Image as a Language: Revisiting Scene Text Recognition via Balanced, Unified and Synchronized Vision-Language Reasoning Network

Jiajun Wei[1], Hongjian Zhan[1]*, Yue Lu[1], Xiao Tu[1], Bing Yin[2], Cong Liu[2], Umapada Pal[3]

[1]Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai, China
[2]iFLYTEK Research, iFLYTEK, Hefei, China
[3]CVPR Unit, Indian Statistical Institute, Kolkata, India
jjwei@stu.ecnu.edu.cn, {hjzhan, xtu, ylu}@cee.ecnu.edu.cn, {bingyin, congliu2}@iflytek.com, umapada@isical.ac.in

Abstract

Scene text recognition is inherently a vision-language task. However, previous works have predominantly focused either on extracting more robust visual features or designing better language modeling. How to effectively and jointly model vision and language to mitigate heavy reliance on a single modality remains a problem. In this paper, aiming to enhance vision-language reasoning in scene text recognition, we present a balanced, unified and synchronized vision-language reasoning network (BUSNet). Firstly, revisiting the image as a language by balanced concatenation along length dimension alleviates the issue of over-reliance on vision or language. Secondly, BUSNet learns an ensemble of unified external and internal vision-language model with shared weight by masked modality modeling (MMM). Thirdly, a novel vision-language reasoning module (VLRM) with synchronized vision-language decoding capacity is proposed. Additionally, BUSNet achieves improved performance through iterative reasoning, which utilizes the vision-language prediction as a new language input. Extensive experiments indicate that BUSNet achieves state-of-the-art performance on several mainstream benchmark datasets and more challenge datasets for both synthetic and real training data compared to recent outstanding methods. Code and dataset will be available at https://github.com/jjwei66/BUSNet.

Introduction

Acquiring the capability to read text from scene images is a crucial objective for artificial intelligence. Early methods (He et al. 2016b; Su and Lu 2017) treat text recognition as a straightforward classification task. Nevertheless, to enhance accuracy in handling challenging factors like occlusion, blur, and noise in text images, researches (Baek et al. 2019; Cheng et al. 2017; Fang et al. 2021; Bautista and Atienza 2022) have shifted towards incorporating language modeling.

However, with previous works mainly treating scene text recognition (STR) as an image-to-text task, researchers of STR mostly focus on extracting more robust visual features (Aberdam et al. 2021; Wang et al. 2021; Zhong et al. 2022) or more appropriate language modeling (Fang et al. 2021; Bautista and Atienza 2022) to improve performance. As illustrated in Figure 1 (a), relying solely on visual features often proves inadequate when handling images with noise or occlusion. Heavy reliance on the language model may cause over-refinement issues like Figure 1 (b) and (c). STR inherently involves vision and language modalities, and how to effectively model the interaction between vision and language to fully leverage complementary information remains an open challenge. Addressing this problem requires exploring the following perspectives:

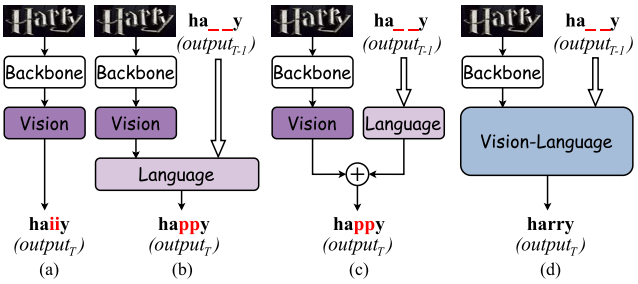

Figure 1: Different modality reliance recognition pipelines. (a) The vision-based model emphasizes vision side. (b) The internal language-based model emphasizes language side. (c) The external language-based model extremely emphasizes vision in vision part and language side in language part. (d) Our vision-language-based model simultaneously emphasizes vision and language side by balanced modeling.

(i) The imbalanced reliance of vision and language: as depicted in Figure 1(a) and Figure 1(b), methods like SVTR (Du et al. 2022) and PARSeq (Bautista and Atienza 2022), which respectively emphasize the vision or language side. Furthermore, methods (Yu et al. 2020; Fang et al. 2021) like Figure 1(c) independently handle each modality, resulting in an extreme imbalance in every single part. The key problem under the imbalance is treating STR as an image-to-text task and assigning distinct roles to two modalities. However, the ultimate target is to obtain a text modality prediction and the information from vision and language should ideally be utilized with no fundamental difference. To overcome the imbalance, a new perspective of handling vision and language in a balanced manner like Figure 1(d) is needed.

(ii) Choice of internal or external modeling: Fang et al (Fang et al. 2021) adopt an approach that decouples the

*Corresponding author.

recognition model into separate vision and language models as illustrated in Figure 1(c). This separation introduces additional computational burden due to the presence of the language model compared to internal modeling methods (Baek et al. 2019) like Figure 1(b), and the refinement process can be prone to errors without vision information. Conversely, external modeling offers the feasibility to leverage extensive prior linguistic knowledge and enables control over how the language model learns character relationships. Achieving a unified approach that harnesses the advantages of both modeling is crucial to benefit from multi-modality fusion.

(iii) Different decoding scheme: linguistic-free methods like CTC-based methods (Liu et al. 2016; Su and Lu 2017) or segmentation-based methods (Liao et al. 2019; Wan et al. 2020) predict all characters in a single time step. In contrast, autoregressive (AR) decoding scheme, where characters are decoded in order, has been the mainstream in linguistic-based methods (Baek et al. 2019; Li et al. 2019; Bautista and Atienza 2022). The presence of diverse decoding schemes poses challenge in effectively fusing vision and language information. Fortunately, recent global linguistic decoding methods (Yu et al. 2020) provides a potential opportunity for parallel language reasoning. Nevertheless, how to design a vision-language reasoning network with synchronized decoding capacity is still unsolved.

