Handwritten Arabic Calligraphy Generation: A Systematic Literature Review

Afnan Sumayli[1], Mohamed Alkaoud[2]
Department of Computer Science, College of Engineering and Computer Science, Jazan University, Jazan 86363, Saudi Arabia[1]
Department of Computer Science, College of Computer and Information Sciences, King Saud University,
Riyadh 12372, Saudi Arabia[2]

***Abstract*—Arabic calligraphy is famous for its distinct artistic style. It is written by skilled calligraphers to highlight the beauty of Arabic letters and represent its rich artistry. Due to the complexity of Arabic text compared to other languages' scripts, Arabic calligraphy writing demands a significant investment of time and effort, as well as the acquisition of high skills from calligraphers to correctly form the curves of Arabic script and accurately represent its various styles. This Systematic Literature Review (SLR) aims to provide a comprehensive analysis of the current state of research in Arabic calligraphy generation using deep learning and generative models. The review follows the PRISMA guidelines and examines 19 primary studies selected from a systematic search of academic databases, with publications spanning from January 2009 to December 2024. The findings indicate that Generative Adversarial Networks (GANs) and their variants are the most commonly used models for generating Arabic calligraphy. Additionally, the review highlights a significant gap in the availability of large, standardized handwritten datasets for model training and evaluation, as most existing datasets are small, custom-made, or privately held. In conclusion, the review offers valuable insights that can help researchers and practitioners advance the field, enabling the generation of high-quality Arabic calligraphy that satisfies both artistic and functional needs.**

Keywords—Arabic calligraphy; deep learning; generative models; handwritten dataset; Generative Adversarial Networks

I. Introduction

Calligraphy represents an artistic way of handwriting. Generally, writing by hand is quite a complicated movement that presents challenges in analyzing and emulating it [1]. However, Arabic script is represented by 28 alphabets written from right to left. Arabic letters are typed in various forms depending on their position in the word: beginning, middle, end, or isolated, as shown in Table I. Furthermore, Arabic calligraphy comes in several writing styles; the six primary styles, also known as "six pens", which are Naskh, Kufic, Diwani, Thuluth, Farsi, and Reqaa [2]; Fig. 1 represents examples of some Arabic calligraphy styles. Thus, creating Arabic calligraphy can be time-consuming and demands professional skills.

Arabic calligraphy is more than just writing; it's a cultural and artistic tradition that has been preserved for centuries. The beauty and complexity of its designs make it a challenging task for computers to replicate. Automating the creation of Arabic calligraphy is important for many reasons. It can help artists and designers create personalized calligraphy, offer tools for teaching Arabic calligraphy in an engaging way, assist in digitizing and preserving historical manuscripts, and lead to the development of new Arabic fonts for digital and print use.

Recent advancements in artificial intelligence (AI) and deep learning have opened new avenues for Arabic calligraphy generation. Techniques such as Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), and transformer-based models have shown promise in generating realistic and diverse calligraphic outputs. Despite these advancements, the field faces several challenges. One major issue is the lack of large, standardized datasets for training models. Most datasets are small, custom-made, or not publicly available, making it hard to share and build on existing research. Arabic calligraphy is also complex because of its intricate letterforms, especially in styles like Diwani and Thuluth as shown in Fig. 1, which are difficult for models to capture accurately. Additionally, the field lacks clear and consistent ways to evaluate the quality of generated calligraphy. These limitations affect the development of models capable of generating high-quality Arabic calligraphy that meets both artistic and functional requirements.

This Systematic Literature Review (SLR) aims to provide a comprehensive analysis of the current state of research in Arabic calligraphy generation and the datasets used for this purpose. Specifically, this review seeks to identify trends, gaps, and future directions in the field. The findings of this review will offer valuable insights for researchers, practitioners, and stakeholders interested in advancing the state of the art in Arabic calligraphy generation. The main contributions of our review are as follows:

- We provide a detailed analysis of the techniques and challenges in Arabic calligraphy generation.
- We critically evaluate existing datasets and their limitations.
- We propose a roadmap for future research, including the development of standardized datasets, the establishment of robust evaluation metrics, and the development of advanced models tailored for handwritten calligraphy generation.

The remainder of this paper is organized as follows: Section II presents the research methodology, including the search strategy, inclusion/exclusion criteria, and data extraction process.

TABLE I EXAMPLES OF SOME ARABIC LETTERS IN DIFFERENT POSITIONS

Isolated	Beginning	Middle	End
ج	جـــ	ــجــ	ـــج
ف	فــ	ـــفـــ	ـــف
ق	قــ	ـــقـــ	ــق
ك	كــــ	ــكـــ	ـــك
ل	لـــ	ـــلـــ	ـــل

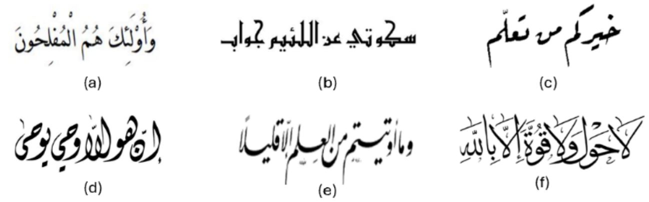

Fig. 1. Illustrative examples of some Arabic calligraphy styles, (a) Naskh (b) Kufic (c) Reqaa (d) Diwani (e) Farsi (f) Thuluth.

