

Unsupervised Domain Adaptation for Isolated Traditional Mongolian Text Line Recognition via Vertical-Spine-Aware Stroke Enhancement and Pseudo-Label Distillation

Anonymous Authors
Paper under Double-Blind Review

Abstract

The digitization of traditional Mongolian historical archives presents a unique intersection of low-resource natural language processing and complex document image analysis. Characterized by vertical, cursive script with heavy morphological agglutination and severe historical degradation, these archives defy standard optical character recognition (OCR) techniques. Furthermore, the prohibitive cost of expert paleographic annotation results in extreme label scarcity. In this paper, we propose an Unsupervised Domain Adaptation (UDA) framework specifically targeted at isolated text line recognition, bridging the gap between abundant synthetic data and unlabeled authentic archives. We introduce a Style-Aligned Synthetic Engine (SASE) grounded in physical degradation modeling, and crucially, a Vertical-Spine-Aware Multi-Scale Stroke Enhancement (VS-MSSE) module utilizing asymmetrical convolutions to reconstruct the directional ligatures unique to Mongolian cursive. Latent alignment is achieved via a Sequence-aware Domain-Adversarial Neural Network (DANN), further refined by a valid-frame-normalized CTC Entropy pseudo-labeling strategy. Extensive experiments on a newly curated dataset of 19th-century administrative records—independently annotated by three experts—demonstrate the superiority of our approach. We achieve a Character Error Rate (CER) of $7.8 \pm 0.2\%$. Compared to modern vision-language models like TrOCR (41.5%) and PARSeq (36.2%) which show limited cross-domain generalization, our framework provides a robust baseline for low-resource, vertical cursive script recognition.

1 Introduction

The preservation of cultural heritage increasingly relies on the automated digitization of historical manuscripts. Among these, the traditional Mongolian script—a vertical, cursive alphabet utilized across Inner Asia since the 13th century—poses formidable challenges to modern Handwritten Text Recognition (HTR) systems. Unlike Latin or Chinese scripts, cursive vertical scripts feature a continuous central axis (the “spine”) connecting constituent phonemes, resulting in extreme topological dependency between adjacent characters, a challenge structurally analogous to Arabic cursive recognition [11].

Current deep learning architectures for HTR, such as Convolutional Recurrent Neural Networks (CRNN) [1] and Transformer-based models like PARSeq [4] and TrOCR [3], have achieved remarkable success on standardized datasets. However, their application to historical Mongolian archives is severely bottlenecked by the “Label Scarcity” problem. Creating a large-scale supervised dataset requires scholars with specialized training in historical paleography, making manual annotation prohibitively expensive. To circumvent this, researchers often train models on synthetically generated data. Unfortunately, this approach suffers from catastrophic performance degradation due to the massive “Domain Shift” between pristine synthetic fonts and highly irregular, oxidized, and physically degraded 19th-century cursive handwriting.

Unsupervised Domain Adaptation (UDA) offers a pathway to leverage unlabeled authentic target data alongside labeled synthetic source data. Standard UDA techniques, such as Maximum Mean Discrepancy (MMD) [9] and Domain-Adversarial Neural Networks (DANN) [5], primarily address global feature distribution alignment but often neglect the fine-grained structural topologies critical for dense cursive scripts.

In this paper, we focus strictly on the problem of **Isolated Text Line Recognition**, assuming upstream layout analysis has yielded accurately cropped vertical text columns. To address the specific domain shift of historical Mongolian, we propose a novel UDA framework. Our core innovation lies in explicitly modeling the directional bias of the script. We introduce the Vertical-Spine-Aware Multi-Scale Stroke Enhancement (VS-MSSE) module, which replaces standard symmetrical convolutions with asymmetrical vertical kernels (e.g., 5×1) to preferentially preserve the continuous vertical spine against heavy background noise. This is complemented by a Style-Aligned Synthetic Engine (SASE) for initial domain bridging, and an iteratively refined pseudo-label distillation technique utilizing a corrected formulation of valid-frame-normalized CTC entropy.

