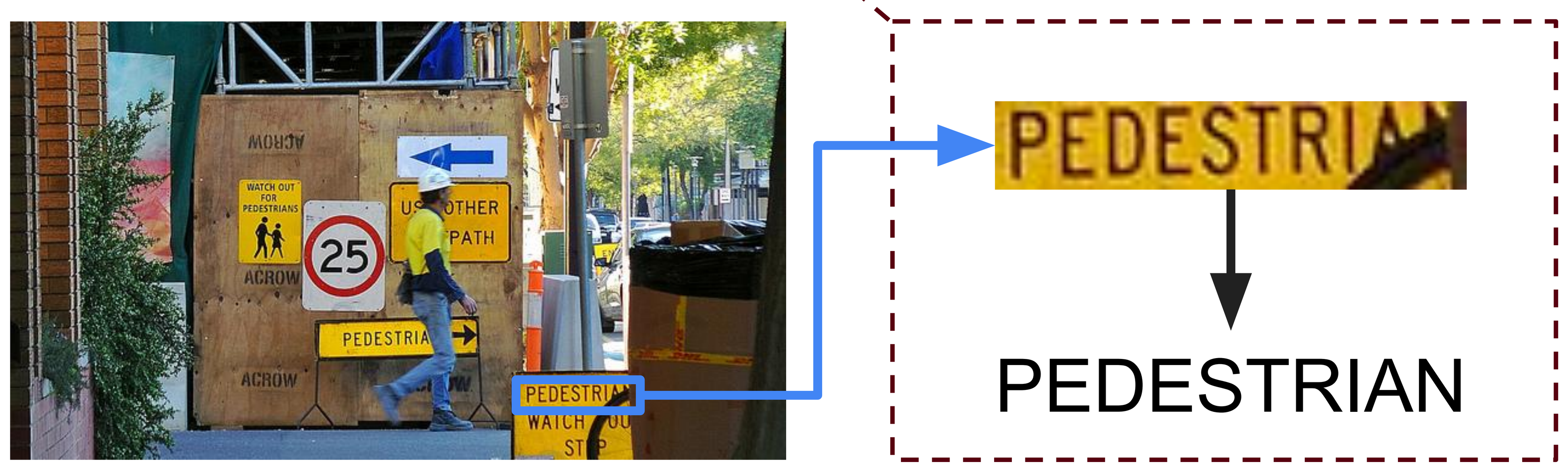
Scene Text Recognition with Permuted Autoregressive Sequence Models

Darwin Bautista and Rowel Atienza
University of the Philippines

gh:baudm/parseq

ECCV TEL AVIV 2022

Problem — Text recognition in natural scenes

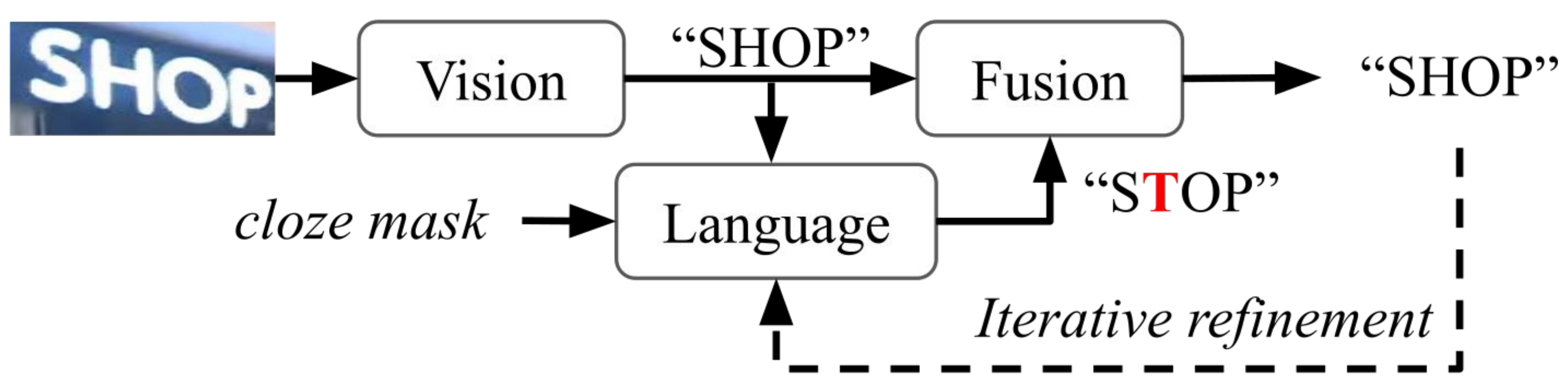


Language context can resolve visual ambiguity

But context-aware STR methods are typically limited to:

- Monotonic decoding (one character at a time)
- Unidirectional context (usually left-to-right)

Recent two-stage approaches are inefficient



ABINet (Fang *et al.* CVPR'21)

Language Model accounts for 35% of the parameters but is:

- Underutilized (uses only 13.65% of total FLOPS)
- Error-prone due to lack of visual context (50.44% word accuracy with ground truth label as input)

Image is the primary input signal in STR, not language context

→ The Language Model should also **consider visual context**

Key Idea — Unify models using AR ensemble

Ensemble of AR models (PARSeq model)

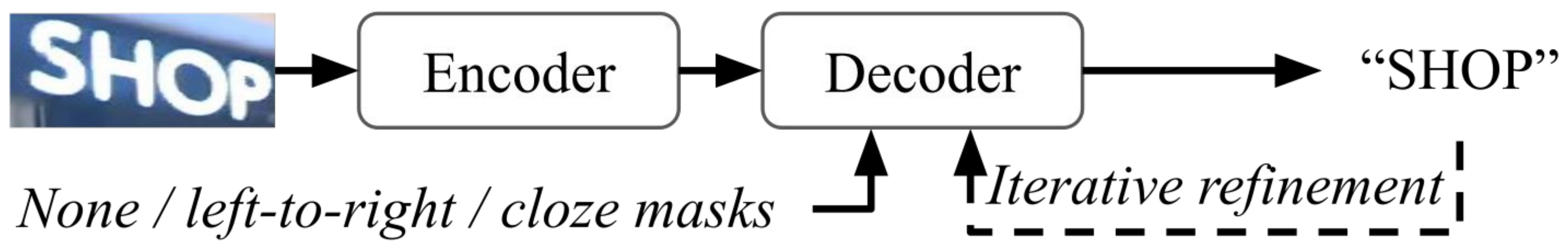
$$P(\mathbf{y}|\mathbf{x})_{[1,2,3]} = P(y_1|\mathbf{x})P(y_2|y_1,\mathbf{x})\boxed{P(y_3|y_1,y_2,\mathbf{x})}$$
$$P(\mathbf{y}|\mathbf{x})_{[3,2,1]} = \boxed{P(y_3|\mathbf{x})}P(y_2|y_3,\mathbf{x})\boxed{P(y_1|y_2,y_3,\mathbf{x})}$$
$$P(\mathbf{y}|\mathbf{x})_{[1,3,2]} = \boxed{P(y_1|\mathbf{x})}P(y_3|y_1,\mathbf{x})\boxed{P(y_2|y_1,y_3,\mathbf{x})}$$
$$\boxed{P(\mathbf{y}|\mathbf{x})_{[2,3,1]}} = \boxed{P(y_2|\mathbf{x})}P(y_3|y_2,\mathbf{x})P(y_1|y_2,y_3,\mathbf{x})$$

$$\prod_{t=1}^{T} P(y_t|\mathbf{y}_{<t},\mathbf{x})$$
Context-aware AR model

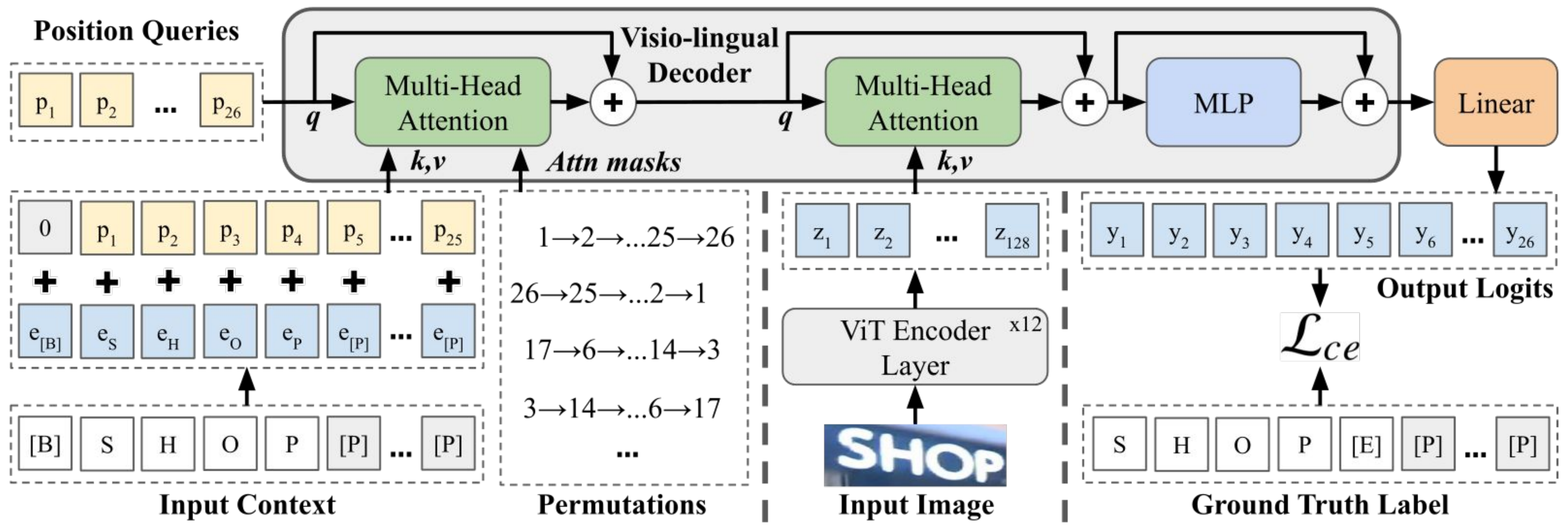
$$\prod_{t=1}^{T} P(y_t|\mathbf{x})$$
Context-free NAR model