Section III discusses the review findings, addressing each research question in detail. Section IV provides a discussion of the key findings and limitations. Section V outlines future research directions. Finally, Section VI concludes the paper with a summary of the contributions and implications for future research.

II. RESEARCH METHODOLOGY

This study follows the PRISMA guidelines [3]. The methodology is divided into four main stages: (1) Research Questions and Objectives, where the scope and goals of the review were defined; (2) Search Strategy, involving a systematic exploration of relevant academic databases; (3) Inclusion and Exclusion Criteria, where studies were evaluated for relevance and quality; and (4) Quality Assessment, focusing on evaluating the methodological rigor of the selected studies. Each stage is described in detail in the following subsections.

A. *Research Questions and Objectives*

The primary goal of this SLR is to provide a comprehensive analysis of the current state of research in Arabic text generation, identify the various techniques applied to generate Arabic handwritten calligraphy and the availability of sufficient datasets for training and evaluating these types of research directions. The main research questions (RQs) were raised to achieve this aim include:

RQ1: What are the key generative models and techniques used to generate Arabic handwritten calligraphy?

RQ2: What is the level of Arabic text generated by the litterateur?

RQ3: What are the main challenges and limitations in the field of Arabic calligraphy generation?

RQ4: What are the standard datasets for Arabic calligraphy generation in literature?

B. *Search Strategy*

A systematic search was conducted to identify all relevant literature on “Arabic handwritten text generation” and “Arabic handwritten text datasets”. The search was performed across five major academic databases: IEEE Xplore, ScienceDirect, SpringerLink, Google Scholar, and ACM Digital Library.

To align with the research objectives and questions, the search keywords were divided into two main categories: (1) Arabic calligraphy generation, focusing on techniques for generating Arabic handwritten text, and (2) Arabic calligraphy datasets, emphasizing datasets used for training and evaluation. The selected keywords are presented in Table II.

Boolean operators were employed to construct search queries. The OR operator was used to combine keywords within each category, while the AND operator was used to concatenate keywords across categories. For example, a sample query was structured as follows:

("Arabic calligraphy generation" OR "Arabic handwritten text generation") AND ("Arabic calligraphy dataset" OR "Arabic handwritten dataset")

The search query was applied to the title, abstract, and keywords of studies published between January 2009 and December 2024. This time frame was chosen to capture the most recent advancements in the field while ensuring a sufficient breadth of literature for analysis. The initial search yielded 269 records, which were subsequently screened for relevance and quality. After extracting the studies from each database, duplicates were removed. In the process of eliminating duplicates, 22 studies were excluded, resulting in 247 unique studies for further screening.

C. *Inclusion and Exclusion Criteria*

The screening process was conducted systematically to ensure the inclusion of studies that align with the research objectives. Initially, the titles and abstracts of the 247 remaining studies were reviewed to assess their relevance. When necessary, the full text of the articles was evaluated to determine their eligibility based on the predefined inclusion and exclusion criteria. This rigorous process resulted in the selection of 19 primary studies for inclusion in this review. The majority of the excluded articles focused on Arabic Handwritten Recognition or Arabic Calligraphy Classification, which fall outside the scope of this study.

The following criteria were used to identify studies relevant to Arabic calligraphy generation and datasets:

- The paper must be a peer-reviewed publication.
- The paper must be published in the English language.
- The paper must be published between January 2009 and December 2024.
- The paper must include an Arabic calligraphy generation model or a dataset for Arabic calligraphy.

Studies were excluded if they met any of the following conditions:

- The paper focused on Arabic handwritten recognition, text segmentation, or classification tasks.
- The dataset was limited to Arabic digits or other non-calligraphy-related tasks.

- The paper lacked sufficient methodological detail or empirical results relevant to Arabic handwritten generation.

All relevant papers were systematically marked on a spreadsheet, downloaded, and organized using Mendeley software. This approach ensured efficient management of the studies and facilitated the extraction and synthesis of data during the review process, Fig. 2 shows a summary of the search process.

D. *Quality Assessment*

A quality assessment (QA) was conducted for the 19 primary studies included in this review to evaluate their credibility, reliability, and methodological rigor. The QA was performed using a set of predefined questions, as shown in Table III, with each question answered as Yes (scored as 1) or No (scored as 0). The first question assessed whether the study clearly stated its objectives, which 94% of the studies answered positively. The second question evaluated the relevance of the studies to Arabic calligraphy generation or datasets. Only 36% of the studies directly addressed this field, while the remaining studies focused on related areas, such as Arabic handwritten text recognition or classification datasets. The third question examined whether the study provided a comprehensive explanation of the approach or methodology used, and 84% of the studies responded positively. The fourth question evaluated the use of appropriate evaluation metrics, with only 47% of the studies answering positively. Finally, the fifth question assessed whether the study clearly stated its findings and contributions, and 79% of the studies met this criterion. The results of the QA, summarized in Fig. 3. However, the quality review process did not rule out any study, as all the studies met the minimum quality threshold based on the assessment questions. Therefore, this review included all 19 studies selected during the screening process.

