MGN-Net: Multigranularity Graph Fusion Network in Multimodal for Scene Text Spotting

Zhengyi Yuan and Cao Shi

***Abstract*—In recent years, scene text spotting approaches have evolved into a multimodal-based framework. Although previous studies have highlighted the crucial importance of the intrinsic synergy between visual and linguistic features, recent advances in multimodal-based methods typically adopt an implicit fusion strategy with single granularity features, which cannot fully exploit the prior contextual relationships embedded in visual and semantic information. We argue that directly integrating visual and semantic features is suboptimal because the multigranularity structure of scene text images is quite different from that of natural images. To address this, we introduce a novel model called the multigranularity visual semantic interactive fusion network (MGN-Net), which comprises a visual semantic multigranularity feature extraction network (VSMN) and a multigranularity graph fusion learning network (MGFN). The VSMN adaptively extracts multigranularity visual and semantic features from the text image, thereby enriching the textual contextual relations. In the MGFN, a cross-modal and cross-hierarchy graph is constructed to align features from different modalities for deep intra- and inter-fusion. This approach also alleviates the inflexibility of the sequential structure when dealing with images of irregularly curved objects. Furthermore, the cross-hierarchy semantic features are designed to facilitate the training of MGN-Net. Experimental results demonstrate that our model significantly outperforms previous state-of-the-art models. The code will be released in MGN-Net.**

***Index Terms*—Computer vision, deep learning, graph convolutional network, multigranularity fusion, scene text spotting.**

I. Introduction

SCENE text spotting aims to detect the location and recognize the semantic content of text instances in natural images, which can then be used in a variety of practical applications within the Internet of Things (IoT) realm. Such applications include autonomous driving, video surveillance, and smart homes, where devices need to understand and interpret text information in their environment. For example, autonomous vehicles may need to recognize road signs or traffic markers, video surveillance systems might need to identify the license plates of vehicles of interest, and smart home devices could be designed to read handwritten notes from users. The diverse font styles and complex curved shapes found in natural scene images pose significant challenges for accurately detecting and recognizing text instances. As a result, scene text spotting has become a focal point in computer vision research and has garnered considerable attention from the research community.

Early approaches to scene text spotting relied on hand-designed features with limited representational capabilities [1], [2]. This limitation led to constrained performance and made it difficult for these methods to generalize and be deployed effectively. With the rapid development of deep learning, novel ideas and solutions have emerged, particularly through region proposal-based [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18] and transformer-based [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36] detection methods. Current research in this field can be broadly categorized into two main types: unimodal methods [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36] and multimodal methods [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54]. Unimodal approaches typically use convolutional neural networks (CNNs) to extract semantic or visual features. They generate region proposals and calculate their intersection ratios with the ground truth to remove redundant proposals. The final prediction bounding box is obtained through postprocessing. However, these unimodal methods may overlook spatial features present in other modes within the image and fail to consider the interaction and complementary effects between different feature types.

Intuitively, the visual and semantic information in an image can help accurately spot text in terms of both content and position. This mutual complementarity and cooperation between visual and semantic information have been termed the multimodal method, which aims to produce a combined effect that is greater than the sum of their separate effects. The workflow of the multimodal approach is illustrated in Fig. 1. For multimodal method, previous works [37], [38], [39], [40], [41] have tried to apply recurrent auto-regressive neural network (RNN) to learn semantic features, and then fused with the output of visual model. However, RNN-based models struggle with the problem of long-distance dependencies between features and suffer from excessively long training times due to their inherently sequential computational properties.

Manuscript received 10 March 2024; revised 9 April 2024; accepted 14 April 2024. Date of publication 18 April 2024; date of current version 9 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61806107, Grant 61702135, and Grant 62201314; in part by the Opening Project of State Key Laboratory of Digital Publishing Technology; and in part by the Shandong Province "Double-Hundred Talent Plan" on 100 Foreign Experts and 100 Foreign Expert Teams Introduction under Grant WST2021020. *(Corresponding author: Cao Shi.)*

The authors are with the School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266100, China (e-mail: yuanzhengyi0224@163.com; caoshi@yeah.net).

Digital Object Identifier 10.1109/JIOT.2024.3390943

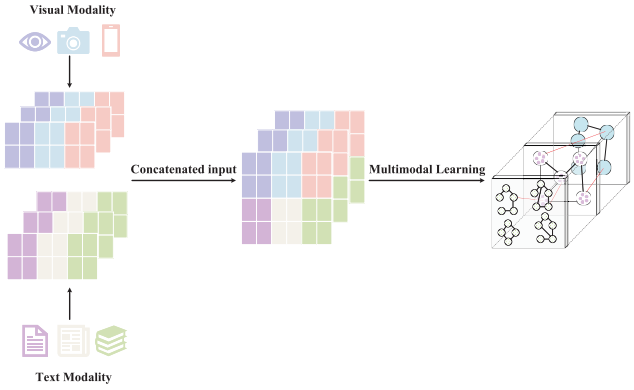


Fig. 1. In the workflow of the multimodal fusion method, the features of each modality are first concatenated in the channel or spatial dimension. Subsequently, the features of different modalities are updated in the neural network.

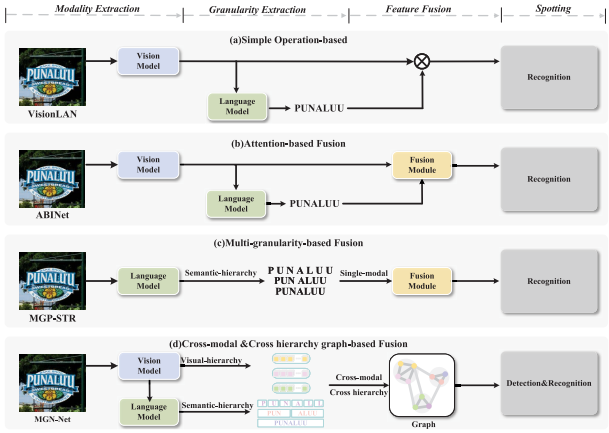


Fig. 2. Overview of some scene text recognition methods closely related to our approach. (a) Simple operations-based fusion methods, adopt a fusion strategy that involves summing or concatenating features from different modalities along either the spatial or channel dimensions. (b) Attention-based fusion methods leverage the self-attention mechanism from the Transformer to perform dot product operations on features from different modalities, thus facilitating network updates. (c) Multigranularity-based methods, execute a learnable fusion on extracted unimodal semantic features across various granularities. (d) Our proposed MGN-Net conduct interactions from cross-modal and cross-hierarchy perspective, hereby unleashing the potential of both visual and semantic features.

More recent work has addressed the sequential computation problem by introducing transformer-based models that utilize self-attention and masking mechanisms. While transformer architectures can effectively learn long-range dependencies in large data sets, they are prone to overfitting on small and medium-sized data sets, such as those comprising scene text images. Additionally, treating natural scene images as sequential structures is not flexible enough to capture the richly irregular and complex objects within them. Moreover, these methods, which are adapted from those used for natural images, only partially explore the unique characteristics of text. Furthermore, the multigranularity levels of information contained in the visual and semantic features are overlooked during the modal fusion process.

Recently, research based on multigranularity has achieved excellent performance in scene text tasks, proposing that images can be viewed as aggregations of features at various levels of granularity. Specifically, features at part-level of granularity combine to form component-level features, which in turn aggregate to create instance-level features. The full potential of visual and semantic features cannot be fully realized through implicit fusion without considering the relationships between features at different levels of the hierarchy. However, the methods that structure features as grids or sequences, as adopted by CNNs or Transformer, lack the flexibility required to handle images with a rich presence of irregular objects [55]. Moreover, these methods overlook the compositional relationships of granularity between different modalities during the cross-modal fusion process.

