

2025.06.10 個人ゼミ

# HTR-JAND: Handwritten Text Recognition with Joint Attention Network and Knowledge Distillation

---

大阪工業大学大学 ロボティクス&デザイン工学部 システムデザイン工学科

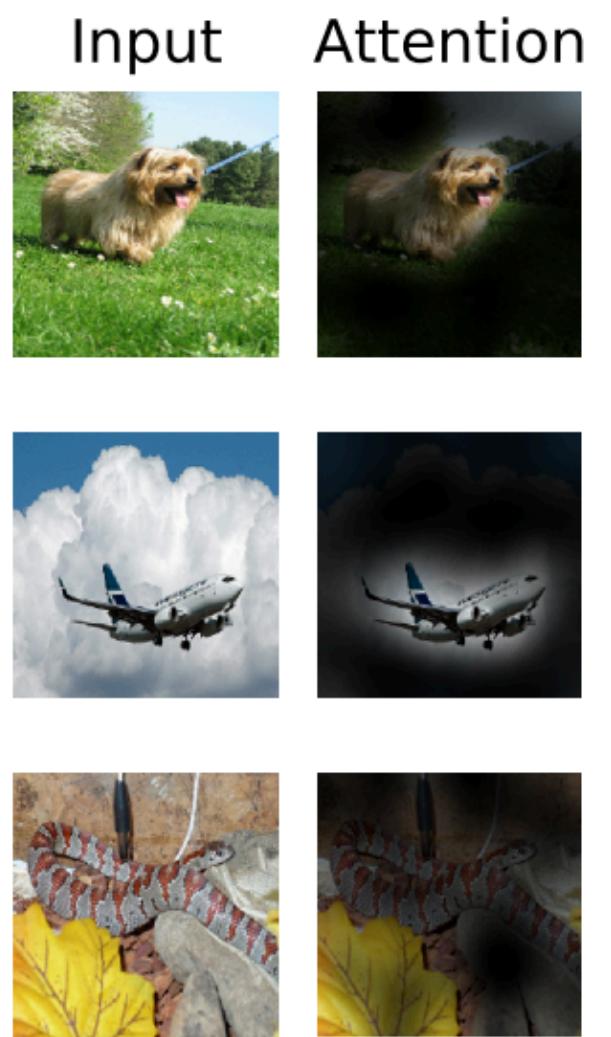
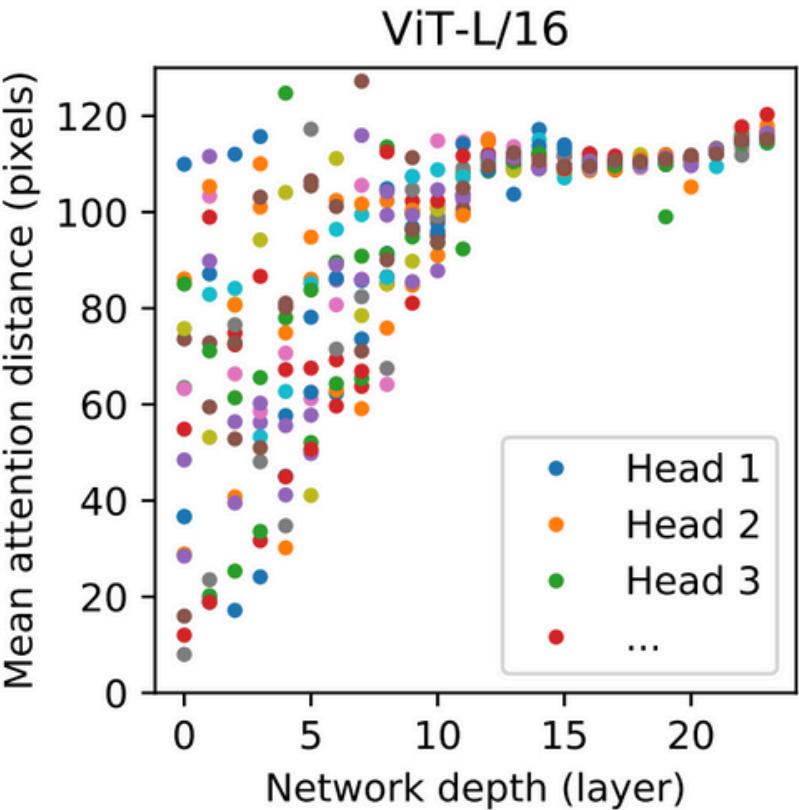
工藤滉青 濑尾昌孝

その前に

Vitで気になったところ

- 9ページ以降は読む必要ない？

この図についてattentionを向ける範囲を広げることでパッチ全体の情報が取れてよいのではないか  
パッチ内の局所領域に重みが大きくなっても割とよくないのではないか？？



## Abstract

## 要旨

現在の手書き文字認識(HTR)課題

- 歴史文書が持つ多様な筆跡スタイル
- 劣化した文字品質
- 複数の言語や時代をまたぐ計算効率

## 対策

- FullGatedConv2d層とSqueeze-and-Excitationブロック
- Proxima AttentionのMulti-Head Self-Attention
- 知識蒸留フレームワーク
  - カリキュラムベースの学習を通じて、精度を維持しつつ効率的なモデル圧縮を可能にする

# INTRODUCTION

## はじめに（1 / 2）

### 深層学習ベースの手法の課題

- 筆跡スタイルや歴史的時代をまたぐ一般化性能の不安定さ
  - 古い文字データセットがあんまないし、その中で学習してデータの多様性を出すのが厳しいから？？(my thought)
- 長いテキストシーケンスの処理の困難さ
  - メモリいっぱい (my thought)
- 実用的なデプロイメントを制限する高い計算コスト
  - 長文タスクによる計算コストの肥大&HTRは文章理解だけでなく画像認識も入っているため (my thought)

### attention 機構の課題

- 認識精度と計算効率のバランスをとるむずかしさ

# INTRODUCTION

## はじめに（2/2）

対策（詳細は後の章で）

- 複数のデータセットにまたがる文字セットの統一と適応的オーバーサンプリングを組み合わせた包括的な前処理パイプライン
- FullGatedConv2d層とSqueeze-and-Excitationブロックを組み合わせたCNNアーキテクチャ
- Multi-Head Self-Attention と Proxima Attention を統合したCombined Attentionメカニズム
- 知識蒸留フレームワーク
- 合成データ生成、アンサンブル学習、マルチタスク学習と組み合わせたカリキュラム学習を用いたトレーニング戦略
- フайнチューニングされたT5モデルを用いた文脈認識型後処理

# RELATED WORK

## 関連研究

以下略語

- GC (Gated Convolution)
- CA (Combined Attention)
- KD (Knowledge Distillation)
- CL (Curriculum Learning)
- SE (Squeeze-and-Excitation Blocks)
- AR (Aspect Ratio Preservation : 縦横比の保持)
- PP (Post-processing : 後処理)

Study	GC	SE	CA	KD	CL	AR	PP
Graves et al. [20]	✓						
Puigcerver [10]							
Bluche [2]	✓						
Chowdhury et al. [8]		✓					
Kang et al. [13]			✓				
Wigington et al. [19]				✓			
Hamdan et al. [15]	✓						
Flor et al. [17]	✓	✓					✓
Retsinas et al. [21]						✓	
(HTR-JAND) this Work	✓	✓	✓	✓	✓	✓	✓

## RELATED WORK

### 関連研究

以前はHMM（Hidden Markov Model）が主流

- 特徴設計に高度な知識が必要
- 長距離依存関係の扱いが困難

GravesらによるCTCの導入

- 未分割の系列データに対するエンドツーエンド学習が可能に
- 出力と入力のアライメントを自動的に学習

HTRを画像からテキストへの「翻訳問題」として捉える手法の登場

- CNN + RNN の統合モデルを用いたアプローチ
- 空間的（画像）+ 時間的（系列）依存関係を同時に捉えることが可能
- 入力・出力の長さが異なるデータにも柔軟に対応

## RELATED WORK

### 関連研究

Puigcerverらの研究

- CNN-LSTM + CTC ロスの有効性を実証
- 実用的なベースラインモデルとして広く採用される

Duttaらの研究

- Spatial Transformer Networks を導入
- 手書きに含まれる幾何学的ゆがみに対応

現在の課題

- 歴史的文書などに見られる手書きの多様性が、依然としてモデルの汎化性能に影響
- ドメイン間（近代 vs 歴史文書など）のギャップにより精度が低下することも

# METHODOLOGY

## 提案手法

### A, データ前処理と拡張

- 文字セットの統一
  - 出現頻度の低い文字を削除
  - すべてのデータセットに共通する103個のユニークな文字セットを構築
- 最大系列長に2のバッファを追加
- 前処理パイプラインは、以下の3つの主要な課題に対応しています：
  - 筆記スタイルの多様性
  - ラベル付きデータの不足
  - 時間的な一貫性 (temporal coherence) の保持