Based on the analyses conducted, we propose a balanced, unified and synchronized vision-language reasoning network (BUSNet). Firstly, after extracting vision and language features, we treat them in a balanced manner by concatenating along the length dimension. This approach considers the images as a form of language with noisy, and reasoning module comprehensively utilizes valid and complementary information of each language. Secondly, we employ masked modality modeling (MMM) to create an ensemble of unified external and internal vision-language models with shared weights. This generates three types of inputs features: single-vision features, single-language features, and vision-language features through two mask tokens $\mathbf{M}_v$ and $\mathbf{M}_l$ and enforces the reasoning module to possess predicting capacity through different combinations. Thirdly, we introduce a novel synchronized vision-language reasoning module (VLRM), which integrates parallel vision modeling with bidirectional language modeling. The VLRM jointly reasons between vision and language in a single time step by specifying the special attention mask. Additionally, we incorporate an iterative reasoning method by utilizing the vision-language prediction as new language input, thereby further enhancing overall performance.

Moreover, through revisiting image as a language, the single-vision recognition procedure acquires linguistic reasoning ability. This capacity can be attributed to the shared decoder between vision and language, which facilitates the acquisition of linguistic relationships. Consequently, even when confronted with occluded images, the reasoning module is still capable of producing accurate predictions.

The contributions of this paper are as follows:

1) We propose a new balanced, unified and synchronized vision-language reasoning network (BUSNet). By revisiting image as a language in a balanced manner, learning an ensemble of unified external-internal modeling and utilizing synchronized vision-language decoding scheme, BUSNet is capable of mutually incorporating both vision features and linguistic knowledge to generate the results while alleviates single modality over-reliance problem.

2) A novel VLRM is introduced, which is designed to reason over the accurate predictions from both noisy vision features and noisy language features in parallel decoding scheme. To the best of our knowledge, the VLRM with synchronized vision parallel decoding and global linguistic language decoding is a novel component, that are natural extensions but have never been explored.

3) Extensive experiments demonstrate that the proposed BUSNet achieves the state-of-the-art (SOTA) on mainstream and recent challenging benchmarks for both synthetic and real training data compared to outstanding methods.

Methodology

Encoder

Image Encoder The 12-layer Vision Transformer (ViT) (Dosovitskiy et al. 2020) without the [class] token and classification head is chosen to be the image encoder. The image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ is reshaped into a sequence of flatten 2D patches $\mathbf{x}_p \in \mathbb{R}^{N_v\times(P_h P_w\cdot C)}$, where (H, W) is the height, width. C is the number of channels, (P_h, P_w) is the resolution of each image patch, and $N_v = HW/P_h P_w$ is the number of patches. Then the patches are flattened and mapped to D dimensions with a trainable patch linear projection layer $\mathbf{W}_p \in \mathbb{R}^{P_h P_w C\times D}$. We refer to the output of $\mathbf{W}_p$ as input tokens $\mathbf{z}_v \in \mathbb{R}^{N_v\times D}$. Fixed cosine position embedding $\mathbf{E}_{pos} \in \mathbb{R}^{N_v\times D}$ of equal dimension, which is used to retain positional information, is added to the tokens before being processed by ViT. The formulation of vision feature $\mathbf{F}_v$ is as below:

$$\mathbf{F}_v = \mathrm{ViT}(\mathbf{z}_v + \mathbf{E}_{pos}) \in \mathbb{R}^{N_v\times D}. \tag{1}$$

Language Encoder Transformer-based or RNN-based encoders are the mainstreams for natural language processing. However, they require interaction among characters and lead to leak information, which results in failing to capture meaningful global linguistic knowledge. In that case, given an one-hot encoding text string $\mathbf{y}_l = (\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_{N_l}) \in \mathbb{R}^{N_l\times c}$ with maximum length N_l and class number c, we opt for a straightforward approach by using a text-embedding layer $\mathbf{W}_T \in \mathbb{R}^{c\times D}$, D is model dimension same as the image encoder. Fixed cosine position embedding $\mathbf{E}_{pos} \in \mathbb{R}^{N_l\times D}$ are also used for retaining positional information. The formulation of language features $\mathbf{F}_l$ is as below:

$$\mathbf{F}_l = \mathrm{Embed}(\mathbf{y}_l) + \mathbf{E}_{pos} \in \mathbb{R}^{N_l\times D}. \tag{2}$$

Vision-Language Reasoning Network

Balanced Vision-Language Modeling Figure 2 illustrates the balanced treatment of $\mathbf{F}_v$ and $\mathbf{F}_l$ achieved through concatenation ($\mathcal{C}$) along the length dimension. The multi-modality features can be formulated as $\mathbf{F}_m = \mathcal{C}(\mathbf{F}_v, \mathbf{F}_l) \in \mathbb{R}^{(N_v+N_l)\times D}$. Our key insight is that instead of treating STR