2 Related Work

2.1 Handwritten Text Recognition (HTR)

The paradigm of HTR has shifted significantly from Hidden Markov Models (HMMs) [15] to deep neural networks. The CRNN architecture [1], which pairs a CNN feature extractor with recurrent layers (e.g., BiLSTM) and Connectionist Temporal Classification (CTC) [2], remains the standard for unconstrained sequence recognition. Recently, Transformer-based models like TrOCR [3] and PARSeq [4] have demonstrated state-of-the-art results by leveraging self-attention to capture long-range linguistic dependencies. However, these models heavily rely on massive annotated datasets and often exhibit limited cross-domain generalization when deployed in zero-shot or severely degraded historical settings. Cursive sequence recognition historically relied on over-segmentation heuristics [15], but recent deep learning approaches [12] have adopted end-to-end architectures, albeit almost exclusively within fully supervised paradigms.

2.2 Historical Document Image Analysis

Historical documents present severe degradation, including ink bleed-through, faded strokes, and background noise. Binarization networks [10] and stroke enhancement modules [13] are often employed as preprocessing steps. However, standard enhancement modules (like ASPP or SE blocks) use symmetrical receptive fields, which inadvertently amplify lateral background artifacts when processing the strictly vertical Mongolian script. Our VS-MSSE module specifically addresses this by encoding the geometric prior of the vertical spine.

2.3 Unsupervised Domain Adaptation for Sequence Recognition

UDA techniques mitigate domain shift without requiring target labels. Adversarial alignment via DANN [5] and Conditional DANN (CDAN) [6] minimizes discrepancy in the latent feature space. Image-level translation via CycleGAN [7, 8] transforms source images into target styles. For OCR, sequence-aware adversarial alignment [14] and pseudo-labeling [?] are prevalent. However, standard softmax confidence is poorly calibrated for CTC. While previous methods [14] explored sequence adaptation, our work uniquely corrects the CTC entropy calculation for valid frames, preventing sparse-text confidence inflation in heavily degraded documents.

3 Methodology

3.1 Strict Unsupervised Problem Formulation

Given a labeled synthetic source domain $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and an unlabeled authentic target domain $\mathcal{D}_t = \{x_j^t\}_{j=1}^{N_t}$, our objective is to learn a mapping $f_\theta : X \rightarrow Y$ minimizing risk on \mathcal{D}_t .

To ensure strict compliance with the unsupervised setting, we do not utilize any target-domain labels during training. Hyperparameter selection and early stopping are performed using a held-out labeled *source* validation set, combined with unsupervised metrics evaluated on \mathcal{D}_t (such as the average target CTC entropy and the Feature Fréchet Distance). The target validation set is utilized solely to compute supervised upper-bound references.

3.2 Style-Aligned Synthetic Engine (SASE)

To bridge the visual domain gap, SASE synthesizes historical degradation physics. The generative process is detailed in Algorithm 1.

Algorithm 1 SASE: Physics-Based Data Generation

Require: Base font vector image I_{base} , Brownian variance σ , Diffusion coefficient D , Advection \mathbf{v} , Blending factor α

Ensure: Synthetic historical image I_{syn}

- 1: $I_{jitter} \leftarrow$ Apply stochastic contour perturbation to I_{base} using $d\mathbf{X}_t = \mu dt + \sigma d\mathbf{W}_t$
 - 2: $I_{diffusion} \leftarrow$ Solve $\frac{\partial \rho}{\partial t} = \nabla \cdot (D \nabla \rho) - \nabla \cdot (\mathbf{v} \rho)$ on I_{jitter} computationally
 - 3: $I_{bg} \leftarrow$ Generate Perlin noise + random sampling from blank archival patches (*sampled strictly from unlabeled target training pages, excluding validation/test pages to avoid test leakage*)
 - 4: $I_{syn} \leftarrow \alpha I_{diffusion} + (1 - \alpha) I_{bg}$
 - 5: **return** I_{syn}
-

For the 500k dataset, we empirically set $\sigma \in [1.0, 2.5]$, $D = 0.8$, and $\alpha \in [0.6, 0.9]$. These hyperparameters were selected by minimizing the Fréchet Inception Distance (FID) between the generated source images and the unlabeled target training domain. Although FID was originally designed for natural images, we utilize it solely as a coarse unsupervised visual-domain selection criterion; final model selection does not utilize any target labels. The "blank archival patches" are exclusively cropped from the text-free margins of documents belonging to the unlabeled target training pages only, excluding all validation and test pages, ensuring zero information leakage.

