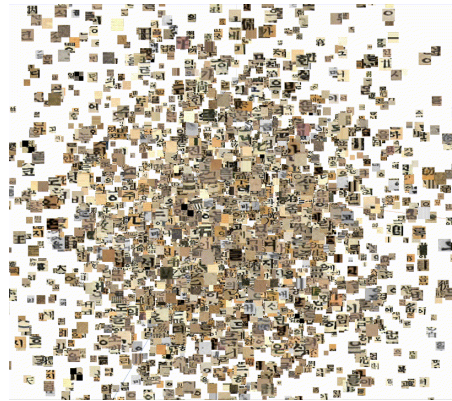


Visual Feature Discovery in Colonial Korean Print using MIL

DH2025



Aron van de Pol, Jelena Prokic, Angus Mol

Leiden University Centre for Digital Humanities & Department of Korean Studies

2025-07-16



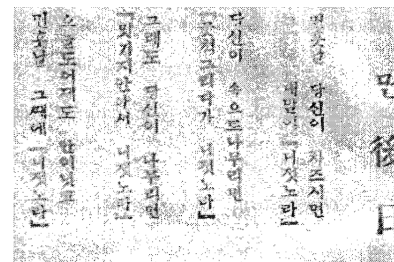
Colonial Korea

- Colonial Korea (1910-1945)
- 1910s no free press
- March 1st Protests incited change
- More liberal press policies in the colony

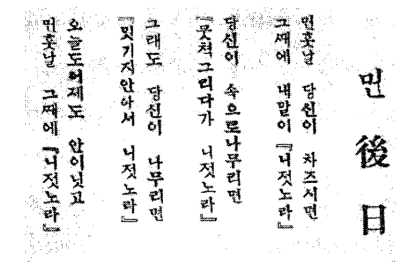


The 1920s

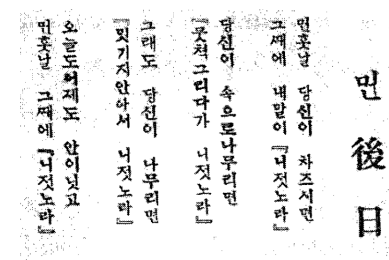
- During the 1920s, choice of 100-200 different print shops.¹
- Choice influenced outcome.
- A good example are poems from Kim So-wŏl's *Chindallaekkot* 진달래꽃 collected works.



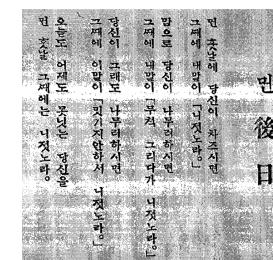
(a) Hansong Toso 漢城圖書 issue of *Chindallaekkot* (collected works)



(b) Chungang Sorim 中央書林 issue of *Chindallaekkot* (collected works)



(c) Haksanggye 學生界 (July 1920) issue of the poem.



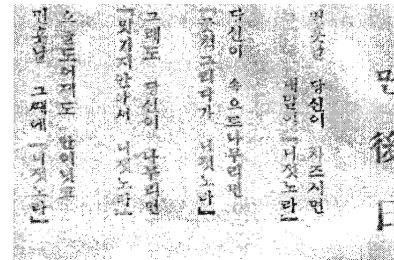
(d) Kaebok 開闢 (August 1922) issue of the poem.

Figure 1: Some Day Long From Now 민 후일 poem from various print runs

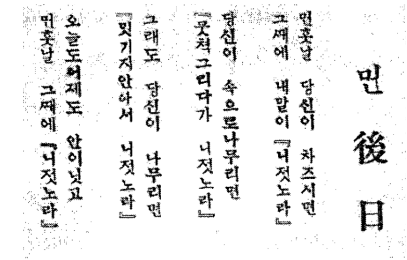


The 1920s

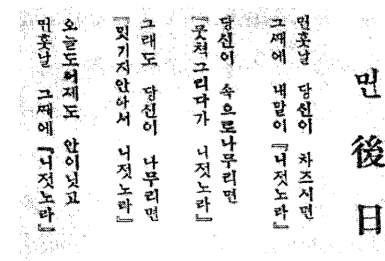
- During the 1920s, choice of 100-200 different print shops¹
- Choice influenced outcome
- A good example are poems from Kim So-wŏl's *Chindallaekkot* *진달래꽃* collected works



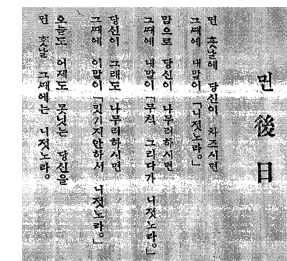
(a) Hansong Toso 漢城圖書 issue of *Chindallaekkot* (collected works)



(b) Chungang Sorim 中央書林 issue of *Chindallaekkot** (collected works)



(c) Haksaenggye 學生界 (July 1920) issue of the poem.



(d) Kaebok 開闢 (August 1922) issue of the poem.

Figure 2: *Some Day Long From Now* 민 후 日 poem from various print runs.



Characteristics?

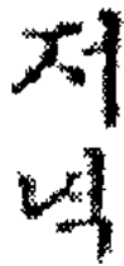
- Work done by De Fremery¹



(a) *Hansong Toso Chusik Hoeisa printed Tang 당*



(b) *Taedong Inswaesong printed Tang 당*



(c) *Hansong Toso Chusik Hoeisa printed Tang 당*



(d) *Taedong Inswaesong printed Tang 당*

Figure 3: Examples of subtle typographic differences. Notice the angle of the stroke on *Tigüt* ㄷ and the differences in the *Chiüt* ㅈ



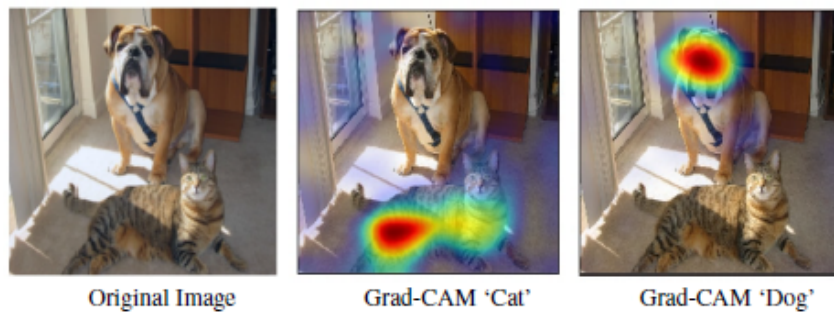
Research

Can neural networks be used to classify historical print shops and identify the specific visual features that distinguish their typographic styles?