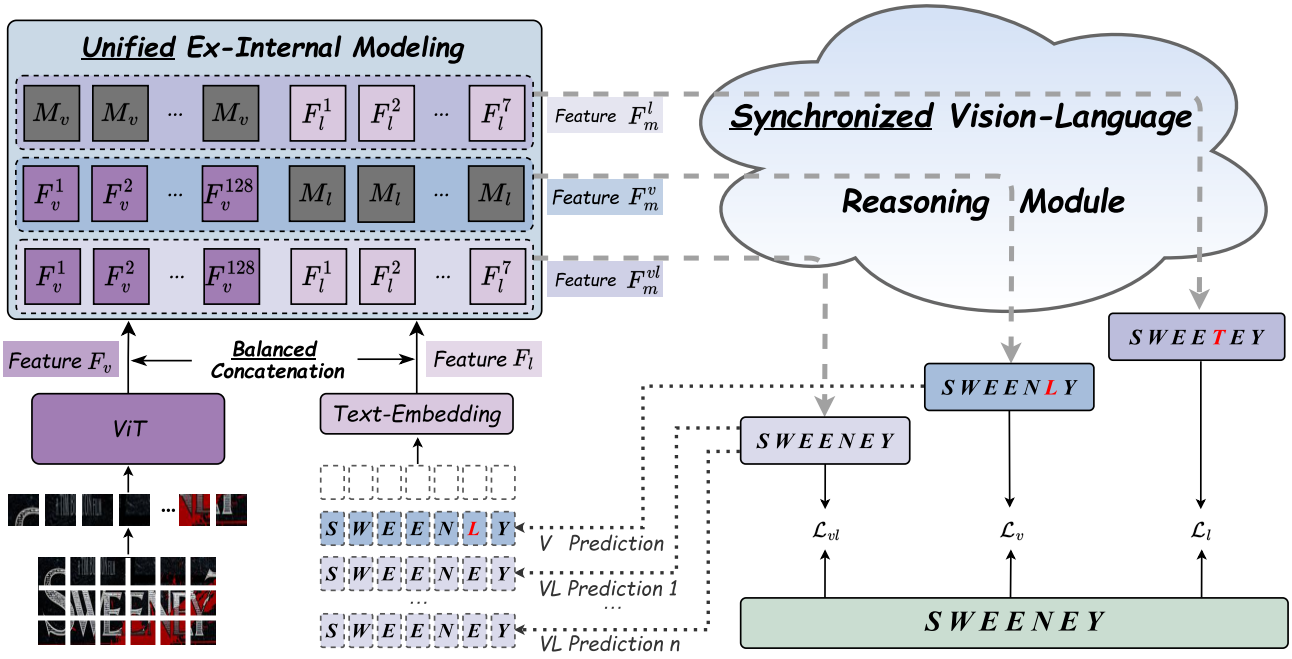

Figure 2: An overview of BUSNet. The input text string is firstly empty and then is provided by vision or vision-language prediction with the increased reasoning times. $\mathbf{F}_v$ and $\mathbf{F}_l$ are concatenated along the length dimension. Mask tokens $\mathbf{M}_v$ and $\mathbf{M}_l$ are appended for modality masking that generates three multi-modality features $\mathbf{F}_m$ for unified ex-internal training. Through reasoning module with synchronized vision-language parallel decoding capacity, the final sequence is output by $\mathbf{F}_m^{vl}$.

as an image-to-text task, we regard it as a special text-to-text task. By concatenating $\mathbf{F}_v$ and $\mathbf{F}_l$, images can be considered as text inputs with noisy like misspelled words, while still carrying valid information for predicting.

Based on balanced modeling, BUSNet excels not only in vision-language reasoning but also mitigates the issues of neglecting linguistic information in visual models and over-refinement of language models. Moreover, as sharing decoder with $\mathbf{F}_l$, the balanced modeling provides linguistic reasoning capacity for handling noisy or blurred $\mathbf{F}_v$, which indicates that vision and language information have no foundation difference in STR.

Unified External-Internal Modeling External modeling methods (Fang et al. 2021; Yu et al. 2020) introduce additional computational and parameter burden, and each single modality unit cannot leverage the information from the other modality. Nevertheless, external modeling enables the utilization of prior knowledge. To harness the advantages of internal and external modeling, we introduce masked modality modeling (MMM) through masking operations.

Specifically, vision and language mask tokens $\mathbf{M}_v$, $\mathbf{M}_l$ are introduced to represent the absence of corresponding modality. When concatenating $\mathbf{F}_v$ and $\mathbf{F}_l$, we selectively replace one modality through expanding ($\mathcal{E}$) the corresponding $\mathbf{M}$ token to the same length. This resulting in three types of multi-modality features $\mathbf{F}_m$: 1) $\mathbf{F}_m^l = \mathcal{C}(\mathcal{E}(\mathbf{M}_v), \mathbf{F}_l)$, which involves single language features to force a model possessing linguistic reasoning ability. 2) $\mathbf{F}_m^v = \mathcal{C}(\mathbf{F}_v, \mathcal{E}(\mathbf{M}_l))$, which includes the vision features alongside language mask tokens to instruct the model to rely on the vision clues alone for predictions. 3) $\mathbf{F}_m^{vl} = \mathcal{C}(\mathbf{F}_v, \mathbf{F}_l)$, which contains the vision and language features, allowing model to simultaneously utilize different modalities information to make predictions.

Under the MMM, BUSNet concurrently benefits from the both modeling. As three types of $\mathbf{F}_m$ share a decoder, the extra computational and parameter burden are not a problem. Besides, this mutually beneficial interaction enhances the unimodality predicting process. For instance, when predicting the whitespace of " S _ O P ", the potential predictions of $\mathbf{F}_m^l$ could be "S T O P" or "S H O P". However, if we provide a scene image of "SHOP", the predicted result of $\mathbf{F}_m^{vl}$ is significantly enhanced by leveraging the complementary information from the vision modality. Furthermore, the utilization of $\mathbf{F}_m^l$ and $\mathbf{F}_m^v$ enables the reasoning module to make prediction using single modality, making it feasible to capture rich prior knowledge through pre-training.

Synchronized Vision-Language Reasoning Module The disparity in decoding schemes has hindered the seamless interaction between vision and language. Benefiting from MMM, global linguistic decoding can be applied in BUSNet. We align vision parallel decoding and language global linguistic decoding in an union framework, which is called vision-language reasoning module (VLRM).

The VLRM consists of several blocks. Each block shares the consistent structure: multi-head attention (MHA)

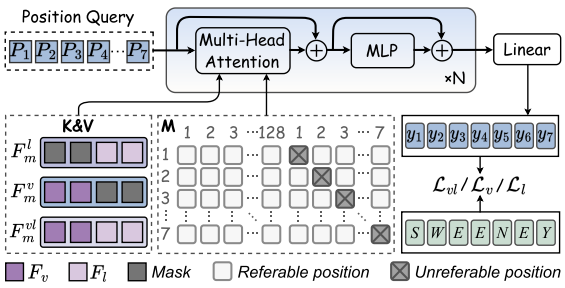

Figure 3: Architecture of VLRM.