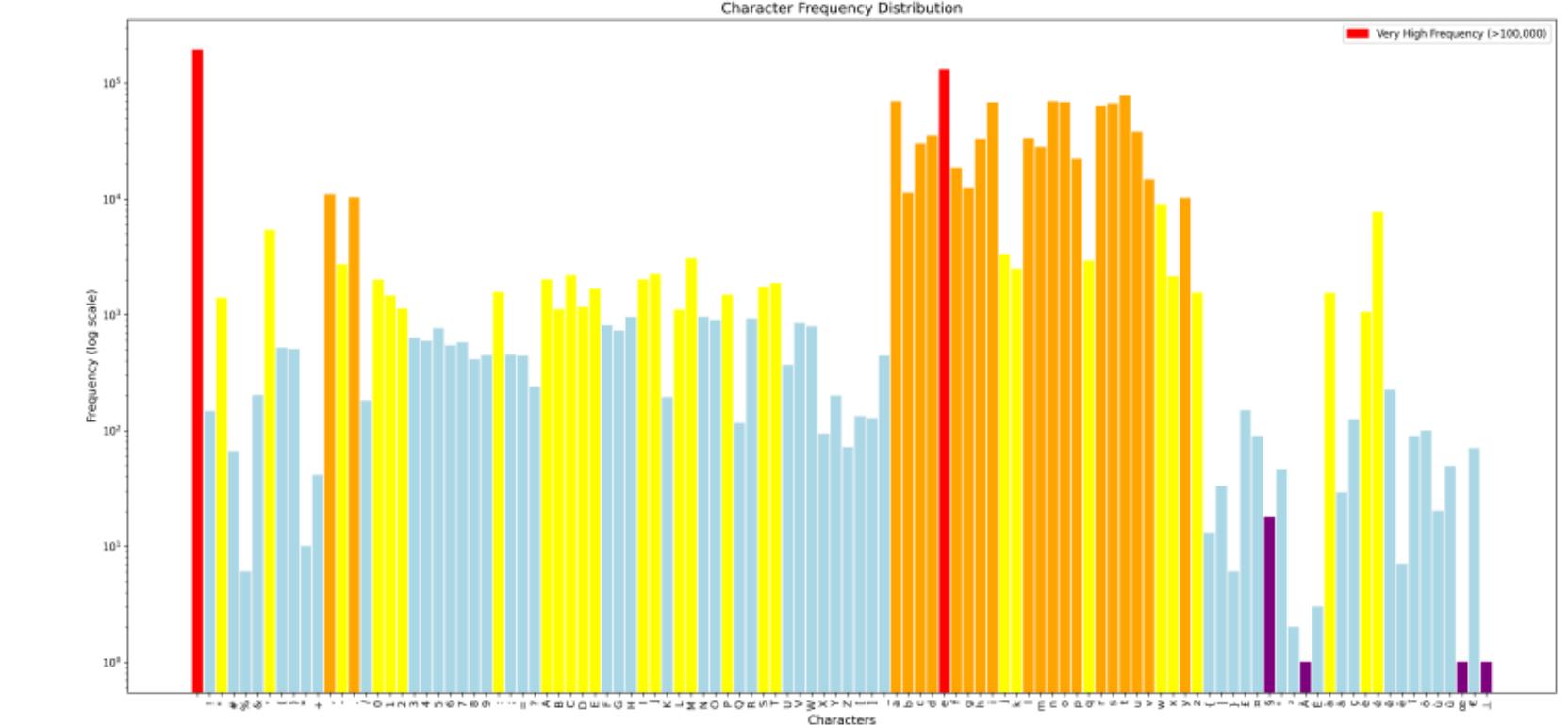


Fig. 2: Distribution of character frequencies across the combined datasets. Note the removal of infrequent characters such as '§', 'À', and 'œ'.

# METHODOLOGY

## 提案手法

A, データ前処理と拡張

以下のアルゴリズムに従って前処理が行われる

---

**Algorithm 1** Preprocessing Pipeline with Synthetic Data (PPS)

---

**Input:**  $D, C, F, r, \alpha$

**Output:**  $D'$

- 1:  $\mathbf{D}_n \leftarrow \text{Normalize}(D)$  {Eq. 1}
  - 2:  $\mathbf{D}_a \leftarrow \text{Augment}(\mathbf{D}_n)$  {Apply transforms}
  - 3:  $\mathbf{D}_s \leftarrow \text{GenerateSynthetic}(C, F, r)$  {Algo 2}
  - 4:  $\mathbf{D}_t \leftarrow \text{Tokenize}(\mathbf{D}_a \cup \mathbf{D}_s, C)$
  - 5:  $\mathbf{D}' \leftarrow \text{BalanceClasses}(\mathbf{D}_t, \alpha)$
  - 6: **return**  $\mathbf{D}'$
- 

---

**Algorithm 2** Synthetic Data Generation (SDG)

---

**Input:**  $C, F, r, D$

**Output:**  $\mathbf{D}_s$

- 1:  $n \leftarrow |D| \cdot r / (1 - r)$
  - 2: **for**  $i = 1$  to  $n$  **do**
  - 3:    $t \leftarrow \text{RandomText}(C)$
  - 4:    $f \leftarrow \text{RandomChoice}(F)$
  - 5:    $I \leftarrow \text{RenderText}(t, f)$
  - 6:    $I_{\text{aug}} \leftarrow \text{Augment}(I)$
  - 7:    $\mathbf{D}_s \leftarrow \mathbf{D}_s \cup \{(I_{\text{aug}}, t)\}$
  - 8: **end for**
  - 9: **return**  $\mathbf{D}_s$
- 

