2022/05/06

Vietnamese OCR: Research and Applications

Nguyen Thanh Hung[1], Dinh Viet Sang[1], Nguyen Phi Le[1], Han-Cheol[2]
[1]School of Information and Communication Technology, HUST
[2]NAVER corporation

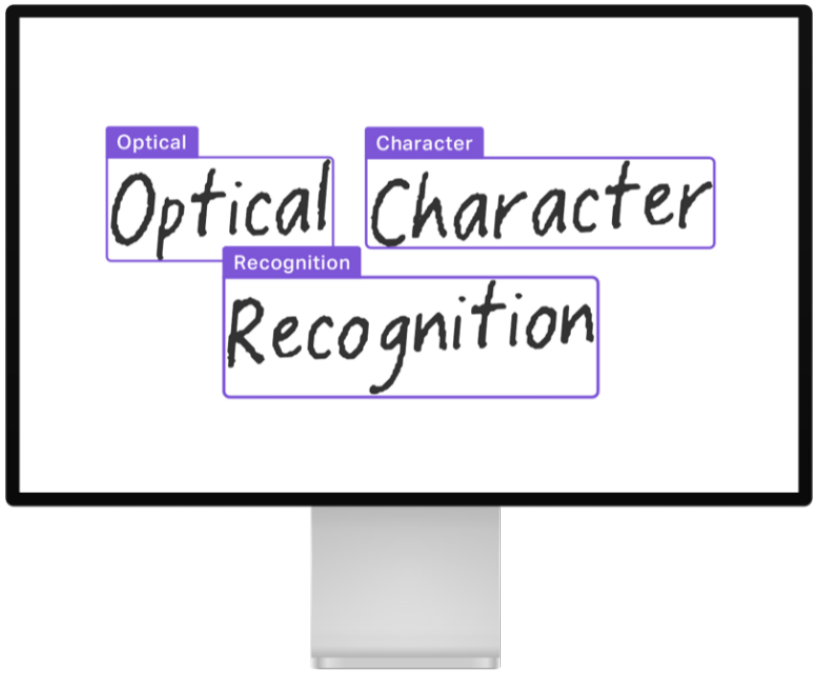
SOICT BKAI NAVER

Agenda

- Motivation
- Our goals
- Our achievements
- Conclusion and Future works

Motivation

- Lack of Vietnamese OCR dataset
- Lack of research on Vietnamese OCR



Credit: https://medium.com/

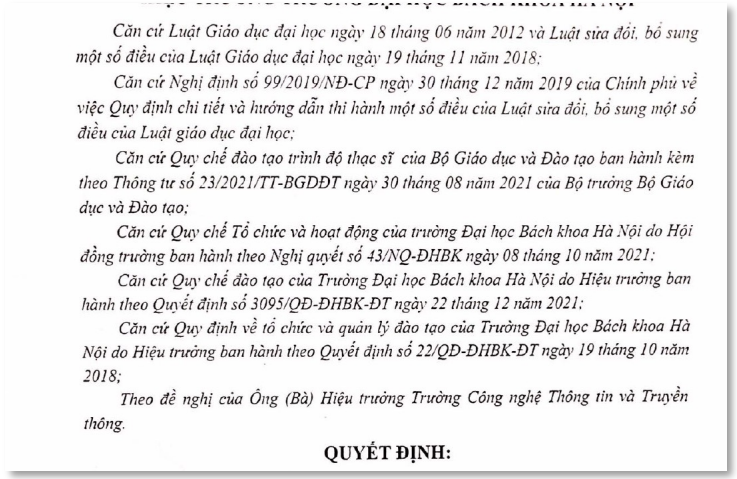
Our goals

- Building Vietnamese OCR dataset
- Studying the performance of SOTA OCR models in Vietnamese
  - Enhancing SOTA OCR models for Vietnamese
- Building a comprehensive Vietnamese OCR system

Building Vietnamese OCR dataset

- Objectives
  - Types of texts
    - Scene text
    - Scanned text
  - Types of data
    - Annotated real data
    - Synthetic data
  - Types of contributions
    - Dataset
    - Labelling tools
    - Synthetic generation library

Scanned text



Căn cứ Luật Giáo dục đại học ngày 18 tháng 06 năm 2012 và Luật sửa đổi, bổ sung một số điều của Luật Giáo dục đại học ngày 19 tháng 11 năm 2018;

Căn cứ Nghị định số 99/2019/NĐ-CP ngày 30 tháng 12 năm 2019 của Chính phủ về việc Quy định chi tiết và hướng dẫn thi hành một số điều của Luật sửa đổi, bổ sung một số điều của Luật giáo dục đại học;

Căn cứ Quy chế đào tạo trình độ thạc sĩ của Bộ Giáo dục và Đào tạo ban hành kèm theo Thông tư số 23/2021/TT-BGDĐT ngày 30 tháng 08 năm 2021 của Bộ trưởng Bộ Giáo dục và Đào tạo;

Căn cứ Quy chế Tổ chức và hoạt động của trường Đại học Bách khoa Hà Nội do Hội đồng trường ban hành theo Nghị quyết số 43/NQ-ĐHBK ngày 08 tháng 10 năm 2021;

Căn cứ Quy chế đào tạo của Trường Đại học Bách khoa Hà Nội do Hiệu trưởng ban hành theo Quyết định số 3095/QĐ-ĐHBK-ĐT ngày 22 tháng 12 năm 2021;