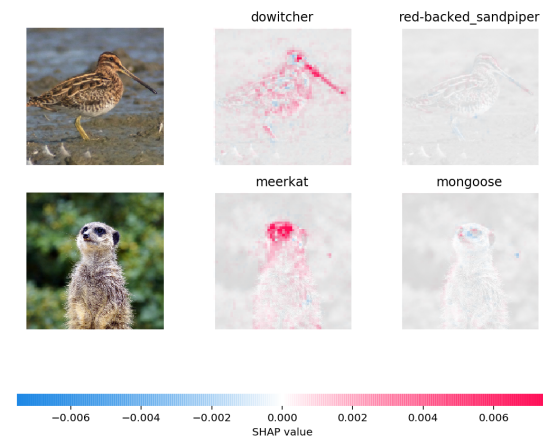


Interpretability

- All studies aim to detect, but *how* a model detects is often neglected
 - For CNNs and Vision Transformers, several interpretability methods have proven successful
1. GradCAM and derivatives¹
 2. SHAP (SHapley Additive exPlanations)²



(a) Example of GradCAM



(b) Example of SHAP



Figure 4: Interpretability methods visualized

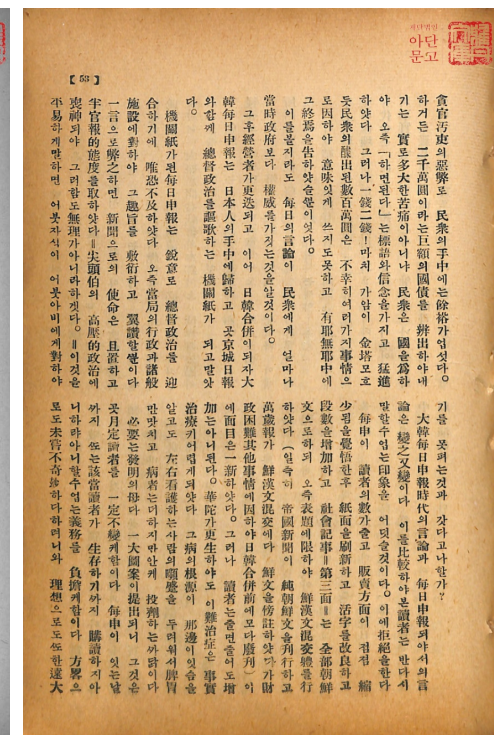
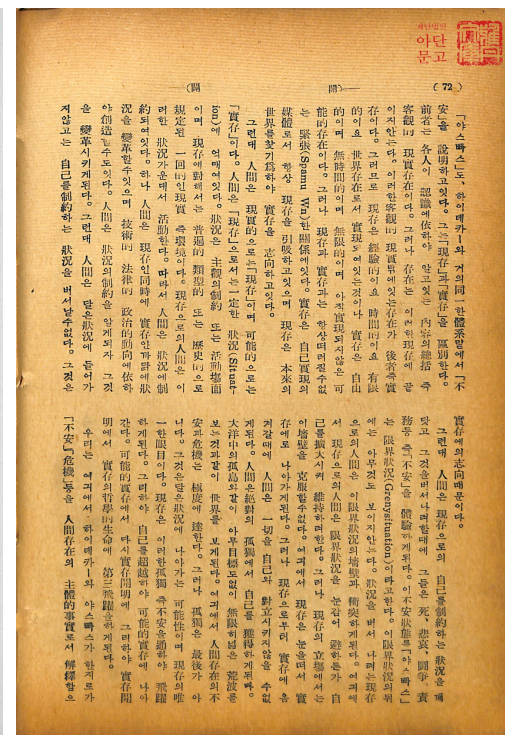


Dataset

Printshop	Printshop (KR)	Pages	Percentage
Taedong Inswaeso	大東印刷所	27,882	42.07%
Hansŏng Toso Chusik Hoeisa	漢城圖書 株式會社	19,244	29.04%
Sinmungwan	新文館	13,050	19.69%
Chosŏn Inswae Chusik Hoeisa	朝鮮印刷株式會社	6,101	9.20%
Total		66,277	100%

- **Class Imbalance:** This distribution reflects real-world production volumes.
- We Chose not to implement class rebalancing techniques.





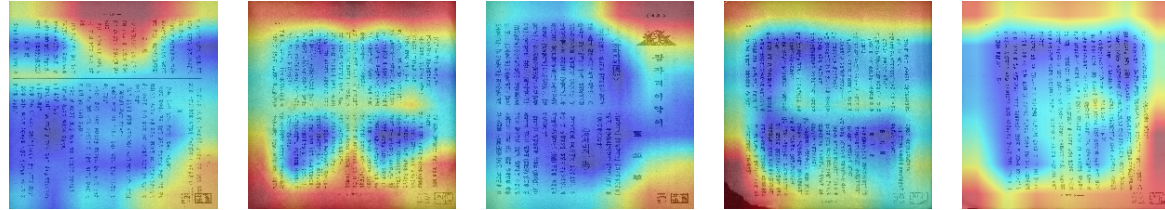
Examples of pages in the dataset.