$$\prod_{t=1}^{T} P(y_t|\mathbf{y}_{\neq t},\mathbf{x})$$
Bidirectional Iterative Refinement model

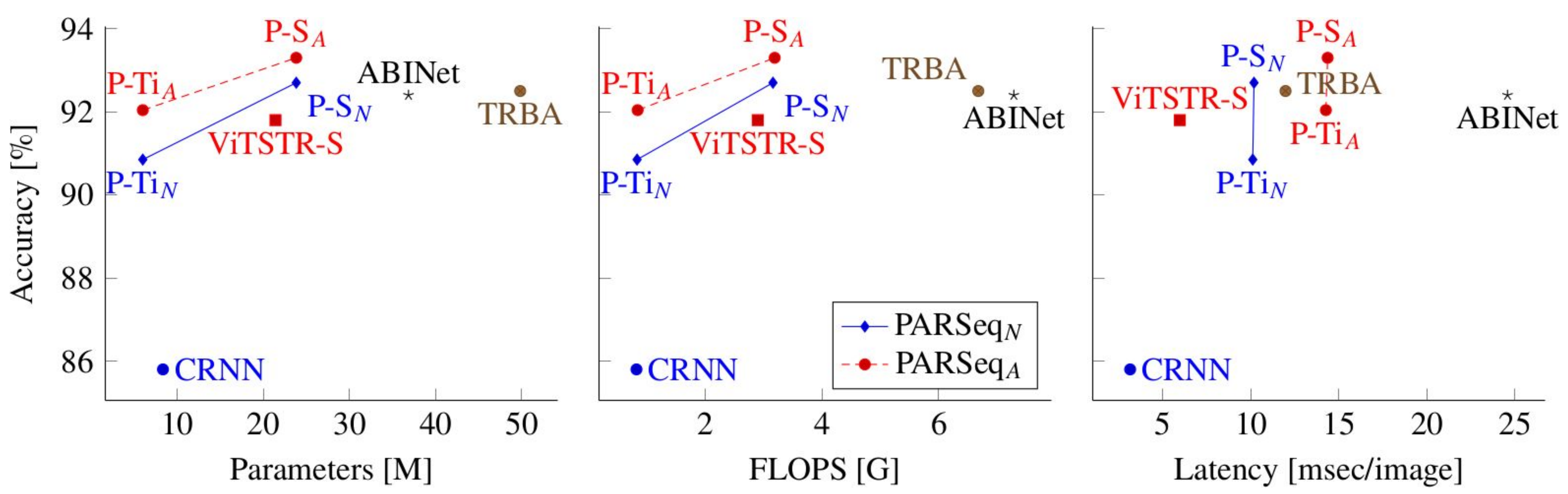
PARSeq: A unified model for STR



Learning PARSeq with Permutation Language Modeling:



PARSeq is flexible, accurate, and efficient

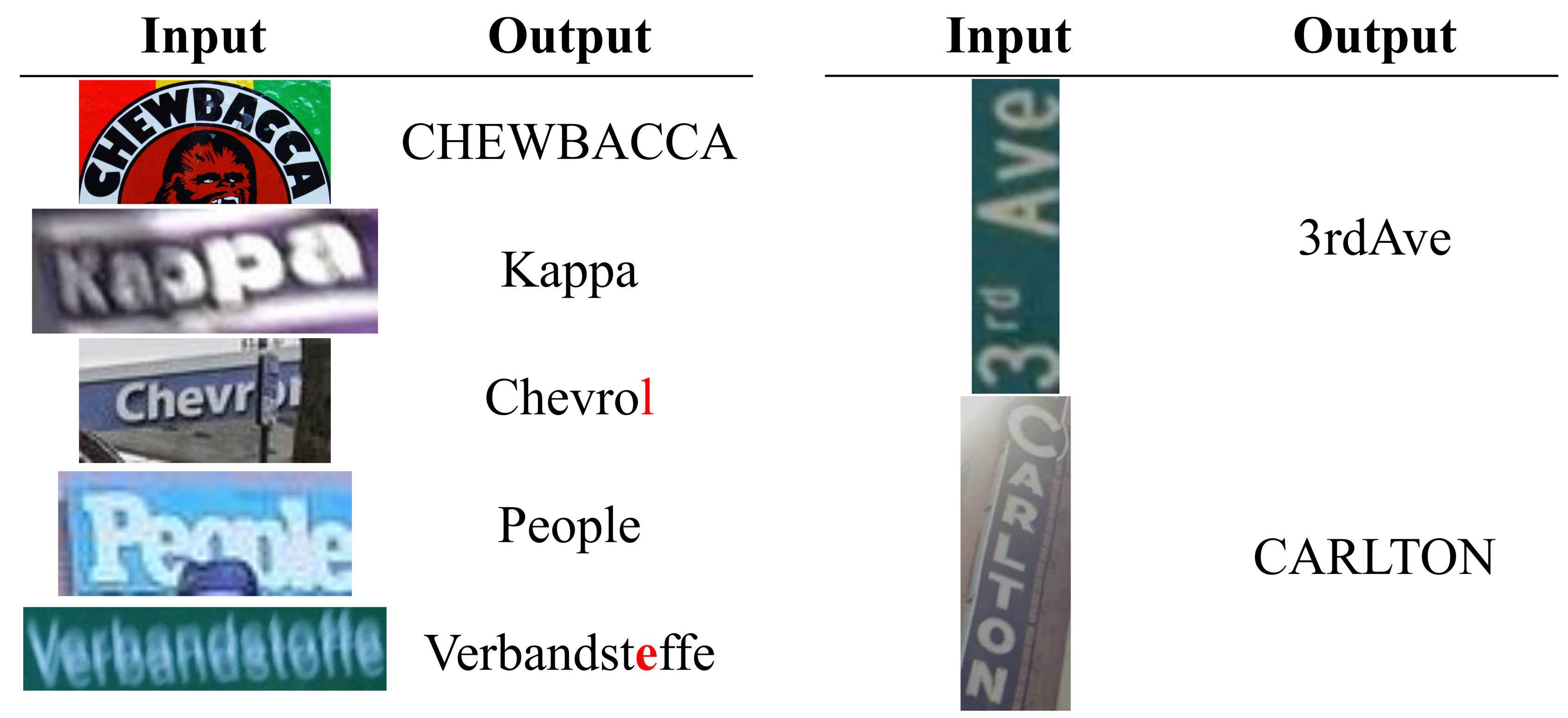


- PARSeq_N for constant-time inference (non-autoregressive)
- PARSeq_A for more accurate results (autoregressive)

Results — SOTA in STR benchmarks

	Method	Conf.	Train data	36-char word acc. 7,248 samples	7,672 samples
Previous Work	SRN	CVPR'20	MJ,ST	90.4	–
	TextScanner	AAAI'20	MJ,ST+	–	91.0
	Bhunia *et al.*	ICCV'21	MJ,ST	–	90.9
	VisionLAN	ICCV'21	MJ,ST	91.2	–
	PREN2D	CVPR'21	MJ,ST	91.5	–
	ABINet	CVPR'21	MJ,ST+	92.7	–
Ours	PARSeq_N		MJ,ST	92.0±0.2	90.7±0.2
	PARSeq_A		MJ,ST	**93.2**±0.2	**91.9**±0.2
	PARSeq_N		*real*	95.7±0.1	95.2±0.1
	PARSeq_A		*real*	**96.4**±0.0	**96.0**±0.0

Robust vs occlusion and arbitrary orientation



Wider gap in more challenging datasets

Method	Train data	36-char word accuracy per dataset: ArT 35,149	COCO 9,825	Uber 80,551	Total 125,525
ViTSTR-S	*real*	81.1±0.1	74.1±0.4	78.2±0.1	78.7±0.1
TRBA	*real*	82.5±0.2	77.5±0.2	81.2±0.3	81.3±0.2
ABINet	*real*	81.2±0.1	76.4±0.1	71.5±0.7	74.6±0.4
PARSeq_N	*real*	83.0±0.2	77.0±0.2	82.4±0.3	82.1±0.2
PARSeq_A	*real*	**84.5**±0.1	**79.8**±0.1	**84.5**±0.1	**84.1**±0.0