Căn cứ Quy định về tổ chức và quản lý đào tạo của Trường Đại học Bách khoa Hà Nội do Hiệu trưởng ban hành theo Quyết định số 22/QĐ-ĐHBK-ĐT ngày 19 tháng 10 năm 2018;

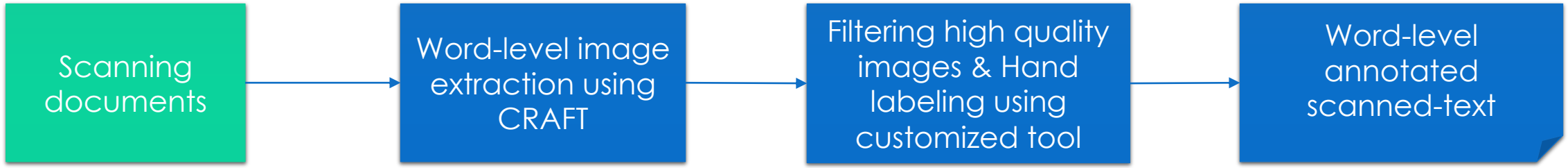
Theo đề nghị của Ông (Bà) Hiệu trưởng Trường Công nghệ Thông tin và Truyền thông.

QUYẾT ĐỊNH:

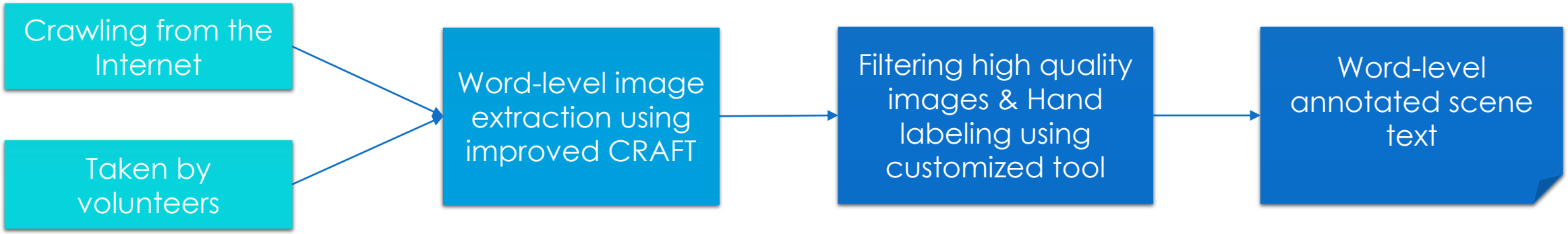
Scene text

Building Vietnamese OCR dataset

- Real data collection processes



For scanned text



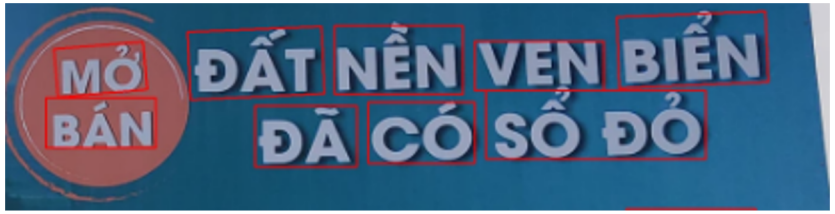
For scene text

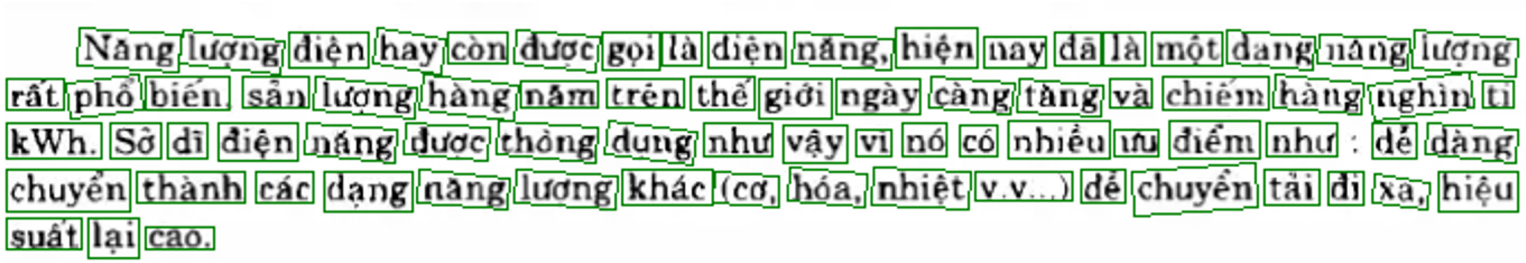
Building Vietnamese OCR dataset

- Real data collection
  - For scanned text
    - Scanning books in various domains
      - 1 philosophy book: 250 pages ~ 90k words
      - 1 novel: 250 pages ~ 90k words
      - Other documents: 200 pages ~ 70k words
  - For scene text
    - Crawling from Internet + real capture
      - Source code: https://github.com/AIoT-Lab-BKAI/ggcrawl
      - Keywords: 62 keywords
        - https://github.com/AIoT-Lab-BKAI/ggcrawl/blob/master/keyword_list.txt
      - Results
        - 4k images
        - Image resolution with both dimensions larger than 100px