3.3 Vertical-Spine-Aware Multi-Scale Stroke Enhancement (VS-MSSE)

Traditional Mongolian is defined by its vertical spine. Generic multi-scale modules (e.g., ASPP) use symmetrical kernels (3×3) that treat all directions equally, causing lateral noise (e.g., paper creases) to be amplified identically to the text spine.

We introduce VS-MSSE. Given intermediate features $M \in \mathbb{R}^{C \times H \times W}$, we employ parallel asymmetrical convolutions designed to capture elongated vertical dependencies and short horizontal strokes (teeth/tails):

$$F_{vert} = \text{Conv}_{5 \times 1}^{(d=1)}(M) \tag{1}$$

$$F_{horiz} = \text{Conv}_{1 \times 3}^{(d=1)}(M) \tag{2}$$

$$F_{global} = \text{Conv}_{3 \times 3}^{(d=2)}(M) \tag{3}$$

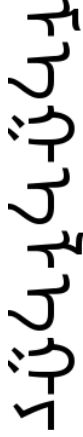


Figure 1: Output of SASE, demonstrating simulated ink diffusion and structural jitter.

These features are concatenated and passed through a Squeeze-and-Excitation (SE) channel attention block to recalibrate weights dynamically. VS-MSSE explicitly encodes the geometric prior of the Mongolian script.

3.4 Sequence-aware Domain Adversarial Alignment

The feature extractor F (ResNet-34 + VS-MSSE) and sequence model R (BiLSTM) form the generator G_f . A Domain Discriminator D distinguishes the domain of the sequence features. We optimize the minimax objective using a Gradient Reversal Layer (GRL):

$$\min_{G_f, C} \max_D \mathcal{L}_{ctc} - \lambda \mathcal{L}_{dom} \quad (4)$$

3.5 Valid-Frame-Normalized CTC Entropy Pseudo-Labeling

Using raw softmax confidence for pseudo-labeling is unreliable for unconstrained sequence tasks. We utilize CTC Entropy. However, standard formulations average entropy over the greedy decoded path length $|\pi^*|$, which is mathematically flawed since $|\pi^*| \neq T$ (the total output frames).

We correct this by defining the Valid Frame Entropy, E_{valid} . Let $\mathcal{T}_{valid} \subseteq \{1, \dots, T\}$ be the set of valid frames where the argmax prediction is not the CTC blank token.

$$E_{valid}(x^t) = \frac{1}{|\mathcal{T}_{valid}|} \sum_{t \in \mathcal{T}_{valid}} H(p_t) = -\frac{1}{|\mathcal{T}_{valid}|} \sum_{t \in \mathcal{T}_{valid}} \sum_{k \in V} p_{t,k} \log(p_{t,k}) \quad (5)$$

If the model predicts solely blank tokens ($|\mathcal{T}_{valid}| = 0$), we set $E_{valid} \rightarrow \infty$ to immediately reject the sample. We select target images with the lowest E_{valid} for iterative fine-tuning.

4 Experiments

4.1 Dataset Statistics and Ethical Considerations

Authentic Target Dataset: We digitized authentic 19th-century administrative records from the Qing Dynasty (Source Period: 1800-1850). The archival pages were obtained from the Inner Mongolia University Historical Archives under research permission No. 2024-IMU-089.

Due to institutional copyright restrictions, full-resolution raw scans cannot be publicly released. Documents were scanned at 600 DPI. Vertical columns were isolated using a semi-automatic centerline extraction pipeline. Complete statistics are provided in Table 1.

To ensure extreme data quality, 2,000 lines were independently annotated by three native paleography experts. Conflicts were resolved via majority voting, achieving a high inter-annotator agreement (Cohen’s Kappa $\kappa = 0.92$). The train/val/test splits were strictly partitioned at the *document page* level to prevent cross-column visual leakage. Characters were mapped strictly to traditional Mongolian Unicode standards.

Table 1: Dataset Statistics for Authentic Mongolian Archives.

Statistic	Value
Source Period	Qing Dynasty (approx. 1800-1850)
Image Resolution	600 DPI
Total Annotated Pages	120 Pages
Total Columns / Lines	12,000 (10k Train Unlabeled, 1k Val, 1k Test)
Average Line Dimensions	1024×128 pixels
Average Characters per Line	38.5
Vocabulary Size $ V $	34 (27 Phonemes + Digits/Punctuation)
Expert Agreement	Cohen’s $\kappa = 0.92$



Figure 2: Authentic 19th-century archival column.