Inspired by the development of graph neural networks and multigranularity feature extraction in the field of computer vision [55], [56], [57], [58], [59], [60], [61], [62], we hypothesize that enabling interaction across multiple modality and levels can facilitate the learning of high-quality representations in a more effective manner. To capture both the local dependencies between features and the global information in scene text images with irregular objects, this article introduces a novel model based on the combination of Graph and Transformer, named multigranularity visual semantic interactive fusion network (MGN-Net), which consists of a visual semantic multigranularity feature extraction network (VSMN) and a multigranularity graph fusion learning network (MGFN). VSMN transforms the extracted visual and semantic multigranularity features into graph nodes to capture context relations at various granular levels. In MGFN, cross-modal and cross-hierarchy fusion processes are employed to fully explore semantic and visual characteristics. Furthermore, the network learns the cross-hierarchical context relationships among text primitives, which enhances the training of the model. Fig. 2 illustrates the differences between MGN-Net and recent related work.

The main contributions of this work are summarized as follows.

1) We propose a novel MGN-Net that combines graph with transformer for scene text spotting, which has achieved superior performance on various publicly available benchmarks.
2) The proposed VSMN extracts visual and semantic contextual relationship at different granularity levels through a shared encoder, which includes character features at the part-level, subword features (e.g., roots and affixes) at the component-level, and word features at the instance-level.
3) The MGFN integrates an intrafusion to accomplish aggregated updates between features of different granularity levels in the same modal space, and interfusion to accomplish the learning of relationships between features at different granularity levels in different modal spaces. The multigranular contextual information contained in the textual primitives is utilized as an auxiliary loss to facilitate the model training. 4) To the best of our knowledge, MGN-Net is one of the first attempts to perform modal fusion across multiple modalities and hierarchies for scene text spotting.

II. Related Work

Scene text spotting refers to the task of predicting textual regions and the semantic contents contained in natural images. Since text instances in natural images often have a variety of font styles, diverse backgrounds, and complex irregular shapes, the scene text spotting task has attracted significant attention in recent years.

Early research [1], [2] was mainly based on traditional region-based features, such as maximally stable extremal regions (MSER) and stroke width Transform, which first obtain numerous character bounding boxes using region-based features or object detection methods. Then, traditional machine learning models are used to associate the character boxes and remove false positives, in order to obtain detection results at the word or text line level. Due to the lack of discriminative power in traditional features and classifiers, traditional detection methods have a cumbersome framework and often require manual tuning based on empirical knowledge. Recently, with the development of deep learning technology, the accuracy and robustness of scene text spotting have rapidly improved and are gradually approaching practical application.

For unimodal methods, previous works [3], [5], [9] have attempted to use a CNN to extract visual feature representations, which are directly fed into the region proposal network (RPN) to generate bounding boxes. The process continues by eliminating redundant boxes through postprocessing to obtain the final prediction, or by converting them into tokens through linear embedding to feed into the end-to-end transformer encoder and decoder. SVTR [3] utilizes a patch-wise image tokenization method to eliminate sequential modeling and incorporates hierarchical stages with global and local mixing blocks, resulting in competitive performance. Song et al. [5] proposed an integrated algorithm that jointly addresses scene text detection and recognition by sharing convolutional feature maps. MAYOR [9] introduced a multilayer perceptron (MLP) decoder in the mask head to alleviate learning confusion, employs instance-aware mask learning, and incorporates adaptive label assignment in the RPN.

Benefiting from the development of Transformer [63], recent research has increasingly adopted Vision Transformer (ViT) [64] or DETR [65] as their foundational framework for the task of text instance detection and recognition. ViTSTR [19] proposes a simple text recognition model for scene images that utilizes a ViT, and adopts the same training scheme as that used for the original ViT. TESTR [66] has designed an end-to-end scene text spotting framework that simultaneously performs detection and recognition through a single encoder and dual decoder, which is also free from Region of Interest (RoI) operations. DeepSolo [67] utilizes a decoder with explicit points to achieve more efficient training while keeping the required annotations less costly. ESTextSpotter [21] emphasizes explicit interaction between task-aware queries for text detection and recognition within a single decoder, which significantly outperforms previous state-of-the-art methods in end-to-end scene text spotting. However, unimodal models might neglect the linguistic knowledge embedded in scene text images.

Multimodal methods typically extract visual and semantic information from natural scene images, performing either simple concatenation operations in the spatial and channel dimensions or leveraging attention mechanisms with dot product calculations for modality fusion, thus generating the final feature representation. Sabir et al. [40] utilized the occurrence probabilities of words and the semantic correlation between scenes and text to improve scene text recognition. oCLIP [42] introduces a weakly supervised pretraining method that enhances optical character recognition tasks by jointly learning and aligning visual and textual information. To better integrate visual and semantic features at a deeper level, JVSR [52] proposes a multistage decoder based on RNN attention, where each decoder samples from visual features as a reference for semantic features. VisionLAN [68] proposes a language-aware visual mask that refers to semantic features to enhance the visual features. MATRN [53] proposes an interactive augmentation network where both visual and semantic features are simultaneously referenced to accomplish bi-directional fusion and stimulate the combination of semantic features into visual features by hiding visual clues related to the characters in the training phase.

Nonetheless, the feature structuring methods employed by CNNs or Transformer, which organize features as grids or sequences, lack sufficient flexibility when handling images with a wealth of irregular objects. Distinct from these approaches, our proposed model maintains the long-range receptive field advantage of transformer for global information aggregation and updating across different modal spaces, while employing message passing between graph nodes to thoroughly excavate the structural relationships among features of various granularity levels within the same modality. Therefore, MGN-Net alleviates the problem transformer encounter due to their lack of inductive bias, which often results in poor performance when confronted with objects that have complex structures of multiple granularities, as shown in Fig. 3. Moreover, these methods overlook the granularity composition relationship between different modalities during the cross-modal fusion process. MGP-STR [57] proposes the adaptive addressing and aggregation module to provide learnable fusion of contextual semantic information at different levels of granularity, which leads to excellent performance in scene text recognition tasks. RCLSTR [69] mitigates overfitting and enhances robustness in the model by employing a contrastive learning methodology and exploring the semantic relationships embedded within textual primitives.

In this article, we propose a novel model for scene text spotting that operates in a multigranularity manner, considering both visual and semantic perspectives. Our model comprises two main components: 1) a VSMN and 2) an MGFN. The VSMN divides visual and semantic features into part-level, component-level, and instance-level granularities. The MGFN constructs cross-modal and cross-hierarchy deep fusion to aggregate and update node features, thereby bridging cross-granularity interactions between different semantic contextual informations and facilitating model training.

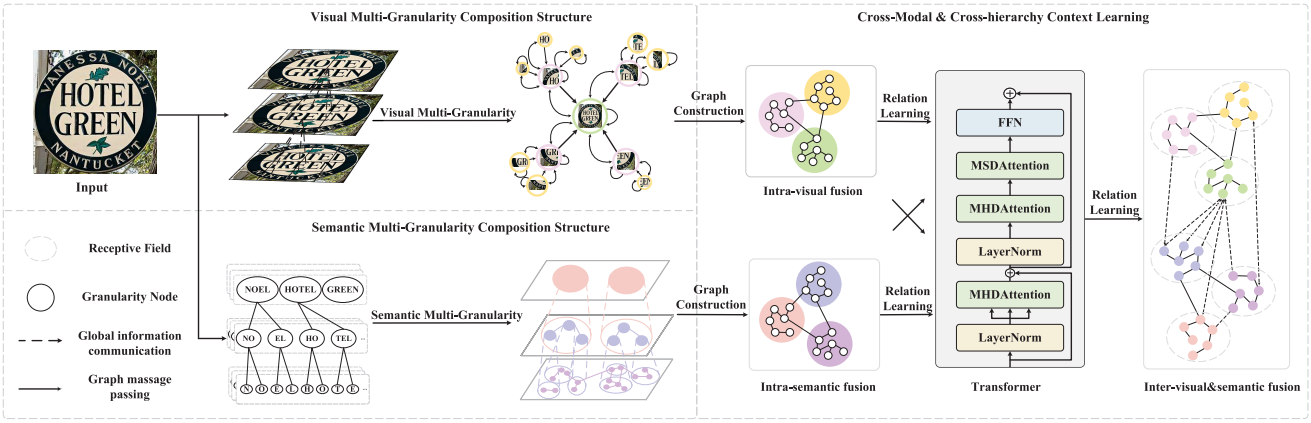


Fig. 3. Overview of the multigranularity composition structure in scene text images is as follows. At the instance level, visual and semantic features can be perceived as aggregations of features at the component level, with each component-level feature composed of a series of part-level features. Features of different granularities from various modalities are fed into the MGFN for cross-modal and cross-hierarchy context learning. MGFN conducts message passing between graph nodes to thoroughly explore the structural relationships among features of various granularity levels within the same modality, achieving local information intrafusion across different hierarchy levels. Additionally, these features are input into a transformer to leverage its long-range receptive field for global information interfusion across the multimodal space. Through this approach, MGN-Net addresses the issue of transformer-based models' poor performance in recognizing scene text with objects that have complex structures of multiple granularities, due to the lack of inductive bias.