Building Vietnamese OCR dataset

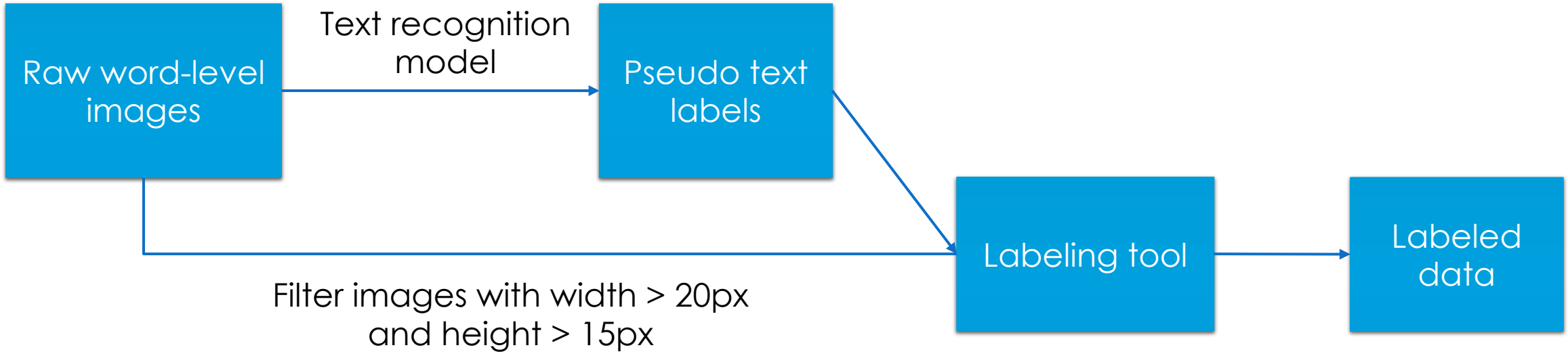
- Real data collection
  - Using CRAFT model (by NAVER) to crop single-word texts





Building Vietnamese OCR dataset

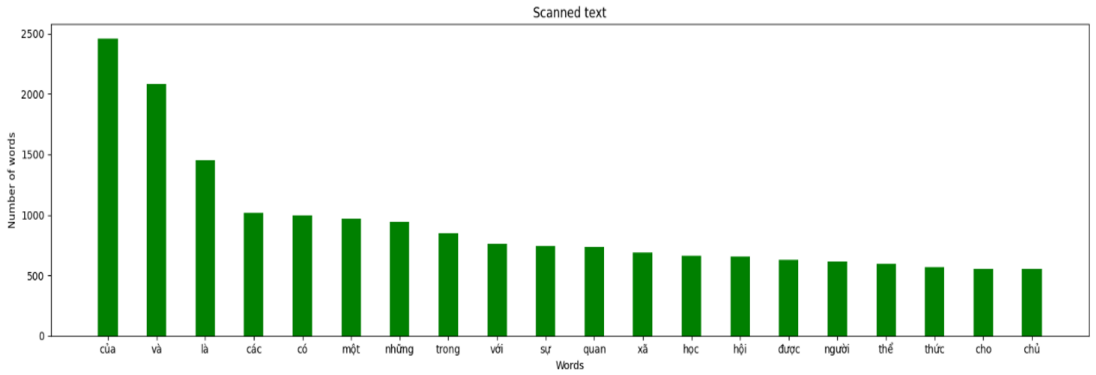
- Real data collection
  - semi-automatically labelling the collected data



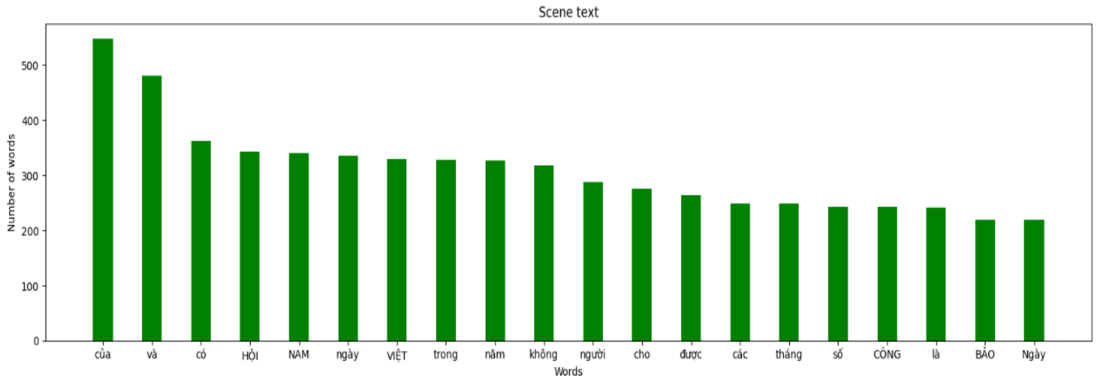
Building Vietnamese OCR dataset

- Real data collection

	Scanned-text images	Scene-text images
Raw data	217k images	140k images
Labeled data	100k images	100k images