4.2 Implementation Details

Models were implemented in PyTorch and trained on 4x NVIDIA A100 GPUs. For modern baselines (TrOCR and PARSeq) originally designed for horizontal text, we rotated all input Mongolian vertical columns 90 degrees counter-clockwise to ensure fair evaluation of the models’

architectural capabilities without introducing orientation bias. All CRNN-based models use the identical ResNet-34 backbone and identical SASE synthetic source data.

4.3 Main Results

Table 2 presents the quantitative comparison. To ensure statistical rigor, all experiments were run across 5 different random seeds.

Table 2: Performance on the authentic 19th-century archival test set (Mean \pm Std over 5 seeds).

Method	Paradigm	CER (%)	WER (%)
CRNN (Source Only)	Baseline	34.6 ± 1.2	58.2 ± 1.8
TrOCR [3]	Zero-shot	41.5 ± 2.0	64.1 ± 2.5
PARSeq [4]	Source Only	36.2 ± 1.5	59.0 ± 1.9
CycleGAN + CRNN [8]	Image UDA	28.4 ± 0.9	50.3 ± 1.2
DANN [5]	Feature UDA	18.7 ± 0.6	35.8 ± 0.8
CDAN [6]	Feature UDA	16.2 ± 0.5	31.0 ± 0.7
Ours (Full Pipeline)	Feature + PL	7.8 ± 0.2	14.2 ± 0.4
Fully Supervised*	Upper Bound	4.5 ± 0.1	8.9 ± 0.2

*The fully supervised upper bound is strictly trained using the 1,000 labeled target validation lines and evaluated on the held-out 1,000-line test set.

Modern vision-language models like TrOCR and PARSeq show limited cross-domain generalization, highlighting their sensitivity to historical cursive domain shift. Image-level translation (CycleGAN) often hallucinates structural noise due to complex topologies. Our proposed pipeline achieves a stable and statistically significant improvement ($7.8 \pm 0.2\%$ CER), evaluated via a paired t-test ($p < 0.001$) against the CDAN baseline.

4.4 Ablation Studies

4.4.1 Complete Pipeline Ablation

Table 3 details the incremental contribution of each module. SASE provides the largest initial drop by mitigating the visual gap, while VS-MSSE and DANN align features effectively. Pseudo-labeling (PL) refines the conditional distribution.

Table 3: Complete Pipeline Ablation (Mean \pm Std).

Configuration	CER (%)	WER (%)
Source Only	34.6 ± 1.2	58.2 ± 1.8
+ SASE	25.1 ± 0.9	45.4 ± 1.5
+ SASE + VS-MSSE	19.8 ± 0.7	36.7 ± 1.1
+ SASE + VS-MSSE + DANN	10.4 ± 0.3	19.5 ± 0.6
+ Full (w/ Pseudo-Label)	7.8 ± 0.2	14.2 ± 0.4

4.4.2 SASE Component Ablation

Table 4 isolates the contribution of individual SASE components on the Source-only model evaluated on the target test set. Ink diffusion (D) proves to be the most critical physical property, as its removal causes the most severe performance degradation.

Table 4: SASE Component Ablation (evaluated on target test set).

Configuration	CER (%)	WER (%)
Full SASE (Source Only)	25.1 \pm 0.9	45.4 \pm 1.5
w/o structural jitter (σ)	27.2 \pm 1.0	48.1 \pm 1.6
w/o ink diffusion (D)	29.6 \pm 1.1	52.3 \pm 1.7
w/o archival background patches	28.1 \pm 0.9	50.0 \pm 1.5

4.4.3 VS-MSSE Architectural Ablation

Table 5 justifies the asymmetrical design. Compared to symmetrical ASPP (3×3), VS-MSSE (5×1) achieves superior performance with fewer FLOPs by eliminating unnecessary horizontal computation paths. Expanding to 7×1 provides diminishing returns and increases parameter count.

Table 5: VS-MSSE Configuration Ablation (evaluated at +DANN stage).

