MSc Artificial Intelligence
Master Thesis

Drawing Insights: Multi-Level Representation Learning in Comics

by
Sam Titarsolej
12206385

June 30, 2024

48 EC
October 2023 - June 2024

Supervisor:
Dr Nanne van Noord

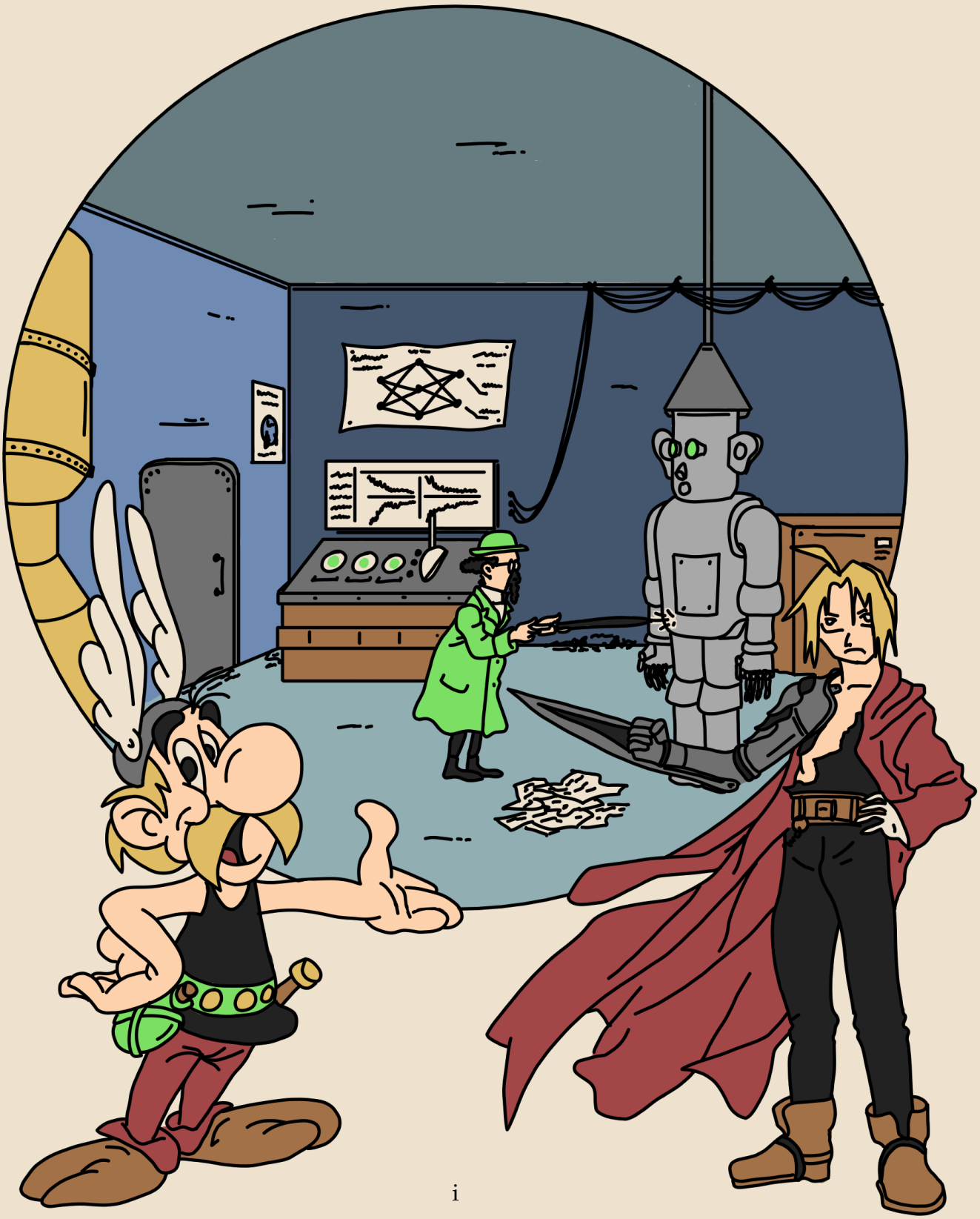
Visual Language Advisor:
Dr Neil Cohn

Examiner:
Dr Yuki M. Asano

Universiteit van Amsterdam

ASTERX AND ELRIC

THE SECRET OF REPRESENTATION LEARNING



Contents

Abstract

This thesis addresses the challenges of large-scale comic analysis by introducing an unsupervised sequential representation learning method tailored specifically for this medium. As a rich source of cultural narratives and visual storytelling, comics present significant complexity due to their intricate artwork and subtle narrative cues. Traditional analysis methods struggle to capture these nuanced visual patterns at scale. The proposed approach leverages the sequential nature of comic panels to learn contextual representations, enabling a deeper understanding of the visual language inherent in comics. Our contributions include the development of ASTERX (A Self-supervised Transformer Encoder for comic panel Representation eXtraction) and ELRIC (contExt normaLisation for Representation learning In Comics). ASTERX captures the continuity and contextual relationships within sequences of comic panels, while ELRIC integrates these contextual representations into existing self-supervised frameworks, enhancing their performance in various downstream tasks. Extensive experiments demonstrate the effectiveness of these methods in tasks such as panel retrieval and character emotion classification, highlighting their potential for machine learning-aided cultural analysis. By providing tools for the relative quantification and comparative analysis of comics, this thesis lays a foundation for systematically exploring visual storytelling across different cultures and artistic styles. Our findings offer insights into the construction of comics and their narrative structures, contributing to the field of comic studies and the broader domain of cultural analysis.

Chapter 1. Introduction

Comics offer insight into cultural narratives, visual storytelling, and the complexities of visual language and human expressiveness [58, 6, 62, 49]. They serve not only as a form of entertainment but also as a reflective mirror of societal constructs and cultural identities. Analysing these constructs at scale requires robust representations that can capture the nuanced visual patterns within comics. Machine learning-aided analysis of comics at scale has already proven an effective method in many aspects of learning specific tasks such as text extraction and character recognition [61, 3, 42]. However, these methods rely heavily on extensive annotations to perform supervised tasks. In an effort to reduce the need for such extensive annotations, while still enabling machine learning-aided analysis of comics at scale, this thesis investigates the possibilities of unsupervised representation learning within comics. To this end, we introduce two novel methods of unsupervised representation learning, specifically tailored for comics. Our first proposed method, ASTERX, leverages the sequential nature of comic panels to encode contextual information transmitted within the panel sequence. Our second method, ELRIC, builds upon the contextual representations learned by ASTERX, to learn representations for specific regions (such as character bounding boxes) within comic panels. For both of these methods, we perform extensive evaluation to understand their respective performance.

Analysing comics on a large scale presents significant challenges due to the nuanced and multifaceted nature of visual patterns within this medium [12, 13, 4]. Comic panels encompass a wide array of stylistic and semantic elements, ranging from intricate artwork to subtle narrative cues. Consequently, systematically decoding and interpreting these elements introduces numerous challenges [12, 13] such as subjectivity of stylistic elements. Our method not only aims to enhance the understanding of the visual language inherent to comics but also lays the groundwork for machine learning-aided cultural analysis of this medium. The task of dissecting comics into consistent stylistic and semantic components is inherently complex due to the integration of art and narrative in diverse and sophisticated ways [12, 13]. Additionally, the portrayal of emotions, settings, and character interactions through these styles adds another layer of complexity, making consistent quantification and annotation a formidable task.

Our approach emphasises relative quantification, developing tools for comparative analysis, such as evaluating how one comic compares to another, rather than relying on absolute measures and tasks as already explored in existing research [6, 3]. This methodology acknowledges the subjective nature of artistic mediums and aims to contextualise comics within a broader spectrum of styles and narratives, providing a comprehensive framework for their analysis.

1.1 Contributions

Our main contributions are the introduction of ASTERX and ELRIC as novel representation learning methods tailored for visual narratives. Furthermore, we explore many aspects of the capabilities of ASTERX and ELRIC through extensive evaluation, ablation studies and qualitative analyses. Finally, we also apply learned representations in a cultural analysis. As such, we summarise the contributions in this thesis as listed below.

1. A novel framework is introduced for learning representations of comic panels using unsupervised sequential representation learning. This framework captures the nuanced visual patterns within comics, enabling a deeper understanding of their multifaceted nature.

2. We propose a method to integrate learned representations into the training processes of existing self-supervised methods. This technique enhances the performance of these methods in various downstream tasks by reducing the contextuality of the representations.

3. Extensive experiments demonstrate an increased performance over existing self-supervised methods in comic representation learning. Our evaluation includes tasks such as panel retrieval and character emotion classification.

4. By systematically analysing the learned representations, we uncover underlying patterns in how comics are constructed and their narrative structures. This analysis extends beyond identifying basic stylistic elements, offering insights into the differences in visual storytelling across cultures. Many of these insights are in line with leading theories in linguistics and inference theory.

Chapter 2. Related Work

This chapter surveys existing methods of quantitative cultural analysis of comics. Then, we explore the applications of machine learning in the context of comics. Finally, we review existing self-supervised representation learning methods.

2.1 Quantitative Approaches in Analysis of Comics

Quantitative approaches to the analysis of comics have greatly enhanced our understanding of visual storytelling across different cultures. Initial studies [12, 20] introduced categorisation methods for scene framing techniques within comic panels, revealing distinct framing practices among comics from various cultural backgrounds. Building on this, [18] extended the analysis to the layout of panel sequences on pages, providing a systematic framework for evaluating page layouts. This study compared comics from six different cultural origins, uncovering significant differences in layout practices and emphasising the cultural specificity of comics.