(Vaswani et al. 2017), feed-forward network (Vaswani et al. 2017), residual connection (He et al. 2016a) as shown in Figure 3. In the VLRM, the multi-modality features $\mathbf{F}_m$ are fed into the MHA layer which enables the module to capture the interactions and dependencies between the vision and language features, facilitating cross-modal reasoning and fusion. An attention mask $\mathbf{M}_a$ is incorporated into the MHA layer to enhance the linguistic ability. The $\mathbf{M}_a$ are designed specifically to prevent the position query from merely focusing on the corresponding position character information. Without the $\mathbf{M}_a$, the VLRM will neglect the other characters and obtain nothing meaningful for semantic reasoning. For the same reason, to avoid information leakage between $\mathbf{F}_m$, no self-attention is applied within the VLRM.

The attention operation inside the MHA layer can be formalized as:

$$\mathbf{M}_a^{ij} = \begin{cases} 0, & i \neq N_v + j \\ -\infty, & i = N_v + j \end{cases}, \tag{3}$$

$$\mathbf{K} = \mathbf{V} = \{\mathbf{F}_m^{vl},\ \mathbf{F}_m^{v},\ \mathbf{F}_m^{l}\}, \tag{4}$$

$$\mathbf{F}_{attn} = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^{\mathbf{T}}}{\sqrt{\mathrm{D}}} + \mathbf{M}_a)\mathbf{V}, \tag{5}$$

where $\mathbf{Q} \in \mathbb{R}^{N_l \times D}$ is the fixed cosine position embeddings $\mathbf{P}$ of character orders in the first layer or the outputs of the last layer otherwise. $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{(N_v+N_l)\times D}$ are obtained from the encoder. $\mathbf{M}_a \in \mathbb{R}^{N_l \times (N_v+N_l)}$ is the matrix of attention masks which effectively make VLRM to possess vision and semantic reasoning ability.

By applying global linguistic decoding, VLRM can handle vision and language in the same decoding manner. The reasoning process of the i_{th} character can be formulated as $y_i = f(\mathbf{F}_v^1, ..., \mathbf{F}_v^{N_v}, \mathbf{F}_l^1, ..., \mathbf{F}_l^{i-1}, \mathbf{F}_l^{i+1}, ..., \mathbf{F}_l^{N_l})$. Blurred or occlusion images can correctly be predicted with the help of semantic knowledge, and the global language reasoning procedure is supervised by visual features.

Iterative Reasoning

Benefiting from MMM, BUSNet takes the advantage of external language and can apply iterative reasoning to boost accuracy in the test phase. Initially, the input text string $\mathbf{y}_l$ is empty and the prediction of $\mathbf{F}_m^{vl}$ and $\mathbf{F}_m^{v}$ are identical. After the first iteration, the prediction of $\mathbf{F}_m^{v}$ is fed, which enables the vision-language reasoning capacity of BUSNet with the inputs of $\mathbf{F}_v$ and $\mathbf{F}_l$.

Typically, two rounds of iterative reasoning are sufficient to achieve satisfactory results. However, as illustrated in Figure 2, we can further exploit the output of $\mathbf{F}_m^{vl}$ as a raw new text string input and further boost accuracy.

Training Objective

BUSNet is trained using the following objectives:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{vl} + \lambda_2 \mathcal{L}_v + \lambda_3 \mathcal{L}_l, \tag{6}$$

where $\mathcal{L}_{vl}$, $\mathcal{L}_v$ and $\mathcal{L}_l$ are cross-entropy losses from $\mathbf{F}_m^{vl}$, $\mathbf{F}_m^{v}$, $\mathbf{F}_m^{l}$, respectively. λ_1, λ_2 and λ_3 are hyper-parameters for each loss.

Experiments

Datasets and Implementation Details

Datasets MJSynth (MJ) (Jaderberg et al. 2014) and Synth-Text (ST) (Gupta, Vedaldi, and Zisserman 2016) are used as synthetic training datasets, and real datasets collected by (Bautista and Atienza 2022) are used as real training datasets. Six mainstream benchmarks include ICDAR 2013 (IC13) (Karatzas et al. 2013), ICDAR 2015 (IC15) (Karatzas et al. 2015), IIIT5k-Words (IIIT5k) (Mishra, Alahari, and Jawahar 2012), Street View Text (SVT) (Wang, Babenko, and Belongie 2011), Street View Text-Perspective (SVTP) (Phan et al. 2013) and CUTE80 (CUTE) (Risnumawan et al. 2014) are used as the testing datasets. In addition to the six benchmark datasets, we also conduct experiments on three more challenging datasets: COCO-Text (9.8k samples; low-resolution, occluded text), Art (Chng et al. 2019) (35.1k samples; curved and rotated text), and Uber-Text (Zhang et al. 2017) (80.6k samples; vertical and rotated text).

The patch height and width are set to 4 and 8. The dimension C of BUSNet$_B$ and BUSNet$_S$ are set to 512 and 384. There are 6 layers in VLRM with 8 and 6 attention head for big and small models in each layer. Balanced factors λ_1, λ_2 and λ_3 are set to 1, 1 and 1. All input images are scaled to 32×128 with data augmentation Adam is selected as the optimizer with initial learning rate $1e^{-4}$ and is decayed to $1e^{-5}$ after 12 epochs. The batch size is set to 192. All experiments are conducted on one NVIDIA 3090 GPU. Specifically, we obtain three outputs V (single-vision), L (single-language) and VL (vision-language) corresponding to $\mathbf{F}_m^{v}$, $\mathbf{F}_m^{l}$ and $\mathbf{F}_m^{vl}$, respectively.

Comparisons with State-of-the-Arts

The comparisons of BUSNet with previous outstanding methods are shown in Table 1.