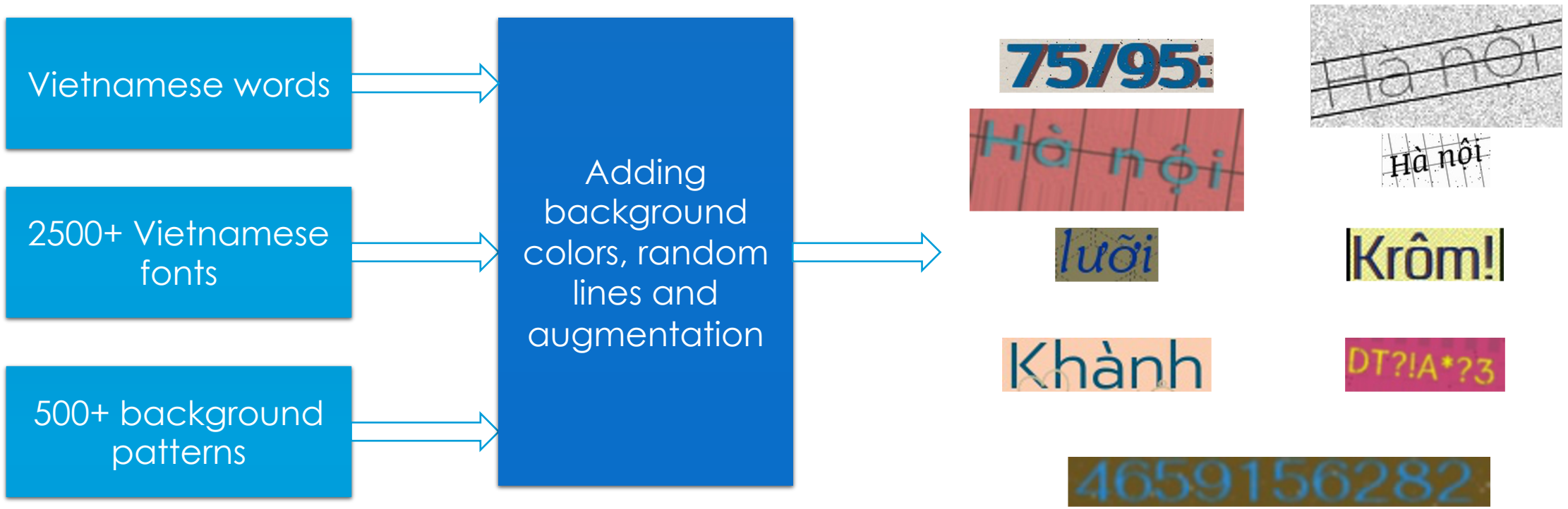
The most common scanned texts



The most common scene texts

Building Vietnamese OCR dataset

- Synthetic data generation



Building Vietnamese OCR dataset

- Synthetic data generation
  - Corpus: Vietnamese wikipedia corpus (an open-source, maintained by Vietnamese NLP Research Group) (https://github.com/undertheseanlp/corpus.viwiki).
  - Splitting the corpus into unigrams, bigrams, and trigrams
  - 2500+ fonts, 500+ background patterns and many augmentation methods
    - Each word can be generated over 1M styles.