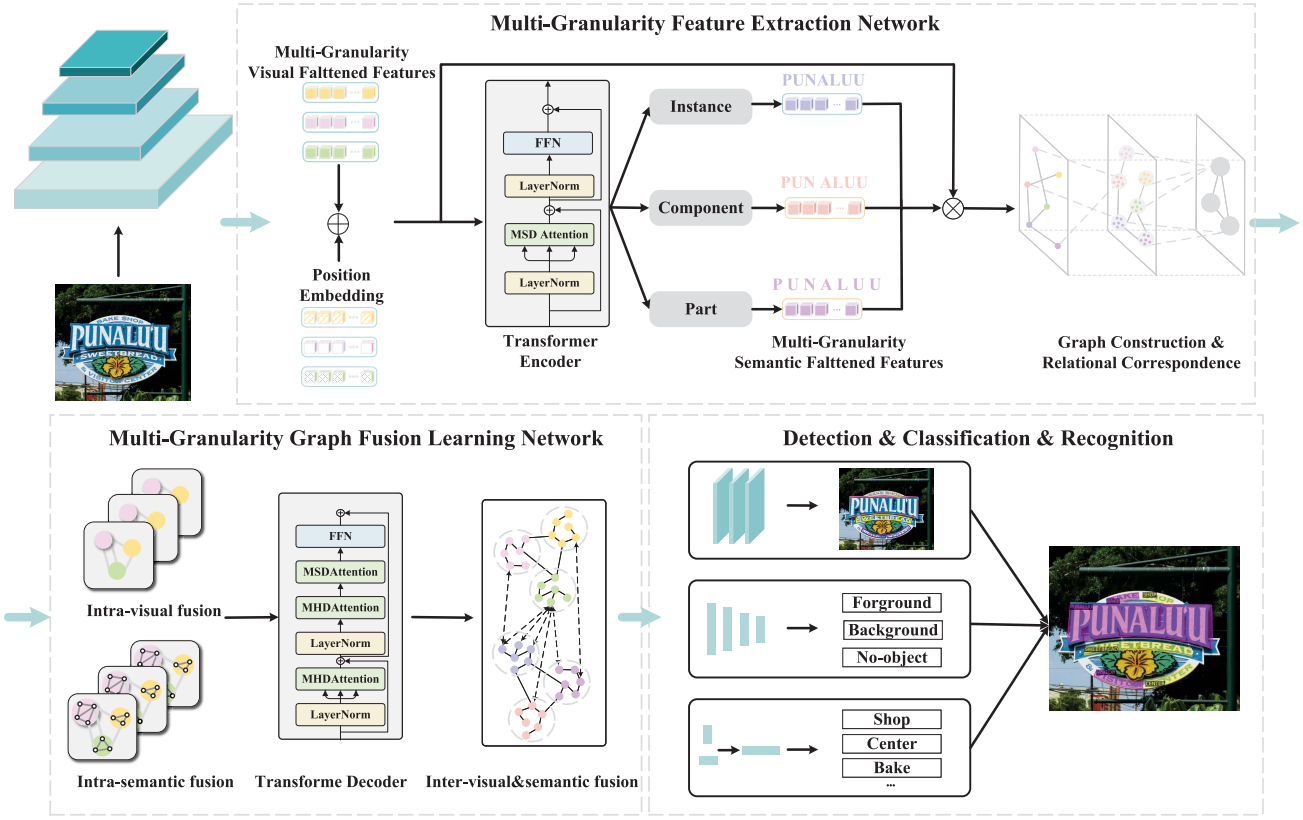


Fig. 4. Overall architecture of MGN-Net consists of two primary components: the VSMN and the MGFN. The VSMN is responsible for extracting visual and semantic features at the instance, component, and part levels from text images. The MGFN performs cross-modal and cross-hierarchical interactive fusion by aggregating and updating information across and within these levels for the graph nodes. This integration process incorporates the different hierarchical levels of contextual relationships into the nodes' features.

III. Method

The overview of the proposed MGN-Net method is depicted in Fig. 4, which consists of three parts: 1) the VSMN; 2) the MGFN; and 3) the detection and recognition module with auxiliary loss. Initially, we employ ResNet-101 as the backbone to extract visual features at different granularity levels, which are then processed by the VSMN. This network extracts visual and semantic features with multigranularity, such as features at the part level, component level, and instance level. The features from different modalities at various levels

of granularity are input into the MGFN for graph construction. In this stage, semantic and visual features of different granularity levels are constructed as graph nodes. Each node undergoes aggregation, exchange, and updating by identifying its nearest neighbors. Following this, the processed features are input into the decoder for detection and recognition tasks. To further enhance the training process, MGN-Net incorporates an auxiliary loss that integrates multigranularity semantic contextual information.

A. Visual Semantic Multigranularity Feature Extraction Network

The task of scene text recognition provides valuable semantic information that can be beneficial to the task of scene text detection. However, the challenge lies in optimally utilizing the semantic information derived from scene text in the detection task. In this study, we introduce a visual semantic multigranularity feature extraction module. This module is designed to thoroughly extract both visual and semantic information at different levels of granularity from the scene text. The extracted semantic information is then utilized to enhance the performance of scene text spotting.

At the highest level, an entire image is considered as a holistic entity for acquiring word representations. Progressing to the middle level, words are dissected into subwords (e.g., roots and affixes), functioning as essential language units from a linguistic perspective. At the lowest level, characters serve as the fundamental building blocks of texts. We propose that exploring relationships across multiple levels within a hierarchical structure can fully exploit semantic information and improve representation learning. We extract multigranularity visual features from ResNet-101. The feature maps *res*3 to *res*5 which can be viewed as $f = [res3, res4, res5]$, and the shape of each layer is $C_i \times H_i \times W_i$. In order to unify the feature dimension, we adopt an 1×1 kernel to convert channel C_i to hidden dimension size D, where $D = 256$. The features are flattened along the final dimension for each layer. Subsequently, the final two dimensions are transposed and these features are concatenated before being sent into transformer encoder with deformable attention to form the sequences S. In this way, a scene text image is represented as a sequence with $N = HW/P^2$, where $P \times P$ is the resolution of each image patch.

In order to sufficiently utilize the semantic information in the scene text images, the sequences S are fed into semantic extraction module. In this module, we turn the sequence S to the T tokens, denoted as $T' = [t_i]_{i=1}^{L}$, each with D dimension, L represents the total number of tokens, and each token corresponds to a character, as illustrated by $t_i = \mathrm{Agg}(S)$. The $\mathrm{Agg}(\cdot)$ function is a special attention mechanism that can learn to select part of the sequence to generate the i_{th} character. The $\mathbf{a_i}$ is an attention mask and $\mathbf{a_i} \in \mathbb{R}^{(N+1)\times D}$. The function $\alpha_i(\cdot)$ represents a group convolution with a single 1×1 kernel, and U is a linear layer to transform features. The details are as follows:

$$\begin{aligned} \mathbf{t_i} &= \mathrm{Agg}_i(S) = \mathbf{a_i^T}\tilde{S} = \mathbf{a_i}(SU)^T \\ \mathbf{a_i} &= \mathrm{softmax}(\alpha_i(S)). \end{aligned} \tag{1}$$

Received the features from aggregation function, the semantic information has been fully utilized. Those features, denoted as $\mathbf{Y}$ will be sent into a classification head built by $\mathbf{G} = \mathbf{YW^T}$, where $\mathbf{W}$ is the weight of linear layer, $\mathbf{G}$ is the classification logits. The process described above pertains to semantic feature extraction at the part-level. Furthermore, to fully harness the potential of linguistic knowledge,we have implemented tokenization at both the component-level and instance-level. The granularity feature extraction processes for component-level and instance-level tokenization are similar to that of part-level, but they operate independently and utilize different parameters. The dimension of the classification heads is determined by the vocabulary size of the respective granularities. This approach allows the model to leverage the grammatical information of words and to integrate semantic information, thereby enhancing its performance. In the process of developing part-level, component-level, and instance-level semantic features, we have employed a Bezier center curve proposal scheme akin to the one utilized in DeepSolo. This approach is adept at fitting and differentiating scene text from each other, and facilitates the use of line annotations. Specifically, following the image feature processing by the encoder, which is fortified with deformable attention, it produces Bezier center curve proposals and their corresponding scores. The proposals with the top-K scores are selected for subsequent processing. Each chosen curve proposal, characterized by four Bezier control points, has N points uniformly sampled along the curve. The coordinates of these points are encoded as positional queries and are merged with learnable content queries to construct composite queries.