またそれぞれの画像に対して以下の変換を行う  $I_{\text{aug}} = t_n(\dots t_2(t_1(I))).$

このパイプラインには、学習の安定性を確保するための3つの主要な戦略が組み込まれています：

1. カリキュラムに基づく合成データ比率の調整
2. 性能に応じて適応的に合成データを統合（初期比率は10%）
3. 強化されたデータ拡張技術

# METHODOLOGY

## 提案手法

### A, データ前処理と拡張

$$w_c = \max\left(1, \frac{\bar{f}}{\epsilon + f_c}\right),$$

$\bar{f}$  : 文字の平均出現頻度

$f_c$  : 文字cの出現頻度

$w_c$  : 文字cに対するサンプリング重み

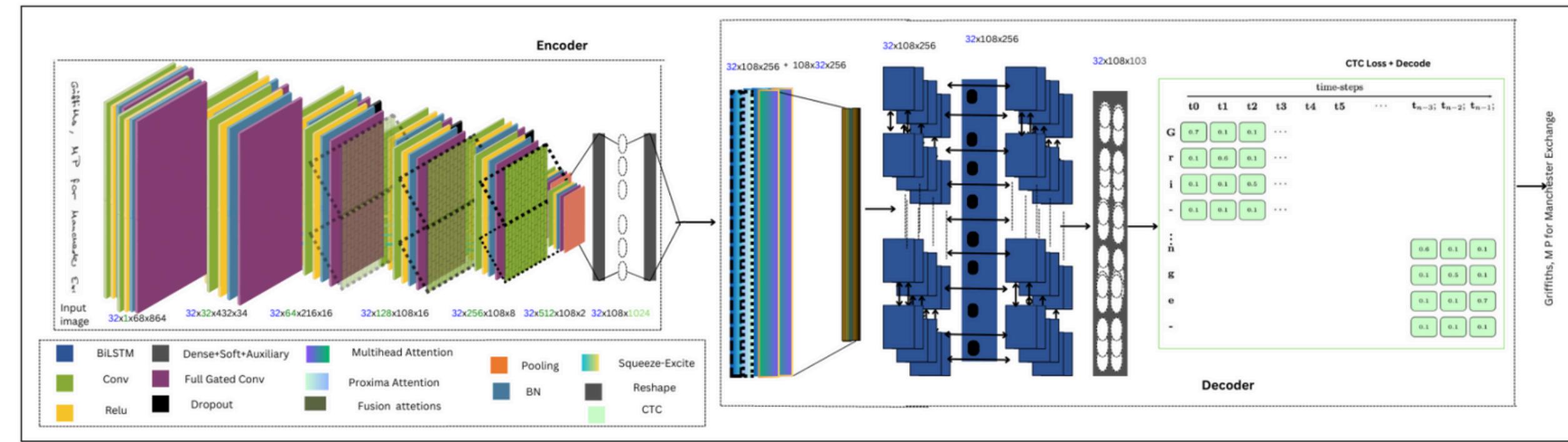
# METHODOLOGY

## 提案手法

### B, モデル ①概要

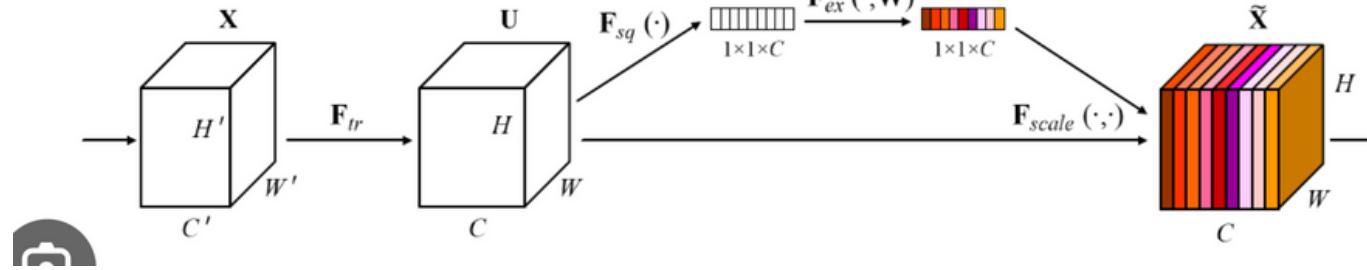
図に示されているように、Teacherモデルは以下の5つの主要な構成要素から成る：

1. Squeeze-and-Excitationモジュール付きのCNNブロック
2. FullGatedConv2dレイヤーによる適応的特徴抽出
3. BiLSTMによる系列モデリング
4. Proxima Attentionと組み合わせたマルチヘッド・セルフアテンション
5. 補助分類を伴うCTCベースのデコーディング



# METHODOLOGY

## 提案手法



B, モデル ② CNNによる特徴抽出

1の処理

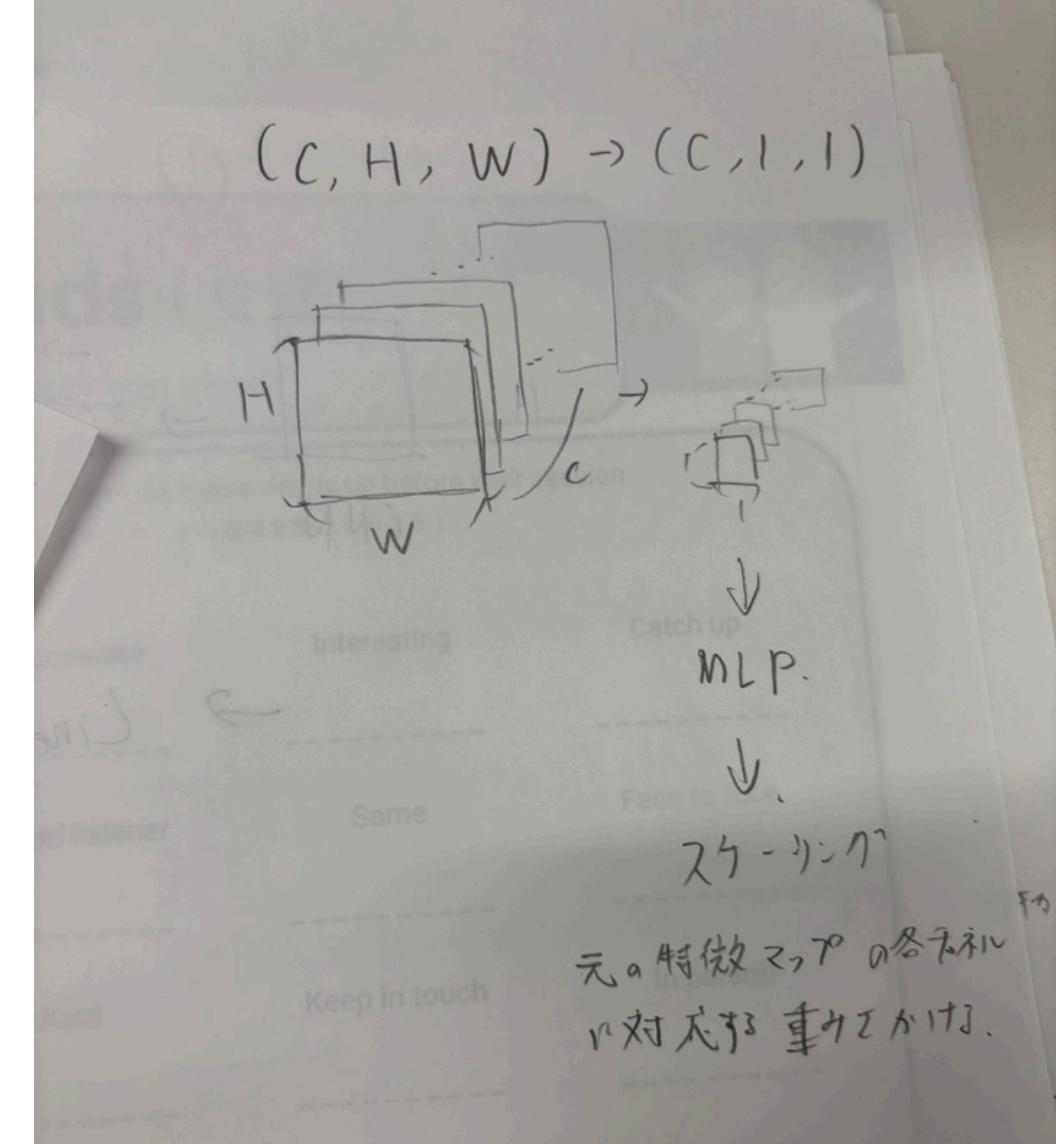
$$\mathbf{f}_l = \text{SE}(\text{MaxPool}(\text{ReLU}(\text{BN}(\mathbf{W}_l * \mathbf{f}_{l-1} + \mathbf{b}_l)))),$$

$$\mathbf{f}_{\text{SE}} = \mathbf{f}_l \cdot \sigma(\mathbf{W}_2 \text{ReLU}(\mathbf{W}_1 \text{GAP}(\mathbf{f}_l))).$$

SEで各チャネルごとの重要度の抽出

2.の処理

$$\text{FullGatedConv2d}(\mathbf{X}) = (\mathbf{W}_1 * \mathbf{X}) \odot \sigma(\mathbf{W}_2 * \mathbf{X}).$$



# METHODOLOGY

## 提案手法

### B, モデル ③BiLSTM

and into four bidirectional LSTM layers for temporal model-

etc.

$$\mathbf{h}_t = [\overrightarrow{\text{LSTM}}(\mathbf{X}_t, \vec{\mathbf{h}}_{t-1}); \overleftarrow{\text{LSTM}}(\mathbf{X}_t, \overleftarrow{\mathbf{h}}_{t+1})], \quad (8)$$

where, each LSTM cell follows:

$$I_t = \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{X}_t] + \mathbf{b}_i) \quad (9)$$

$$f_t = \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{X}_t] + \mathbf{b}_f) \quad (10)$$

$$o_t = \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{X}_t] + \mathbf{b}_o) \quad (11)$$

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c[\mathbf{h}_{t-1}, \mathbf{X}_t] + \mathbf{b}_c) \quad (12)$$

$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + I_t \odot \tilde{\mathbf{c}}_t \quad (13)$$

$$\mathbf{h}_t = o_t \odot \tanh(\mathbf{c}_t). \quad (14)$$

(9) 現在の情報、主張

(10) 過去の：

(11)

(12) 現在の説明文

(13) 文脈文

(14) 出力

$$\mathbf{h}_t = [\overrightarrow{\text{LSTM}}(\mathbf{X}_t, \vec{\mathbf{h}}_{t-1}); \overleftarrow{\text{LSTM}}(\mathbf{X}_t, \overleftarrow{\mathbf{h}}_{t+1})],$$

where, each LSTM cell follows:

$$I_t = \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{X}_t] + \mathbf{b}_i)$$

$$f_t = \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{X}_t] + \mathbf{b}_f)$$

$$o_t = \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{X}_t] + \mathbf{b}_o)$$

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c[\mathbf{h}_{t-1}, \mathbf{X}_t] + \mathbf{b}_c)$$

$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + I_t \odot \tilde{\mathbf{c}}_t$$

$$\mathbf{h}_t = o_t \odot \tanh(\mathbf{c}_t).$$

# METHODOLOGY

## 提案手法

B, モデル ④ 注意機構

通常の multi-head attention

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}. \quad \text{MultiHead}(\mathbf{X}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)\mathbf{W}^O$$

multi-head attention with Proximal Attention

動的なクエリ更新

$$\mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V$$

出力の結合

$$\mathbf{O}_{\text{combined}} = \text{LayerNorm}(\mathbf{W}_f[\mathbf{O}_{\text{MHA}}; \mathbf{O}_{\text{Proximal}}] + \mathbf{X})$$