TABLE II KEYWORDS USED FOR SEARCH PROCESS

Group 1"Arabic calligraphy generation"	Arabic calligraphy generation Deep learning for Arabic calligraphy Arabic handwritten generation Generative models for Arabic calligraphy
Group 2 "Arabic calligraphy dataset"	Arabic calligraphy dataset Arabic handwritten dataset

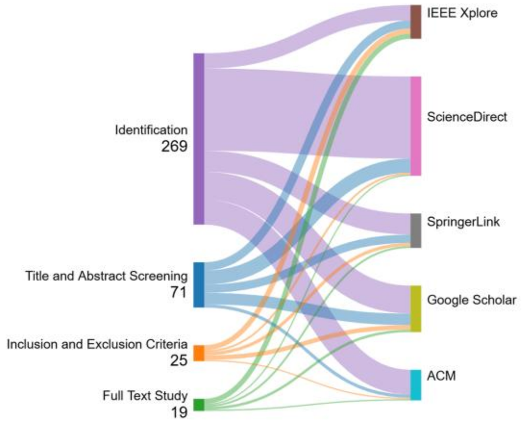

Fig. 2. A brief summary of the search process.

TABLE III QUALITY ASSESSMENT QUESTIONS

AQ	Assessment Question
1	Are the research objectives clearly stated?
2	Does the study directly address Handwritten Arabic calligraphy generation or the creation of Arabic calligraphy datasets?
3	Does the study clearly describe the research methodology?
4	Are the results supported by appropriate evaluation metrics
5	Are the findings and contributions of the study explicitly stated?

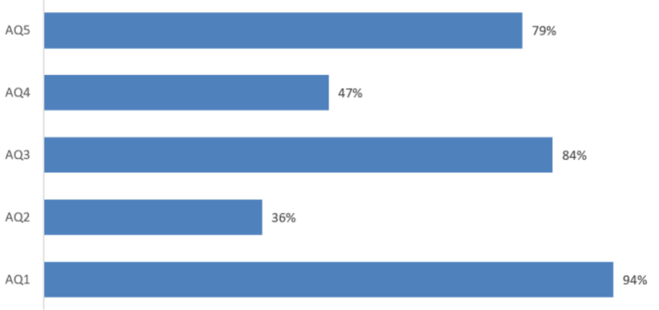

Fig. 3. The percentage-based quality assessment scores of the studies.

III. REVIEW FINDINGS

This section presents review findings by analyzing the 19 primary studies from four aspects, namely the generation approach, the text generation level, the dataset and metric used.

A. *RQ1: What are the Key Generative Models and Techniques used to Generate Arabic Handwritten Calligraphy?*

The literature identifies Generative Adversarial Networks (GANs) as the primary deep learning approach for Arabic calligraphy generation, owing to their ability to produce realistic and diverse outputs. Several variants of GANs, including Pix2Pix, Deep Convolutional GANs (DCGAN), CycleGAN, and Vector Quantized GAN (VQ-GAN), have been applied to this domain, each addressing specific challenges and use cases. For example, Ahmed et al. [4] and Chebouat [5] utilized DCGAN to generate Arabic calligraphy letters from handwritten images. These studies incorporated architectural modifications, such as adding Gaussian noise and altering activation functions, to enhance model performance. Similarly, CycleGAN has been employed for style transfer, transforming handwriting into specific calligraphic (e.g. Naskh and Thuluth). However, as noted in Ahmed et al. [4], CycleGAN struggles with accurately mapping complex geometric features, limiting its effectiveness for intricate styles. Another study proposed by Hadj Azzem et al. [6] explored the use of pix2pix and CycleGAN for image-to-image translation, specifically converting computer fonts (e.g. Arial) into three Arabic calligraphy styles (Diwani, Reqaa, and Farsi). While pix2pix preserved the shape of the ground truth, it introduced noticeable noise, whereas CycleGAN produced visually appealing results but faced challenges in accurately mapping certain features. Additionally, Bagido [7] demonstrated the creative potential of VQ-GAN by combining Arabic calligraphy with Rawashin art (wooden windows of Hijazi buildings), producing high-quality artistic designs. A summary of the reviewed studies, categorized by the GAN methods used, is presented in Table IV.

B. RQ2: What is the Level of Arabic Text Generated by Literature?

The level of Arabic text generated by the literature varies, with most studies focusing on letter-level generation and a smaller subset addressing word-level generation. However, the quality of the generated letters in many studies [4], [5], [7] was reported to be suboptimal, with outputs often being unclear or of poor quality. For instance, while Ahmed et al. [4] and Chebouat [5] utilized DCGAN to generate Arabic calligraphy letters, the results were inconsistent, with some letters being poorly formed or unrecognizable. Similarly, Bagido [7] employed VQ-GAN to create artistic designs combining Arabic calligraphy with Rawashin art, but the generated letters were often unclear, limiting their practical applicability. In contrast, the work that has been done by Hadj Azzem et al. [6] stands out as the best in terms of quality, producing clear and accurate Arabic text. However, it did not focus on handwritten text generation, instead generating printed or stylized text. A limited number of studies, including Ahmed et al. [4] and Hadj Azzem et al. [6], explored a word-level generation, which presents additional challenges such as maintaining contextual coherence and geometric consistency across multiple letters. While these studies represent a step forward, the overall quality of generated text remains a significant limitation, particularly for handwritten calligraphy. In summary, the level of Arabic text generation in the reviewed literature is moderate to low, with persistent issues in clarity and quality at both the letter and word levels. This highlights the need for further research to improve the robustness and accuracy of generative models, particularly for handwritten Arabic calligraphy.