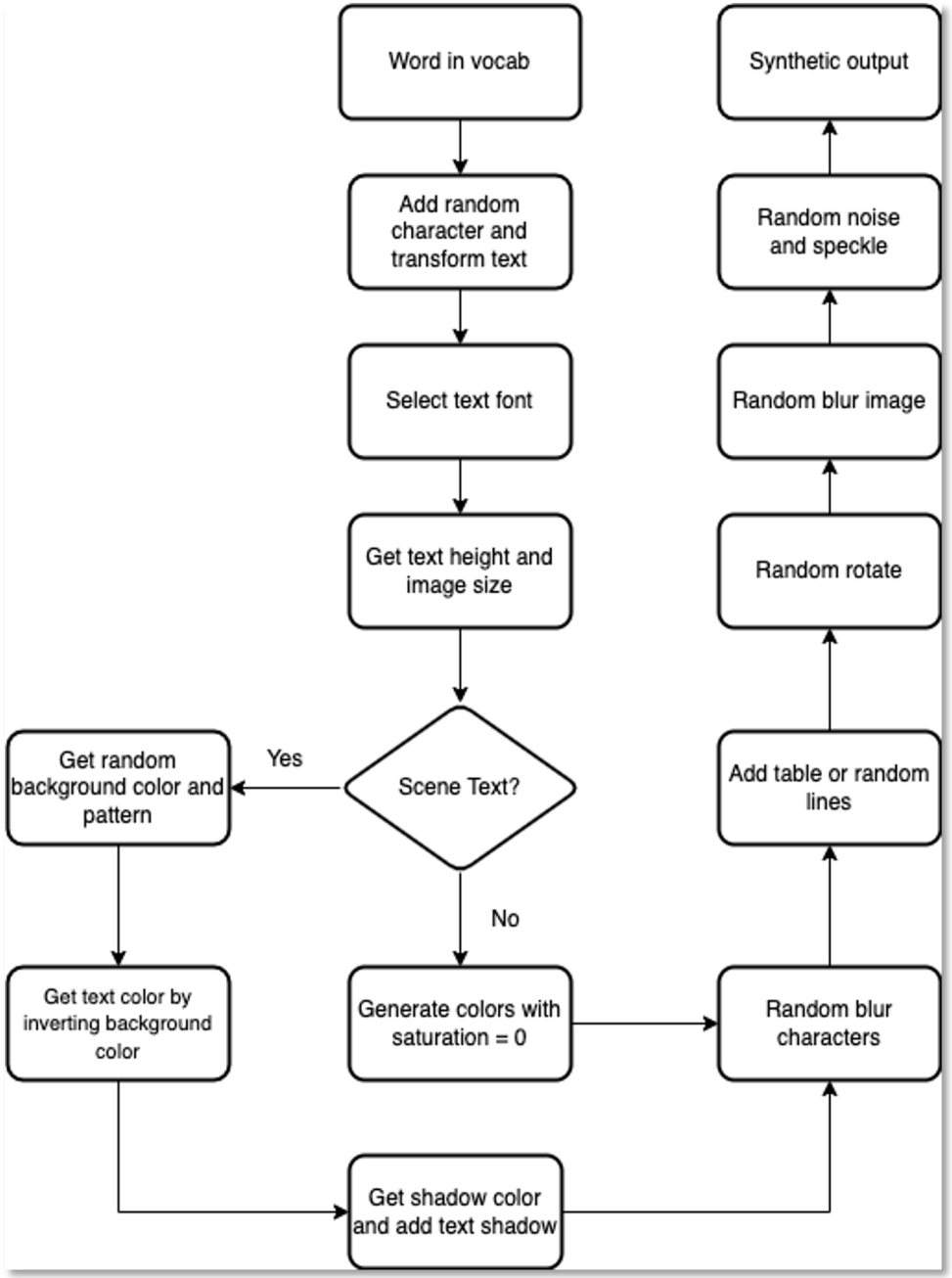
4.232.180
unigram words

2.875.307
bigram words

7.312.079
trigram word

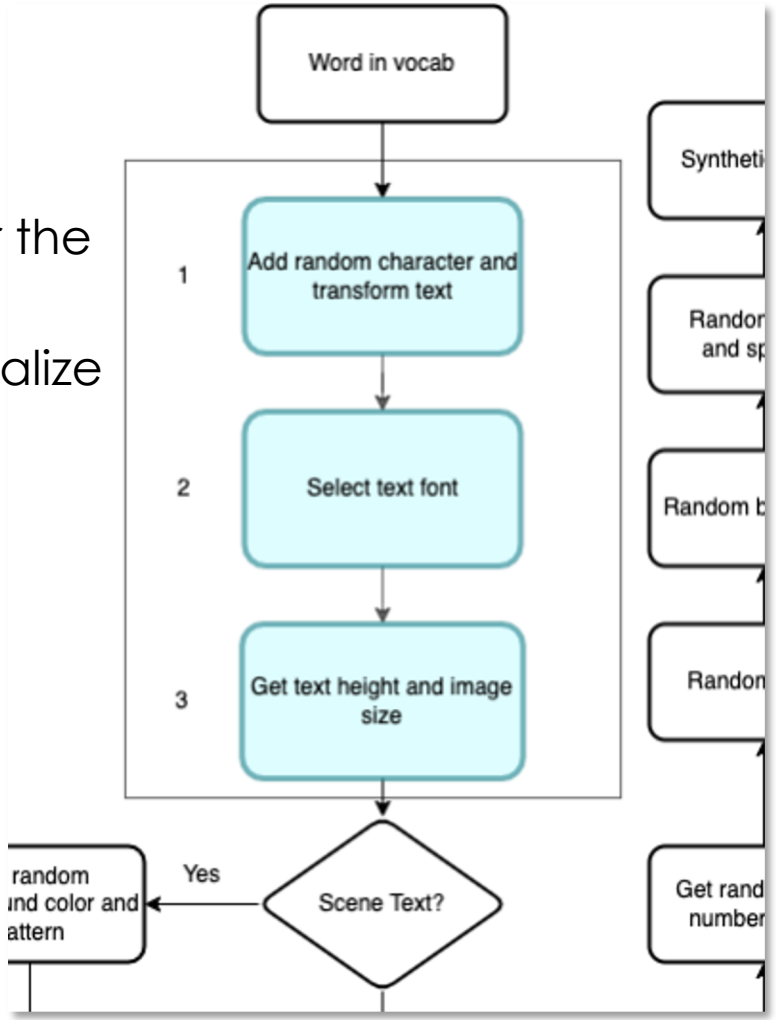
Building Vietnamese OCR dataset

- Synthetic data generation process



Building Vietnamese OCR dataset

- Synthetic data generation process
  - In the first step:
    - add random punctuations and digits into the beginning or the end of the word with a probability of 20%
    - randomly transform text to uppercase, lowercase, or capitalize
  - In the next steps:
    - choose a random font from the font set
    - randomly select a text size in the range of 30-56 pt
    - perform padding to a random background size.