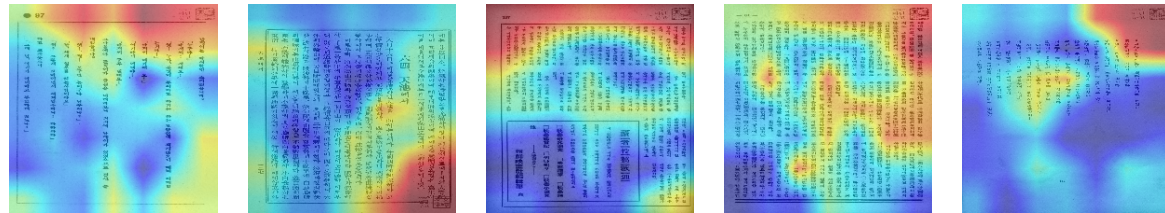
Results Approach 1

ConvNext Base architecture - 98% Accuracy (F1=0.98)

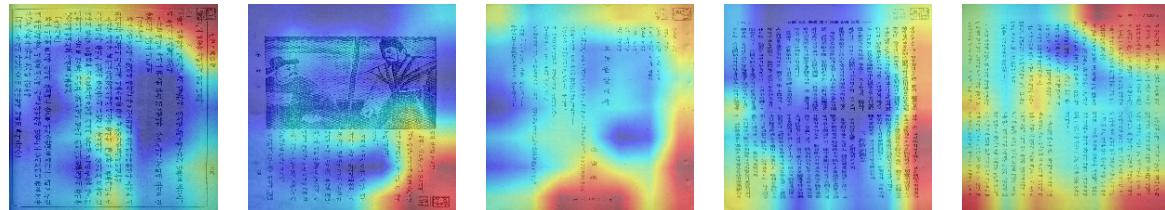
Chosŏn Inswae Chusik Hoeisa



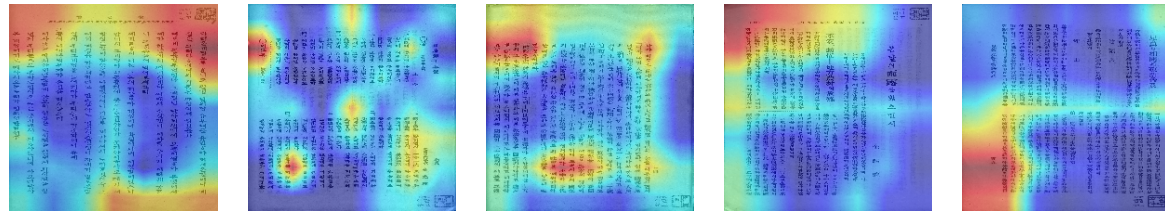
Hansŏng Toso Chusik Hoeisa



Taedong Inswaesŏ



Sinmungwan

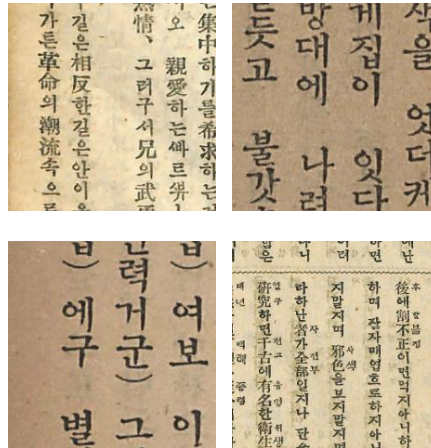


Can neural networks be used to classify historical printshops and identify the specific visual features that distinguish their typographic styles?



Approach 2

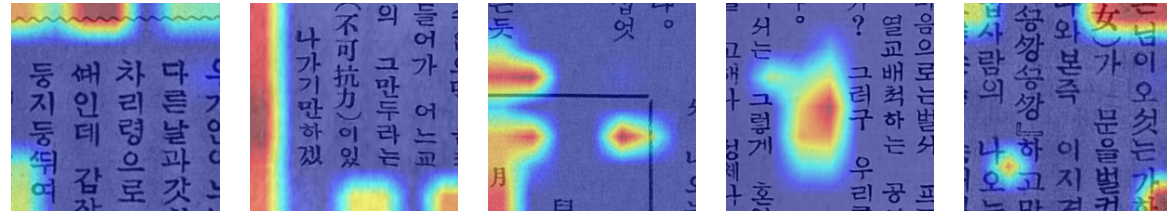
Following idea of Seuret et al.¹ a page is cut into 4 random cutouts, while reducing overlap to max 30%