C. RQ3: What are the main Challenges and Limitations in the Field of Arabic Calligraphy Generation?

The field of Arabic calligraphy generation faces several significant challenges and limitations, as highlighted by the reviewed literature. One of the most pressing issues is the lack of standardized datasets. While Hadj Azzem et al. [6] introduced a private dataset named Arabic Calligraphy Generation-3 (ACG-3), consisting of 14,908 pairs of images, Ahmed et al. [4], Chebouat [5], and Bagido [7] relied on small and custom datasets, which limit the generalizability and reproducibility of results. This is particularly problematic for deep learning models, which require large amounts of high-quality data to achieve optimal performance. Another major challenge is the complexity of Arabic calligraphy styles, such as Diwani, Thuluth, and Kufic, which require precise geometric accuracy and artistic variation. While advanced models like GANs have shown promise, they often struggle with capturing intricate geometric features, as seen in the case of CycleGAN [4].

Additionally, evaluation metrics remain a significant limitation. Many studies relied on subjective human judgment or qualitative assessments, which lack objectivity and consistency. For example, Ahmed et al. [4] and Bagido [7] used surveys and visual inspections to evaluate their results. While these methods capture subjective aspects such as aesthetic quality, they do not provide standardized or quantifiable measures of accuracy or performance. A notable example of addressing this limitation is the work by Hadj Azzem et al. [6], who employed Fréchet Inception Distance (FID) scores to quantitatively evaluate the performance of pix2pix and CycleGAN models. FID measures the similarity between generated and real images by comparing their feature distributions, providing a more objective measure of model performance. Similarly, other studies have used precision, recall, and F1-score to evaluate the accuracy of calligraphy recognition systems, particularly for tasks like character or style classification. For example, Kaoudja et al. [8] and Allaf et al. [9], utilized these metrics to assess the performance of their calligraphy style classification model, achieving high accuracy across multiple styles. Despite these advancements, the field still lacks a unified framework for evaluating Arabic calligraphy generation, as most evaluation techniques focus on specific aspects (e.g., image quality or recognition accuracy) rather than providing a holistic assessment of both artistic and functional qualities.

Furthermore, computational resource requirements pose a barrier, as training advanced models like GANs and transformer-based architectures demands significant computational power and time. Finally, while some studies [6] have achieved high-quality results, they often focus on printed or stylized text rather than handwritten calligraphy, leaving a gap in the literature for generating realistic handwritten Arabic text. These challenges highlight the need for standardized datasets, improved evaluation metrics, more efficient models, and a greater focus on handwritten text generation to advance the field. Table V summarizes the evaluation metrics used by studies present in this literature.

D. QR4. What are the Standard Datasets for Arabic Calligraphy Generation in Literature?

The literature reveals a variety of datasets used for Arabic handwritten and machine-generated text, each with unique characteristics and applications. These datasets can be broadly categorized into three types: handwritten text datasets, calligraphy datasets, and machine-generated text datasets.

1) Handwritten text datasets: The KHATT Dataset [10] is one of the most widely used datasets in Arabic handwritten text research. It consists of 6,712 lines and words written by 1,000 writers across 18 countries. The dataset is publicly available and includes annotations in text and XML files. However, it primarily focuses on non-artistic Arabic text, as illustrated in Fig. 4, which limits its application for Arabic calligraphy generation that requires more artistic and stylized features. Another notable dataset is Arabic Handwritten Letters Dataset proposed by [11], which includes 2,800 images of Arabic letters written by 10 native Arabic writers. Each letter is written ten times, providing valuable data for letter recognition tasks. However, this dataset lacks the diversity necessary for word-level or sentence-level calligraphy generation, limiting its use for training more complex models.

2) Calligraphy datasets: Several datasets are dedicated to Arabic calligraphy and can be used for training models focused on artistic styles and handwritten calligraphy generation.

The Arabic Calligraphic Letters (ACL) Dataset [12] contains 3,467 images of individual Arabic letters, categorized into 32 classes. While it is publicly available, the dataset is limited to isolated characters, as shown in Fig. 5. The dataset compiled by

Allaf et al. [9] is a private collection of 267 text images across three calligraphy styles (Reqaa, Thuluth, Kufic). These images are manually segmented into 71-word images per style, making it a valuable resource for training models focused on specific calligraphic styles. However, the small size and private access limit its widespread use. The dataset proposed by Kaoudja et al. [8] is another significant resource, comprising 1,685 high-resolution images of Arabic text in nine calligraphic styles (Naskh, Reqaa, Diwani, Thuluth, Parsi, Kufic, Square-Kufic, Maghribi, Mohakek).