VisionLAN and ABINet respectively emphasize that the visual and textual semantic information can be utilized for STR. They can be considered as two special condition of BUSNet. As can be seen from the comparison, when trained on synthetic datasets, BUSNet outperforms VisionLAN with 2.6%, 3.8%, 0.4%, 3.5%, 5.8% and 2.8% on IC13, SVT, IIIT5k, IC15,SVTP and CUTE datasets and outperforms ABINet 0.9%, 2.0%, 1.2%, 2.5% and 2.1% on IC13, SVT, IC15, SVTP and CUTE datasets, respectively. Observing the

Methods	Train Data	IC13 857	IC13 1015	SVT 647	IIIT5k 3000	IC15 1811	IC15 2077	SVTP 645	CUTE 288	Param ($\times 10^6$)	Time (ms)
SRN (Yu et al. 2020)	S	95.5	-	91.5	94.8	82.7	-	85.1	87.8	54.7	14.1
RobustScanner (Yue et al. 2020)	S	-	94.8	88.1	94.8	-	77.1	79.5	90.3	-	-
TextScanner (Wan et al. 2020)	S	-	92.9	90.1	93.9	-	79.4	84.3	83.3	57.0	56.8
PREN2D (Yan et al. 2021)	S	96.4	-	94.0	95.6	83.0	-	87.6	91.7	59.0	61.6
VisionLAN (Wang et al. 2021)	S	95.7	-	91.7	95.8	83.7	-	86.0	88.5	32.8	16.8
ViTSTR† (Atienza 2021)	S	93.2	92.4	87.7	88.4	78.5	72.6	81.8	81.3	23.8	19.1
TRBA (Baek et al. 2019)	S	-	93.1	88.9	92.1	-	74.7	79.5	78.2	21.4	8.0
ABINet† (Fang et al. 2021)	S	97.4	-	93.5	96.2	86.0	-	89.3	89.2	36.7	29.6
SGBANet (Zhong et al. 2022)	S	95.1	-	89.1	95.4	78.4	-	83.1	88.2	-	-
LevOCR (Da, Wang, and Yao 2022)	S	96.8	-	92.8	96.6	86.4	-	88.0	91.6	92.6	60.5
PARSeq$^\dagger_A$ (Bautista and Atienza 2022)	S	97.0	96.2	93.6	**96.8**	86.5	82.9	88.9	**92.2**	23.8	21.1
PARSeq$^\dagger_N$ (Bautista and Atienza 2022)	S	96.3	95.5	92.6	95.7	85.1	81.4	87.9	91.4	23.8	14.1
MATRN* (Na, Kim, and Park 2022)	S	97.2	94.7	93.5	96.4	86.4	82.3	88.5	90.9	44.2	45.6
BUSNet$^\dagger_S$ (Ours)	S	97.3	**96.3**	**94.9**	96.5	**87.1**	**83.3**	**91.2**	89.6	32.0	20.1
BUSNet$^\dagger_B$ (Ours)	S	**98.3**	**97.4**	**95.5**	96.2	**87.2**	**83.3**	**91.8**	91.3	56.8	23.6
ViTSTR (Atienza 2021)	R	97.6	97.7	95.8	98.1	88.4	87.1	91.4	96.1	23.8	19.1
CRNN (Baek et al. 2019)	R	94.1	94.5	90.7	94.6	82.0	78.5	80.6	89.1	8.4	4.5
TRBA* (Baek et al. 2019)	R	97.6	97.5	96.7	98.5	89.8	88.7	92.3	96.5	21.4	8.0
ABINet* (Fang et al. 2021)	R	97.5	97.4	96.9	98.6	90.1	88.0	92.4	95.8	36.7	29.6
PARSeq$_A$ (Bautista and Atienza 2022)	R	98.3	98.4	97.9	**99.1**	90.7	89.6	95.7	98.3	23.8	21.1
PARSeq$_N$ (Bautista and Atienza 2022)	R	98.0	98.1	97.5	98.3	89.6	88.4	94.6	97.7	23.8	14.1
BUSNet$_S$ (Ours)	R	97.6	97.6	97.4	98.0	**91.1**	**89.8**	95.0	**98.3**	32.0	20.1
BUSNet$_B$ (Ours)	R	**98.3**	**98.5**	**98.5**	98.0	**91.3**	**90.2**	**96.3**	98.0	56.8	23.6

Table 1: Accuracy comparison with the State-of-the-Art Methods on mainstream benchmarks. 'S' and 'R' denote the synthetic and real datasets. '*' indicates that the models are reproduced by us. '†' means the inference time is estimated using one NVIDIA 3090 GPU. 'N' and 'A' are NAR and AR decoding schemes. '$_B$' and '$_S$' are big and small models.

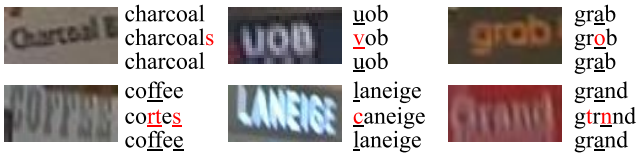

Figure 4: Qualitative examples where ABINet fails but BUSNet gives correct results. From top to bottom are ground truth, ABINet prediction and our prediction.

qualitative results from Figure 4 we present that hard examples with confusing fonts and blurred appearance can be recognized by BUSNet but can not be recognized by ABINet. Besides, compared with recent SOTA and competitive work PARSeq, especially, BUSNet has prominent superiority on SVT, IC15 and SVTP. As depicted in the Table 1, when trained with real datasets, we can observe that some regular benchmarks have been saturated. However, BUSNet still presents superiority on irregular datasets like IC15 and SVTP. But BUSNet doesn't achieve the best on the IIIT5k and we will discuss it in the late error analysis section.

Comparisons on More Challenging Datasets

Since the performances on mainstream benchmark datasets are close to saturation, we further compare BUSNet with previous SOTA methods on challenging datasets. As shown in Table 2, BUSNet outperforms competitive NAR methods like ABINet and PARSeq 7.3% and 2.8% on average when trained on synthetic datasets. When trained on real datasets, though all methods get great improvement, BUSNet still gains 8.4% and 0.9% improvement on average compared to ABINet and PARSeq. The results demonstrate that BUSNet consistently performs well across all benchmarks, revealing the robustness and effectiveness of BUSNet.