Meanwhile, a three-layer MLP is utilized to predict offsets to the four Bezier control points, determining a curve that represents a single text instance. Let j represents a pixel from visual features at different stage $f = [res3, res4, res5]$ with 2-D normalized coordinates $\hat{c}_i = (\hat{c}_{ix}, \hat{c}_{iy}) \in [0, 1]^2$, and its corresponding Bezier control points $BC_i = \{\bar{bc}_{i_0}, \bar{bc}_{i_1}, \bar{bc}i_2, \bar{bc}_{i_3}\}$, where $j \in \{0, 1, 2, 3\}$, and σ is the sigmoid function. The MLP head only predicts the offsets $\Delta p_i\{x_0, y_0, \ldots, x_3, y_3\}$. Furthermore, we use a linear layer for the text or not text classification. The top-K scored curves are selected as the proposals

$$\overline{bc}_{i_j} = \left(\sigma\left(\Delta p_{ix_j} + \sigma^{-1}(\hat{p}_{ix})\right), \sigma\left(\Delta p_{iy_j} + \sigma^{-1}(\hat{p}_{iy})\right)\right). \tag{2}$$

B. Multigranularity Graph Fusion Learning Network

In this section, we propose a graph fusion learning network. As one of the most versatile data structures, graphs can represent both grid architectures like CNN and sequential architectures like Transformer. In a graph, nodes can flexibly interact with their neighboring nodes. Graph neural networks possess powerful capabilities for information interaction, which can be harnessed to learn visual and semantic features embedded in a scene text image. Therefore, we utilize a graph-based approach to facilitate information exchange between visual and semantic features. During the update process, the graph fusion learning module aligns semantic features with visual features at various levels of granularity. The graph structure selects the most associated nodes as neighbors for

each node, enabling the nodes to update themselves and achieve effective information exchange.

Upon receiving the features of vision and semantics, these features are view as N tokens. All the features are transformed into $X = [x_1, x_2, \ldots, x_N]$, where $x_i \in \mathbb{R}^D$ and i is the index of tokens ranging from 1 to N, and D is the dimension of each token. These tokens correspond to a set of unordered nodes of a graph, denoted as $N = n_1, n_2, \ldots, n_n$. Next, we employ the K-nearest neighbors method to generate K neighbors for each node. We first calculate a distance matrix $\mathbf{d_{ij}} \in D$, where $i, j = 1, 2, \ldots, n$. $\mathbf{d_{ij}}$ is the distance from n_i to n_j. Then, for node n_i, we select K nearest nodes from remaining $n-1$ nodes $\mathcal{V}\{v_i\}$ and add an edge e_{ji} from n_j to n_i for all $v_j \in \mathcal{N}\{n_i\}$. This process defines the graph $G = (N, E)$, where E represents all the edge in the graph. The graph construction process can be denoted as $G = \mathcal{G}(X)$. By employing this graph representation, we utilize the advantages of the graph data structure to deeply fuse cross-modal and cross-hierarchical features using a GNN.

Graph neural networks typically involve two steps: 1) aggregation and 2 updating. During the aggregation step, a node gathers information from its neighbors. Subsequently, in the updating step, the node's information is updated based on the aggregated data. Specifically, we employ the max-relative graph convolution method for both aggregating and updating information. The aggregation operation entails subtracting the feature values of a node from those of all its neighbors and then selecting the maximum value to concatenate with the node's own features. We convert the feature from VSMN into a graph $G = (N, E)$ through the building process $G = \mathcal{G}(x)$. Function $F(\cdot)$ represents the graph convolution operation, while W_{agg} and W_{update} are the learnable weights for the aggregation and update operations, respectively. For each node, the graph first gathers information from its neighboring edges using W_{agg}. Then, the gathered information is used to update the nodes N. The graph convolution process can be summarized as follows:

$$\begin{aligned} \mathcal{G}' &= F(G, W) \\ &= \text{Update}\big(\text{Aggregate}(G, W_{\text{agg}}), W_{\text{update}}\big) && (3) \\ X_i' &= h\big(x_i, g\big(x_i, \mathcal{N}(x_i), W_{\text{agg}}\big), W_{\text{update}}\big) && (4) \\ g(\cdot) &= x_i'' = \big[x_i, \max(\{x_j - x_i | j \in \mathcal{N}(x_i)\})\big] && (5) \\ h(\cdot) &= x_i' = x_i'' W_{\text{update}}. && (6) \end{aligned}$$

The process described above can be represented as $X' = \mathit{GraphConv}(X)$. Furthermore, to prevent the over-smoothing phenomenon, we incorporate an additional linear layer following the graph convolution operation.

Meanwhile, we select the top-K proposals, denoted as $BC_k(k \in \{0, 1, \ldots, K-1\})$, and uniformly sample N points on each curve based on the Bernstein polynomials to obtain the normalized point coordinates, denoted as *Coords*. The point positional queries $\mathbf{P_q}$ are generated by the following process:

$$\mathbf{P_q} = \text{MLP}(PE(\mathit{Coords})) \tag{7}$$

where PE represents the sinusoidal positional encoding function. Furthermore, we initialize point content queries $\mathbf{C_q}$, using learnable embeddings. Finally, the composite queries $\mathbf{Q_q}$, are computed by

$$\mathbf{Q_q} = \mathbf{C_q} + \mathbf{P_q}. \tag{8}$$

The input to the Transformer decoder consists the composite queries $\mathbf{Q_q}$. To focus on the internal relationships between object queries, we first calculate the intragroup self-attention among the point queries, with keys identical to the queries $\mathbf{Q_q}$ and values derived solely from the content queries $\mathbf{C_q}$. Subsequently, an intergroup self-attention is performed across the K instances to capture the relationships between different instances. Finally, the updated composite queries are processed through multiscale deformable cross-attention to extract text features.

C. Detection and Recognition Module With Auxiliary loss

Similar to the method employed by DeepSolo, we also adapt the Bezier center curve to fit scene text of arbitrary shape using a fixed number of control points. In MGN-Net, features from the encoder are utilized to generate Bezier curve proposals, along with their corresponding scores. The top-K proposals are selected, and four Bezier control points are extracted as reference points. Following predefined rules, N points are then sampled from these control points. These sampled points, considered as positional queries, are combined with learnable content queries to form composite queries. The composite queries are fed into the decoder to aggregate text information. The final queries can generate results about text, coordinates, and classes through parallel prediction heads.

Parallel Prediction Head: Upon receiving the updated queries from the Transformer decoder, we employ four parallel prediction heads to perform different tasks.

1) Instance classification, where a linear layer predicts whether the object is text or background. A fixed threshold, set as the average value of N scores, is used for classification.
2) Character classification, which also uses a linear layer to determine whether each query, transformed from a Bézier point, represents a character or background.
3) Center curve offsets, where a three-layer MLP predicts the coordinate offsets relative to the original coordinates generated by the encoder. The final coordinate results are obtained by adding these offsets to the original coordinates.
4) Boundary offsets, which are predicted in a similar manner to the center curve coordinates. A three-layer MLP is used to estimate the offsets relative to the ground truth points.