TABLE IV ARABIC TEXT GENERATION APPROACHES

Model used	Studies
DCGAN	[4], [5]
vanilla GAN, VQ GAN	[7]
CycleGAN	[4], [6]

TABLE V EVALUATION METRICS FOR ARABIC TEXT GENERATION USED IN THE LITERATURE

Evaluation metric	studies
Human Assessment	[4] [5], [7], [6]
FID score	[6]

This dataset is publicly available and provides a rich source of data for Arabic calligraphy style recognition. The dataset proposed by Belila and Gasmi [13] was created by segmenting sentences collected by Kaoudja et al. [8], with the cropping process designed to preserve the calligraphy features in the generated images. 100 sentences were equally selected from the nine Arabic styles, producing high-resolution images. The cropped images maintain features that capture correlations between segmented images. However, some images in this dataset suffer from background noise, which can interfere with model performance. Fig. 6 illustrates sample images from one class of the dataset.

The CALLIAR Dataset [14] is another publicly available resource with 2,500 images and 45,000 strokes across multiple styles (Diwani, Thuluth, Kufic, and Farsi). Fig. 7 illustrates a sample from the CALLIAR Dataset. Although it is a valuable resource for calligraphy generation, its relatively small size limits its utility for training deep learning models. The HICMA Dataset [15], the largest publicly accessible calligraphy dataset, includes over 5,000 images across five styles (Kufic, Naskh, Diwani, Thuluth, and Mohakek). However, like Belila and Gasmi [13] dataset, some images contain background noise, which may hinder model performance. The KERTAS Dataset [16] contains 2,000 images from manuscripts spanning 14 Islamic centuries. While publicly available and focusing on historical Arabic manuscripts, it lacks the diversity of calligraphy styles required for modern calligraphy generation tasks. Fig 8 illustrates a sample from the KERTAS Dataset. Other notable datasets include the study proposed by Alrehali et al. [17], a private dataset of 5,240 images from 7th and 8th-century manuscripts, and the dataset proposed by Khayyat et al. [18], which includes 2,653 images from 37 manuscripts covering six styles. Both datasets are not publicly available, limiting their contribution to the field.

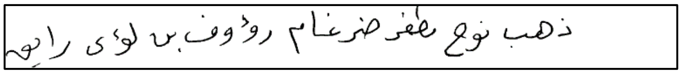

Fig. 4. Sample of line text from the KHATT [10] dataset.

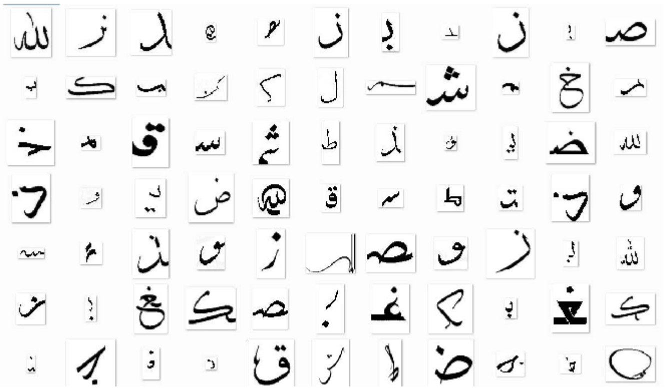

Fig. 5. Sample of images from the ACL dataset [12].

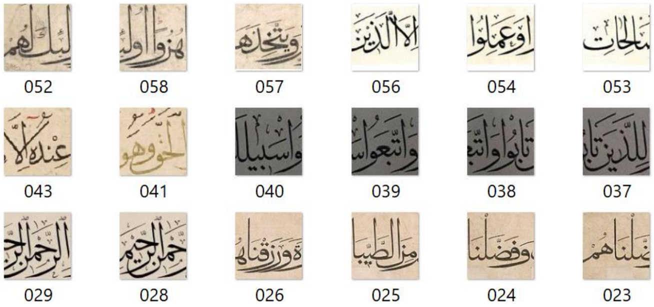

Fig. 6. Sample images of one calligraphy style from the Belila and Gasmi [13] dataset.

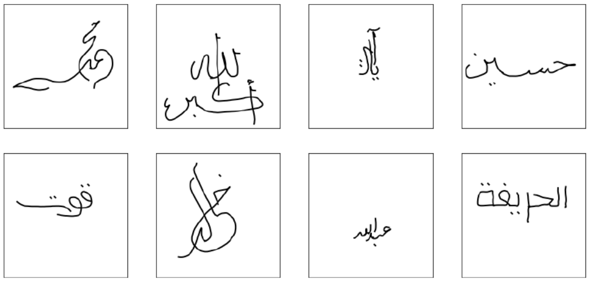

Fig. 7. Sample of word images from the CALLIAR dataset [14].

3) Machine-generated text datasets: The APTI Dataset [19] is one of the largest resources for printed Arabic text, containing over 45 million images of 113,284 words. While publicly available, the dataset focuses on machine-written text and lacks the artistic qualities necessary for training models in Arabic calligraphy generation. Fig. 9 presents sample word images from the APTI Dataset. The PATDB dataset [20] is another dataset with 6,954 pages collected from books, chapters, advertisements, and newspapers. While it is freely available, it is not specifically designed for calligraphy generation, focusing more on printed Arabic text.