Furthering this line of inquiry, [4] introduced a framework for analysing the continuity of panel sequences, focusing on dimensions of time, space, and character continuity. This approach enabled detailed comparisons of narrative structures across comics from different cultural origins, highlighting variations in continuity practices.[35] introduced a novel method for discretising continuity in panel sequences, quantifying changes in time, space, and character dimensions across panels. By systematically categorising these changes, the study conducted a comparative analysis of transitions in comics from three continents, revealing significant cultural differences. Furthermore, quantitative image analysis techniques have been applied to comics to extract and analyse visual patterns. For instance, [54] utilised digital image analysis to study the brightness and visual details in Tintin comics, highlighting how these visual properties can be quantified and compared across different works.

The exploration of cross-cultural differences in visual narratives has provided valuable insights into how cultural contexts influence the depiction of scenes and the use of backgrounds in comics. [2] found that East Asian comics tend to convey contextual information more implicitly compared to Western comics, which aligns with cognitive patterns of attention and distinct graphic styles.

These studies demonstrate the value of quantitative methods in the cultural analysis of comics. They highlight the utility of such approaches and suggest the potential of computational techniques to scale up the analysis. Automation of these methods could facilitate the examination of larger datasets, offering deeper insights into the cultural variances in comic storytelling.

2.2 Applications of Machine Learning in Comics

Computational approaches to comic processing revolve around supervised machine learning techniques. Both [61] and [3] offer extensive overviews of the state-of-the-art applications of machine learning in computational analysis of comics. Their work encompasses a range of approaches from basic pattern recognition in comic art to complex narrative structure analysis. A substantial portion of existing work in computational comic analysis can be attributed to advancements in computer vision techniques, particularly in the automated extraction of key comic elements such as panels, speech bubbles, and characters.

Panel Extraction is crucial for quantitative analysis of comics and has been addressed by [33, 53, 32] among others. Their methods leverage various computer vision techniques to identify and segment comic book panels, thereby facilitating the analysis of narrative sequences within comics. These approaches all leverage annotated comic panels to train object detection models for panel extraction.

Text Block Extraction is an essential first step for text analysis within comics. Methods introduced in [57] and [24] attempt to segment text blocks in comics through supervised segmentation. Their supervised methods rely on rich and extensive annotations.

Character Detection in comics poses unique challenges due to the varied artistic styles and the dynamic nature of character appearances. Methods for character detection have been introduced in [64] [50] [26]. Furthermore, an extensive analysis of the robustness of various character detection methods across different styles of comics is introduced in [46]. Finally, [66] introduced a method for cross-style unsupervised character detection, aiming to overcome domain shifts in comics, a common issue when dealing with diverse artistic styles.

Emotion Detection of characters within comics through machine learning has been explored in a competition described in [51], in which various approaches to supervised character emotion detection are proposed. An automated framework for detecting emotions in comics through graphical cues is introduced in [59], which takes into account various aspects of human expression including facial expressions, body language, background effects, and onomatopoeia. This holistic approach marks a significant stride towards understanding the emotional depth and narrative techniques employed in comics.

In summary, the exploration of computationally-aided analysis of comics has predominantly focused on supervised machine learning methods, as detailed by [61, 3, 42]. While these developments mark significant achievements in the field, the reliance on supervised learning highlights a dependency on extensive, annotated datasets. This requirement often limits the scalability of analysis and the adaptability to the vast stylistic diversity inherent in comics.

2.3 Self Supervised Representation Learning

The landscape of self-supervised representation learning offers a rich array of methods that extract meaningful representations from data without the need for labeled datasets. In this section, several key strategies that have significantly advanced the field are explained.

Contrastive learning techniques have become fundamental in self-supervised representation learning, playing a pivotal role in the development of robust and invariant feature representations. One of the pioneering methods in this domain is Momentum Contrast (MoCo) [30]. MoCo innovates by employing a dynamic queue and a momentum-updated encoder. This approach addresses the challenge of maintaining a large and consistent set of negative samples, which is crucial for effective contrastive learning. The dynamic queue allows for a diverse and extensive set of negative examples, while the momentum encoder ensures stability and consistency in the representation space over different training iterations. By contrasting positive pairs—differently augmented views of the same image—against a large pool of negative pairs, MoCo effectively learns invariant features that are robust to data augmentations. Building on the foundation laid by MoCo, MoCov2 [9], further refines this paradigm. MoCov2 enhances the original framework by incorporating advanced data augmentation techniques and improving the representation projection head. These enhancements lead to better feature representations and improved performance on downstream tasks. The core idea remains the same: leverage a large, diverse set of negative samples to learn robust feature representations through contrastive loss. However, MoCov2's improvements in augmentation strategies and the representation head significantly boost the model's performance, demonstrating the importance of fine-tuning the details in contrastive learning frameworks.

Self-distillation methods such as Bootstrap Your Own Latent (BYOL) [28] introduce an approach to self-supervised representation learning in which the use of negative samples is completely avoided altogether. Instead, BYOL employs a student-teacher setup where the student model learns to predict the output of the teacher model. This is achieved by using differently augmented views of the same image, ensuring that both the student and teacher networks receive diverse yet related inputs. The teacher model's parameters are updated with an exponential moving average of the student parameters, fostering a stable learning process. This setup allows BYOL to learn effective representations by focusing solely on positive pairs, where both views of the same image are encouraged to produce similar representations, thereby avoiding the need for explicit negative samples. This innovative strategy addresses the complexities and computational burdens associated with negative sample mining, proving highly effective in learning robust features. Similarly, DINO (Self-Distillation with No Labels) [8] and its successor DINOv2 [52] leverage the same self-distillation technique to enhance representation learning, specifically utilizing Vision Transformers (ViT) [69]. DINO operates by passing different augmentations of the same image through a student and a teacher network, aiming to align their output feature distributions. This alignment is achieved through self-distillation, where the student network is trained to match the teacher's outputs. To ensure the representations do not collapse to trivial so-

lutions, DINO employs techniques such as sharpening and centering. Sharpening adjusts the output distribution to be more confident, while centering normalises the feature distributions to maintain diversity in the learned representations. DINOv2 builds on these principles by refining the approach and further improving the robustness and quality of the learned features.

Reconstruction-Based methods such as Masked Autoencoders (MAE) [30] represent a significant advancement in self-supervised learning by focusing on the reconstruction of masked parts of an image to learn robust representations. This technique draws inspiration from masked language modelling [22] in natural language processing, where parts of the input are hidden during training and the model learns to predict these masked segments. In the context of images, MAE addresses the challenge of pixel-wise image reconstruction such as seen in traditional AutoEncoders, which often fail to capture meaningful high-level representations due to the variability and complexity of individual pixel values in images. Instead of reconstructing individual pixels, MAE reconstructs larger image patches. This approach helps the model to focus on higher-level structures and semantics within the image, rather than getting stuck on fine-grained pixel details that may not contribute significantly to the overall understanding of the image content. By masking substantial portions of the image and then learning to reconstruct these patches, MAE forces the model to infer the global context and spatial relationships, leading to more meaningful and coherent feature representations.

2.4 Residual Learning

Finally, our work is inspired by related research in residual learning. Residual learning is introduced in [31]. Their work enables training of deep neural networks through the introduction of residual networks (ResNet). Residual layers are able to reference their input, which allows for optimisation of such deep networks without information loss. In [43], Residual Continual Learning (ResCL) is introduced. In many sequential learning setups, networks suffer from catastrophic forgetting, where networks are unable to transfer knowledge in domain shifts [67]. (ResCL) addresses these issues by combining layers from networks before and after (i.e. fine-tuned) a domain shift takes place. Our work takes inspiration from both works in a residual sequential learning task.

Chapter 3. Learning from Sequences

In this chapter, we introduce **A** **S**elf-supervised **T**ransformer **E**ncoder for comic panel **R**epresentation e**X**traction (**ASTERX**, named after the comic character *Asterix*). ASTERX leverages the sequential nature of comic panels to learn general and contextual representations of comic panels. Inspired by the principles of optimisation found in masked language modelling [22], our approach leverages narrative continuity embedded within adjacent comic panels. This enables facilitation of the transmission of contextual information - from scene composition to character expression. As discussed in [18] [4] [35], the continuity of such aspects in sequences of panels in comics is of great significance to cultural analysis in comics. Our proposed method of representation learning also serves as a starting point for downstream. First, we will provide an in-depth description of ASTERX and our proposed optimisation scheme, then we will describe our experimental setup including a description of used datasets and evaluation methods. Finally, we will discuss the performance of ASTERX in comparison to existing representation learning methods that do not leverage the sequential nature of comics.

3.1 ASTERX

ASTERX models sequences of comic panels with a transformer [69] encoder (Figure 3.1). To enable analysis at both panel-level and sequence-level a special [CLS] token is prepended to each sequence, with one input token per individual panel. The [CLS] token captures the sequence-level information, while panel-level representations are encoded by the matching output tokens for each input panel. To optimise these representations we train ASTERX with two tasks inspired by masked language modelling.

3.1.1 Pretext Tasks

During training, the encoder is optimised through two tasks: panel retrieval and order classification. To facilitate these tasks the encoder receives two panel sequences of the same length as input, separated by a [SEP] token. For the panel retrieval task, a single input token (panel) is masked using the special [MASK] token in the first panel sequence, and the encoder is optimised to produce a representation that matches the masked panel. The second input sequence is used in an order classification task. The order of the two input sequences is shuffled at random to facilitate learning of longer panel distance relations. A linear classification layer takes the output embedding at the position of the [CLS] token and is optimised to classify whether the sequences are in the correct order.