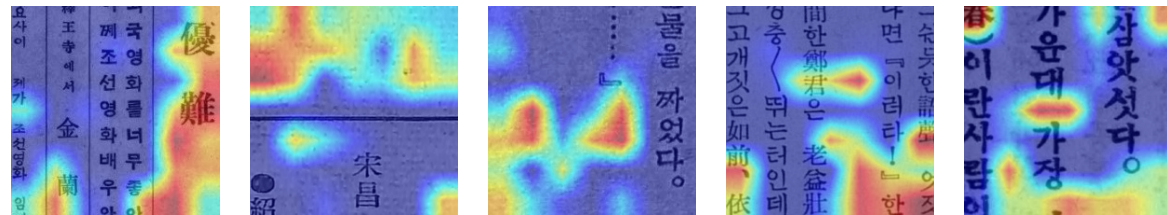
Approach 2 Results

99.8% Accuracy (F1=0.99) Swin S3 Base-224

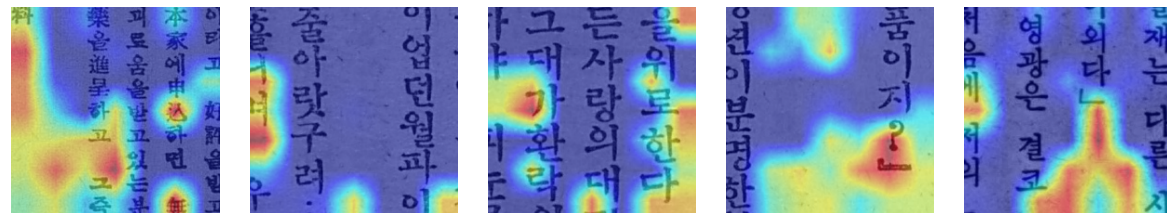
Chosŏn Inswae Chusik Hoeisa



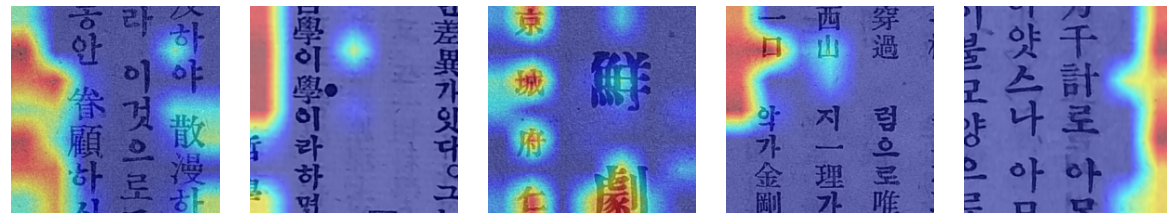
Hansŏng Toso Chusik Hoeisa



Taedong Inswaesŏ



Sinmungwan



Can neural networks be used to classify historical printshops and identify the specific visual features that distinguish their typographic styles?



MIL

- *Multi Instance Learning*.¹
- Used in the field of medical imagery.²
- Similar issues faced by humanists:
 - Retrieve model's decision making
 - Interpretably decision making
- We follow the *AttriMIL* implementation of Cai et al.³

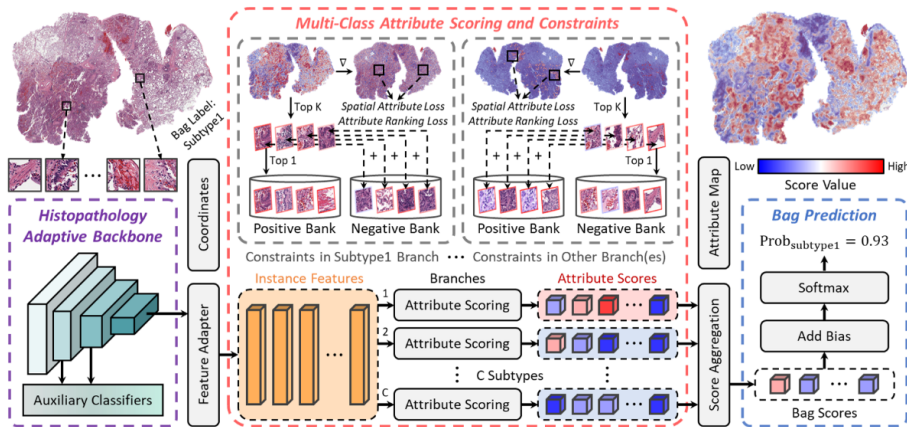


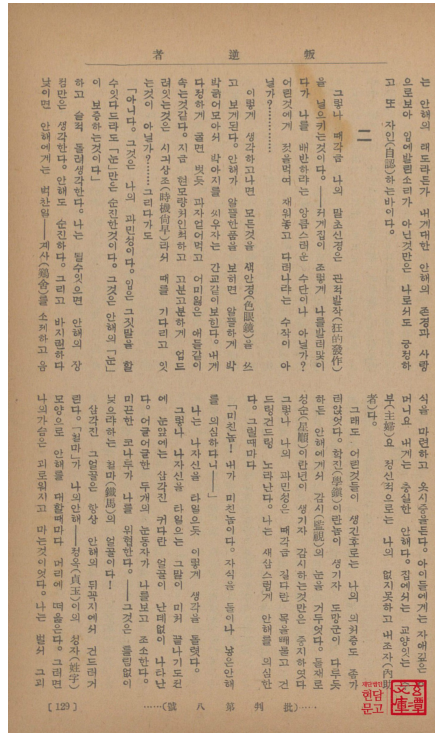
Figure 5: MIL visualization.⁴

1. Maron and Lozano-Pérez, "A Framework for Multiple-Instance Learning".
 2. Papadopoulos, Topouzis, and Delopoulos, "An Interpretable Multiple-Instance Approach for the Detection of Referable Diabetic Retinopathy from Fundus Images"; Javed et al., "Additive MIL"; Yang et al., "HAMIL"; Deng et al., "Cross-Scale Multi-Instance Learning for Pathological Image Diagnosis"; Waqas et al., "Exploring Multiple Instance Learning (MIL)"; Gadermayr and Tschuchnig, "Multiple Instance Learning for Digital Pathology".

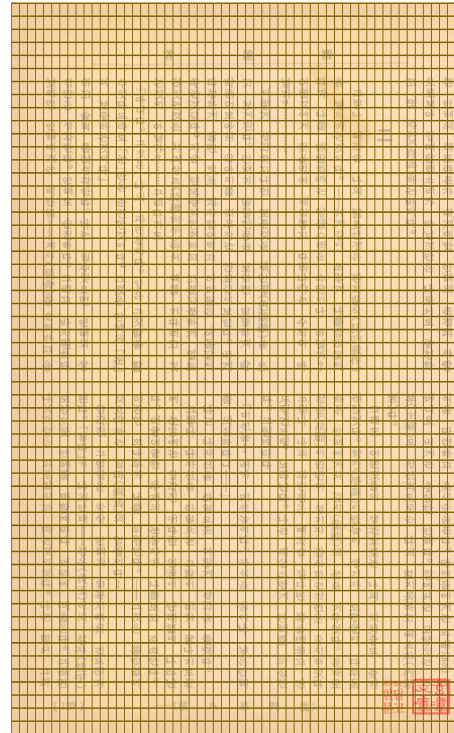
3. "Rethinking Attention-Based Multiple Instance Learning for Whole-Slide