# METHODOLOGY

## 提案手法

B, モデル ⑤student model architecture

項目	Teacher モデル	Student モデル	差分の概要
CNNブロック数	5	3	<b>2層削減</b>
チャネル数の初期値	32	16	<b>半分に縮小</b>
アテンションヘッド数	2	1	<b>ヘッド数を1つに削減</b>
LSTMの隠れ次元数	128	64	<b>次元数を半分に縮小</b>

この設計で、パラメータ数が約48%削減され（1,504,544 → 750,654）、それでも知識蒸留によって認識性能は維持

# METHODOLOGY

## 提案手法

### C.知識蒸留

知識蒸留手法は、高性能なTeacherモデルで学習した表現を小型のStudentモデルに転送することで、効率的なモデルデプロイを可能にする

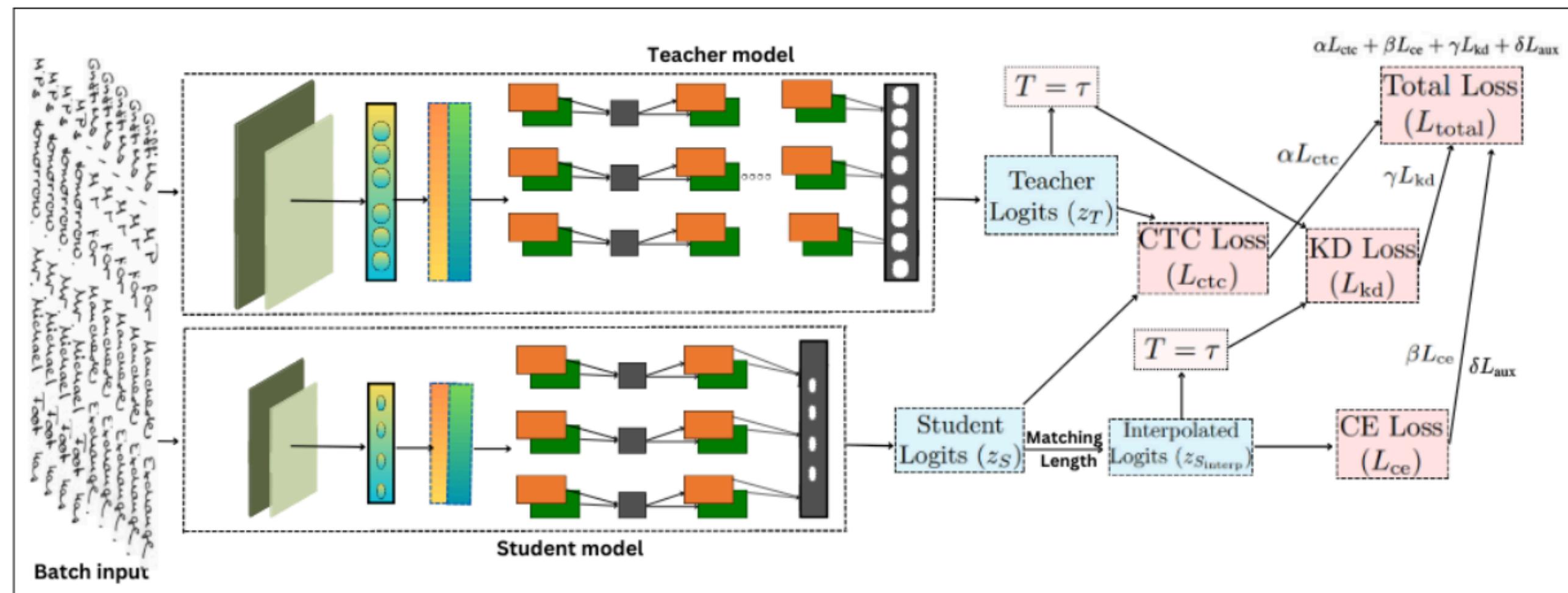


Fig. 4: Overview of our proposed knowledge distillation framework for handwritten text recognition (HTR).

# METHODOLOGY

## 提案手法

D,損失関数の設計

- CTC損失

以下の確率の最大化しこれに負をかけ対数をとったものが損失

$$p(\mathbf{y}|\mathbf{X}) = \sum_{\pi \in \mathcal{B}^{-1}(\mathbf{y})} p(\pi|\mathbf{X}), \quad \mathcal{L}_{\text{ctc}} = -\log(p(\mathbf{y}|\mathbf{X})).$$

$\pi$  : blankトークンや重複文字を含む整列パターン

$\mathcal{B}^{-1}(\mathbf{y})$  : 「ブランクや繰り返し文字を除いたときに  $\mathbf{y}$  に整形されるすべての整列」の集合

$\mathbf{X}$  : 画像フレーム

$\mathbf{y}$  : 目標系列であるテキスト

# METHODOLOGY

## 提案手法

D,損失関数の設計

- CTC損失

$\mathcal{B}^{-1}(y)$  例

$$\beta^{-1}("cat") = \left\{ \begin{array}{l} "c a <\text{blank}> t", \\ "<\text{blank}> <\text{blank}> c a t", \\ "<\text{blank}> c <\text{blank}> a t", \\ \dots \end{array} \right\}$$

これらがcatであるということの確率の最大化したい  
つまり、このような文字列からcatという単語を推測できるように学習  
したい

# METHODOLOGY

## 提案手法

### D,損失関数の設計

- クロスエントロピー損失

$$\mathcal{L}_{\text{ce}} = - \sum_i y_i \log(\hat{y}_i),$$

$y_i$  : 正解ラベル

$\hat{y}_i$  : クラス*i*に対する予測確率

すべてのクラスを均等に扱うことを目的

図に示されているようなクラス不均衡問題への対応と、文字レベルの教師の補完

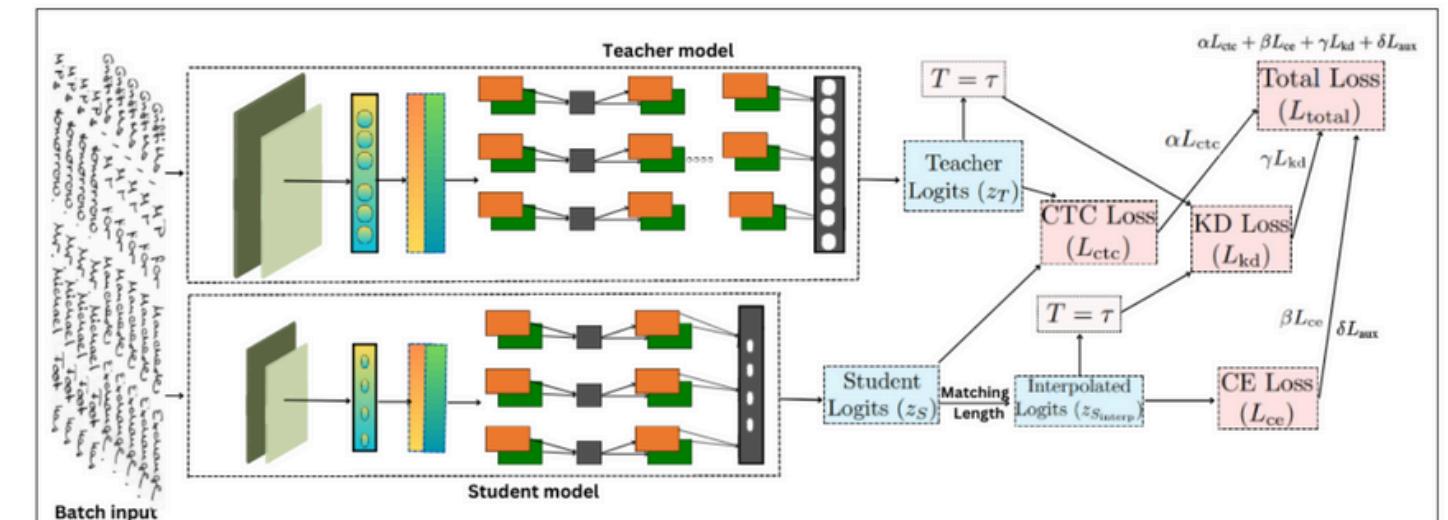


Fig. 4: Overview of our proposed knowledge distillation framework for handwritten text recognition (HTR).

# METHODOLOGY

## 提案手法

D,損失関数の設計

- 知識蒸留損失

損失関数は以下である

$$\mathcal{L}_{\text{kd}} = \text{KL}(\text{softmax}(\mathbf{z}_T / \tau), \text{softmax}(\mathbf{z}_S / \tau)),$$

$\mathbf{z}_T, \mathbf{z}_S$  : teacher, student の logit(??)