Out of the 15 datasets reviewed, 10 are publicly available (69%) [8], [10], [11], [12], [13], [14], [15], [16], [19], [21], while

5 are private (31%) [9], [17], [18], [20], [22]. The largest dataset is the APTI Dataset [19], with over 45 million images, followed by the HICMA Dataset [15] with 5,031 images and the CALLIAR Dataset [14] with 2,500 images. Smaller datasets, such as Allaf et al. [9] dataset (267 images) and HAMCDB [22] (1,560 images), are limited in scope. The most commonly used Arabic calligraphy styles include Naskh [8], [15], [18], [21], Thuluth [8], [14], [15], [18], Diwani [8], [14], [15], [18], Kufic [8], [14], [15], [21], and Reqaa [9], [18]. These styles are featured across several datasets and are central to the recognition and generation of Arabic calligraphy applications. Fig. 10 illustrates the dataset publication years, highlighting the trends in dataset development over time.

There are various sample types across the datasets, typically containing images at different levels: Character-level samples [11], [12], [21], [22], Word-level samples [8], [9], [18], Line-level samples [10], [14], [15], [17], or Page-level samples [16], [18], [20], [21]. The variation in sample types reflects the intended tasks for each dataset. Character- and word-level samples are more suited for Arabic text generation tasks, where the focus is on individual letters or words. On the other hand, line- and page-level samples are more appropriate for broader tasks such as text recognition or the generation of full-length, artistic calligraphy. Table VI illustrates a comparative analysis of the mentioned images datasets based on five criteria namely number of samples, data type, size of images, number of Arabic calligraphy styles, and whether the data was accessible or not.

IV. Discussion

The analysis of the reviewed datasets reveals several important findings and highlights key limitations that need to be addressed for the effective development of Arabic handwritten calligraphy generation models. Despite the availability of multiple datasets, none are sufficiently large or diverse to fully support the training of deep learning models in this domain. One of the primary limitations is the lack of diversity across datasets.

Fig. 8. Sample of images from the KERTAS Dataset [16].

Image_10_Arabic Transparent_46
Image_10_Arabic Transparent_47
Image_10_Arabic Transparent_48
Image_10_Arabic Transparent_49
Image_10_Arabic Transparent_38
Image_10_Arabic Transparent_43
Image_10_Arabic Transparent_44
Image_10_Arabic Transparent_35
Image_10_Arabic Transparent_36
Image_10_Arabic Transparent_37
Image_10_Arabic Transparent_29
Image_10_Arabic Transparent_30
Image_10_Arabic Transparent_31
Image_10_Arabic Transparent_32
Image_10_Arabic Transparent_33
Image_10_Arabic Transparent_23
Image_10_Arabic Transparent_24
Image_10_Arabic Transparent_16
Image_10_Arabic Transparent_17
Image_10_Arabic Transparent_18

Fig. 9. Samples of word images from the APTI [19] dataset.

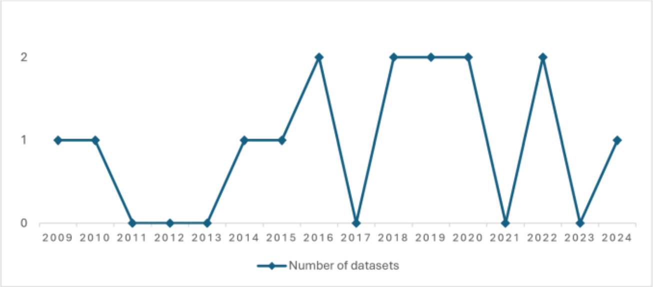

Fig. 10. Datasets publication year-wise summary.

TABLE VI SUMMARY OF THE PRESENTED DATASETS IN THE LITERATURE

studies	Sample No.	Data Type	Image size	Style No.	Publicly Available
APTI [19] (2009)	45.31 million, Machine written text	Word	Varied sizes	10	Yes
KHATT [10] (2014)	2,000/9,327	Paragraph/Line	-	Free style	Yes
Allaf et al. [9] (2016)	267	Line / Word	Different sizes, optimized by Genetic Algorithm	3	No
CALLIAR[14] (2016)	2,500	Sentence	64 × 64	*	Yes
KERTAS [16] (2018)	2,000	Page	50 × 50	-	Yes
ACL [12] (2018)	3,467	Character	64 × 64	-	Yes
Kaoudja et al. [8] (2019)	1,685	Line	-	9	Yes
Alrehali et al. [17] (2020)	5,240	Character	30 × 30	1, Naskh	No
Khayyat et al. [18] (2020)	2,653	Page	224 × 224	6	No
Belila and Gasmi [13] (2022)	900	Word	100 × 100	9	Yes
HICMA[15] (2024)	5,031	Line	-	5	Yes
AlKhateeb [11] (2015)	2,800	Character	-	Free style	Yes
BADAM[21] (2019)	400	Page	-	4	Yes
PATDB [20] (2010)	6954	Page	Varied sizes	-	No
HAMCDB [22] (2022)	1560	Character	-	1, Maghrebi	No