Our approach follows the ideas behind masked language modelling introduced for BERT

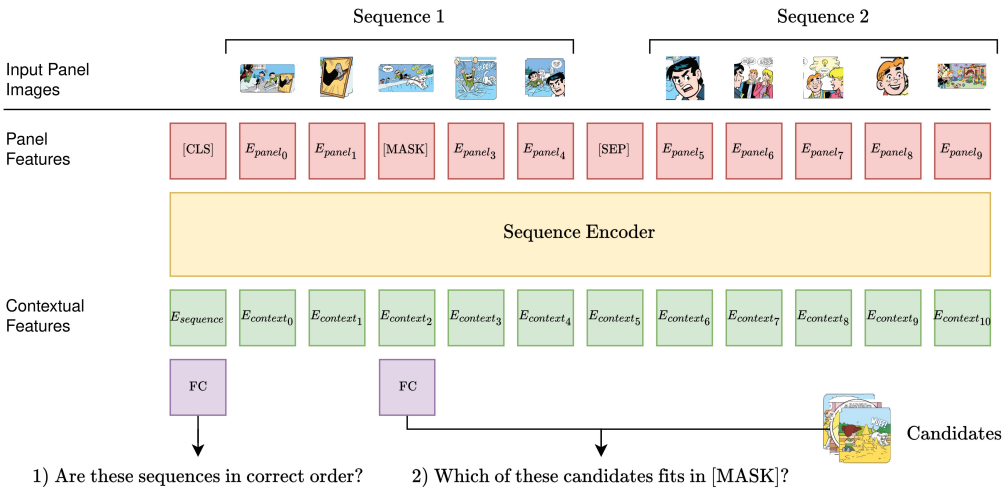


Figure 3.1: Overview of ASTERX. Individual panels are encoded with a backbone, after which the sequence of panels is encoded by our sequence encoder. This sequence encoder is optimised for both panel retrieval and panel order classification.

[22]. However, a key difference between natural language and visual language is the open-set nature of visual language. Where natural language can be modelled through a discrete set of words or tokens (i.e., the vocabulary), visual language is continuous and thus cannot be directly modelled in the same closed-set fashion as masked language modelling. Even though for example framing of characters in comics [12, 33] often follow patterns across comics that can be discretised, the underlying data format (i.e. images) makes pixel-wise reconstruction - in comparison to token-wise reconstruction in masked language modelling - an infeasible task as discussed in [29, 70].

We address the challenges that come from the open-set nature of visual language by casting it as a retrieval task. For the panel retrieval task, we construct a collection of "candidate" panels. These panels are selected from the training set based on a predefined sampling method (section 3.1.2). With these candidates, the sequence encoder is not only optimised to minimise the distance between the output embedding at the [MASK] position and the masked input token; it is also honed to maximise the distance from all incorrect candidate panel embeddings. This configuration allows us to optimise the sequence encoder by learning to select the correct panel. The probability assigned by the sequence encoder for an individual candidate is formulated as:

$$\hat{\boldsymbol{y}}_i = \frac{\exp f(\boldsymbol{x}, \boldsymbol{c}_i)}{\sum_{\boldsymbol{c} \in C} \exp f(\boldsymbol{x}, \boldsymbol{c})}. \tag{3.1}$$

Here, $f(\boldsymbol{x}, \boldsymbol{c})$ denotes a distance metric used to gauge the similarity between two embeddings. $\boldsymbol{x}$ represents the output embedding at the [CLS] token position, and C encompasses the set of candidate tokens including the authentic masked token. When expanded in the batch dimension, each x is paired with a unique C constructed according to the candidate sampling strategy. Employing this formulation, the sequence encoder is optimised using the cross-entropy loss, where the ground truth label aligns with the index of the masked input token within the candidate set.

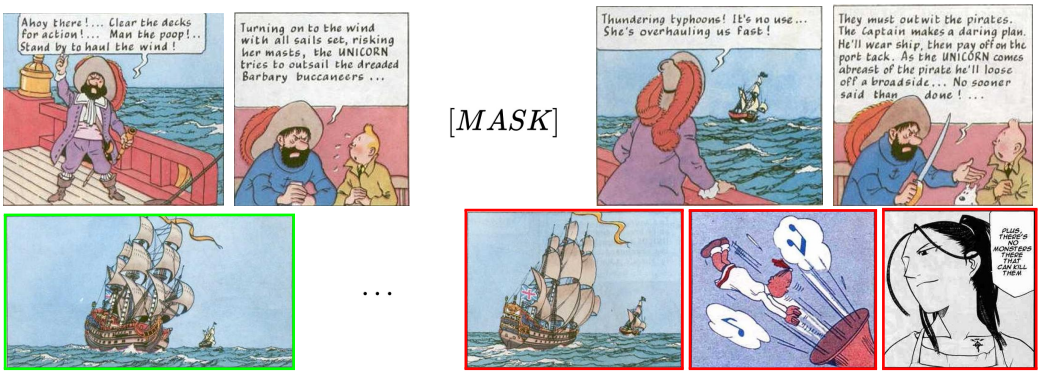


Figure 3.2: Sequence of panels from *Tintin: The Secret of the Unicorn* (first row) along with two candidate samples from the same comic and two from *Suske en Wiske* and *Fullmetal Alchemist* respectively (second row). The correct candidate is highlighted in green, and the incorrect in red.

3.1.2 Candidate Sampling

The candidate sampling strategy intricately shapes the embedding space of the panel sequence encoder. Consequently, we propose four sampling strategies: **1) Random sampling**: Candidates are randomly selected from the training dataset without any specific criteria. **2) Pixel-intensity based sampling:** Candidate panels are chosen within a range of similar pixel-value intensities as the masked panel, based on histogram bins. **3) Panel shape-based sampling:** Candidates are selected within a range of comparable panel shapes, defined by the height-to-width ratio, akin to the masked panel. **4) Same comic sampling:** Candidates are exclusively sourced from within the same comic as the input sequence. Crucially, all properties used in these sampling strategies do not require any annotation. For each strategy, two variants exist: a pure sampling strategy and a mixed strategy. For the pure variants, panels are only sampled according to the criteria described, while in the mixed variant half of the panels are sampled according to the criteria and the other half are sampled completely randomly. In each of these strategies, the true target panel is always included in the set of candidate panels.

3.2 Experiments

Our experiments involve a comparison of our method with various methods on two evaluation tasks: training a linear classifier using learned representations for a specific classification task and panel retrieval. We compare our method to the following existing works: 1) supervised ResNet-50 [31] trained to perform only style classification, 2) OpenCLIP [10] trained on the DataComp-1B [27] dataset, 3) DINO [8] and DINOv2 [52] trained on ImageNet [21] and 4) DINO trained on the COMICS [36] dataset first, and then on the TINTIN corpus [7]. The OpenCLIP, DINO and DINOv2 methods are also evaluated in a "Pooled" variant, where the features of neighbouring panels surrounding the target panel are mean-pooled to obtain a contextualised representation.

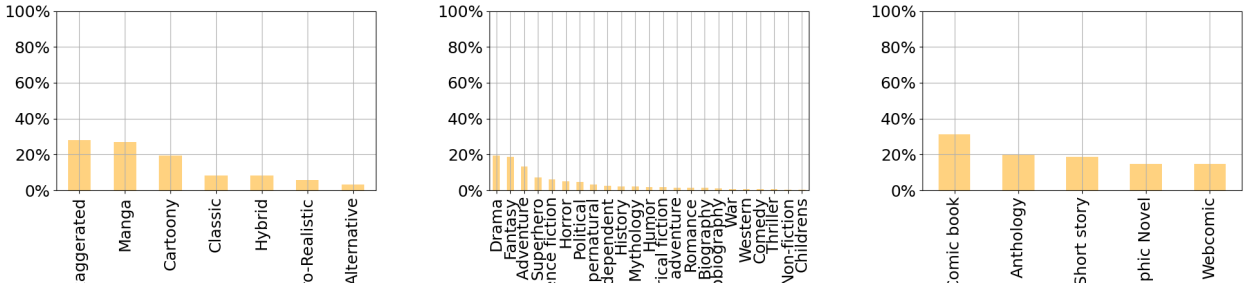


Figure 3.3: Distribution of *style*, *genre* and *format* attributes of comics in the TINTIN [7] corpus.

3.2.1 Data

For our experiments, we use two datasets, the COMICS dataset [36] and the TINTIN corpus [7]. The COMICS dataset consists of 1.2 million panel images and features American comics from the golden age (i.e., 1940s and 1950s). We use the COMICS dataset as a pre-training starting point due to the large quantity of available data within the domain of comics, despite its larger uniformity in comic style.

The TINTIN corpus features more diversity, with over 1,000 comics from over 144 countries/territories, and spanning more than 77,000 panels. The TINTIN corpus is also accompanied by various types of comic-level annotations such as style, format and genre. For the style attribute, the majority class covers 24.5% of a total of 13 classes with classes such as "Manga" "Superhero" and "Cartoon". The format attribute contains classes such as "Comic book", "Webcomic", and "Graphic Novel" with a total of 8 classes and the majority class covering 31.2%. Lastly, the genre attribute contains a total of 91 classes such as "Supernatural", "Action", and "Political commentary", with the majority class spanning 17.4%. All quantitative evaluation of our approach is performed on the TINTIN corpus, as it encapsulates more varied data and provides annotations as ground truth information. The TINTIN corpus is split into an 80% training split, a 10% validation split and a 10% test split for evaluation.