$\tau$  : tempreture パラメーター

KullbackLeibler ダイバージェンス

$$\text{KL}(\mathbf{P} \parallel \mathbf{Q}) = \sum_i P(i) \log \left( \frac{P(i)}{Q(i)} \right),$$

P, Qが等しくなるように学習(ほんと??)

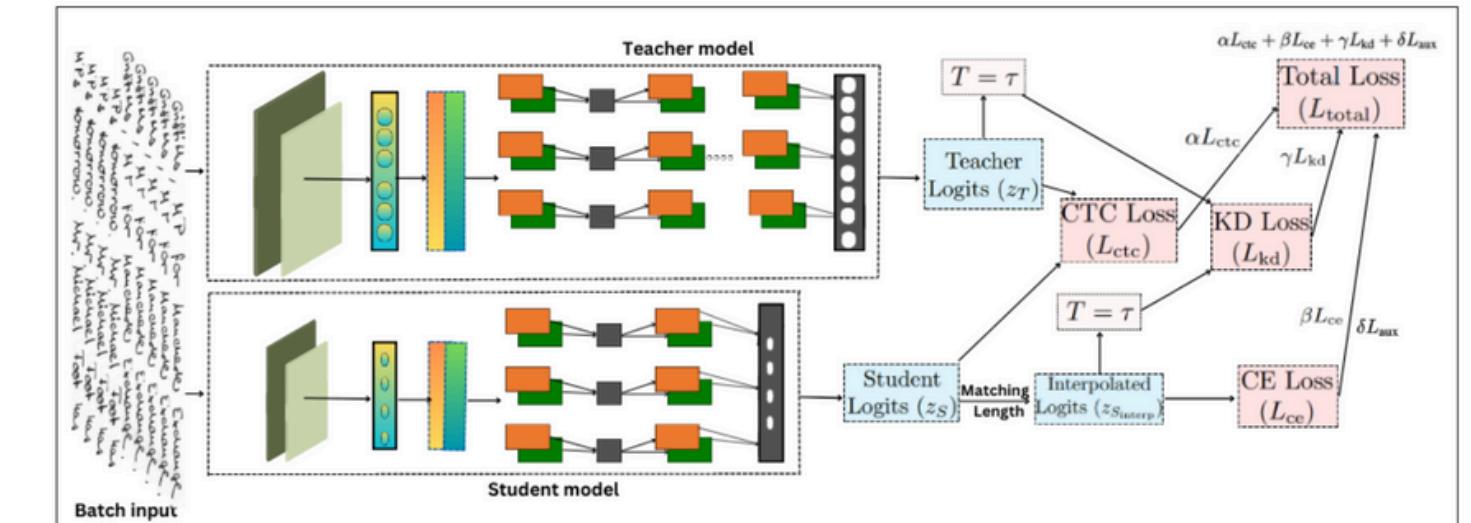


Fig. 4: Overview of our proposed knowledge distillation framework for handwritten text recognition (HTR).

# METHODOLOGY

## 提案手法

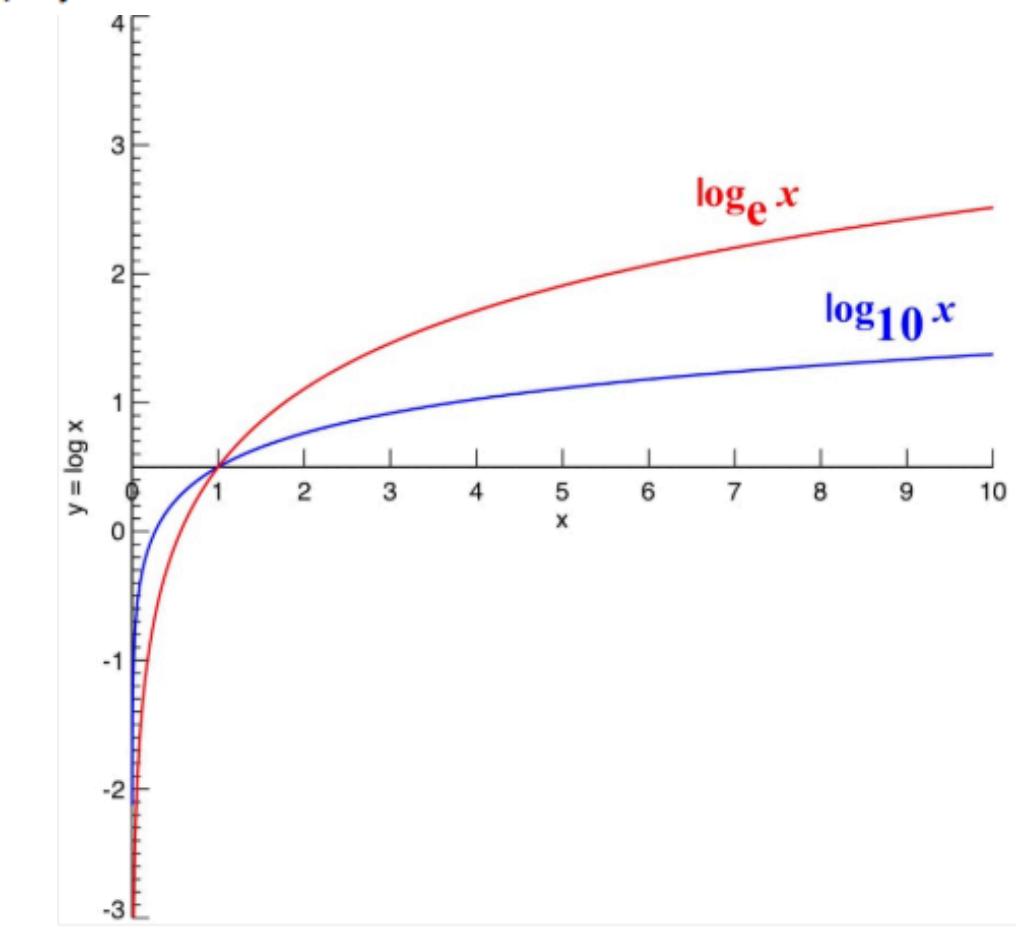
D,損失関数の設計

- 知識蒸留損失

$$\mathcal{L}_{\text{kd}} = \text{KL}(\text{softmax}(\mathbf{z}_T / \tau), \text{softmax}(\mathbf{z}_S / \tau)),$$

$$\text{KL}(\mathbf{P} \parallel \mathbf{Q}) = \sum_i P(i) \log \left( \frac{P(i)}{Q(i)} \right),$$

Lの最小化→KLの最小化したい→  
logの性質より Pの最小化、 Qの最大化なのでは？？



# METHODOLOGY

## 提案手法

D,損失関数の設計

- Auxiliary Classifier

$$\mathcal{L}_{\text{aux}} = - \sum_i y_i \log(\hat{y}_{\text{aux},i}),$$

$\hat{y}_{\text{aux},i}$  : クラス  $i$  に対する補助分類器の予測確率

# METHODOLOGY

## 提案手法

### IV, 高度な学習戦略

- A. 学習プロセスの概要

カリキュラム学習は以下の進行比率に従う

$$r_s(e) = \min(r_{\max}, r_0 + \frac{e}{E}(r_{\max} - r_0)),$$

$r_0=0.1$  ,  $r_{\max} = 0.4$  , e:現在のエポック数 , E:トータルのエポック数

この進行比率に従って、実データのみから始めて徐々に実データと合成データのバランスを取る形にしていく

---

#### Algorithm 2 Synthetic Data Generation (SDG)

---

**Input:**  $C, F, r, D$

**Output:**  $D_s$

```
1:  $n \leftarrow |D| \cdot r / (1 - r)$ 
2: for  $i = 1$  to  $n$  do
3:    $t \leftarrow \text{RandomText}(C)$ 
4:    $f \leftarrow \text{RandomChoice}(F)$ 
5:    $I \leftarrow \text{RenderText}(t, f)$ 
6:    $I_{\text{aug}} \leftarrow \text{Augment}(I)$ 
7:    $D_s \leftarrow D_s \cup \{(I_{\text{aug}}, t)\}$ 
8: end for
9: return  $D_s$ 
```

---

# METHODOLOGY

## 提案手法

### IV, 高度な学習戦略

- A. 学習プロセスの概要

TeacherとStudentモデルのアーキテクチャが異なることに対応するために、ロジット整合機構を以下のように行う

$$z_{S_{\text{interp}}} = \text{Interpolate}(z_S, \text{len}(z_T)).$$

カリキュラム進行の適応的管理は下の図のアルゴリズムに従う

**Algorithm 3** Adaptive Curriculum Progression (ACP)

**Input:**  $M, S_0, T, \Delta_T$

**Output:**  $M^*$

- 1:  $S \leftarrow S_0$  {Stage initialization}
- 2: **while**  $S < S_{max}$  **do**
- 3:   Train  $M$  on stage  $S$  data
- 4:   Evaluate  $M$  on validation set
- 5:   **if** Performance  $> T$  **then**
- 6:      $S \leftarrow S + 1$  {Advance stage}
- 7:      $T \leftarrow T + \Delta_T$  {Adjust threshold}
- 8:   **end if**
- 9: **end while**
- 10: **return**  $M^*$

