panduranga, S. liberata Ullo *et al.*, “Analysis and computation of encryption technique to enhance security of medical images,” in *IOP conference series: materials science and engineering*, vol. 925, no. 1. IOP Publishing, 2020, p. 012028.
- [55] A. Potier, R. Galicher, P. Baudoz, E. Huby, J. Milli, Z. Wahhaj, A. Boccaletti, A. Vigan, M. N’Diaye, and J.-F. Sauvage, “Increasing the raw contrast of vlt/sphere with the dark hole technique-i. simulations and validation on the internal source,” *Astronomy & Astrophysics*, vol. 638, p. A117, 2020.
- [56] L. Mahon and T. Lukasiewicz, “Minimum description length clustering to measure meaningful image complexity,” *Pattern Recognition*, vol. 144, 2023.