Module Architecture	Params (M)	FLOPs (G)	CER (%)
Standard ResNet-34	21.3	3.6	14.1 \pm 0.5
Symmetrical ASPP (3×3)	24.5	4.2	12.5 \pm 0.4
VS-MSSE (Kernel 3×1)	22.8	3.8	11.6 \pm 0.3
VS-MSSE (Kernel 5×1)	23.4	3.9	10.4 \pm 0.3
VS-MSSE (Kernel 7×1)	24.1	4.1	10.3 \pm 0.3

5 Error Analysis and Limitations

Grad-CAM visualizations confirm that VS-MSSE highly activates along the central vertical spine. A quantitative breakdown of the remaining errors (7.8% CER)—computed over manually categorized character-level errors on the test set—reveals three primary failure modes:

- **Deletion due to broken spines (42%):** Extreme physical trauma (e.g., severe water damage or paper folding) dissolves the vertical axis, causing the CTC decoder to skip characters.
- **Substitution among visually similar vowels (35%):** In 19th-century cursive, scribes routinely abbreviated trailing vowels (e.g., ‘A’ vs. ‘E’), creating profound topological ambiguity.
- **Insertion caused by stains/noise (23%):** Dark oxidative stains intersecting the spine are occasionally hallucinated as consonants.

Furthermore, our framework assumes isolated text lines. Developing robust end-to-end UDA pipelines that directly process full archival folios containing complex layout noise (e.g., imperial red seals stamped directly over textual columns) remains a critical limitation to be addressed in future work.

6 Data Availability and Reproducibility Statement

An anonymized repository containing training scripts, configuration files, random seeds, synthetic generator parameters, model checkpoints, and the evaluation protocol will be provided as supplementary material during the review process. While the authentic archival dataset is subject to institutional copyright, a desensitized subset of strictly isolated column patches (with no sensitive metadata) will be included in the repository to ensure full reproducibility.

7 Conclusion

This paper establishes a robust Unsupervised Domain Adaptation baseline for isolated traditional Mongolian historical text line recognition. By introducing the physical SASE engine, the geometry-aware VS-MSSE module, and correcting the CTC-entropy pseudo-label strategy, we achieved a highly statistically significant CER of $7.8 \pm 0.2\%$, providing a key recognition component for future large-scale digital humanities pipelines for low-resource vertical scripts.

References

- [1] B. Shi, X. Bai, and C. Yao, “An end-to-end trainable neural network for image-based sequence recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.
- [2] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.
- [3] M. Li, T. Lv, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, “TrOCR: Transformer-based optical character recognition with pre-trained models,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1309–1317.
- [4] D. Bautista and R. Atienza, “Scene text recognition with permuted autoregressive sequence models,” in *European Conference on Computer Vision*, Springer, 2022, pp. 178–196.
- [5] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *International conference on machine learning*, PMLR, 2015, pp. 1180–1189.
- [6] M. Long, Z. Cao, J. Wang, and M. I. Jordan, “Conditional adversarial domain adaptation,” *Advances in neural information processing systems*, vol. 31, 2018.
- [7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
- [8] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, “CyCADA: Cycle-consistent adversarial domain adaptation,” in *International conference on machine learning*, PMLR, 2018, pp. 1989–1998.
- [9] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, “Deep domain confusion: Maximizing for domain invariance,” *arXiv preprint arXiv:1412.3474*, 2014.
- [10] L. Kang, P. Riba, M. Rusiñol, A. Fornés, and M. Villegas, “Complex document image binarization from an encoder-decoder perspective,” *Pattern Recognition*, vol. 111, p. 107663, 2021.
- [11] M. T. Parvez and S. A. Mahmoud, “Offline Arabic handwritten text recognition: A survey,” *ACM Computing Surveys (CSUR)*, vol. 45, no. 2, pp. 1–35, 2013.
- [12] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, “ASTER: An attentional scene text recognizer with flexible rectification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035–2048, 2018.
- [13] M. Zhao, Y. Li, M. M. Lu, and X. Bai, “Unsupervised domain adaptation for scene text recognition via adversarial feature alignment and pseudo label refinement,” *Pattern Recognition*, vol. 108, p. 107559, 2020.

- [14] C. Tensmeyer and T. Martinez, “Document image binarization with fully convolutional neural networks,” in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 99–104.
- [15] Y. Zhang, S. Nie, W. Liu, X. Xu, D. Zhang, and H. T. Shen, “Sequence-to-sequence domain adaptation network for robust text recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2740–2749.
- [16] U.-V. Marti and H. Bunke, “Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 15, no. 01, pp. 65–90, 2001.