3.2.2 Experimental Setup

We use DINO-trained [8] ViT [23] features for the individual panel representations that serve as input tokens for the sequence encoder. To explore the influence of the backbone we experiment with variations of training the ViT feature extractor on different datasets, and with different ViT settings. For the main experiments the backbone - pre-trained on ImageNet - is first fine-tuned on the COMICS dataset, then fine-tuned on the TINTIN corpus. All models are optimised using the AdamW optimiser [44], with batch size 256 and learning rate 0.0001 for 50 epochs. Unless specified otherwise, we sample 64 candidate panels during training. The architecture of the panel sequence encoder consists of four attention heads and four transformer encoder layers. Each layer has a dropout probability of 0.4. The input embedding size is 384, and the output embedding size is 768. Each training and evaluation sequence consists of 5 panels, of which the center panel is masked.

3.2.3 Linear Fine-tuning

We follow a common approach for the quantitative evaluation of self-supervised training methods. This approach involves training a linear classifier on the feature space derived from the self-supervised model for a specific classification task. In our configuration, we leverage diverse comic-level annotations extracted from the TINTIN corpus. Specifically, we employ style, format, and genre annotations to train a linear classifier. Subsequently, this classifier's performance is evaluated based on classification accuracy, which provides insight into how well the feature space captures specific information. For this analysis we compare models in three different settings: (1) Fully Supervised, here the ResNet-50 models are pre-trained for style and then fine-tuned on Format and Genre. (2) Linear Fine-tuned, which utilises pre-trained backbone on image datasets, and a linear classifier that is fine-tuned for each classification task. OpenCLIP is pre-trained on DataComp-1B, and DINO and DINOv2 are pre-trained on ImageNet. In (3) the Fully Fine-tuned section all methods are first pre-trained on the COMICS dataset, and then fine-tuned on the TINTIN corpus.

Table 3.1 shows the classification performance of a linear classifier trained to classify three aspects of a comic in the TINTIN corpus using the features of various methods. The ResNet-50 trained to classify style outperforms all other methods for style classification, yet generalises poorly to format and genre. Our method outperforms all unsupervised methods, as well as the ResNet-50 on format and genre classification, indicating that supervised pre-training does not result in a model that generalises well.

Method	ViT	Style	Format	Genre
Fully Supervised				
ResNet-50 [31]	-	85.6	53.2	47.7
ResNet-50 Pooled	-	**87.3**	**56.1**	**49.8**
Linear Fine-tuned				
OpenCLIP [10]	B/16	35.6	38.1	25.4
DINO [8]	S/16	34.8	37.7	23.5
DINOv2 [52]	S/14	33.4	36.6	21.7
OpenCLIP Pooled	B/16	**36.2**	**38.4**	**26.0**
DINO Pooled	S/16	35.4	38.2	23.9
DINOv2 Pooled	S/14	34.0	36.7	22.3
Fully Fine-tuned				
DINO	B/16	54.8	51.4	42.6
DINO Pooled	B/16	56.2	54.8	43.6
ASTERX (ours)	B/16	**65.6**	**63.2**	**51.2**

Table 3.1: Linear classification results on various attributes of the TINTIN corpus [7]. The ResNet-50 [31] model is trained for style classification, and the models in the Not Fine-tuned section are trained on DataComp-1B [27] and ImageNet [21]. The methods in the Fully Fine-tuned section are first fine-tuned on the COMICS dataset [36] and then Fine-tuned on the TINTIN corpus.

3.2.4 Panel Retrieval

To further evaluate our method, we pose a retrieval task that aims to retrieve a missing panel within a sequence. To determine how well the model handles large diversity in style and content we use the entire test set, consisting of approximately 7.700 panel images, as the pool of candidate panels. We compute Recall@K across different values of k to gauge the retrieval performance, and thus the ability of the model to encode contextual information within panel sequences.

The panel retrieval performance of the compared methods is shown in Table 3.2. Across the board, we see that this is a highly challenging task. The ResNet-50 model that has been trained in a supervised manner for style classification performs poorly in this retrieval setting, failing to generalise to retrieval. Similarly, the pre-trained OpenCLIP and DINO backbones do not perform well, with OpenCLIP outperforming DINO and DINOv2, which may presumably be due to the presence of comic(-like) images in the larger pre-training set. Fine-tuning DINO on comic data boosts the performance somewhat, but presumably more data is required to obtain good performance. Comparatively, our method performs notably better, thereby showcasing an increased contextual understanding of panels.

Comparing the retrieval performance in Table 3.2 to the classification performance in Table 3.1, there is an obvious gap in the performance of the ResNet-50 where the strong classification performance does not translate to retrieval performance. This gap showcases the reality of the generalisability of features learned with supervision, as good classification performance does not guarantee good retrieval performance. Comparing the retrieval performance with the classification performance for the unsupervised methods, we observe much more balance in performance between these varied tasks, indicating that more general representations are learned.

Method	ViT	R@1	R@5	R@10
Fully Supervised				
ResNet-50 [31] Pooled	-	**0.1**	**0.1**	**0.2**
Not Fine-tuned				
OpenCLIP Pooled [10]	B/16	**0.5**	**4.5**	**15.6**
DINO [8] Pooled	S/16	0.1	0.5	8.1
DINOv2 [52] Pooled	S/14	0.1	0.4	7.6
Fully Fine-tuned				
DINO Pooled	B/16	1.5	19.7	37.1
ASTERX (ours)	B/16	**15.7**	**72.3**	**89.6**

Table 3.2: Panel retrieval performance on the TINTIN corpus [7]. The ResNet-50 [31] model is trained for style classification, the models in the Not Fine-tuned section are trained on DataComp-1B [27] and ImageNet [21]. The methods in the Fully Fine-tuned section are first fine-tuned on the COMICS dataset [36] and then Fine-tuned on the TINTIN corpus.

	Linear Classifier ↑			Retrieval ↑		
Method	Style	Format	Genre	R@1	R@5	R@10
Random	56.8	55.0	43.7	2.3	45.5	60.0
Pure Sampling						
Intensity	57.9	56.8	45.4	4.1	52.7	67.9
Ratio	57.8	57.0	45.7	3.6	51.2	67.5
Comic	**60.9**	**59.5**	**47.0**	**4.5**	**51.6**	**68.7**
Mixed Sampling						
Intensity + Random	59.0	58.2	45.9	11.3	67.2	82.6
Ratio + Random	59.3	58.6	46.4	11.5	67.1	82.9
Comic + Random	**63.8**	**62.5**	**47.6**	**12.3**	**68.8**	**84.2**

Table 3.3: Classification and retrieval performance of different candidate sampling strategies. All mixed sampling strategies outperform pure sampling strategies, and the mixed comic sampling strategy performs best overall.

3.3 Analyses

To understand the performance of ASTERX beyond the comparisons made in the previous sections, this section showcases the results of a comprehensive ablation study of the individual components of ASTERX. We also showcase various success and failure cases of ASTERX in the retrieval task. Finally, we use T-SNE [68] to project the learned representations of ASTERX and analyse the results.

3.3.1 Ablations

We perform three ablations to increase understanding of our method. First, Table 3.3 demonstrates both the linear classifier performance and the retrieval performance of our method trained using various candidate sampling strategies. All mixed sampling strategies outperform the pure sampling counterpart, indicating that a diverse mix of sampling candidates is beneficial in all evaluation aspects. Furthermore, in both mixed and pure sampling categories sampling candidates from the same comic outperforms the other sampling strategies.

Backbone	ASTERX	R@1	R@5	R@10
COMICS [36]	COMICS	0.3	13.8	32.4
COMICS	TINTIN [7]	0.6	29.3	54.7
COMICS → TINTIN	TINTIN	**12.3**	**68.8**	**84.2**
TINTIN	TINTIN	10.8	65.3	79.7

Table 3.4: Comparison of the retrieval performance for the backbone and ASTERX trained on COMICS [36] and/or TINTIN [7], and subsequently evaluated on COMICS or TINTIN to evaluate cross-dataset performance. We observe that cross-dataset pre-training works best.

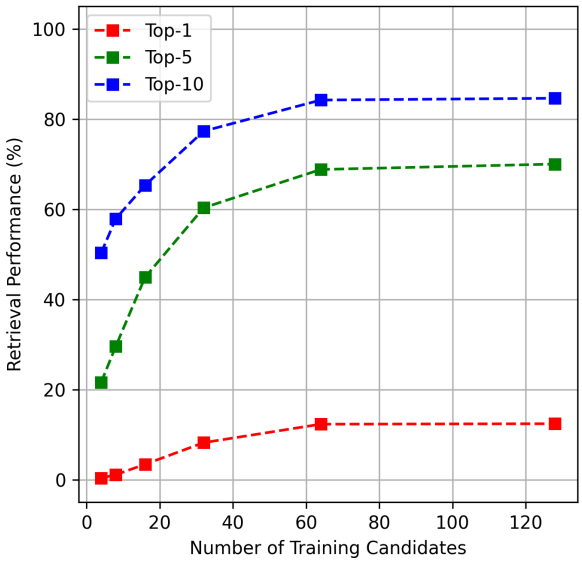


Figure 3.4: Retrieval performance of our method against the number of candidates sampled during training.

The fully random sampling strategy is outperformed by all other sampling strategies.