- Unspecified. * The exact number of styles was not explicitly stated in Alyafeai et al. [14], but it mentioned Diwani, Thuluth, Kufi, Farsi and more styles

Many of the existing datasets are concentrated on specific calligraphy styles [17], [22] or applications [11], [20], [21], [22], which restricts their applicability for broader calligraphy generation tasks. For example, some datasets focus on isolated characters or single styles [11], [12], [21], [22], limiting their generalizability for more complex tasks like generating full-page calligraphy or working with multiple styles simultaneously. Another significant challenge is the background noise present in certain datasets. Datasets such as those from [13], [15] suffer from background noise that can compromise the performance of models trained on them. The presence of noise in the images makes it difficult for models to learn the intricacies of calligraphy, potentially leading to less accurate results when generating Arabic calligraphic text. Additionally, there is a problem with limited accessibility for several datasets. Some datasets [17], [18] are not publicly accessible, which creates barriers to reproducibility and benchmarking within the research community. Lastly, the small size of some datasets [9], [13], [21] poses a limitation for training robust deep learning models. Small datasets are insufficient for developing models that can generalize well and perform accurately across diverse Arabic calligraphy styles. This is particularly important in the context of deep learning, where larger datasets are essential to ensure that models can learn complex patterns and handle real-world variations in data.

V. Future Research Directions

To address existing limitations in Arabic calligraphy generation, several key directions for future research can be identified. First, there is a need for larger and more diverse datasets that cover a broader range of Arabic calligraphy styles. These datasets should also include comprehensive annotations, such as stroke-level data, to facilitate the training of more accurate and versatile models. Ensuring the public accessibility of datasets is also essential for fostering reproducibility and collaboration within the research community. Open access to high-quality datasets would enable the development of standardized models, encouraging global contributions and improvements. Public datasets would also facilitate the dissemination and replication of new research findings, promoting more robust and reliable results. Another critical step is to standardize evaluation metrics for Arabic calligraphy generation. Currently, the lack of consistent benchmarks makes it difficult to compare model performance across different datasets. Establishing standardized metrics would allow researchers to more effectively assess model strengths and weaknesses, streamlining the development of accurate calligraphy generation systems. Finally, further datasets should be designed with real-world applications in mind, such as supporting artistic calligraphy generation and aiding in the preservation of historical Arabic manuscripts, to address challenges like digital preservation and the development of artistic tools for Arabic calligraphy. A notable gap in the literature is the lack of research on handwritten word calligraphy generation. While several studies focus on generating individual letters or printed text, to the best of our knowledge, there is virtually no work on generating complete handwritten words or sentences in artistic calligraphy styles.

This gap limits the applicability of existing models for real-world applications, such as personalized calligraphy design or educational tools. Future work should focus on developing models capable of generating complete handwritten words and sentences in artistic calligraphy styles. This requires addressing challenges such as maintaining geometric consistency across letters and ensuring contextual coherence.

VI. Conclusion

Generative modelling has seen remarkable progress in recent years, with the emergence of GAN Networks. This innovative approach in generative modelling has shown massive potential across various domains, particularly in image generation. This Systematic Literature Review (SLR) provides a comprehensive analysis of the investigated 19 relevant papers (2009 to 2024) in Arabic calligraphy generation, addressing four key research questions. The review highlights the dominance of deep learning models, particularly GANs networks, in generating Arabic text. However, significant challenges remain, including the lack of standardized datasets, the absence of research on handwritten calligraphy generation, and the need for robust evaluation metrics. By addressing these challenges and exploring the proposed future directions, researchers can develop more robust models that meet both artistic and functional requirements. This will advance the state-of-the-art in Arabic calligraphy generation. The insights from this review provide a foundation for future research and collaboration in this interdisciplinary field.

References

[1] A. Ahmadian, K. Fouladi, and B. N. Araabi, “Model-based Persian calligraphy synthesis via learning to transfer templates to personal styles,” International Journal on Document Analysis and Recognition (IJDAR), vol. 23, no. 3, pp. 183–203, Sep. 2020, doi: 10.1007/s10032-020-00353-1.

[2] R. Al-Hmouz, “Deep learning autoencoder approach: Automatic recognition of artistic Arabic calligraphy types,” Kuwait Journal of Science, vol. 47, no. 3, 2020.

[3] M. J. Page et al., “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews,” BMJ, vol. bmj, no. 372, Mar. 2021, doi: 10.1136/bmj.n71.

[4] M. A. Ahmed, M. Ali, J. A. Jassim, and H. M. Al-Ammal, “Generative Adversarial Networks (GAN) for Arabic Calligraphy,” in 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT 2021, Institute of Electrical and Electronics Engineers Inc., Sep. 2021, pp. 652–657. doi: 10.1109/3ICT53449.2021.9581388.

[5] A. CHEBOUAT, “Generating Arabic Letters using Generative Adversarial Networks (GANs),” Thesis, UNIVERSITY KASDI-MERBAH OUARGLA, Ouargla, Algeria, 2018.