Multigranularity Auxiliary Loss: The total loss function consists of three parts: 1) encoder loss; 2) decoder loss; and 3) our multigranularity auxiliary loss. The encoder loss comprises two components: 1) the classification loss and 2) the coordinate loss, which are denoted as follows:

$$\mathcal{L}_{enc} = \sum_i \Big(\lambda_{cls}\mathcal{L}_{cls}^{(i)} + \lambda_{coord}\mathcal{L}_{coord}^{(i)}\Big). \tag{9}$$

The decoder loss function consists of four parts: 1) the instance classification loss; 2) the character classification loss;

TABLE I
PERFORMANCE COMPARISON ON TOTAL-TEXT DATA SET

Mehtod	Backbone	P	R	F1	None	NULL
TextDragon [70]	VGG-16	85.6	75.7	80.3	48.8	74.8
SRSTS [71]	ResNet-50	92.0	83.0	87.2	78.8	86.3
CharNet [72]	ResNet-50-Hourglass57	88.6	81.0	84.6	63.6	-
TextPerceptron [32]	ResNet-50-FPN	88.8	81.8	85.2	69.7	78.3
Boundary [11]	ResNet-50-FPN	88.9	85.0	87.0	65.0	76.1
Mask TextSpotter v3 [73]	ResNet-50-FPN	-	-	-	71.2	78.4
PGNet [74]	ResNet-50-FPN	85.5	86.8	86.8	63.1	-
MANGO [75]	ResNet-50-FPN	-	-	-	72.9	83.6
ABCNet v2 [25]	ResNet-50-FPN	90.2	84.1	87.0	70.4	78.1
GLASS [26]	ResNet-50-FPN	90.8	85.5	88.1	79.9	86.2
SwinTextSpotter [23]	Swin-T-FPN	-	-	88.0	74.3	84.1
UNITS [76]	Swin-B	-	-	89.8	78.7	86.0
TESTR [66]	ResNet-50	93.4	81.4	86.9	73.3	83.9
TTS [77]	ResNet-50	-	-	-	78.2	86.3
SPTS v2 [30]	ResNet-50	-	-	-	75.5	84.0
ESTextSpotter [21]	ResNet-50	92.0	**88.1**	**90.0**	80.8	87.1
DeepSolo [67]	ResNet-50	93.2	84.6	88.7	82.5	**88.7**
MGN-Net	ResNet-50	**93.6**	84.9	89.1	**82.9**	87.5

3) the center Bézier coordinate loss; and 4)the boundary loss. Specifically, the character classification loss utilizes the CTC (connectionist temporal classification) loss to account for any mismatches between the lengths of the ground truth and the predictions. The decoder loss function is formulated as follows:

$$\mathcal{L}_{dec} = \sum_{k} \left(\lambda_{cls} \mathcal{L}_{cls}^{(k)} + \lambda_{text} \mathcal{L}_{text}^{(k)} + \lambda_{cd} \mathcal{L}_{cd}^{(k)} + \lambda_{bd} \mathcal{L}_{bd}^{(k)} \right). \quad (10)$$

Furthermore, we incorporate CTC loss into VSMN to predict text labels using three levels of semantic contextual features, as illustrated below

$$\mathcal{L}_{mgp} = \sum_{j} \left(\lambda_{ILS} \mathcal{L}_{ILS}^{(j)} + \lambda_{CLS} \mathcal{L}_{CLS}^{(j)} + \lambda_{PLS} \mathcal{L}_{PLS}^{(j)} \right). \quad (11)$$

The above hyperparameters λ_{cls}, λ_{cd}, λ_{text}, and λ_{bd} are adopted from our baseline DeepSolo [67]. In contrast, λ_{ILS}, λ_{CLS}, and λ_{PLS} are designed by us, and their effectiveness is validated through ablation experiments presented in Table V. Overall, the total loss function is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{enc} + \mathcal{L}_{dec} + \mathcal{L}_{mgp}. \quad (12)$$

IV. EXPERIMENTS

To validate the effectiveness of our proposed MGN-Net, we perform a series of experiments on three publicly available data sets. Additionally, we conduct a series of ablation studies to evaluate the impact of various factors. These studies include assessing the effectiveness of the proposed modules, examining the influence of different visual-to-semantic node ratios on model performance, and testing the effect of incorporating various granularity levels of semantic information into the loss computation on the model's performance.

A. Public Data Sets

ICDAR 2015 Incidental Text (ICDAR15) data set [1] is designed for scene text detection and recognition tasks, comprising 1000 training images and 500 test images. These 1500 images are sourced from real-world scenes, including streets, shops, and billboards, and are captured by various means such as street patrols, car cameras, and pedestrian mobile phones, featuring a considerable amount of irregular text.

SCUT-CTW1500 [79] is a data set tailored for curved text detection, comprising 1000 training images and 500 test images. It encompasses various text shapes, including curved, twisted, and rotated, and features both chinese and english text instances.

Total-Text [80] is a comprehensive data set widely used for detecting arbitrary-shaped scene text. It contains 1255 training images and 300 test images, with each text instance annotated at the word level using a polygon.

B. Implementation Details

Our MGN-Net is implemented using the PyTorch 1.7 framework, with ResNet-101 serving as the backbone neural network for feature extraction. We use six layers in both the encoder and decoder, each with a hidden dimension of 256. During the training phase, MGN-Net is trained using Adaptive Moment Estimation with a weight decay of 0.0001. The model is trained on 4 2080Ti GPUs with a batch size of 2. Initially, our model is pretrained on the SynthText and MLT2017 data sets for 375k iterations, with an initial learning rate of 0.01. Subsequently, MGN-Net is fine tuned on the training set of the target data set for another 375k iterations, also starting with a learning rate of 0.01. The ratio of the number of visual to semantic modal nodes is set at 2:1. The loss weights for L_{enc}, L_{dec}, and L_{mgp} are all set to 1.0.

C. Comparisons With State-of-the-Art Methods

In this section, we compare MGN-Net with state-of-the-art methods on benchmarks for arbitrarily shaped and oriented scene text. To validate the effectiveness of MGN-Net, we employ the experimental setup described earlier to benchmark against prevailing detection methods. For fairness, all methods listed in Table I use ResNet as their backbone network,

TABLE II
PERFORMANCE COMPARISON ON ICDAR2015 DATA SET

Method	P	R	F1	S	W	G
TextDragon [70]	92.5	83.8	87.9	82.54	78.34	65.15
PAN [80]	82.9	77.8	80.3	-	-	-
Text Perceptron [32]	91.6	81.8	86.4	78.2	74.5	63.0
ABCNetV2 [25]	90.4	86.0	88.1	82.7	78.5	73.0
MANGO [79]	-	-	-	80.3	77.8	66.1
TESTR [66]	90.3	**89.7**	90.0	85.2	79.4	73.6
SPTS [30]	-	-	-	77.5	70.2	65.8
SwinTextSpotter [23]	-	-	-	83.9	77.3	70.5
DeepSolo [67]	92.8	87.4	90.0	86.8	81.9	76.9
ESTextSpotter [21]	92.5	89.6	**91.0**	88.1	**82.9**	**77.9**
MGN-Net	**92.7**	88.5	90.4	**88.6**	82.2	76.8

TABLE III
PERFORMANCE COMPARISON ON CTW1500 DATA SET

Method	P	R	F1	None	Full
TextDragon [70]	84.5	82.8	83.6	39.7	72.4
PAN [80]	86.4	81.2	83.7	-	-
Text Perceptron [32]	88.7	78.2	83.1	48.6	-
ABCNetV2 [25]	85.6	83.8	87.0	57.5	77.2
MANGO [75]	-	-	-	58.9	78.7
TESTR [66]	89.7	83.1	86.3	53.3	79.9
SPTS [30]	-	-	-	63.6	83.8
SwinTextSpotter [23]	-	-	-	57.5	77.2
DeepSolo [67]	-	-	-	60.1	78.4
ESTextSpotter [21]	91.5	**88.6**	**90.0**	**66.0**	**83.6**
MGN-Net	**91.7**	87.4	89.4	60.7	79.5

except for SwinTextSpotter and UNITS, which utilize the Swin Transformer and TextDragon, which employs VGG-16. The abbreviations P, R, and F1 represent precision, recall, and F1-score, respectively. "None" indicates lexicon-free results, "Full" refers to results using a lexicon that includes all words in the test data set, and E2E denotes end-to-end spotting results.