Methods	Train Data	ArT 35149	COCO 9825	Uber 80551
CRNN (Baek et al. 2019)	S	57.3	49.3	33.1
ViTSTR (Atienza 2021)	S	66.1	56.4	37.6
TRBA (Baek et al. 2019)	S	68.2	61.4	38.0
ABINet (Fang et al. 2021)	S	65.4	57.1	34.9
PARSeq (Bautista 2022)	S	69.1	60.2	39.9
BUSNet (Ours)	S	**70.3**	**65.4**	**43.1**
CRNN (Baek et al. 2019)	R	66.8	62.2	51.0
ViTSTR (Atienza 2021)	R	81.1	74.1	78.2
TRBA (Baek et al. 2019)	R	82.5	77.5	81.2
ABINet (Fang et al. 2021)	R	81.2	76.4	71.5
PARSeq (Bautista 2022)	R	83.0	77.0	82.4
BUSNet (Ours)	R	**83.4**	**79.4**	**83.2**

Table 2: Comparisons on More Challenging Datasets

Outputs	IC13	SVT	IIIT5k	IC15	SVTP	CUTE
V^{wo}	96.7	95.4	96.7	88.7	91.0	94.1
V^{w}	**97.9**	**97.5**	**97.7**	**90.1**	**94.9**	**97.9**
L^{wo}	67.7	68.0	55.7	57.1	67.3	57.3
L^{w}	**70.3**	**68.6**	**57.0**	**59.0**	**66.7**	**59.4**
VL	**98.3**	**98.5**	**98.0**	**91.3**	**96.3**	**98.0**

Table 3: Ablation study of balanced modeling. Same kind outputs are compared under different training strategies. V^{w} or V^{wo} means single-vision output trained with or without a balanced language part. L^{w} or L^{wo} means single-language output trained with or without a balanced vision part.

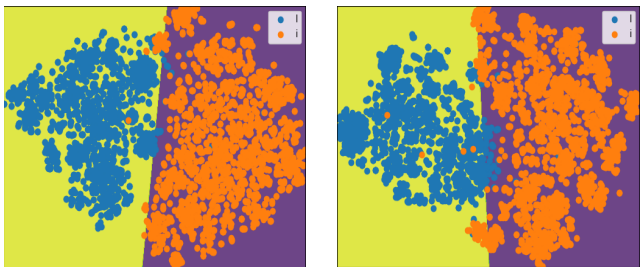

Figure 5: T-SNE plot of characters 'i' and 'l' features before the last classification linear layer of VLRM. The left and right are the V^{w} and V^{wo} output features.

Ablation Study

The effectiveness of Balanced Modeling From the results in Table 3 we can observe: 1) VL outperforms V^{wo} and L^{wo} with 2.2% and 36.5%. This indicates extremely unbalanced modeling and unimodal reliance bring a big drop in performance. 2) As balanced modeling means vision and language are both utilized in the same manner, which provides more "vision form language" training examples compared to single-language training, we observe improved accuracy with 1.3% for L. 3) BUSNet is trained with a focus on reasoning ability in the "text language" phase, we still observe performance improvements of 1.6% on V. Some V^{wo} and V^{w} predictions of confusing images are presented in Figure 6. Typically, even only with occluded or blurred visual features, the visual prediction can be correctly refined by semantic reasoning capacity in balanced modeling. For example, the incomplete character 'y' can be inferred from the visual clues of rest characters 'beaut' in Figure 6.

To emphasize image as a language, we present a visual feature of two visually similar characters 'i' , 'l' in test datasets in Figure 5. From the T-SNE plot, we can observe: 1) the features of the two visually similar characters 'i' and 'l' are noticeably separated to a greater extent than trained without the balanced language part. 2) A decreased example of wrongly prediction. 3) The features trained in a balanced manner tend to be divided by a linear classification layer.

The effectiveness of Unified External-Internal Modeling As depicted in Table 4, we adopt the models trained with or without MMM and observe the following: 1) the accuracy

Figure 6: Qualitative examples where V^{wo} fails but V^{w} succeeds. From top to bottom are ground truth, V^{wo} and V^{w}.

Outputs	IC13	SVT	IIIT5k	IC15	SVTP	CUTE
V^{wo}	2.5	3.1	2.9	2.3	2.5	3.5
V^{w}	**97.9**	**97.5**	**97.7**	**90.1**	**94.9**	**97.9**
L^{wo}	0.0	0.0	0.7	0.0	0.0	0.4
L^{w}	**70.3**	**68.6**	**57.0**	**59.0**	**66.7**	**59.4**
VL^{wo}	97.1	95.7	97.5	89.0	91.6	96.2
VL^{w}	**98.3**	**98.5**	**98.0**	**91.3**	**96.3**	**98.0**

Table 4: Ablation study of masked modality modeling. X^{w} or X^{wo} means corresponding accuracy of X trained with or without $\mathbf{F}_m^v$ and $\mathbf{F}_m^l$.

of V and L with MMM all outperform the accuracy without MMM. Besides, the 2.7% accuracy of V^{wo} and 0.4% accuracy of L^{wo} indicate that the capacity of merely utilizing unimodality information to predict is destroyed. 2) The 1.6% accuracy growth of VL indicates that the better understanding unimodality assists model combining information from different modality. 3) The improved performance of 0.6% and 35.2% of VL^{w} compared to V^{w} and L^{w} further demonstrates that learning to effectively combine information from vision-language to predict more accurate results by MMM.

To demonstrate unimodality masking effectiveness to the final accuracy, we compare the VL under different masking settings. From the Table 5, the results are improved 0.6% or 0.7% when we apply $\mathbf{F}_m^v$ or $\mathbf{F}_m^l$. The nearly same promotion further demonstrates that the balanced function of vision and language. Besides, the performance with both modality masking is further improved 1.6%.