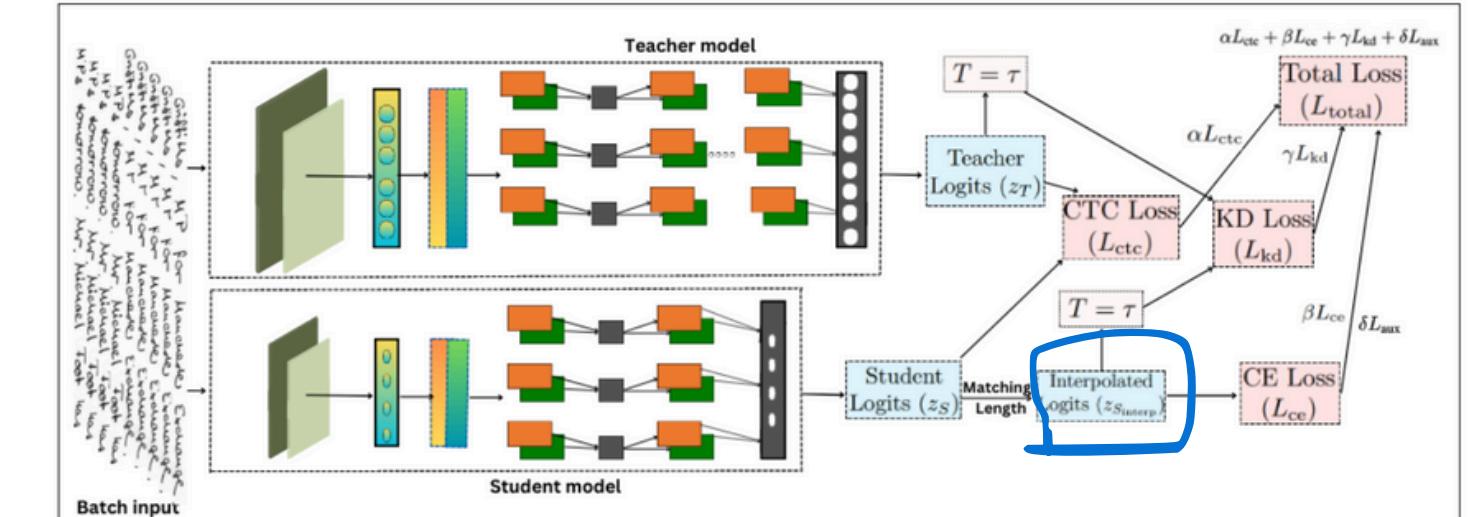


Fig. 4: Overview of our proposed knowledge distillation framework for handwritten text recognition (HTR).

# METHODOLOGY

## 提案手法

IV, 高度な学習戦略

A. 学習プロセスの概要

multi-component loss framework

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{ctc}} + \beta \mathcal{L}_{\text{ce}} + \gamma \mathcal{L}_{\text{kd}} + \delta \mathcal{L}_{\text{aux}},$$

$\delta$  : 定数 (0.1)

$\beta$  :  $1 - (\alpha + \gamma + \delta)$

学習の初期段階 :  $\alpha = 0.7$   $\gamma = 0.2$  により文字レベルの学習の強調

学習の最終段階 :  $\alpha = 0.4$   $\gamma = 0.5$

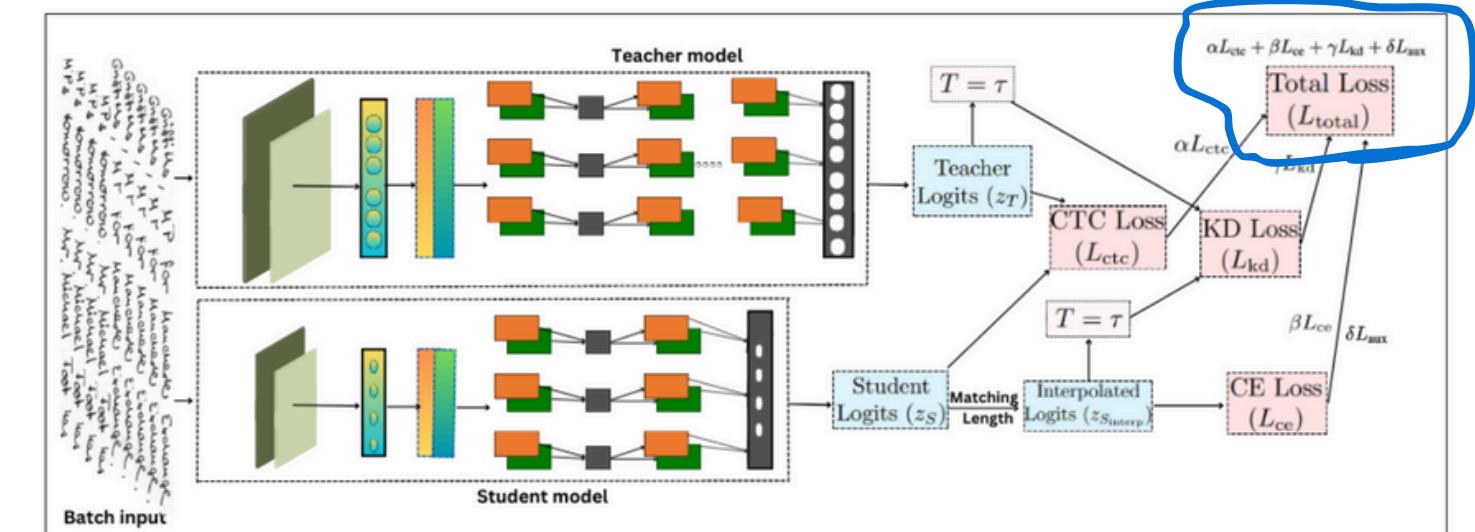


Fig. 4: Overview of our proposed knowledge distillation framework for handwritten text recognition (HTR).

# METHODOLOGY

## 提案手法

### IV, 高度な学習戦略

#### A. 学習プロセスの概要

マルチタスクの場合は以下のようにロスをとる

$$\mathcal{L}_{\text{multi-task}} = \sum_{k=1}^K \lambda_k \mathcal{L}_k,$$

これにより異なる歴史的時代や筆記スタイルをまたいだ効果的な知識伝達が可能

アーリーストッピングは以下の条件で適用される：

- patience : 10エポック (10回連続で改善がなければ停止)
- 最小改善閾値 : バリデーション損失が0.001未満しか改善しない場合は停止

# METHODOLOGY

## 提案手法

### IV, T5を用いた後処理(post-processing)

#### 1) モデル選定と適応

T5-small（パラメータ数6000万）モデルの使用、言語の違いや、筆記の時代的・スタイル的な差を多く含むデータセットの使用

#### 2) トークン化とテキスト正規化

トークン化にはSentencePieceを用た。処理の内容は以下の通りです：

- 歴史的な表記揺れや略語に対応したサブワードトークン化
- レイアウト保持のための特殊トークンの挿入
- 一貫した文字表現のためのUnicode正規化
- 手書き文に見られる不規則なスペース処理のための空白の標準化

#### 3) 学習データの準備

学習は、知識蒸留後のモデル予測結果を使って、予測と正解のペアを生成することで進めた。

具体的には：

- 最初に、我々のモデルで各データセットに対する予測を行う
- その予測結果を言語や時代ごとに分析し、エラーパターンを抽出
- 抽出されたパターンに基づき体系的なエラーを人工的に導入
- それらを含む文脈ウィンドウを構築して、訂正精度を高める

#### 4) 統合パイプライン

右図に示された我々のT5後処理フレームワークは、以下のような多層的な訂正戦略を取る：

- 文脈認識によるエラー検出
- 信頼度に基づく訂正の適用
- 各データセットに合わせた形式保持

---

#### Algorithm 5 T5 Post-Processing Pipeline (T5P)

---

**Input:**  $P, T_f, \theta, D$

**Output:**  $C$

```
1: Initialize  $C \leftarrow \emptyset$ 
2: Train SentencePiece on  $D$ 
3: for each batch  $B$  in  $P$  do
4:    $S \leftarrow \text{Segment}(B)$ 
5:    $\text{ctx} \leftarrow \text{BuildContext}(S)$ 
6:   for  $s$  in  $S$  do
7:      $\text{err} \leftarrow \text{DetectErrors}(s, D)$ 
8:     if  $\text{err} \neq \emptyset$  then
9:        $t \leftarrow \text{TokenizeSP}(s, \text{ctx})$ 
10:       $\text{cand} \leftarrow T_f(t, \text{ctx})$ 
11:       $\text{scr} \leftarrow \text{Confidence}(\text{cand})$ 
12:      if  $\text{scr} > \theta$  then
13:         $s \leftarrow \text{ApplyCorrection}(s, \text{cand})$ 
14:      end if
15:    end if
16:     $C \leftarrow C \cup \text{Format}(s)$ 
17:  end for
18: end for
19: return  $C$ 
```