Table I showcases MGN-Net's performance on the TOTAL-TEXT data set, with our model achieving a precision of 93.6%, outperforming DeepSolo and TESTR by 0.4% and 0.2%, respectively. Unlike single-modality networks such as TextDragon and CharNet, MGN-Net performs deep fusion of visual and semantic features across multiple granularities. ESTextSpotter, which uses a single encoder to synergize detection and classification tasks, benefits from the complementarity between these tasks. Similarly, our model leverages this synergy in multimodal fusion by using a shared encoder to extract semantic and visual features at various granularities, enabling a deeper integration of different modalities. Furthermore, our approach differs from others in that we treat context granularity features as graph nodes. Each node utilizes the K-nearest neighbors algorithm to find neighbors for feature aggregation and updating. Consequently, our model achieves a lexicon-free (None) result of 82.9%, surpassing other methods.

Tables II and III showcases MGN-Net's performance on the ICDAR2015 and CTW1500 data sets. On the CTW1500 data set, our model achieves a precision that is 0.2% higher than that of ESTextSpotter in the detection task. This improvement is attributed to MGN-Net's effective extraction of contextual granularity information and structural compositional relationships in scene text images through cross-modal and cross-hierarchical methods. On the ICDAR2015 data set, the abbreviations "S," "W," and "G" denote strong, weak, and generic lexicons, respectively. Our model surpasses the current leading model in precision metrics by 0.2% for the detection task and by 0.5% for the strong lexicon in the recognition task. Unlike ESTextSpotter, our approach employs a graph network to facilitate feature aggregation and updating across different modalities and granularities within the same modality. This approach overcomes the limitations of sequential structures in handling irregular and curved text images. The visualization of MGN-Net's performance is provided in Fig. 5.

TABLE IV
ABLATION EXPERIMENTS ON THE TOTAL-TEXT DATA SET

Baseline	+VSMN	+MGFN	P	R	F1	None	Full	Speed(ms/img)
Baseline	-	-	93.1	84.7	88.9	82.3	88.5	61
+VSMN	✓	-	93.3	84.6	88.8	82.4	87.4	63
+MGFN	-	✓	93.5	84.4	88.9	81.7	87.2	69
MGN-Net	✓	✓	93.6	84.9	89.1	82.9	87.5	71

D. Ablation Study

To validate the effectiveness of each proposed module, we conduct several ablation studies on the Total-Text data set. We also assess the impact of different visual-to-semantic node ratios on the model's performance. Additionally, we examine how varying levels of semantic contextual granularity affect the model's overall performance. Our baseline framework is DeepSolo, with ResNet-101 as the backbone, and both the encoder and decoder are configured with six layers each. The training is carried out on the SynthText, MLT2017, and Total-Text data sets. The visualization of the ablation study is provided in Fig. 6.

We conducted experiments to assess the enhancements to the baseline model provided by the VSMN and the MGFN modules, as shown in Table IV. Incorporating contextual information from different modalities at various granularity levels, the baseline augmented with VSMN achieves a 0.2% and 0.1% increase in detection precision and lexicon-free spotting results, respectively. The enhancement observed implies that the task of recognition is advantaged by the incorporation of semantic features, potentially owing to the synergistic effect of features originating from diverse modalities during the training phase. The MGFN module, designed to transform visual and semantic features into graph nodes, addresses the shortcomings of sequential structures when dealing with images containing irregular objects. As anticipated, the baseline model enhanced with MGFN achieves a 0.4% increase in precision, reaching 93.5%. Furthermore, when trained with both VSMN and MGFN, MGN-Net shows a 0.2% improvement in F1 score (89.1%) and a 0.6% increase in the lexicon-free (None) result (82.9%), confirming that the synergy between the two modules positively impacts the model's performance.

To thoroughly assess the impact of semantic contextual features at various granularity levels on the model, we carried out experiments to explore how different levels of semantic contextual features affect MGN-Net's loss computation. The results, presented in Table V, indicate that the abbreviations PLS, CLS, and ILS correspond to part-level semantic loss,

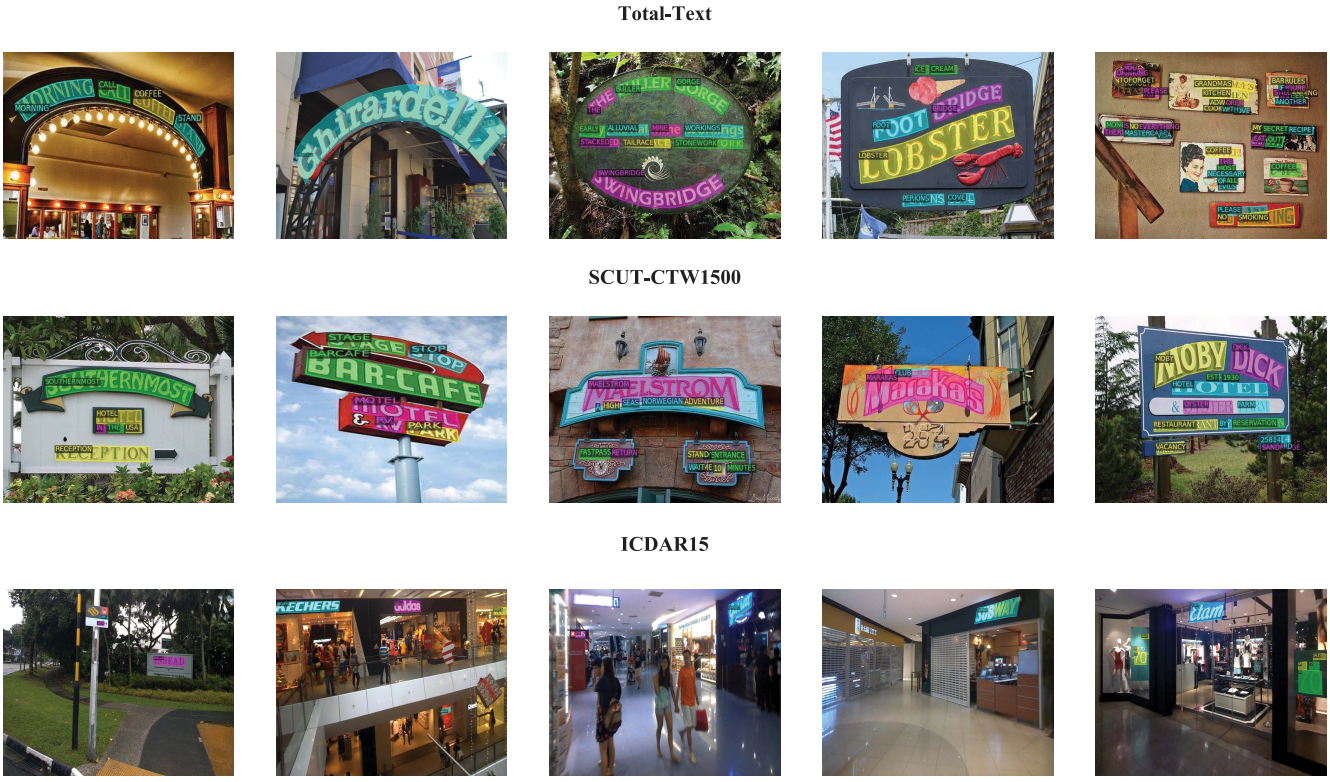
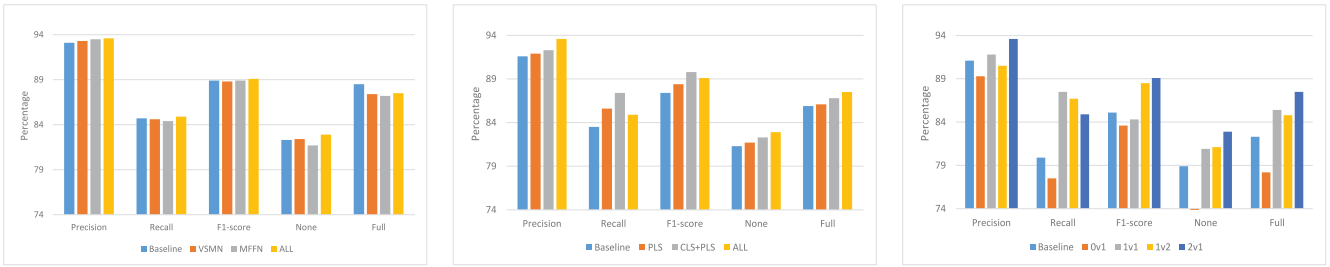


Fig. 5. Qualitative spotting results of MGN-Net, which consists of the Total text, CTW1500, and ICDAR data sets.