The effectiveness of Synchronized VLRM For global linguistic reasoning, a crucial operation is masking the corresponding position character in reading order. As shown in Table 6, the accuracies of V^{wo}, L^{wo} and VL^{wo} are nearly the same. Besides, not utilizing the attention mask respectively results in decreased accuracy 1.2% and 1.6% for V and VL and a big margin improvement between L^{wo} and L^{w} occurs. The above observations suggest that VLRM degrades to a pure vision reasoning module and makes no difference between V and VL without the attention mask $\mathbf{M}_a$. The accuracy drop of V and VL also demonstrates the importance of a decoder possessing semantic reasoning ability. Meanwhile, the nearly same accuracy between L^{wo}, V^{wo} and VL^{wo} indicates that VLRM learns nothing meaningful for semantic reasoning but merely outputs the same as V.

To gain a deeper understanding of the VLRM, a visualization of the attention weights in Figure 8 is presented by using the word "THEATRE" as an example. For instance, the sec-

$\mathbf{F}_m^v$	$\mathbf{F}_m^l$	IC13	SVT	IIIT5k	IC15	SVTP	CUTE
-	-	97.1	95.7	97.5	89.0	91.6	96.2
✓	-	97.6	96.6	97.5	90.3	92.8	96.2
-	✓	97.7	97.1	97.6	90.0	94.1	96.9
✓	✓	**98.3**	**98.5**	**98.0**	**91.3**	**96.3**	**98.0**

Table 5: Ablation study of each modality masking influence to VL prediction accuracy.

Outputs	IC13	SVT	IIIT5k	IC15	SVTP	CUTE
V^{wo}	96.4	96.6	97.2	88.7	91.3	96.5
V^{w}	**97.9**	**97.5**	**97.7**	**90.1**	**94.9**	**97.9**
L^{wo}	**96.4**	**96.6**	**97.4**	**89.0**	**91.5**	**96.5**
L^{w}	70.3	68.6	57.0	59.0	66.7	**59.4**
VL^{wo}	96.4	96.6	97.4	89.0	91.6	96.5
VL^{w}	**98.3**	**98.5**	**98.0**	**91.3**	**96.3**	**98.0**

Table 6: Ablation study of attention mask of VLRM. X^{w} or X^{wo} means corresponding accuracy of X with or without the attention mask $\mathbf{M}_a$.

ond position query "[2]" concentrates on the corresponding 'H' visual area, as well as the other characters 'T', 'E', 'A', 'T', 'R', and 'E'. This dual focus allows VLRM to leverage linguistic information to recognize noisy visual areas, while also guides the language reasoning procedure through visual clues. This mutual aid reasoning approach empowers VLRM with robust capabilities.

Besides, we compare the results obtained by implementing with varying numbers of VLRM blocks. As shown in Table 7, VLRM implemented with 6 units which has the stronger reasoning capability but at cost of parameter and speed. Furthermore, the performance begins to plateau when the number of blocks is increased to 8.

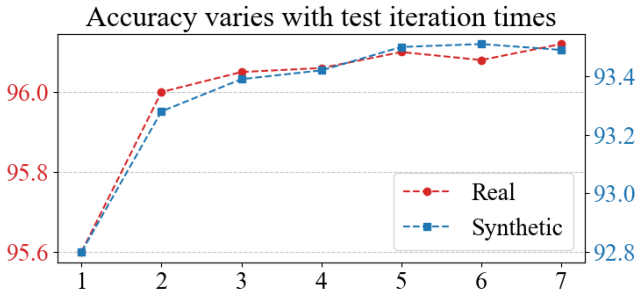

Figure 7: Accuracy of iterative reasoning times under real and synthetic training data.

The effectiveness of Iterative Reasoning Iterative reasoning is applied in the testing phase. As shown in Figure 7, real and synthetic datasets trainings both benefit from iterative reasoning. Besides, 2 iterations bring the most improvement and then model enters saturation after 3 or 4 times.

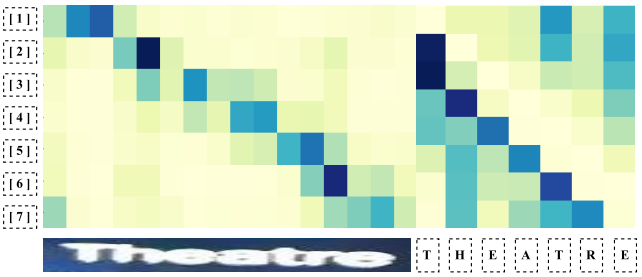

Figure 8: Examples of attention map in VLRM.

Methods	IC13 IC15	SVT SVTP	IIIT5k CUTE	Avg	Param ($\times 10^6$)	Time (ms)
2-VLRM	97.3 90.0	97.2 94.6	97.6 97.2	95.4	44.2	18.4
4-VLRM	97.9 91.1	97.8 95.7	97.7 97.2	95.9	50.5	20.3
6-VLRM	98.3 91.3	98.5 96.3	98.0 98.0	96.2	56.8	23.6
8-VLRM	98.1 91.6	97.8 96.1	97.9 98.3	96.2	63.1	25.8

Table 7: Ablation study of VLRM configuration. "N-VLRM" means the VLRM has N number of blocks.

Qualitative Error Analysis

These wrongly predicted images of IIIT5k in Figure 9 are basically clear and easy to recognize. However, characters with superscripts may be omitted by the BUSNet as the network considers them as meaningless symbols. Besides, these words are mostly not belong to English words, so the semantic reasoning capacity is also limited. However, the visual clues reasoning ability is also why BUSNet shows superiority on confusing benchmarks like IC15 and SVTP.

Figure 9: Examples of recognition errors of BUSNet.