---

# METHODOLOGY

## 提案手法

### IV RESULTS AND DISCUSSION

#### A. 教師モデルと生徒モデルの性能

文字誤り率 (CER)、単語誤り率 (WER)、文誤り率 (SER) で性能をはかる。

- 教師モデルと生徒モデルでの性能差が顕著(特に複雑なデータセット)

TABLE III: Performance Comparison of Teacher and Student Models

Model	Metric %	IAM	RIMES	Bentham	Saint Gall	Washington	Combined
Teacher	CER	2.34	2.21	3.12	4.01	4.76	2.89
	WER	8.22	7.11	6.98	11.33	13.30	7.88
	SER	80.12	75.76	78.90	71.33	68.22	82.45
Student	CER	4.59	6.22	5.13	4.23	6.99	12.91
	WER	18.54	21.99	17.01	24.78	22.11	28.45
	SER	91.45	94.01	89.33	94.55	92.11	95.90

# METHODOLOGY

## 提案手法

### IV RESULTS AND DISCUSSION

#### Bモデル予測の分析

モデルの性能において、T5後処理の有効性がみられた

- ・時制、大文字小文字の区別の誤りなどの検知

予測過程：アテンションの可視化を行った図

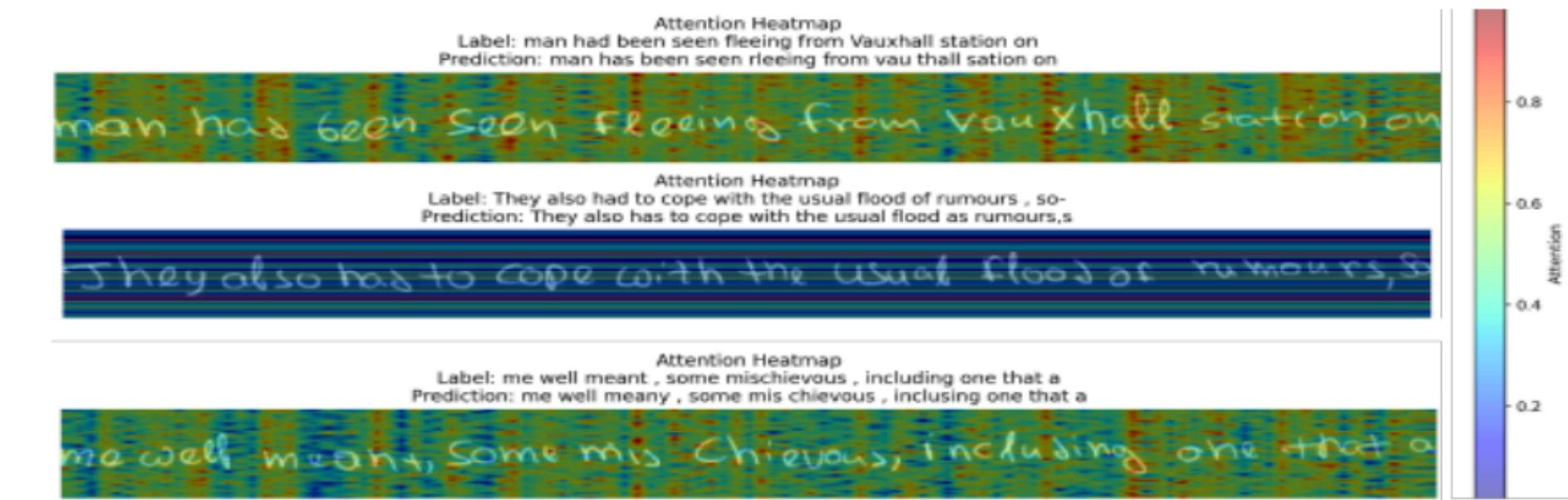


Fig. 5: Visualization of the model's attention heatmaps for the sample predictions. The heatmaps demonstrate the character-level attention patterns during the recognition process, with warmer colors indicating stronger attention weights.

TABLE IV: Comparison of Ground Truth, Initial Predictions, and T5-Corrected Output

Ground Truth	Initial Prediction	T5-Corrected Prediction
1. "man had been seen fleeing from Vauxhall station on"	"man has been seen rleeing from vauxhall station on"	"man had been seen fleeing from Vauxhall station on"
2. "They also had to cope with the usual flood of rumours, so-"	"They also has to cope with the usual flood as rumours,s"	"They also had to cope with the usual flood of rumours, so-"
3. "me well meant, some mischievous, including one that a"	"me well meany, some mis chievous, inclusing one that a"	"me well meant, some mischievous, including one that a"

# METHODOLOGY

## 提案手法

### IV RESULTS AND DISCUSSION

#### C , Ablation Study

知識蒸留、マルチタスク学習、アンサンブル学習、カリキュラム学習、辞書ベース補正がモデル性能に及ぼす影響を示した。

これらの組み合わせによる相乗効果が見られた

TABLE V: Comprehensive Ablation Study Results

Dataset	Metric	Baseline	+KD	+CL	+EL	+LBC
IAM	CER	12.21	4.59	2.34	2.02	<b>1.23</b>
	WER	28.32	18.54	8.22	5.22	<b>3.78</b>
	SER	95.34	91.45	80.12	78.12	<b>19.22</b>
RIMES	CER	15.34	6.22	2.21	1.89	<b>1.02</b>
	WER	31.45	21.99	7.11	5.43	<b>2.45</b>
	SER	94.10	94.01	75.76	68.78	<b>12.45</b>
Bentham	CER	20.11	5.13	3.12	3.12	<b>2.02</b>
	WER	36.89	17.01	6.98	6.11	<b>4.23</b>
	SER	97.00	89.33	78.90	76.53	<b>21.67</b>
Saint Gall	CER	7.56	4.23	4.01	3.81	<b>2.21</b>
	WER	18.12	24.78	11.33	9.27	<b>6.89</b>
	SER	89.32	94.55	71.33	68.17	<b>15.54</b>
Washington	CER	8.44	6.99	4.76	3.12	<b>2.98</b>
	WER	20.12	22.11	13.30	15.32	<b>6.34</b>
	SER	91.56	92.11	68.22	63.14	<b>11.22</b>

# METHODOLOGY

## 提案手法

### IV RESULTS AND DISCUSSION

D, state-of-the-artとの比較

我々のアプローチは、最先端の性能を達成しており、IAMおよびRIMESデータセットの両方で既存手法を大きく上回る結果を示した。

Method	Metric	IAM	RIMES
Ours (+LBC)	CER	<b>1.23</b>	<b>1.02</b>
	WER	<b>3.78</b>	<b>2.45</b>
Retsinas et al. [40]	CER	4.55	3.04
	WER	16.08	10.56
Yousef et al. [12]	CER	4.9	-
	WER	-	-
Tassopoulou et al. [11]	CER	5.18	-
	WER	17.68	-
Michael et al. [9]	CER	5.24	-
	WER	-	-
Wick et al. [14]	CER	5.67	-
	WER	-	-
Dutta et al. [5]	CER	5.8	5.07
	WER	17.8	14.7
Puigcerver [41]	CER	6.2	2.60
	WER	20.2	10.7
Chowdhury et al. [8]	CER	8.10	3.59
	WER	16.70	9.60

# METHODOLOGY

## 提案手法

### IV RESULTS AND DISCUSSION

#### E, アテンションの可視化解析

図6では、予測された文字と正解ラベル（Ground Truth）文字に対して、モデルがどのようにアテンションを配分しているかを示すクラス確率ヒートマップが下の図  
対角線上のアライメントが強調されし視覚的に似ているものにattentionが分散される  
 $\alpha$ と $\alpha$ など