Secondly, Table 3.4 provides insight into the generalisability of the data domain. Three configurations of fine-tuning both the feature extraction backbone and the sequence encoder on different datasets are shown. All retrieval evaluation results are on the TINTIN corpus validation set. Unsurprisingly, fine-tuning both models on the target dataset results in the best performance. However, only training both models on the COMICS dataset already approaches similar performance of DINO trained on the target dataset (as in Table 3.2).

Finally, Figure 3.4 displays the evaluation retrieval performance of our method based on the number of candidates used during training. The graph clearly depicts a stark increase in performance between 0 and 50 candidates, and flattens beyond 50 candidates, indicating an optimal value of around 64 candidates.

3.3.2 Success and Failure Cases

To form a qualitative understanding of the learned representations of our method in comparison to other methods, two retrieval results are presented in Figure 3.5. In both cases, the original comic page with the masked panel of interest is shown, as well as retrieval results from ASTERX (ours), DINO and the supervised ResNet-50. In both cases, only ASTERX retrieved the correct target panel. In the left case, DINO retrieved a panel from the same comic, in which one of the main characters of this comic is portrayed. The supervised ResNet-50 retrieves a visually similar panel but is unable to retrieve a panel from the correct comic. On the right, all results are close in visual style, but DINO returns a panel from a different edition of the same comic series, whereas ResNet-50 selects a panel from a different series altogether. Interestingly, all three results are approximately the right shape.

Figure 3.6 provides insight into two cases in which ASTERX is unable to retrieve the correct panel. In the left case, the retrieved panel is very similar (both visually and seman-

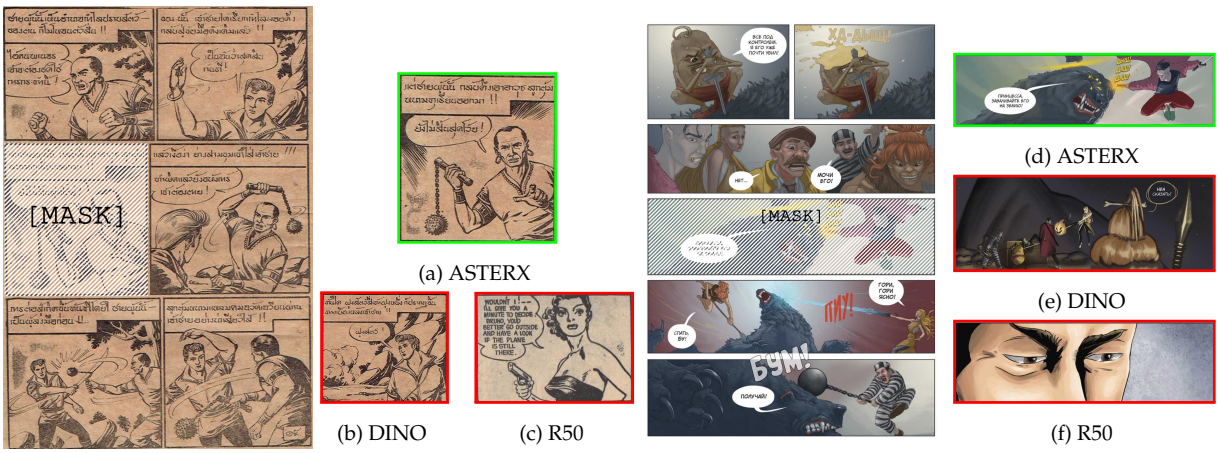


Figure 3.5: Two examples of retrieval results of our method (ASTERX) compared to a supervised ResNet-50 [31] and the unsupervised DINO [8] method. In both cases, our method can retrieve the correct panel while the other methods cannot.

tically) to one of the neighbouring panels. Interestingly, DINO makes the same mistake in this case. In the right case, once again, the panel ASTERX retrieves is very similar to the correct panel in many aspects. This time DINO can retrieve the correct panel, while the supervised ResNet-50 retrieves a panel that is similar in terms of color distribution but is from a different comic than the neighbouring panels altogether.

3.3.3 Feature Space Exploration

In this section, we perform a qualitative analysis of feature spaces derived from ASTERX and DINO. By leveraging t-SNE for dimensionality reduction, we can visualize the high-dimensional feature vectors in a two-dimensional space. This visualisation helps in understanding the structural properties and organisation of the learned feature spaces.

Figures 3.8 and 3.8 illustrate the t-SNE projections of ASTERX and Pooled DINO feature

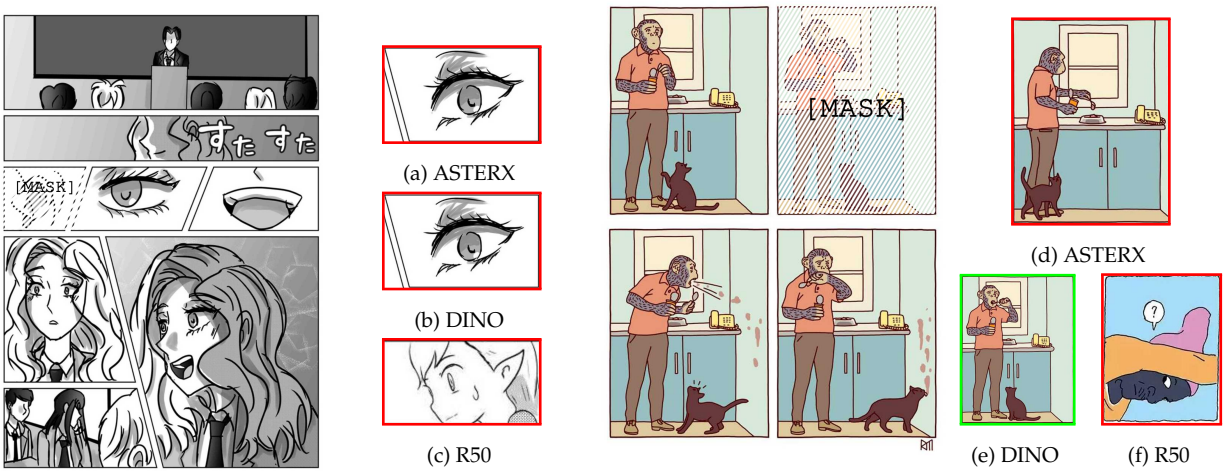


Figure 3.6: Two examples of retrieval results of our method (ASTERX) compared to a supervised ResNet-50 [31] and the unsupervised DINO [8] method. In both cases, our method was unable to retrieve the correct panel.

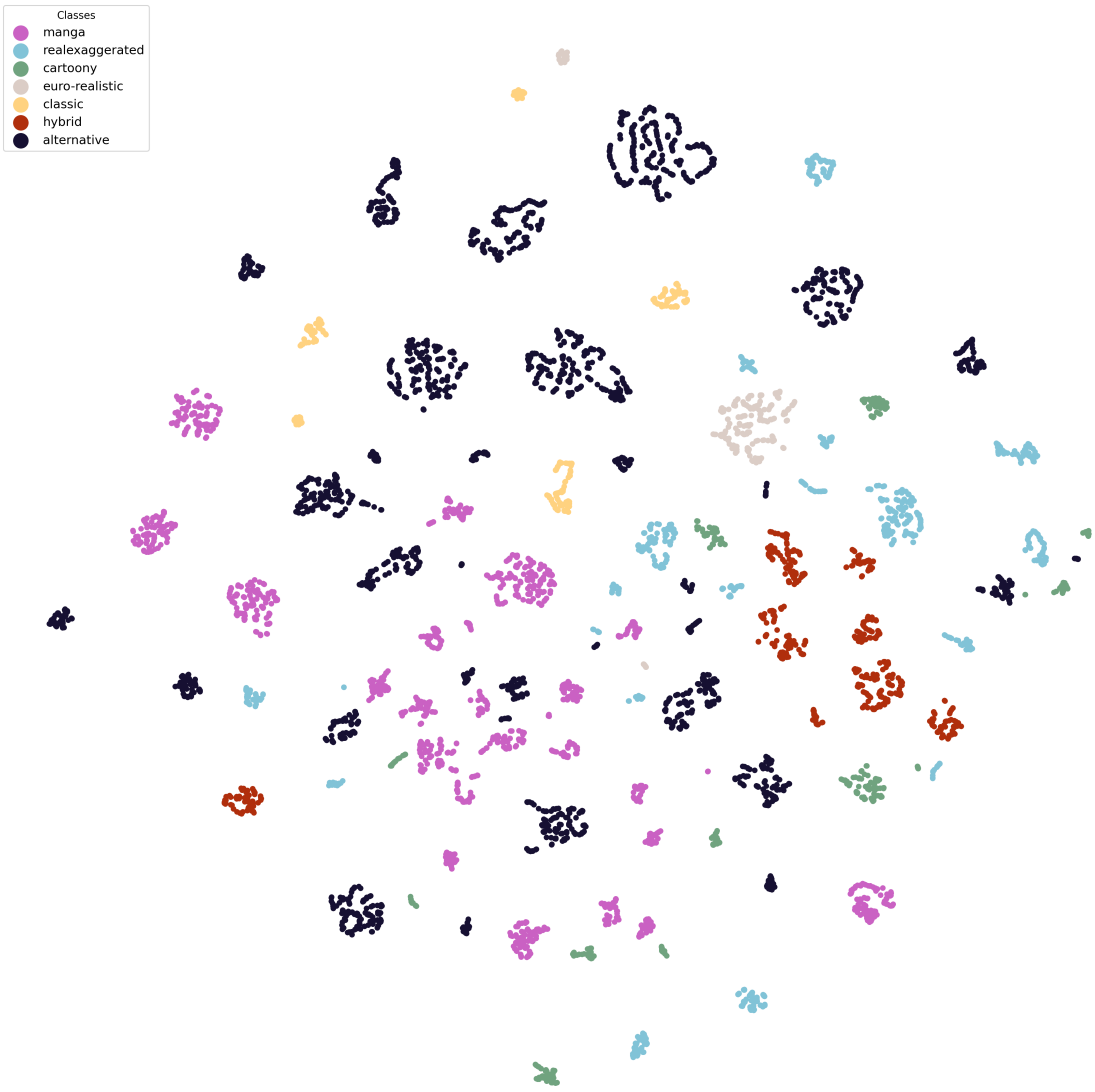
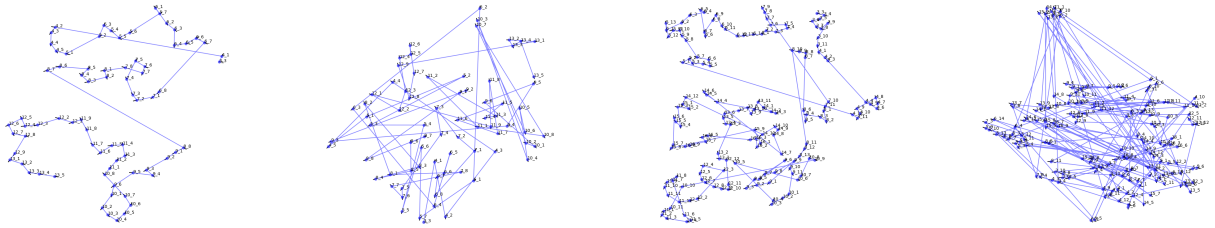