Conclusion

In this paper, we propose BUSNet which explores approach for effectively mitigating unimodally over-reliant issue in scene text recognition. By treating image as a language, the unbalanced issue of vision and language is alleviated. Besides, an ensemble of unified internal and external vision-language model is learned by MMM. Furthermore, we propose a novel VLRM with synchronized vision-language decoding capacity. Additionally, we can take an advantage of the iterative reasoning for more accurate prediction. Experiment results on mainstream benchmarks demonstrate the superiority of BUSNet.

Acknowledgments

This work was jointly supported by the National Key Research and Development Program of China under Grant No.2020AAA0107903, the National Natural Science Foundation of China under Grant No.62176091.

References

Aberdam, A.; Litman, R.; Tsiper, S.; Anschel, O.; Slossberg, R.; Mazor, S.; Manmatha, R.; and Perona, P. 2021. Sequence-to-sequence contrastive learning for text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15302–15312.

Atienza, R. 2021. Vision transformer for fast and efficient scene text recognition. In *International Conference on Document Analysis and Recognition*, 319–334. Springer.

Baek, J.; Kim, G.; Lee, J.; Park, S.; Han, D.; Yun, S.; Oh, S. J.; and Lee, H. 2019. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4715–4723.

Bautista, D.; and Atienza, R. 2022. Scene Text Recognition with Permuted Autoregressive Sequence Models. In *European Conference on Computer Vision*, 178–196. Springer.

Cheng, Z.; Bai, F.; Xu, Y.; Zheng, G.; Pu, S.; and Zhou, S. 2017. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE international conference on computer vision*, 5076–5084.

Chng, C. K.; Liu, Y.; Sun, Y.; Ng, C. C.; Luo, C.; Ni, Z.; Fang, C.; Zhang, S.; Han, J.; Ding, E.; et al. 2019. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1571–1576. IEEE.

Da, C.; Wang, P.; and Yao, C. 2022. Levenshtein OCR. In *European Conference on Computer Vision*, 322–338. Springer.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, Y.; Chen, Z.; Jia, C.; Yin, X.; Zheng, T.; Li, C.; Du, Y.; and Jiang, Y.-G. 2022. Svtr: Scene text recognition with a single visual model. *arXiv preprint arXiv:2205.00159*.

Fang, S.; Xie, H.; Wang, Y.; Mao, Z.; and Zhang, Y. 2021. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7098–7107.

Gupta, A.; Vedaldi, A.; and Zisserman, A. 2016. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2315–2324.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, P.; Huang, W.; Qiao, Y.; Loy, C. C.; and Tang, X. 2016b. Reading scene text in deep convolutional sequences. In *Thirtieth AAAI conference on artificial intelligence*.

Jaderberg, M.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*.

Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V. R.; Lu, S.; et al. 2015. ICDAR 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, 1156–1160. IEEE.

Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Bigorda, L. G.; Mestre, S. R.; Mas, J.; Mota, D. F.; Almazan, J. A.; and De Las Heras, L. P. 2013. ICDAR 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, 1484–1493. IEEE.

Li, H.; Wang, P.; Shen, C.; and Zhang, G. 2019. Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8610–8617.

Liao, M.; Zhang, J.; Wan, Z.; Xie, F.; Liang, J.; Lyu, P.; Yao, C.; and Bai, X. 2019. Scene text recognition from two-dimensional perspective. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8714–8721.

Liu, W.; Chen, C.; Wong, K.-Y. K.; Su, Z.; and Han, J. 2016. Star-net: a spatial attention residue network for scene text recognition. In *BMVC*, volume 2, 7.

Mishra, A.; Alahari, K.; and Jawahar, C. 2012. Top-down and bottom-up cues for scene text recognition. In *2012 IEEE conference on computer vision and pattern recognition*, 2687–2694. IEEE.

Na, B.; Kim, Y.; and Park, S. 2022. Multi-modal text recognition networks: Interactive enhancements between visual and semantic features. In *European Conference on Computer Vision*, 446–463. Springer.

Phan, T. Q.; Shivakumara, P.; Tian, S.; and Tan, C. L. 2013. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, 569–576.

Risnumawan, A.; Shivakumara, P.; Chan, C. S.; and Tan, C. L. 2014. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18): 8027–8048.

Su, B.; and Lu, S. 2017. Accurate recognition of words in scenes without character segmentation using recurrent neural network. *Pattern Recognition*, 63: 397–405.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wan, Z.; He, M.; Chen, H.; Bai, X.; and Yao, C. 2020. Textscanner: Reading characters in order for robust scene text recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12120–12127.

Wang, K.; Babenko, B.; and Belongie, S. 2011. End-to-end scene text recognition. In *2011 International conference on computer vision*, 1457–1464. IEEE.

Wang, Y.; Xie, H.; Fang, S.; Wang, J.; Zhu, S.; and Zhang, Y. 2021. From two to one: A new scene text recognizer with visual language modeling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14194–14203.

Yan, R.; Peng, L.; Xiao, S.; and Yao, G. 2021. Primitive representation learning for scene text recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 284–293.

Yu, D.; Li, X.; Zhang, C.; Liu, T.; Han, J.; Liu, J.; and Ding, E. 2020. Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12113–12122.

Yue, X.; Kuang, Z.; Lin, C.; Sun, H.; and Zhang, W. 2020. Robustscanner: Dynamically enhancing positional clues for robust text recognition. In *European Conference on Computer Vision*, 135–151. Springer.

Zhang, Y.; Gueguen, L.; Zharkov, I.; Zhang, P.; Seifert, K.; and Kadlec, B. 2017. Uber-text: A large-scale dataset for optical character recognition from street-level imagery. In *SUNw: Scene Understanding Workshop-CVPR*, volume 2017, 5.

Zhong, D.; Lyu, S.; Shivakumara, P.; Yin, B.; Wu, J.; Pal, U.; and Lu, Y. 2022. SGBANet: semantic GAN and balanced attention network for arbitrarily oriented scene text recognition. In *European conference on computer vision*, 464–480. Springer.