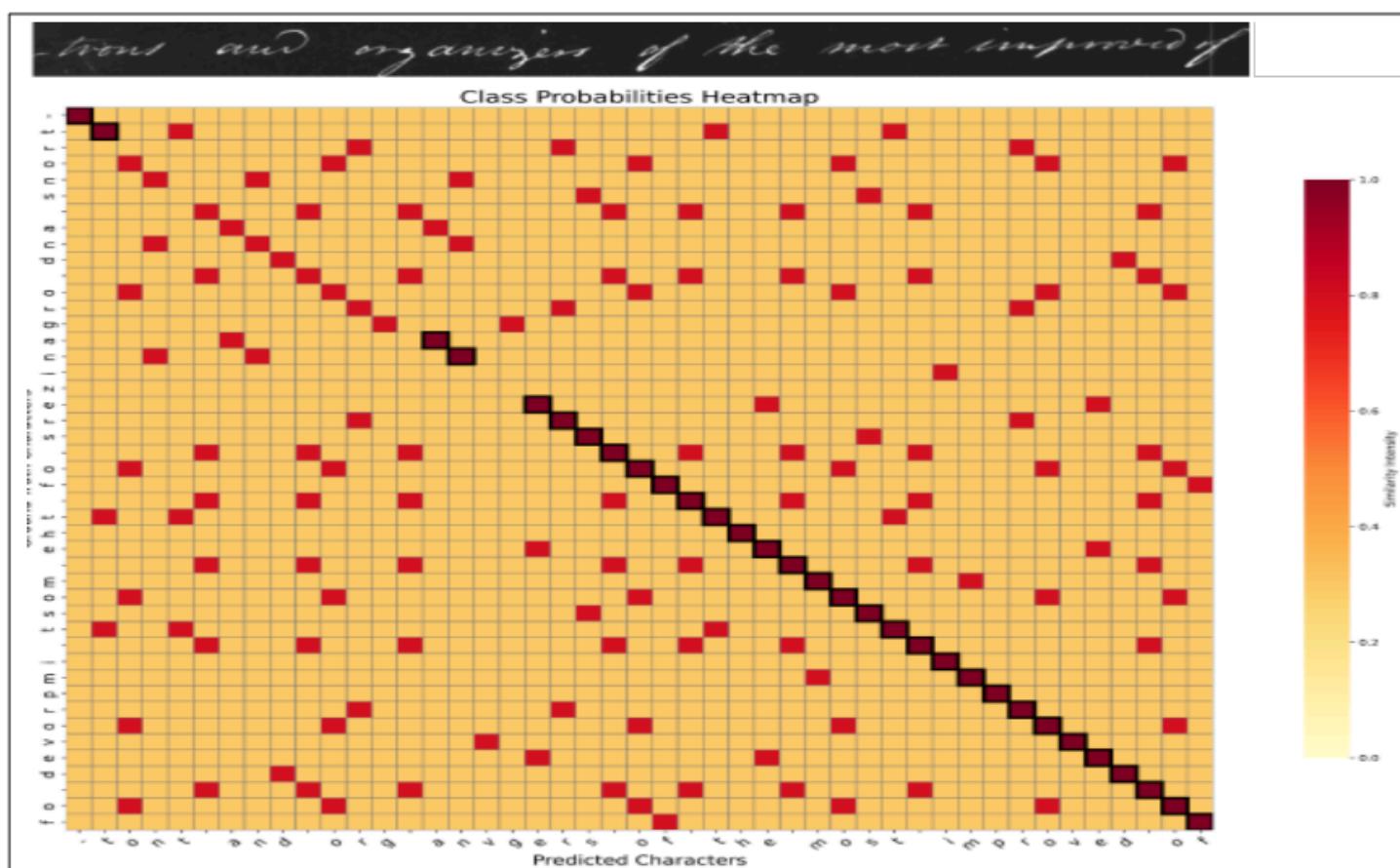


Fig. 6: Class probabilities heatmap for character alignment in the Rimes dataset. Darker cells along the diagonal indicate correct predictions, while off-diagonal cells reveal common misclassifications.

# METHODOLOGY

## 提案手法

### IV RESULTS AND DISCUSSION

#### F, 計算効率の分析

Testing : 1行あたりの推論時間

Teacherモデルは、1.50Mパラメータで最先端の性能

Studentモデルでは、パラメータ数を0.75Mまで削減しつつ、BlucheやFlorのモデルと同程度のサイズで推論時間は短く、性能はそれらよりも優れてい

Model	Params (M)	Testing(ms/line)	CER/IAM (%)
Our Teacher (+CL)	1.50	58	2.34
Our Student (+CL)	0.75	28	4.12
Puigcerver [41]	9.4	81	4.94
Bluche [2]	0.7	32	6.60
Flor [17]	0.8	55	3.72

# METHODOLOGY

## 提案手法

### IV 今後の課題 課題

- 図の混同行列の分析では、歴史文書において視覚的に類似した文字の区別が難しいことが依然として課題であると示された。
- 歴史的文脈に特有の語彙（専門用語など）に対しては、未知語処理の精度が限定的であることも明らかになりました。
- Studentモデルは軽量化には成功したが、まだどんな環境でも最適に動作するとは限らないのでさらに工夫の必要性

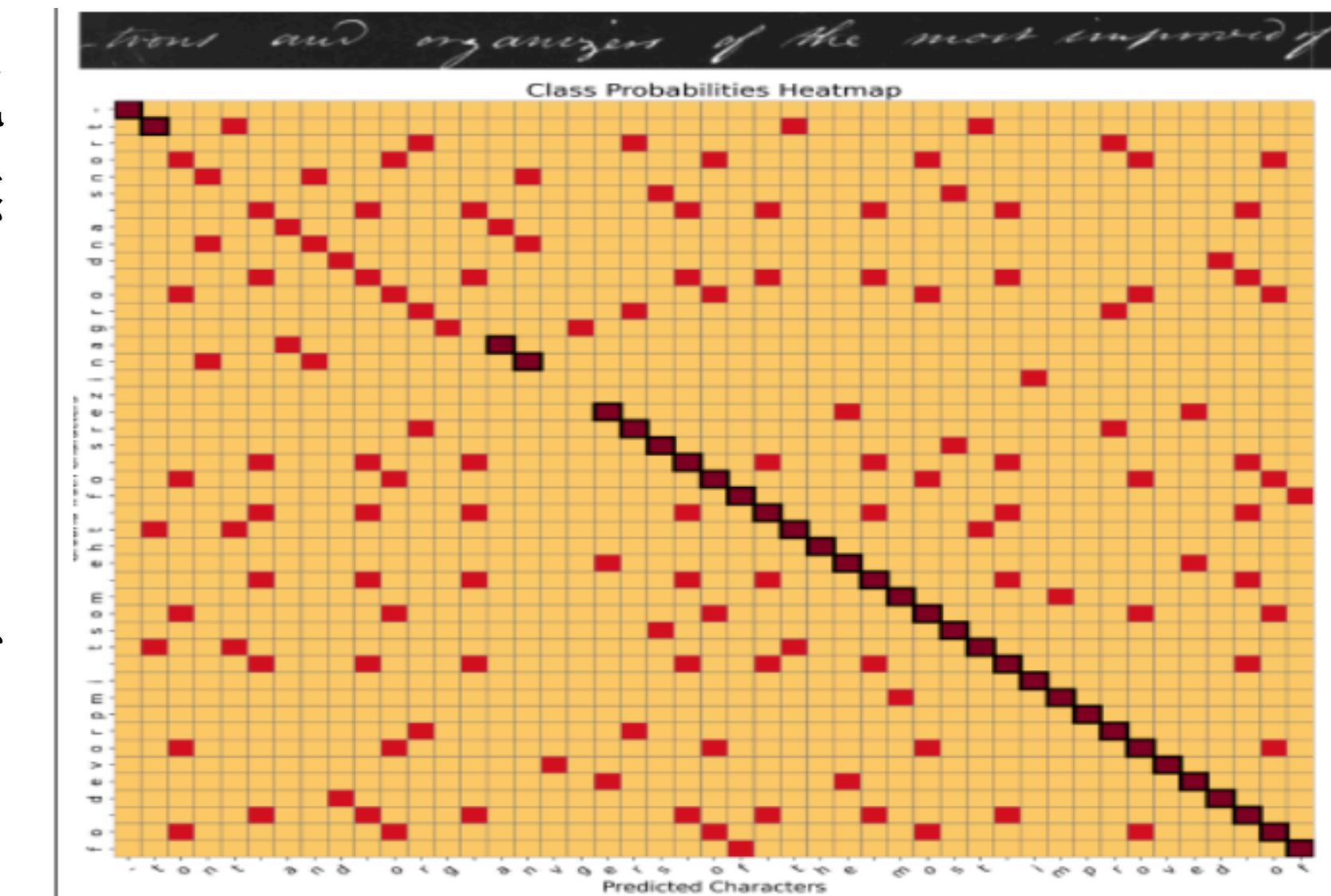


Fig. 6: Class probabilities heatmap for character alignment in the Rimes dataset. Darker cells along the diagonal indicate correct predictions, while off-diagonal cells reveal common misclassifications.