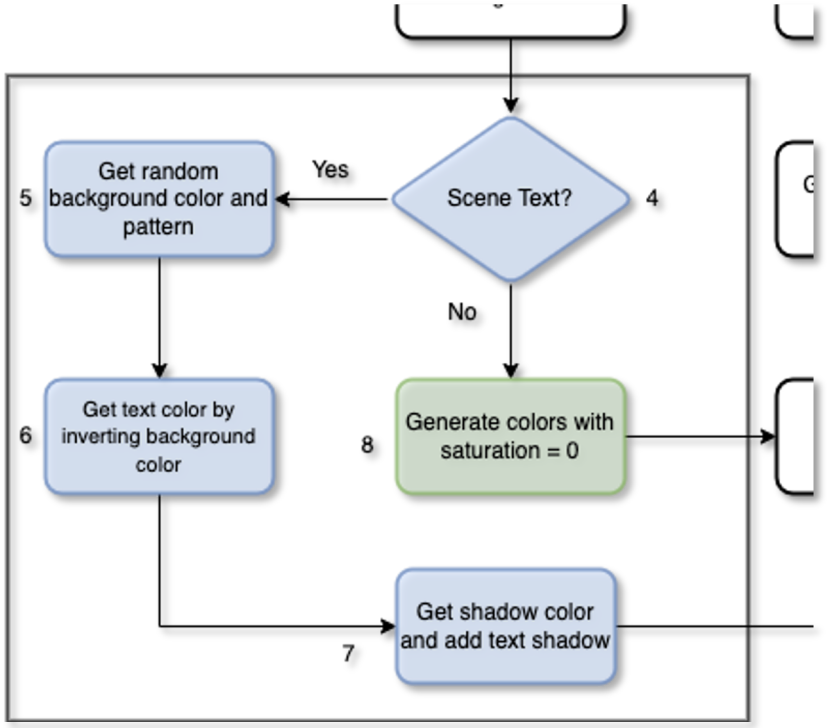


Building Vietnamese OCR dataset

- Synthetic data generation process
  - For scene text image
    - randomly choose a background color and pattern
    - get text color by inverting median background color (always add random factors).
    - add shadow to text
  - For scanned text image:
    - choose a grayscale color for text and background

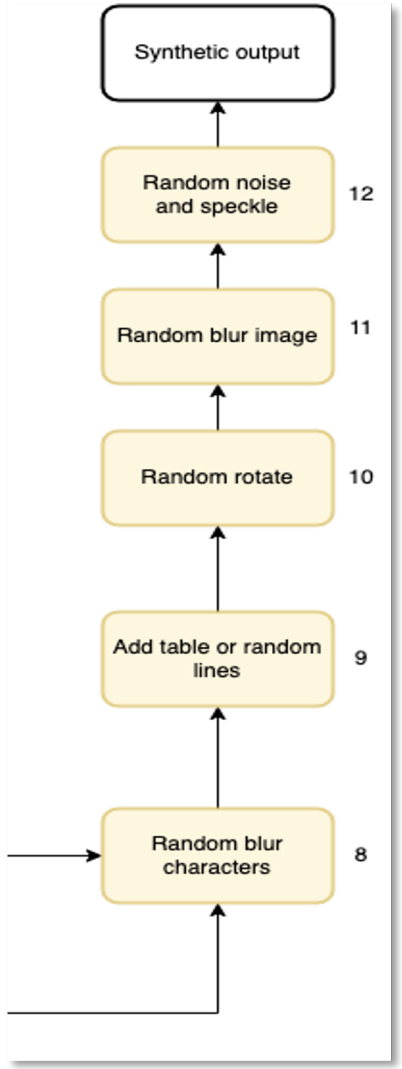
an example of generated Scene Text

an example of generated Scanned Text



Building Vietnamese OCR dataset

- Synthetic data generation process
  - randomly blur 0-2 characters
    - Guaranteeing the readability
  - add table or lines
    - The probability for generating horizontal lines > probability for generating vertical lines > probability for generating diagonal lines
  - apply image distortion such as: rotating, adding noise, adding speckle, skewing
    - Guaranteeing the readability



Scene text with vertical lines

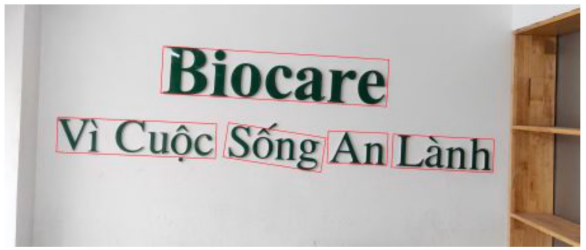
Scanned text with horizontal lines

Building Vietnamese OCR dataset

- Generated synthetic data
  - Scan text synthetic images:
    - Unigram: 250M
    - Bigram: 150M
    - Trigram: 100M
  - Scene text synthetic images:
    - Unigram: 250M
    - Bigram: 150M
    - Trigram: 100M

SOTA OCR models in Vietnamese

- Targeted models
  - Text localization
    - CRAFT (by NAVER)
  - Text recognition
    - VGG, Transformer, Seq2Seq
- Findings
  - Difficulty in capturing punctuations
    - The models have not been trained with Vietnamese
    - Punctuations are separated from other characters
- Solutions
  - Generating Vietnamese synthetic data with punctuations and transfer learning CRAFT
  - Customizing CRAFT's weakly supervised learning algorithm