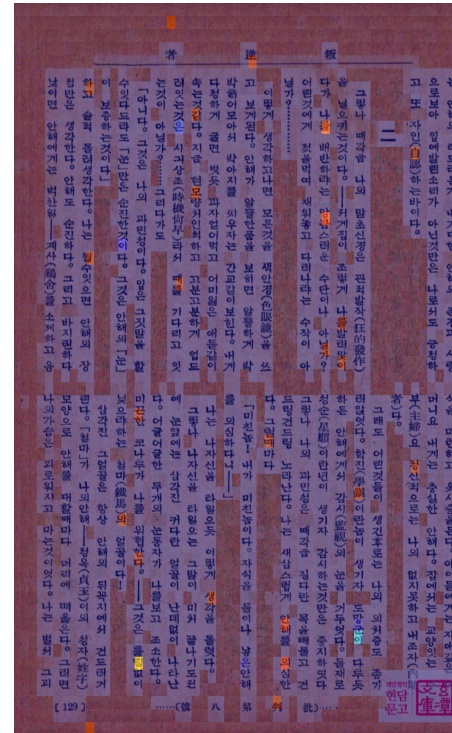
MIL Applied



(a) Original



(b) Patched



(c) Attention displayed

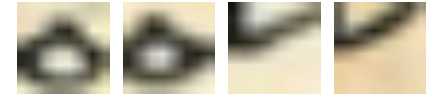


Figure 6: Applied MIL



Embeddings Space

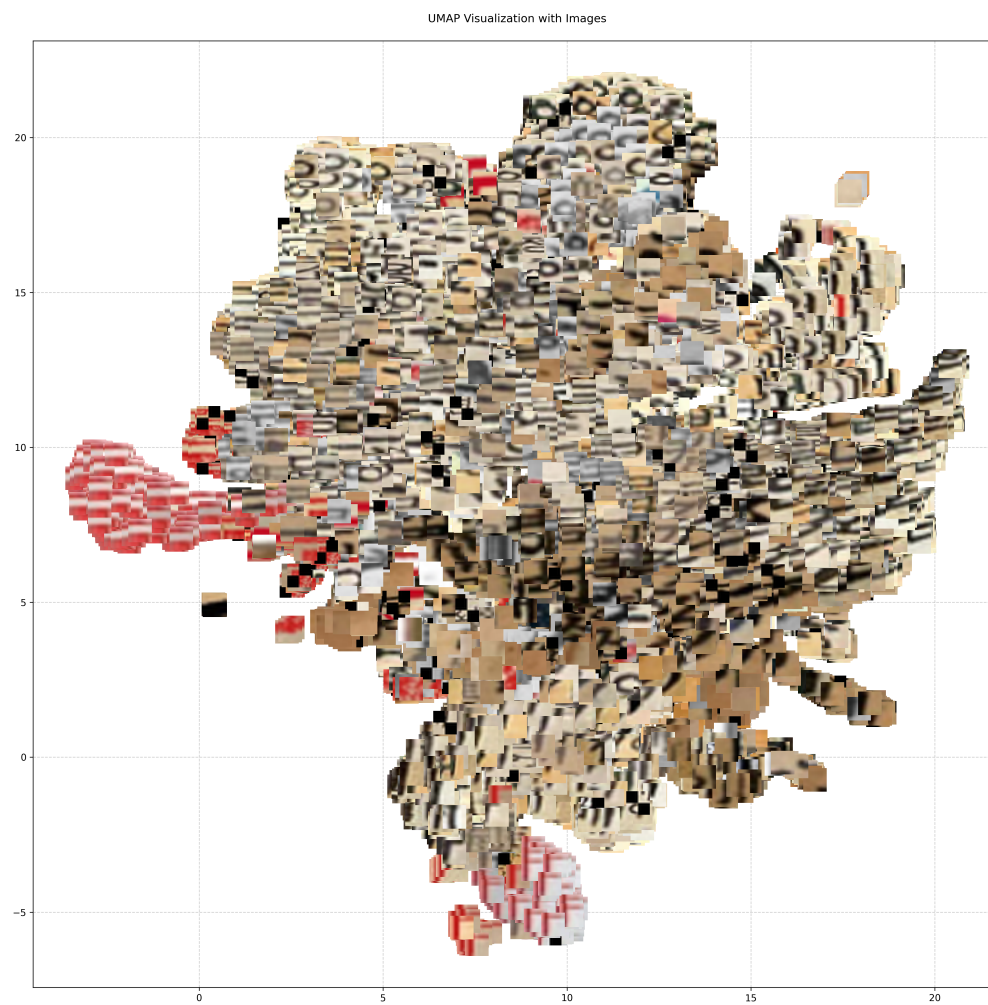


Figure 7: UMAP Visualisation of MIL Embeddings





Embeddings Space

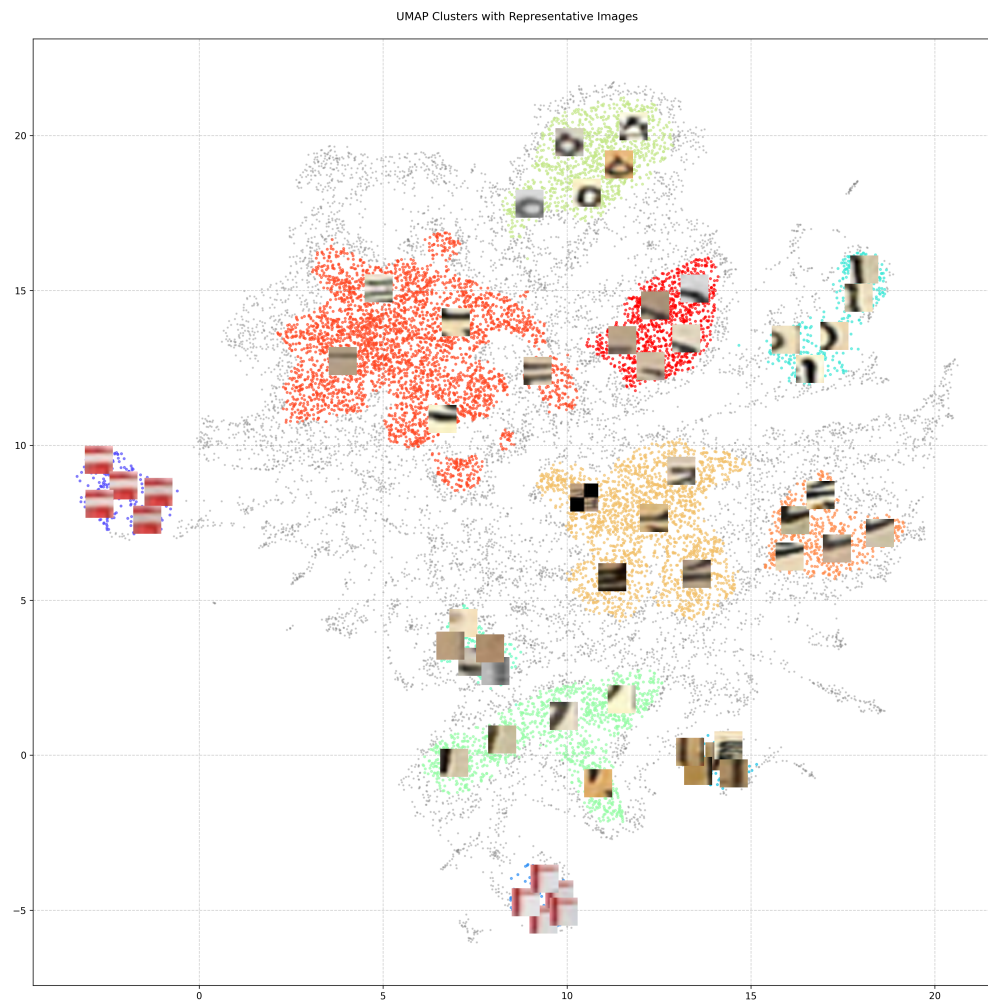


Figure 8: UMAP Visualisation of MIL Embeddings





Sampling clusters



Cluster 3 Composition
Chosŏn Insaew Chusik Hoelisa: 45.5% (105 samples)
Taedong Insaewo: 36.7% (76 samples)
Hansŏng Toso Chusik Hoelisa: 11.4% (24 samples)
Simmungwan: 6.7% (12 samples)

Chosŏn Insaew Chusik Hoelisa



Taedong Insaewo



Hansŏng Toso Chusik Hoelisa



Simmungwan



(a) Cluster3

Cluster 6 Composition
Taedong Insaewo: 54.8% (675 samples)
Hansŏng Toso Chusik Hoelisa: 18.4% (187 samples)
Chosŏn Insaew Chusik Hoelisa: 17.3% (149 samples)

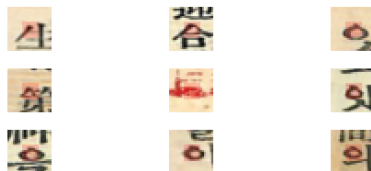
Taedong Insaewo



Hansŏng Toso Chusik Hoelisa



Simmungwan



Chosŏn Insaew Chusik Hoelisa



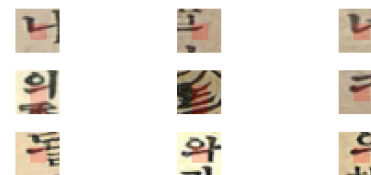