Figure 3.7: t-SNE [68] projections of ASTERX features of the entire validation set. Even in these projections, clear clusters of comic panels within comics and clusters of comics within syles are captured.

vectors, respectively, extracted from the panels of a single comic. These figures provide a comparative view of how both methods capture and organise comic panels that follow each other in sequence. The t-SNE projection of ASTERX feature vectors demonstrates a distinct sequential pattern that aligns with the nature of comic panels. Panels that are temporally adjacent in the comic sequence tend to form tight clusters in the t-SNE space. This suggests that ASTERX effectively captures the continuity and progression inherent in the comic narrative, preserving the contextual relationships between consecutive panels. In contrast, the t-SNE projection of the Pooled DINO feature vectors is scattered, without much pattern in terms of adjacent panels. These DINO features do not encode the sequential dependencies between panels. Instead, the features appear more independently distributed, reflecting a less cohesive representation of contextual information.



ASTERX DINO ASTERX DINO

(a) *The Bluecoats* (b) *Suske en Wiske*

Figure 3.8: t-SNE [68] projections of ASTERX features in comparison to t-SNE projections of DINO [8] features of two comics.

3.3.4 Oracle Robustness Studies

Training ASTERX requires bounding box annotations for panels. Within the TINTIN corpus and the COMICS datasets, annotated bounding boxes are provided. However, methods for automated panel bounding box extraction exist but are not as perfect as human annotations. Hence why we perform a robustness study of ASTERX concerning the permutation of panel bounding boxes, to simulate how ASTERX would perform when an automated method is applied. Figure 3.9 shows the results of this study. Each graph displays the results of training ASTERX with permuted bounding boxes (translated, up-scaled and down-scaled) by a specified severity (percentage of page size). For all types of permutations, the performance of ASTERX decreases as the severity increases. Severity beyond 25% seems to be a breakpoint for most types of permutation.

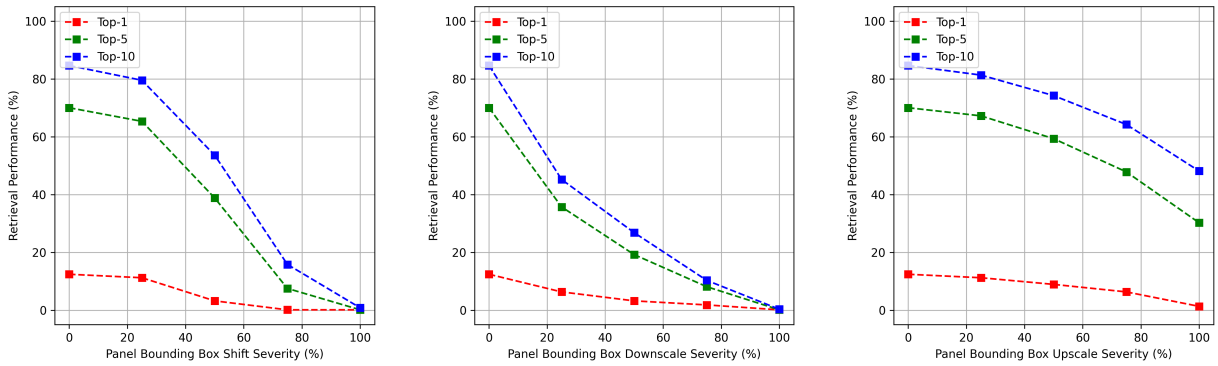


Figure 3.9: Panel retrieval performance of ASTERX when different kinds of permutations are applied to the panel bounding boxes.

Chapter 4. Context Normalisation

Our second experiment focuses on the challenging task of unsupervised character-level representation learning in comics. This complexity arises from the diverse designs of comic characters, their varied framing, posing, and expressions [49, 11]. Additionally, character drawings in comics are often influenced by preceding panels, making the task even more intricate. For example, if a character is fighting in one panel, it is likely that the character will also be fighting in the next panel. Such inferences are made by readers but are hard to explicitly capture in computational approaches [14, 15, 17, 39]. Our approach aims to isolate and exclude contextual elements such as style, design, and sequence information, thereby concentrating exclusively on aspects like posing, framing, and expressions of the character instance. This method builds on the sequence representations discussed in Chapter 3, applying these foundational insights to enhance our character-level representation learning.

4.1 ELRIC

We introduce cont**E**xt norma**L**isation for **R**epresentation learning **I**n **C**omics (**ELRIC**, named after the *Elric* brothers in *Fullmetal Alchemist*). ELRIC integrates the contextual representations learned by ASTERX — capturing information from adjacent comic panels — into the training process of existing self-supervised learning frameworks. By doing so, ELRIC learns context-independent features for a specific panel or even a specific region within a panel. This context-normalised representation makes ELRIC highly suitable for panel or region-specific downstream tasks such as emotion character matching and emotion classification. ELRIC is inspired by related works on residual learning [31, 37, 43].

4.1.1 Context Injection

Within the ELRIC framework, frozen context-rich features are combined with learnable features before the loss function computation in a residual connection. This setup forces the learnable features to learn aspects which are not encoded in the context of the region of interest (ROI):

$$\boldsymbol{v}' = \frac{f(\boldsymbol{v}) + \boldsymbol{c}}{\boldsymbol{m}}, \tag{4.1}$$

in which $\boldsymbol{v}$ represents the learnable feature vector, $\boldsymbol{c}$ denotes the frozen context feature vector and $\boldsymbol{m}$ is a learnable normalisation vector. The function f is a learnable linear projection of $\boldsymbol{c}$ into the dimensionality of $\boldsymbol{c}$. By incorporating context-rich features into the learning process, ELRIC effectively normalises context-dependent variations. This enables the model to focus on region-specific attributes, enhancing its performance in tasks that require detailed

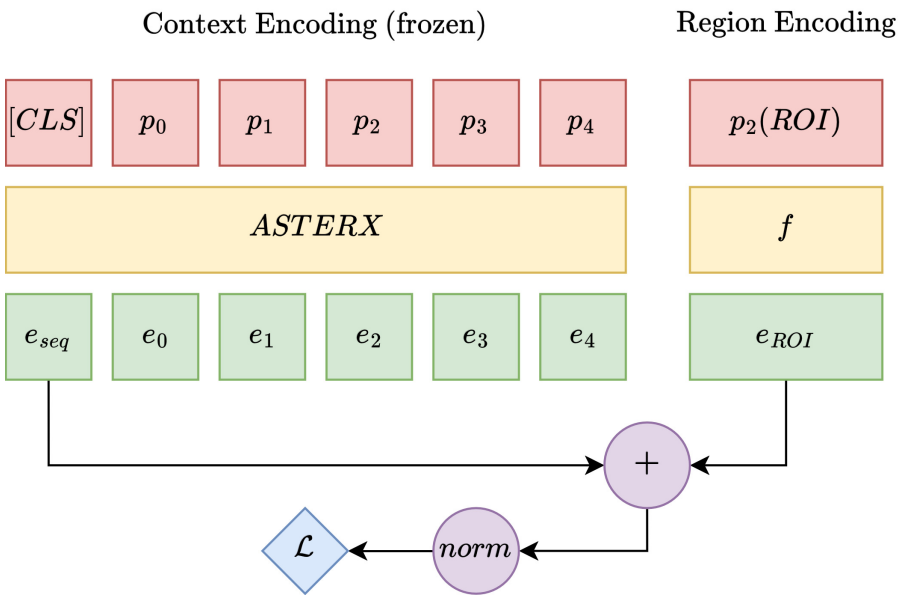


Figure 4.1: Diagram of the ELRIC framework, in which ASTERX context features are used to enhance available information during training. As a result, region-specific (ROI: Region of Interest) features become less context-dependent during inference.

character-level analysis. The result is a robust framework capable of isolating essential character traits, thereby facilitating more accurate and context-independent character representation learning in comics. Figure 4.1 showcases the ELRIC framework diagrammatically.