(a) (b) (c)

Fig. 6. Detailed analysis of MGN-Net. (a) Ablation study to demonstrate the effectiveness of the VSMN and MGFN modules. (b) Ablation study to evaluate the impact of different-level semantic features. (c) Ablation study to assess the performance of various modality ratio configurations. None denotes lexicon-free. Full denotes the inclusion of all words present in the test data set.

TABLE V
ABLATION EXPERIMENTS ON TOTAL-TEXT DATA SET WITH DIFFERENT SEMANTIC LEVELS

	PLS	CLS	ILS	P	R	F1	None	Full	Speed(ms/img)
MGN-Net+	-	-	-	91.6	83.5	87.4	81.3	85.9	61
MGN-Net+	✓	-	-	91.9	85.6	88.4	81.7	86.1	65
MGN-Net+	-	✓	-	91.5	84.7	88.0	81.2	86.2	60
MGN-Net+	-	-	✓	90.8	84.2	87.5	80.8	85.8	62
MGN-Net+	✓	-	✓	91.3	85.2	88.1	81.5	86.1	64
MGN-Net+	-	✓	✓	91.9	85.5	88.6	81.9	86.4	69
MGN-Net+	✓	✓	-	92.3	87.4	89.8	82.3	86.8	68
MGN-Net+	✓	✓	✓	93.6	84.9	89.1	82.9	87.5	71

TABLE VI
ABLATION EXPERIMENTS ON TOTAL-TEXT DATA SET WITH DIFFERENT MODALITY RATION CONFIGURATION

V:S	P	R	F1	None	Full
1:0	91.1	79.9	85.1	78.9	82.3
0:1	89.3	77.5	83.6	71.6	78.2
1:1	91.8	87.5	84.3	80.9	85.4
1:2	90.5	86.7	88.5	81.1	84.8
2:1	93.6	84.9	89.1	82.9	87.5

component-level semantic loss, and instance-level semantic loss, respectively. As indicated in Table V, the findings reveal that MGN-Net's performance benefits from the integration of semantic contextual information across both coarse and fine granularity levels.

Additionally, Table VI presents experiments that incorporate various proportions of visual and semantic features, demonstrating the importance of complementary effects between different modalities for both detection and recognition tasks. It is noteworthy that relying solely on visual or semantic context information for graph node aggregation and updating results in suboptimal performance in the baseline model.

V. Conclusion

This article proposes a novel framework for scene text spotting: the MGN-Net. To take advantage of linguistic prior information embedded in text image, we argue that the compositional relationships among multiple granularities within textual instances can be regarded as contextual information, and text characteristics should be fully explored. In this framework, the correspondences between visual and semantic modality features at different hierarchical levels are thoroughly investigated. The visual semantic multigranularity feature extraction module is proposed to enrich the contextual relationships of textual primitives, while the multigranularity graph fusion learning module is employed to facilitate cross-modal and cross-hierarchy interactions among multimodality features, thereby achieving intra- and inter-deep fusion. Additionally, the semantic features at various granularity levels are employed to facilitate the model training. Extensive experimental results demonstrate the superior performance of our MGN-Net.

References

[1] D. Karatzas et al., "ICDAR 2015 competition on robust reading," in *Proc. 13th Int. Conf. Doc. Anal. Recognit. (ICDAR)*, 2015, pp. 1156–1160.

[2] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2963–2970.

[3] Y. Du et al., "SVTR: Scene text recognition with a single visual model," 2022, *arXiv:2205.00159*.

[4] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *Int. J. Comput. Vis.*, vol. 116, pp. 1–20, Jan. 2016.

[5] Z. Song, H. Zhang, and P. Cui, "Towards end-to-end scene text spotting by sharing convolutional feature map," in *Proc. IEEE 5th Int. Conf. Comput. Commun. (ICCC)*, 2019, pp. 1814–1820.

[6] W. Feng, F. Yin, X.-Y. Zhang, W. He, and C.-L. Liu, "Residual dual scale scene text spotting by fusing bottom-up and top-down processing," *Int. J. Comput. Vis.*, vol. 129, pp. 619–637, Mar. 2021.

[7] B. Zou, W. Yang, K. Li, E. Huang, and S. Liu, "Character flow detection and rectification for scene text spotting," in *Proc. 38th Comput. Graph. Int. Conf.*, 2021, pp. 288–299.

[8] Y. Gao, Z. Huang, Y. Dai, K. Chen, J. Guo, and W. Qiu, "Wacnet: Word segmentation guided characters aggregation net for scene text spotting with arbitrary shapes," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2019, pp. 3382–3386.

[9] X. Qin et al., "Mask is all you need: Rethinking mask R-CNN for dense and arbitrary-shaped scene text detection," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 414–423.

[10] W. Wang et al., "Pan++: Towards efficient and accurate end-to-end spotting of arbitrarily-shaped text," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5349–5367, Sep. 2022.

[11] H. Wang et al., "All you need is boundary: Toward arbitrary-shaped text spotting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 12160–12167.

[12] S. Qin, A. Bissacco, M. Raptis, Y. Fujii, and Y. Xiao, "Towards unconstrained end-to-end text spotting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4704–4714.

[13] H. Li, P. Wang, and C. Shen, "Towards end-to-end text spotting with convolutional recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5238–5246.

[14] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, "Mask TextSpotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 67–83.

[15] Q. Wang, Y. Zheng, and M. Betke, "A method for detecting text of arbitrary shapes in natural scenes that improves text spotting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 540–541.

[16] J. Jiang et al., "An end-to-end text spotter with text relation networks," *Cybersecurity*, vol. 4, no. 1, pp. 1–13, 2021.

[17] M. Liao, Z. Zou, Z. Wan, C. Yao, and X. Bai, "Real-time scene text detection with differentiable binarization and adaptive scale fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 919–931, Jan. 2023.

[18] P. Dai, S. Zhang, H. Zhang, and X. Cao, "Progressive contour regression for arbitrary-shape scene text detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7393–7402.

[19] R. Atienza, "Vision transformer for fast and efficient scene text recognition," in *Proc. Int. Conf. Doc. Anal. Recognit.*, 2021, pp. 319–334.

[20] Y. Zhai, X. Zhang, X. Qin, S. Zhao, X. Dong, and J. Shen, "TextFormer: A query-based end-to-end text spotter with mixed supervision," 2023, *arXiv:2306.03377*.

[21] M. Huang et al., "ESTextSpotter: Towards better scene text spotting with explicit synergy in transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 19495–19505.

[22] Y. Su, "Efficient and accurate scene text detection with low-rank approximation network," 2023, *arXiv:2306.15142*.

[23] M. Huang et al., "SwinTextSpotter: Scene text spotting via better synergy between text detection and text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4593–4603.

[24] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang, "ABCNet: Real-time scene text spotting with adaptive Bezier-curve network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9809–9818.

[25] Y. Liu et al., "ABCNet v2: Adaptive Bezier-curve network for real-time end-to-end text spotting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8048–8064, Nov. 2022.

[26] R. Ronen, S. Tsiper, O. Anschel, I. Lavi, A. Markovitz, and R. Manmatha, "Glass: Global to local attention for scene-text spotting," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 249–266.

[27] P. Lu, H. Wang, S. Zhu, J. Wang, X. Bai, and W. Liu, "Boundary TextSpotter: Toward arbitrary-shaped scene text spotting," *IEEE Trans. Image Process.*, vol. 31, pp. 6200–6212, 2022.

[28] C. Zeng, Y. Liu, and C. Song, "Rwin-FPN++: Rwin transformer with feature pyramid network for dense scene text spotting," *Appl. Sci.*, vol. 12, no. 17, p. 8488, 2022.

[29] D. Peng et al., "SPTS: Single-point text spotting," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 4272–4281.