(b) Cluster 6

Cluster 8 Composition
Taedong Insaewo: 54.8% (675 samples)
Simmungwan: 28.7% (105 samples)
Hansŏng Toso Chusik Hoelisa: 11.4% (24 samples)
Chosŏn Insaew Chusik Hoelisa: 1.2% (14 samples)

Taedong Insaewo



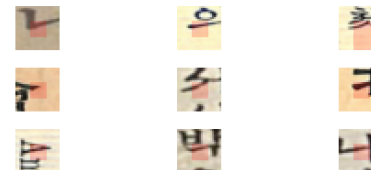
Simmungwan



Hansŏng Toso Chusik Hoelisa



Chosŏn Insaew Chusik Hoelisa



(c) Cluster 8

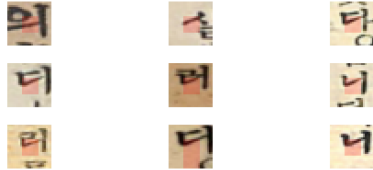
Figure 9



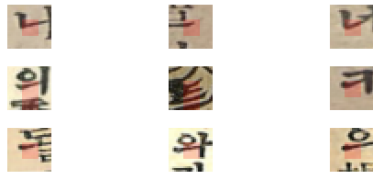
Compared

Cluster II Composition
 Taedong Inswaesŏ 14.7% (133 samples)
 Sinmungwan 24.5% (125 samples)
 Hansong Toso Chusik Hoeisa 11.2% (134 samples)
 Chosŏn Inswae Chusik Hoeisa 5.2% (14 samples)

Taedong Inswaesŏ



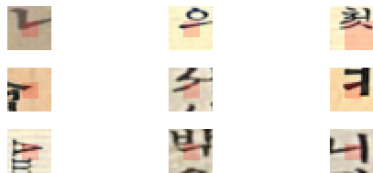
Sinmungwan



Hansong Toso Chusik Hoeisa



Chosŏn Inswae Chusik Hoeisa



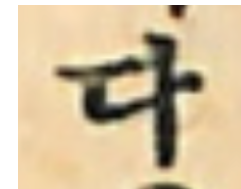
(a) Chosŏn Inswae Chusik Hoeisa



(b) Hansong Toso Chusik Hoeisa



(c) Taedong Inswaesŏ



(d) Sinmungwan

Figure 10: Sharp Stroke of ㄷ *Tigūt* as feature for Taedong Inswaesŏ / Sinmungwan. Also prevalent in ㄴ *Nieun* and ㄹ *Lieul*