[6] Y. C. Hadj Azzem, A. Moussaoui, and M. Berrimi, “Arabic Calligraphy Generation Through Image-to-Image Translation Using Generative Adversarial Networks (GANs),” in 2nd International Engineering Conference on Electrical, Energy, and Artificial Intelligence, EICEEAI 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/EICEEAI60672.2023.10590292.

[7] R. Bagido, “Generating New Arabic Letters-Rawashin Design using GAN,” in Proceedings of 2022 5th National Conference of Saudi Computers Colleges, NCCC 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 186–192. doi: 10.1109/NCCC57165.2022.10067330.

[8] Z. Kaoudja, M. L. Kherfi, and B. Khaldi, “An efficient multiple-classifier system for Arabic calligraphy style recognition,” in Proceedings - ICNAS 2019: 4th International Conference on Networking and Advanced Systems, Institute of Electrical and Electronics Engineers Inc., Jun. 2019. doi: 10.1109/ICNAS.2019.8807829.

[9] S. R. Allaf and R. Al-Hmouz, “Automatic Recognition of Artistic Arabic Calligraphy Types,” JKAU: Eng. Sci, vol. 27, no. 1, pp. 3–17, 2016, doi: 10.4197/Eng.

[10] S. A. Mahmoud et al., “KHATT: Arabic offline Handwritten Text Database,” Pattern Recognition, ScienceDirect, vol. 47, pp. 1096–1112, 2014, doi: 10.1109/ICFHR.2012.224.

[11] J. H. AlKhateeb, “A Database for Arabic Handwritten Character Recognition,” Procedia Comput Sci, vol. 65, pp. 556–561, 2015, doi: 10.1016/j.procs.2015.09.130.

[12] S. AlSalamah and R. King, “Towards the Machine Reading of Arabic Calligraphy: A Letters Dataset and Corresponding Corpus of Text,” in 2nd IEEE International Workshop on Arabic and Derived Script Analysis and Recognition, ASAR 2018, 2018. doi: 10.1109/ASAR.2018.8480228.

[13] S. Belila, Y. Gasmi, B. Khaldi, and S. Euchi, “Arabic Calligraphy Recognition Using The Intrinsic Cues of Styles Members of jury,” University of Kasdi Merbah Ouargla, Ouargla, Algeria, 2022.

[14] Z. Alyafeai, M. S. Al-shaibani, M. Ghaleb, and Y. A. Al-Wajih, “Calliar: An Online Handwritten Dataset for Arabic Calligraphy,” arXiv preprint, vol. arXiv:2106.10745, Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.10745.

[15] A. Ismail, Z. Kamel, and R. Mahmoud, “HICMA: The Handwriting Identification for Calligraphy and Manuscripts in Arabic Dataset,” in Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP), Computational Linguistics, Dec. 2023, pp. 24–32. Accessed: Mar. 22, 2024. [Online]. Available: https://hicma.net.

[16] K. Adam, A. Baig, S. Al-Maadeed, A. Bouridane, and S. El-Menshawy, “KERTAS: dataset for automatic dating of ancient Arabic manuscripts,” International Journal on Document Analysis and Recognition, vol. 21, no. 4, pp. 283–290, Dec. 2018, doi: 10.1007/s10032-018-0312-3.

[17] B. Alrehali, N. Alsaedi, H. Alahmadi, and N. Abid, “Historical Arabic Manuscripts Text Recognition Using Convolutional Neural Network,” in Proceedings - 2020 6th Conference on Data Science and Machine Learning Applications, CDMA 2020, Institute of Electrical and Electronics Engineers Inc., Mar. 2020, pp. 37–42. doi: 10.1109/CDMA47397.2020.00012.

[18] M. Khayyat and L. Elrefaei, “A deep learning based prediction of arabic manuscripts handwriting style,” International Arab Journal of Information Technology, vol. 17, no. 5, pp. 702–712, Sep. 2020, doi: 10.34028/iajit/17/5/3.

[19] F. Slimane, R. Ingold, S. Kanoun, A. M. Alimi, and J. Hennebert, “A new Arabic printed text image database and evaluation protocols,” in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2009, pp. 946–950. doi: 10.1109/ICDAR.2009.155.

[20] H. Bouressace and J. Csirik, “Printed Arabic Text Database for Automatic Recognition Systems,” in Proceedings of the 2019 5th International Conference on Computer and Technology Applications, New York, NY, USA: ACM, Apr. 2010, pp. 107–111. doi: 10.1145/3323933.3324082.

[21] B. Kiessling, D. S. Ben Ezra, and M. T. Miller, “BADAM: A Public Dataset for Baseline Detection in Arabic-script Manuscripts,” in Proceedings of the 5th International Workshop on Historical Document Imaging and Processing, New York, NY, USA: ACM, Sep. 2019, pp. 13–18. doi: 10.1145/3352631.3352648.

[22] S. Djaghbellou, A. Attia, A. Bouziane, and Z. Akhtar, “Local features enhancement using deep auto-encoder scheme for the recognition of the proposed handwritten Arabic-Maghrebi characters database,” Multimed Tools Appl, vol. 81, no. 22, pp. 31553–31571, Sep. 2022, doi: 10.1007/s11042-022-13032-6.