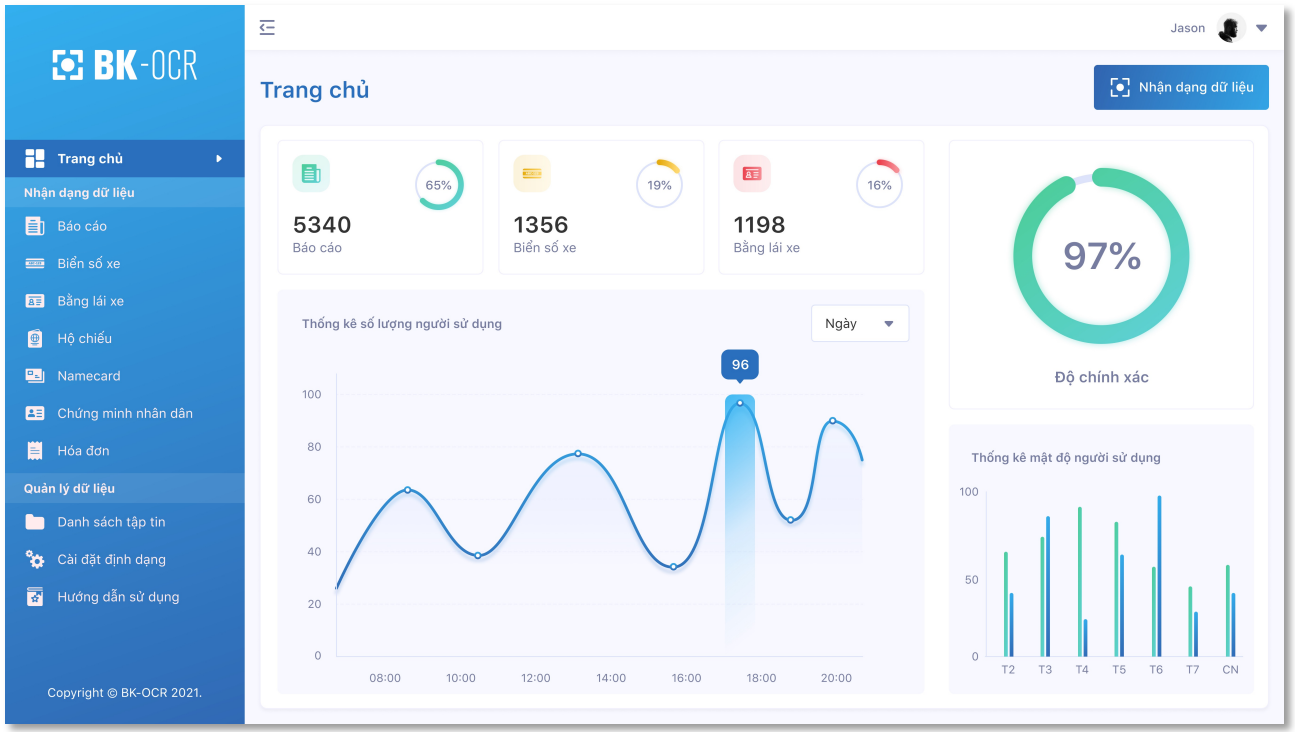
SOTA OCR models in Vietnamese

- Enhanced results

Type	Pretrained models		Ours (Transfer-learning)	
	VGG-Transformer	VGG-Seq2seq	VGG-Transformer	VGG-Seq2seq
Scan text (32047 samples)	92.07%	88.28%	**97.20%** (+5,13%)	**97.28%** (+9%)
Scene text (29999 samples)	82.99%	83.28%	**92.87%** (+9,88%)	**92.57%** (+9,27%)

BK-OCR:
a comprehensive Vietnamese OCR system

- Text localization and recognition, especially for Vietnamese
- Build-in functions
  - Template recognition
  - Information extraction
  - Template definition
- URL
  - https://bkai.ai/solution/ocr/