Features over Time

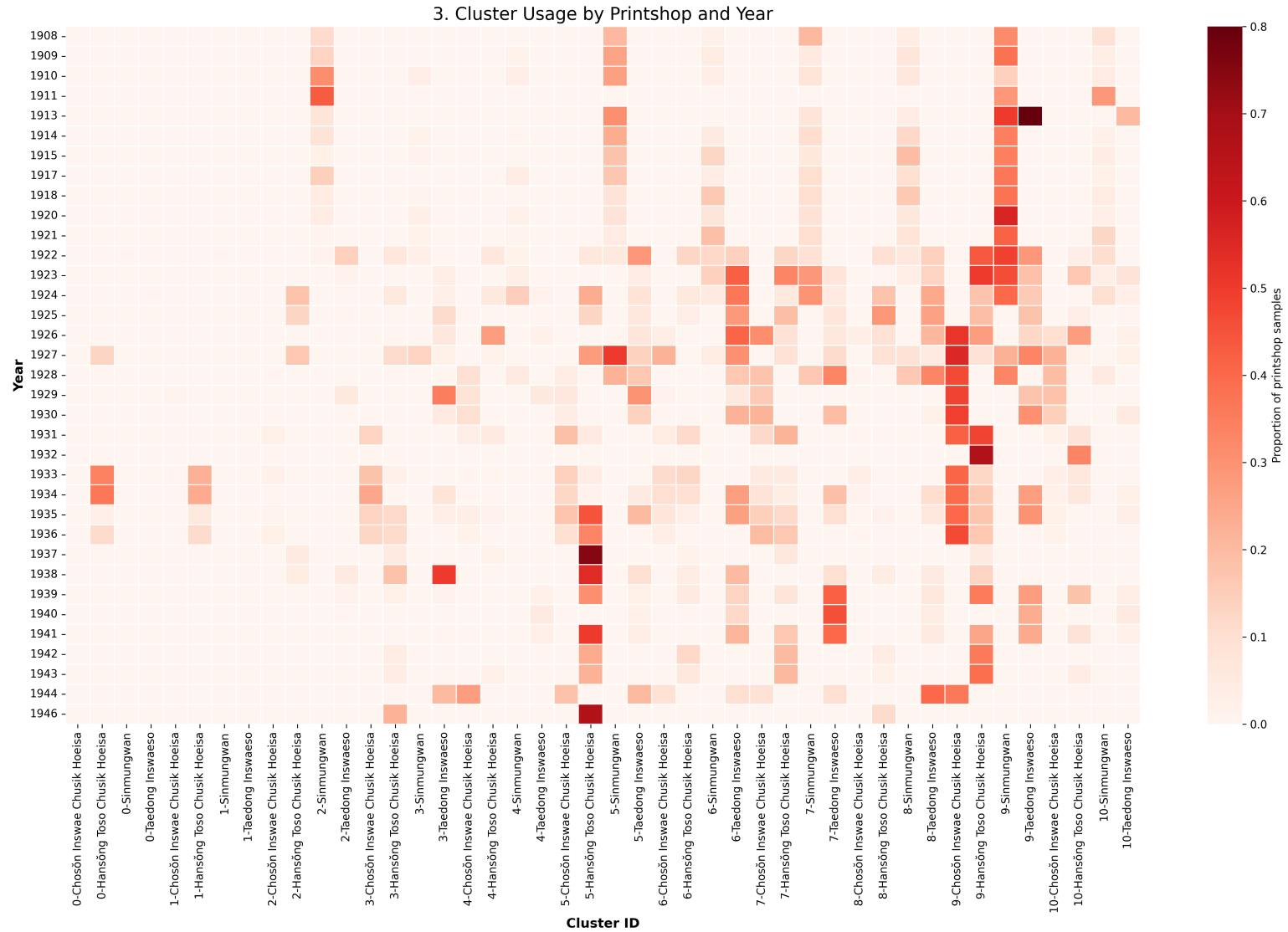


Figure 11: Heatmap of clusters division over time.