4.2 Experiments

We evaluate the performance of ELRIC in two settings: character re-occurrence matching using a clustering approach and emotion classification. In the character matching setting, we compare the classification performance of ELRIC with DINO [8] trained on the COMICS [36] dataset first, and then on the characters images in the TINTIN [7] corpus and ASTERIX trained on the panel images of the TINTIN corpus. For emotion classification, we extend

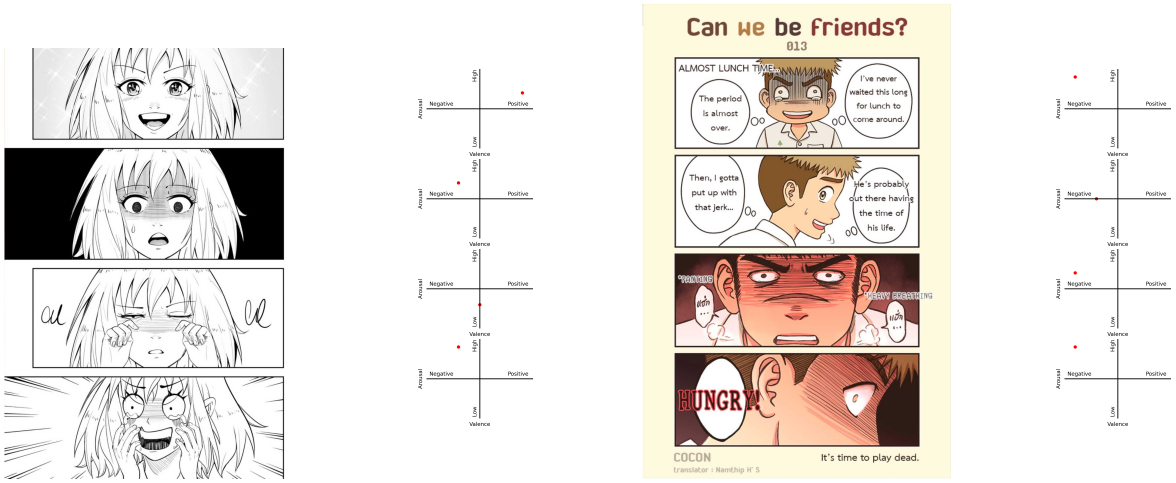


Figure 4.2: Two examples of valence arousal decomposition in comic panels.

our comparison by including a supervised ResNet-50 [31]. This ResNet-50 model is trained on the opposite dimension of emotion relative to the evaluation dimension. Specifically, for arousal evaluation, the ResNet-50 is trained on valence classification, and for valence evaluation, the ResNet-50 is trained on arousal classification. Through this setup, we gain an understanding of to which extent emotion classification generalises between these two dimensions.

4.2.1 Data

For our experiments with ELRIC, we utilise two datasets: the COMICS dataset [36] and the TINTIN corpus [7]. The COMICS dataset is only used for pre-training purposes. The TINTIN corpus is split into 80% training, 10% validation, and 10% test sets. In the context of the emotion classification task, TINTIN corpus annotation which describe two key dimensions are utilised. These two dimensions are valence and arousal. Valence represents the positivity or negativity of an emotion, ranging from unpleasant to pleasant. For instance, emotions like sadness and anger have low valence, while happiness and excitement have high valence. Arousal, on the other hand, measures the intensity of the emotion, from calm to excited. Low-arousal emotions include relaxation and boredom, whereas high-arousal emotions encompass fear and exhilaration. Both dimensions are discretised into 5 levels (classes) within the TINTIN corpus.

The character-level annotations in the TINTIN corpus, which include bounding boxes of character occurrences and character names, allow us to focus on specific regions within panels and evaluate character-matching performance. We filter the dataset to include only characters that appear at least 25 times. This filtering ensures that our model is trained on a robust and representative subset, consisting of about 100 characters and about 3.000 training images.

4.2.2 Experimental Setup

Our experiments with ELRIC build upon the experiments conducted for ASTERX, utilising the same experimental setup and model parameters for the ASTERX components as stated in section 3.2.2. We introduce a fifth candidate sampling method in ASTERX training tailored for ELRIC, where candidates are sampled based on the comic characters that appear in a panel. This method ensures that only panels containing the same comic character as

Method	Accuracy↑	Precision↑	Recall↑	F1↑	Silhouette↑
DINO [8]	78.7	80.7	78.8	78.8	0.602
ASTERX	61.2	63.5	62.4	63.1	0.355
ELRIC (ours)	**87.4**	**87.9**	**87.4**	**87.9**	**0.772**

Table 4.1: Results of linear fine-tuning on ELRIC features for the character matching task. ELRIC outperforms all baseline methods (ASTERX and DINO [8]) by significant margins. The results of ASTERX indicate the negative effect of contextual information on single instance classification.

Method	Accuracy↑	Precision↑	Recall↑	F1↑	Silhouette↑
Valence Classification					
ResNet-50 [31] (Arousal)	32.3	31.5	33.2	32.9	-0.138
DINO [8]	59.2	53.3	59.2	55.3	0.349
ASTERX	37.5	37.9	36.4	36.4	0.032
ELRIC (ours)	**68.3**	**64.9**	**63.5**	**64.4**	**0.571**
Arousal Classification					
ResNet-50 [31] (Valence)	31.7	31.8	31.2	31.3	-0.192
DINO [8]	54.4	52.8	54.4	53.0	0.243
ASTERX	36.4	38.6	35.4	37.4	0.010
ELRIC (ours)	**63.7**	**60.4**	**62.2**	**62.0**	**0.533**

Table 4.2: Performance of ELRIC in comparison to baseline methods for the emotion classification tasks. ELRIC outperforms all baseline methods (ASTERX, DINO [8] and a ResNet-50 [31] trained on the orthoganl dimension of emotion). The results of the ResNet-50 indicate that emotion classification does not generalise smoothly across dimensions.

the target panel are sampled. In our experiments, we focus on integrating ELRIC into the DINO optimisation framework. We optimise a ViT S/16 model using the DINO framework, applying ELRIC context injection before the DINO head is used. The ViT S/16 model is optimised with the AdamW optimiser, using a batch size of 256, and a learning rate of 0.0001, over 25 epochs. The ASTERX context features are projected from 768 dimensions down to 384 dimensions, matching the dimensionality of the ViT B/16 features.

4.2.3 Character matching

The TINTIN corpus offers detailed character annotations, which we leverage in our first evaluation task for ELRIC. In this task, ELRIC features are used as input for a k-nearest neighbours (kNN) classifier with $k = 5$. The kNN classifier is trained to perform character classification, providing insight into how well ELRIC constructs character clusters in feature space. To obtain a holistic view of both classification and clustering performance, we compute several metrics: classification accuracy, precision, recall, f1-score, and the cluster silhouette score. Table 4.1 presents the performance of ELRIC compared to various baseline methods based on these metrics. ELRIC significantly outperforms all other methods in every metric in the character classification task, indicating that the injection of contextual features enhances the character-specific richness of the learned features.

4.2.4 Valence and Arousal Classification

In addition to character matching, we evaluate ELRIC's performance in emotion classification through linear probing. This task involves predicting five classes of valence and arousal, which represent the sentiment and the intensity of emotions.

For this evaluation, we train a linear classifier using ELRIC features as input. The performance of this classifier is then assessed in terms of the same classification and clustering

metrics mentioned in section 4.2.3. We compare ELRIC against several baseline methods: a supervised ResNet-50 trained on the orthogonal dimension of emotion (i.e., for arousal evaluation, the ResNet-50 is trained on valence classification and vice versa), DINO, and ASTERX.

Table 4.2 showcases the performance of ELRIC and the baseline methods across various metrics. ELRIC consistently outperforms the other methods in emotion classification, demonstrating higher accuracy and better cluster cohesion. The results indicate that ELRIC's ability to isolate context-independent features significantly enhances its performance in distinguishing and classifying emotional states. Notably, the supervised ResNet-50 does not generalise well into the other dimension of emotion.

4.3 Analyses

In this section, we aim to understand ELRIC through several ablation studies and qualitative results analysis of ELRIC in character matching and emotion classification settings.

4.3.1 Ablations

Two ablation studies are showcased in this section. The first ablation study focuses on the effect of ASTERX candidate sampling methods on the performance of ELRIC. Secondly, we show the effect of training all required components (backbone, ASTERX and ELRIC) on different variations of the dataset to understand the robustness of our method with respect to domain shifts. Table 4.3 showcases the result of our first ablation study, where various sampling techniques of ASTERX are shown, as well as the performance of the resulting ELRIC performance on the character matching and emotion classification tasks. Similar to the performance of ASTERX, using mixed sampling strategies demonstrate increased perfor-

	Character		**Valence**		**Arousal**	
Method	Accuracy	F1	Accuracy	F1	Accuracy	F1
Random	79.9	79.5	60.4	61.0	59.6	59.3
Pure Sampling						
Intensity	81.4	81.9	62.1	60.4	59.8	59.8
Ratio	81.0	80.7	62.6	62.1	60.4	59.5
Comic	**85.5**	**85.0**	**65.3**	**65.6**	63.3	62.0
Character	83.2	83.5	64.9	64.3	**63.5**	**62.3**
Mixed Sampling						
Intensity + Random	82.1	83.5	63.5	62.8	61.3	62.4
Ratio + Random	82.4	80.2	63.6	63.4	60.9	60.9
Comic + Random	**87.4**	**87.3**	**68.3**	**67.3**	**63.7**	**63.6**
Character + Random	85.8	83.8	64.2	64.5	61.5	60.7

Table 4.3: Results of the ASTERX sampling methods ablation study where the effect of ASTERX sampling methods on the performance of ELRIC in all three tasks is measured.