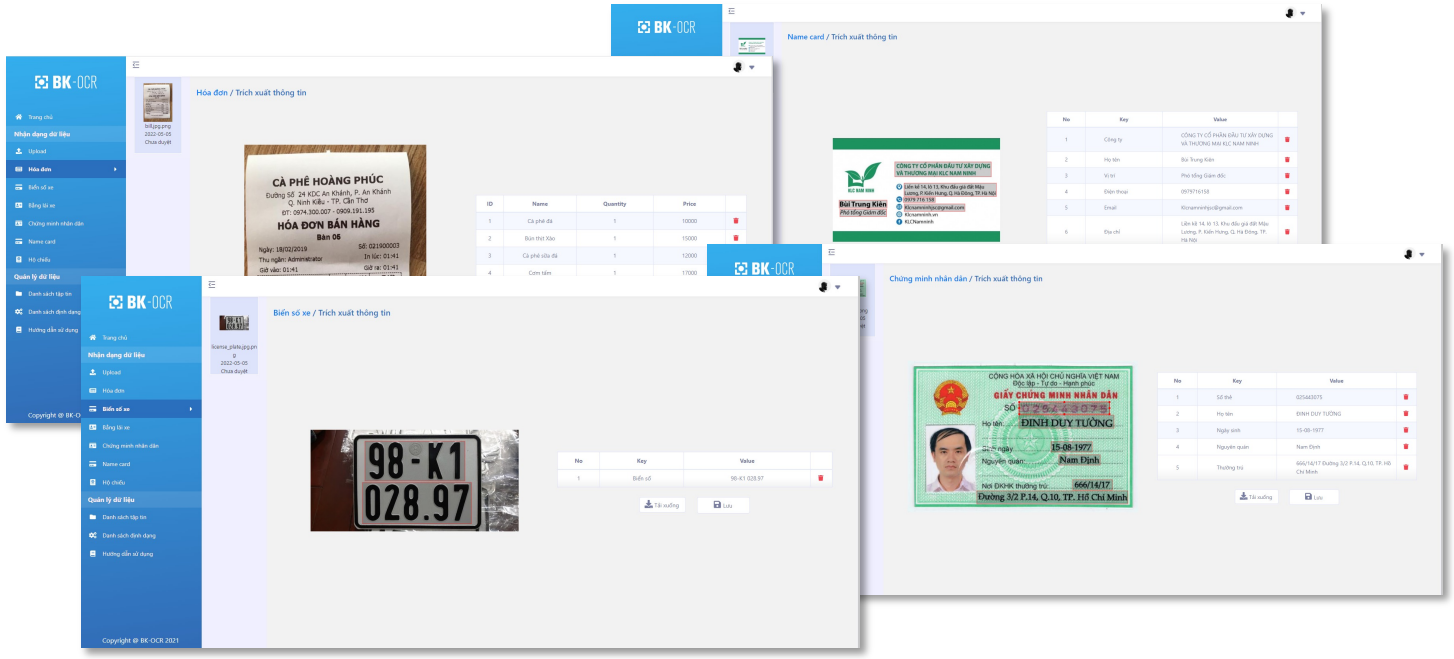


BK-OCR:
a comprehensive Vietnamese OCR system

- Text localization and recognition, especially for Vietnamese
- Build-in functions
  - Template recognition
    - Automatically recognize the structure of uploaded images
  - Information extraction
  - Template definition

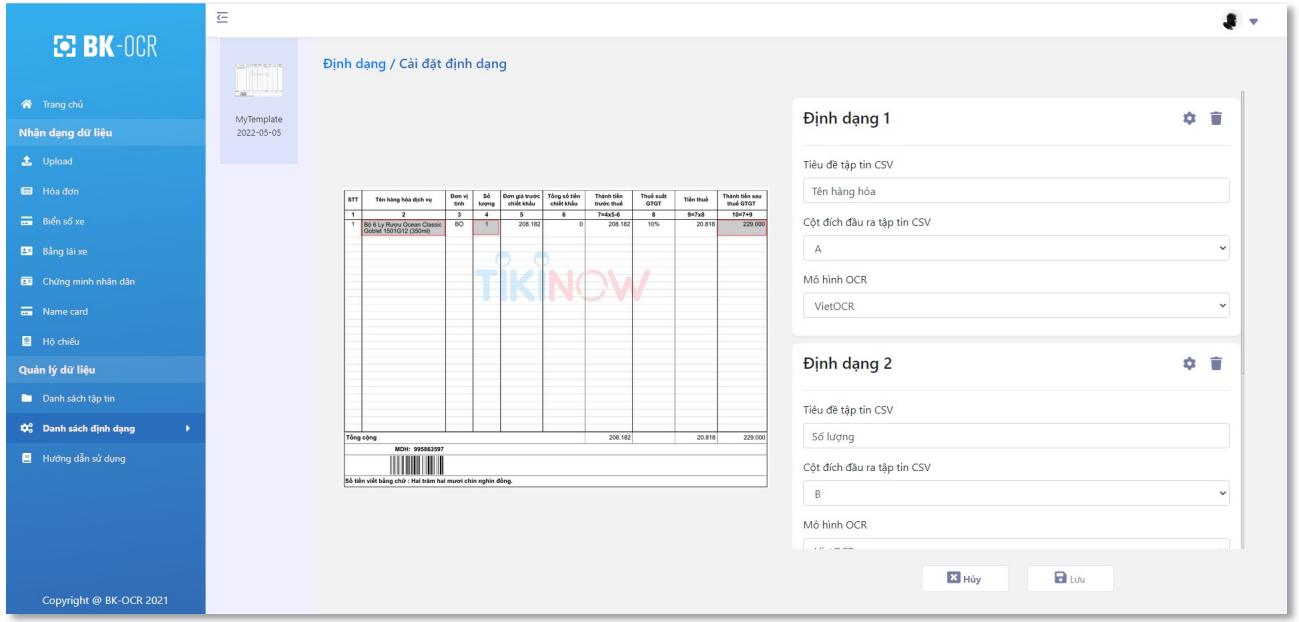
BK-OCR:
a comprehensive Vietnamese OCR system

- Text localization and recognition, especially for Vietnamese
- Build-in functions
  - Template recognition
  - Information extraction: automatically extract information of predefined fields from predefined templates
    - Invoice
    - Driving license
    - ID card
    - Name card
    - Passport
    - Number plate
  - Template definition



BK-OCR:
a comprehensive Vietnamese OCR system

- Text localization and recognition, especially for Vietnamese
- Build-in functions
  - Template recognition
  - Information extraction
  - Template definition
    - Allow users to freely defined new templates



BK-OCR:
a comprehensive Vietnamese OCR system

- demo

Conclusion and future works

- What we have done so far
  - Constructing a process to
    - collect and annotate real data
    - generate synthetic data

    for Vietnamese OCR
  - Improving the SOTA OCR models for Vietnamese
  - Building a comprehensive Vietnamese OCR system
- Future works
  - Releasing the data/pretrained models for community
  - Focusing on research for Vietnamese OCR


Thank You!