Shifts in Features

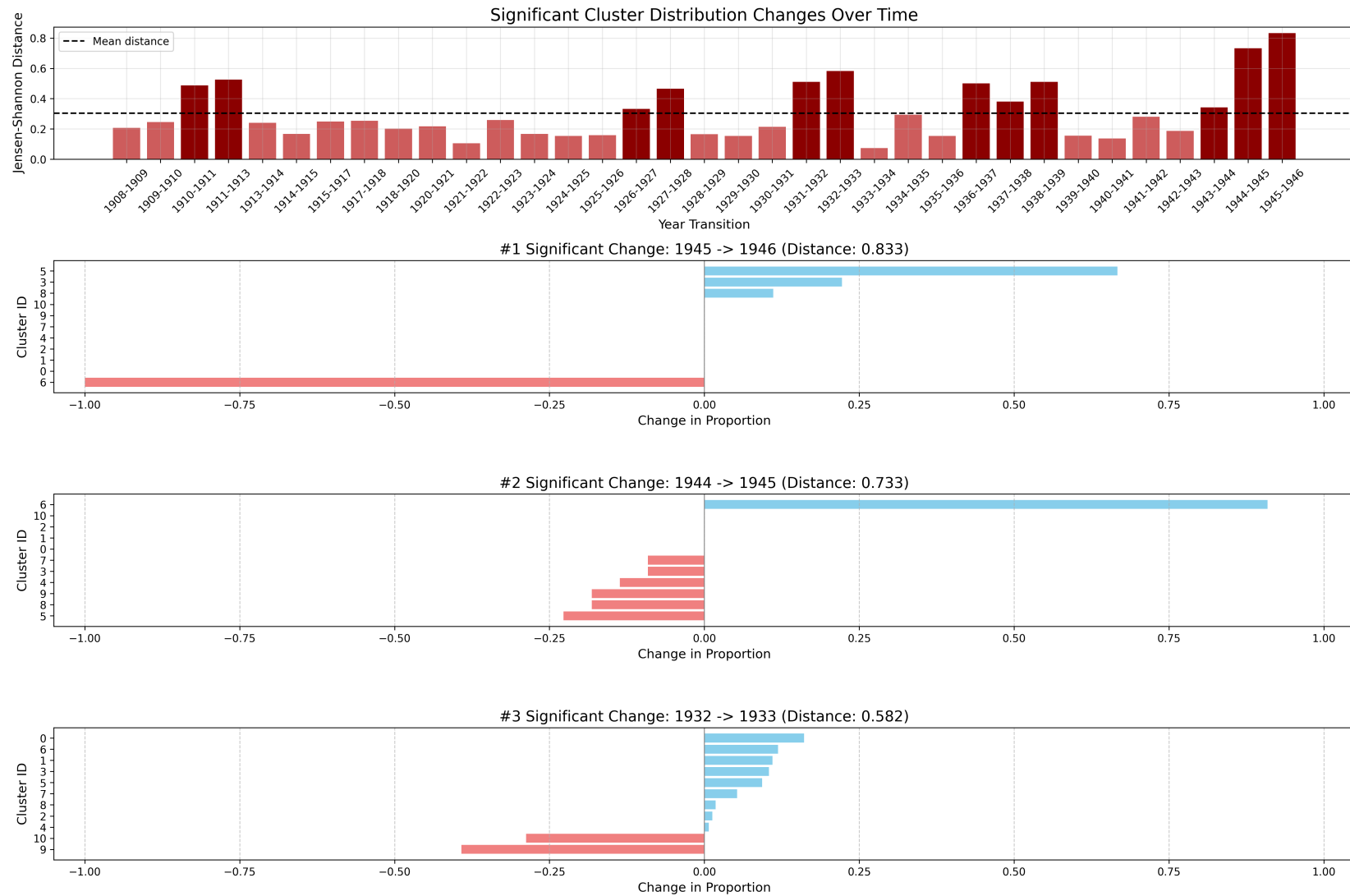


Figure 12: Cluster Shifts



Further

- Bags of Patches as feature, not singular patch
- Improvements on Clustering
- Move to typology of printshop



Thank you

Referenced Works

- Cai, Linghan, Shenjin Huang, Ye Zhang, Jinpeng Lu, and Yongbing Zhang. "Rethinking Attention-Based Multiple Instance Learning for Whole-Slide Pathological Image Classification: An Instance Attribute Viewpoint." arXiv, March 2024. <https://arxiv.org/abs/2404.00351>.
- De Fremery, Peter Wayne. "How Poetry Mattered in 1920s Korea." PhD thesis, Harvard University, 2011.
- Deng, Ruining, Can Cui, Lucas W. Remedios, Shunxing Bao, R. Michael Womick, Sophie Chiron, Jia Li, et al. "Cross-Scale Multi-Instance Learning for Pathological Image Diagnosis." *Medical Image Analysis* 94 (May 2024): 103124. <https://doi.org/10.1016/j.media.2024.103124>.
- Gadermayr, Michael, and Maximilian Tschuchnig. "Multiple Instance Learning for Digital Pathology: A Review of the State-of-the-Art, Limitations & Future Potential." *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society* 112 (March 2024): 102337. <https://doi.org/10.1016/j.compmedimag.2024.102337>.
- Hyundam Mun'go Foundation. "Hyundam Mun'go Collection." Archive, 2021.
- Javed, Syed Ashar, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. "Additive MIL: Intrinsically Interpretable Multiple Instance Learning for Pathology." arXiv, October 2022. <https://doi.org/10.48550/arXiv.2206.01794>.
- Lundberg, Scott M, and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems* 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–74. Curran Associates, Inc., 2017.
- Maron, Oded, and Tomás Lozano-Pérez. "A Framework for Multiple-Instance Learning." *Advances in Neural Information Processing Systems* 10 (1997).
- Papadopoulos, Alexandros, Fotis Topouzis, and Anastasios Delopoulos. "An Interpretable Multiple-Instance Approach for the Detection of Referable Diabetic Retinopathy from Fundus Images." *Scientific Reports* 11, no. 1 (July 2021): 14326. <https://doi.org/10.1038/s41598-021-93632-8>.
- Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization." *International Journal of Computer Vision* 128, no. 2 (February 2020): 336–59. <https://doi.org/10.1007/s11263-019-01228-7>.
- Seuret, Mathias, Saskia Limbach, Nikolaus Weichselbaumer, Andreas Maier, and Vincent Christlein. "Dataset of Pages from Early Printed Books with Multiple Font Groups." In *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing*, 1–6. HIP '19. New York, NY, USA: Association for Computing Machinery, 2019. <https://doi.org/10.1145/3352631.3352640>.
- Waqas, Muhammad, Syed Umaid Ahmed, Muhammad Atif Tahir, Jia Wu, and Rizwan Qureshi. "Exploring Multiple Instance Learning (MIL): A Brief Survey." *Expert Systems with Applications* 250 (September 2024): 123893. <https://doi.org/10.1016/j.eswa.2024.123893>.
- Yang, Yang, Yanlun Tu, Houchao Lei, and Wei Long. "HAMIL: Hierarchical Aggregation-Based Multi-Instance Learning for Microscopy Image Classification." *Pattern Recognition* 136 (April 2023): 109245. <https://doi.org/10.1016/j.patcog.2022.109245>.



Dataset

- We scraped the Hyundam Mun'go for magazines & paperbacks dating between 1900-1950.¹
- 177.101 Images of Pages.
- 14.597 Publications.
- Contributions of 202 unique print shops.
- 2552 Publishers.
- 787 Distribution outlets.



Note: These are uniquely named entities. The Publishers/
Distribution Outlets need more cleaning for a definitive
answer.