[30] Y. Liu et al., "SPTS v2: Single-point scene text spotting," 2023, *arXiv:2301.01635*.

[31] P. Wang, H. Li, and C. Shen, "Towards end-to-end text spotting in natural scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 7266–7281, Oct. 2022.

[32] L. Qiao et al., "Text perceptron: Towards end-to-end arbitrary-shaped text spotting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 11899–11907.

[33] Z. Huang et al., "ADATS: Adaptive RoI-align based transformer for end-to-end text spotting," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2023, pp. 1403–1408.

[34] R. Bagi, T. Dutta, and H. P. Gupta, "Cluttered TextSpotter: An end-to-end trainable light-weight scene text spotter for cluttered environment," *IEEE Access*, vol. 8, pp. 111433–111447, 2020.

[35] R. Yoshihashi, T. Tanaka, K. Doi, T. Fujino, and N. Yamashita, "Context-free TextSpotter for real-time and mobile end-to-end text detection and recognition," in *Proc. 16th Int. Conf. Doc. Anal. Recognit.*, 2021, pp. 240–257.

[36] A. Singh, G. Pang, M. Toh, J. Huang, W. Galuba, and T. Hassner, "TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2021, pp. 8802–8812.

[37] J. Baek et al., "What is wrong with scene text recognition model comparisons? Dataset and model analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4715–4723.

[38] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "An attentional scene text recognizer with flexible rectification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2035–2048, Sep. 2019.

[39] W. Wang et al., "AE TextSpotter: Learning visual and linguistic representation for ambiguous text spotting," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 457–473.

[40] A. Sabir, F. Moreno-Noguer, and L. Padró, "Visual re-ranking with natural language understanding for text spotting," in *Proc. 14th Asian Conf. Comput. Vis.*, 2019, pp. 68–82.

[41] A. Sabir, F. Moreno-Noguer, and L. Padró, "Enhancing text spotting with a language model and visual context information," in *Artificial Intelligence Research and Development*. Amsterdam, The Netherlands: IOS Press, 2018.

[42] C. Xue, W. Zhang, Y. Hao, S. Lu, P. H. Torr, and S. Bai, "Language matters: A weakly supervised vision-language pre-training approach for scene text detection and spotting," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 284–302.

[43] J. Lee, S. Park, J. Baek, S. J. Oh, S. Kim, and H. Lee, "On recognizing texts of arbitrary shapes with 2D self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 546–547.

[44] F. Sheng, Z. Chen, and B. Xu, "NRTR: A no-recurrence sequence-to-sequence model for scene text recognition," in *Proc. Int. Conf. Doc. Anal. Recognit. (ICDAR)*, 2019, pp. 781–786.

[45] L. Yang, P. Wang, H. Li, Z. Li, and Y. Zhang, "A holistic representation guided attention network for scene text recognition," *Neurocomputing*, vol. 414, pp. 67–75, Nov. 2020.

[46] Z. Qiao et al., "PIMNet: A parallel, iterative and mimicking network for scene text recognition," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 2046–2055.

[47] S. Fang, Z. Mao, H. Xie, Y. Wang, C. Yan, and Y. Zhang, "Abinet++: Autonomous, bidirectional and iterative language modeling for scene text spotting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7123–7141, Jun. 2023.

[48] M. Cheng et al., "ViSTA: Vision and scene text aggregation for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5184–5193.

[49] S. Song et al., "Vision-language pre-training for boosting scene text detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15681–15691.

[50] D. Yu et al., "Towards accurate scene text recognition with semantic reasoning networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12113–12122.

[51] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7098–7107.

[52] A. K. Bhunia, A. Sain, A. Kumar, S. Ghose, P. N. Chowdhury, and Y.-Z. Song, "Joint visual semantic reasoning: Multi-stage decoder for text recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14940–14949.

[53] Y. Wang, H. Xie, S. Fang, J. Wang, S. Zhu, and Y. Zhang, "From two to one: A new scene text recognizer with visual language modeling network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14194–14203.

[54] B. Na, Y. Kim, and S. Park, "Multi-modal text recognition networks: Interactive enhancements between visual and semantic features," in *Proc. 17th Eur. Conf. Comput. Vis.*, 2022, pp. 446–463.

[55] K. Han, Y. Wang, J. Guo, Y. Tang, and E. Wu, "Vision GNN: An image is worth graph of nodes," in *Proc. 36th Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 8291–8303.

[56] H. Bi et al., "HGR-Net: Hierarchical graph reasoning network for arbitrary shape scene text detection," *IEEE Trans. Image Process.*, vol. 32, pp. 4142–4155, 2023.

[57] C. Da, P. Wang, and C. Yao, "Multi-granularity prediction with learnable fusion for scene text recognition," 2023, *arXiv:2307.13244*.

[58] X. Canhui, L. Yuteng, S. Cao, Z. Honghong, B. Hengyue, and C. Yinong, "HiM: Hierarchical multimodal network for document layout analysis," *Appl. Intell.*, vol. 53, pp. 24314–24326, Jul. 2023.

[59] C. Shi et al., "Graph-based convolution feature aggregation for retinal vessel segmentation," *Simul. Model. Pract. Theory*, vol. 121, Dec. 2022, Art. no. 102653.

[60] C. Shi, C. Xu, H. Bi, Y. Cheng, Y. Li, and H. Zhang, vol. 71, no. 1, pp. 1–10, Aug. 2022. "Lateral feature enhancement network for page object detection," *IEEE Trans. Instrum. Meas.*,

[61] C.-h. Xu, C. Shi, and Y.-n. Chen, "End-to-end dilated convolution network for document image semantic segmentation," *J. Central South Univ.*, vol. 28, no. 6, pp. 1765–1774, 2021.

[62] C. Xu et al., "A page object detection method based on mask R-CNN," *IEEE Access*, vol. 9, pp. 143448–143457, 2021.

[63] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[64] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[65] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[66] X. Zhang, Y. Su, S. Tripathi, and Z. Tu, "Text spotting transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9519–9528.

[67] M. Ye et al., "DeepSolo: Let transformer decoder with explicit points solo for text spotting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19348–19357.

[68] Y. Wang, H. Xie, S. Fang, J. Wang, S. Zhu, and Y. Zhang, "From two to one: A new scene text Recognizer with visual language modeling network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 14174–14183.

[69] J. Zhang, T. Lin, Y. Xu, K. Chen, and R. Zhang, "Relational contrastive learning for scene text recognition," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 5764–5775.

[70] W. Feng, W. He, F. Yin, X.-Y. Zhang, and C.-L. Liu, "TextDragon: An end-to-end framework for arbitrary shaped text spotting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9076–9085.

[71] J. Wu, P. Lyu, G. Lu, C. Zhang, K. Yao, and W. Pei, "Decoupling recognition from detection: Single shot self-reliant scene text spotter," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1319–1328.

[72] H. Zhang et al., "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," 2022, *arXiv:2203.03605*.

[73] M. Liao, G. Pang, J. Huang, T. Hassner, and X. Bai, "Mask TextSpotter v3: Segmentation proposal network for robust scene text spotting," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 706–722.

[74] P. Wang et al., "PGNet: Real-time arbitrarily-shaped text spotting with point gathering network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 2782–2790.

[75] L. Qiao et al., "MANGO: A mask attention guided one-stage scene text spotter," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 2467–2476.

[76] T. Kil, S. Kim, S. Seo, Y. Kim, and D. Kim, "Towards unified scene text spotting based on sequence generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15223–15232.

[77] Y. Kittenplon, I. Lavi, S. Fogel, Y. Bar, R. Manmatha, and P. Perona, "Towards weakly-supervised text spotting using a multi-task transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4604–4613.

[78] W. Wang et al., "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8440–8449.

[79] Y. Liu, L. Jin, S. Zhang, C. Luo, and S. Zhang, "Curved scene text detection via transverse and longitudinal sequence connection," *Pattern Recognit.*, vol. 90, pp. 337–345, 2019.

[80] C.-K. Ch'ng, C. S. Chan, and C.-L. Liu, "Total-text: Toward orientation robustness in scene text detection," *Int. J. Doc. Anal. Recognit. (IJDAR)*, vol. 23, no. 1, pp. 31–52, 2020.