Backbone	ASTERX	ELIRC	Accuracy	Precision	Recall
COMICS [36]	COMICS	COMICS	10.4	12.6	13.3
COMICS	COMICS	TINTIN [7]	34.8	37.2	36.7
COMICS	TINTIN	TINTIN	74.2	73.7	76.7
COMICS → TINTIN	TINTIN	TINTIN	**87.4**	**87.9**	**87.4**
TINTIN	TINTIN	TINTIN	43.8	45.5	44.3

Table 4.4: Results of the dataset ablation study, where the performance of ELRIC on the character matching task is measured when different components of the ELRIC pipeline are trained on different datasets.

mance compared to pure sampling strategies. Interestingly, the same comic-based sampling method outperforms the character-based sampling method in all evaluation tasks by a small margin. These results further indicate that a balance between specificity and variation is required for strong representations.

Secondly, table 4.4 provides insight into the effect of different datasets on the performance of ELRIC. We show the dataset used to train the ASTERX backbone, ASTERX and ELRIC as well as the performance of ELRIC on the character-matching task. Unsurprisingly, the configuration where the domain shift is smallest shows the best performance. The ASTERX backbone however requires a larger amount of data (in the form of the COMICS dataset) for the best performance.

4.3.2 Success and Failure Cases

The following results sections provide examples of success and failure cases of ELRIC in both the character matching task, as well as the emotion classification task.

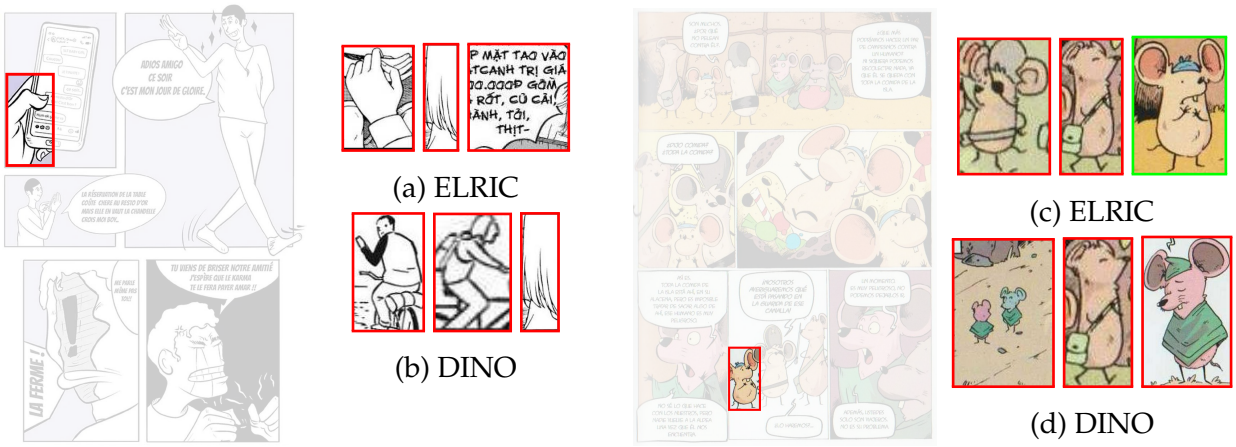
(a) ELRIC

(b) DINO

(c) ELRIC

(d) DINO

Figure 4.3: Two cases in which ELRIC successfully classified the same character. On the left, a page from a comic is shown along with a bounding box for a character of interest, while the right images show the top 3 most similar characters using ELRIC and DINO [8].



(a) ELRIC

(b) DINO

(c) ELRIC

(d) DINO

Figure 4.4: Two cases in which ELRIC failed to classify the same character. On the left, a page from a comic is shown along with a bounding box for a character of interest, while the right images show the top 3 most similar characters using ELRIC and DINO [8].

Character Matching

In the character matching task, ELRIC clusters characters by identifying the most similar character images in feature space. Figure 4.3 showcases two examples where ELRIC correctly matched characters, whereas DINO failed. In the left example, ELRIC's top three retrieved samples exhibit a high degree of similarity to the target character in terms of framing, posing, and clothing. In the right example, despite notable changes in facial expression across the samples, ELRIC still correctly identified images of the target character. In contrast, while DINO's retrieved samples in the right example exhibit similar posing to the target character, they fail to match the correct character, indicating a limitation in DINO's character recognition capabilities. In the right example, the characters retrieved by DINO are very similar in terms of framing - arguably more so than the characters retrieved by ELRIC.

In the two samples in figure 4.4, we examine two cases where ELRIC faces difficulties in the character matching task. These examples highlight the inherent challenges of character matching in comics. In the first example, only the hand of the character is visible, making it almost impossible to identify the character. Interestingly, ELRIC's most similar retrieval is another hand. In the second example, the character of interest is depicted as a small image. ELRIC retrieves characters with similar posing and framing, but only the third retrieval is the correct character. For both examples, DINO fails to retrieve the correct character.

Figure 4.5: Two qualitative examples of ELRIC in the emotion classification task, where ELRIC successfully classified both dimensions of emotion.

Figure 4.6: Two qualitative examples of ELRIC in the emotion classification task, where ELRIC failed to classify both dimensions of emotion.

Valence and Arousal Classification

The arousal and valence analysis in this study provides a deeper understanding of ELRIC's performance in emotion classification tasks. Figure 4.6 showcases two instances where ELRIC successfully classified both valence and arousal, while DINO failed. These examples highlight the inherent challenges of valence and arousal classification. In the left case, part of the subject's face is covered, making it difficult to assess emotions accurately. In the right case, the challenge lies in capturing the low arousal level, which ELRIC successfully identified. Figure 4.5 presents two scenarios where ELRIC did not perform as expected, failing either in both dimensions (left) or only in the arousal dimension (right). In both cases, DINO also failed in at least one dimension. In the right case, the failure is attributed to the difficulty in discerning subtle differences in low arousal states, which underscores the nuances involved in emotion classification tasks.

4.3.3 Oracle Robustness Studies

Similar to how ASTERX relies on panel bounding boxes for training, our experiments with ELRIC rely on character bounding box annotations. To this end, we perform an oracle robustness study similar to the oracle robustness study for ASTERX in section 3.3.4. The annotated character bounding boxes in the TINTIN corpus are translated, down-scaled and up-scaled with a certain severity when training ELRIC. Figure 4.7 shows the effect of these permutations on the character-matching performance of ELRIC. Unsurprisingly, as the severity

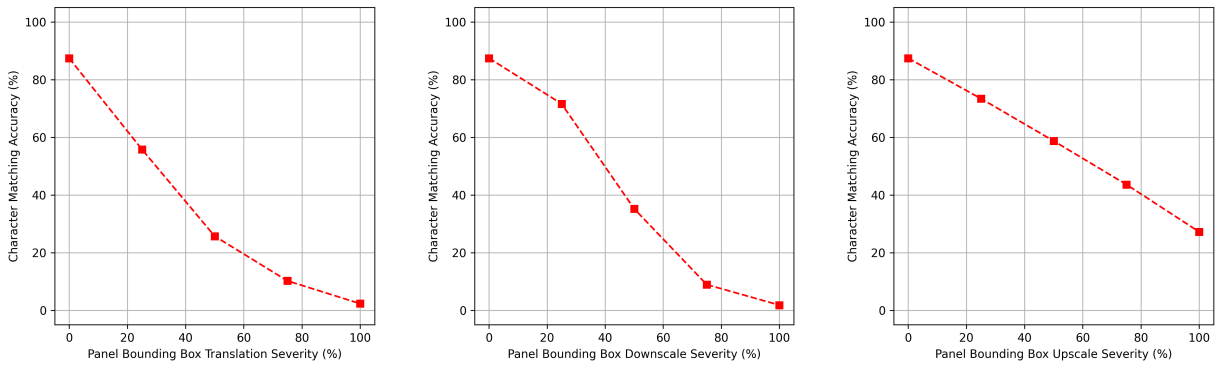


Figure 4.7: Character matching performance of ELRIC when different kinds of permutations are applied to the character bounding boxes.

of permutations increases, performance decreases. Compared to ASTERX however, ELRIC performance decreases faster for the upscaling permutation.

Chapter 5. Cultural Analysis

In this chapter, we explore how ASTERX and ELRIC can contribute to the field of cultural analysis in comics. This chapter is divided into two sections: one focusing on ASTERX and the other on ELRIC.

5.1 ASTERX Studies - Panel Continuity

In this section, we explore the contributions of ASTERX to cultural analysis, with a focus on the continuity aspects of comic panels. We examine how different genres, styles, and regions of origin vary in terms of continuity. We decompose continuity into two facets: euclidean distance and angle. The angle is calculated as the cosine distance between the difference vectors of two consecutive pairs of panel embeddings (i.e., how much the angle changes between three consecutive panels).

Figure 5.1 shows the distributions of Euclidean distance and angle differences across the entire TINTIN corpus. The distance distribution is slightly skewed towards longer distances, roughly following a gamma distribution. The angle distribution is dense around an angle difference of 0 and drops off swiftly between $\frac{1}{2}\pi$ and π radians.

5.1.1 Genre, Region of origin and Style

Figures 5.3 illustrate the distance distributions per category of specific attributes in the TINTIN corpus: genre, region of origin, and style, respectively. Notable outliers in terms

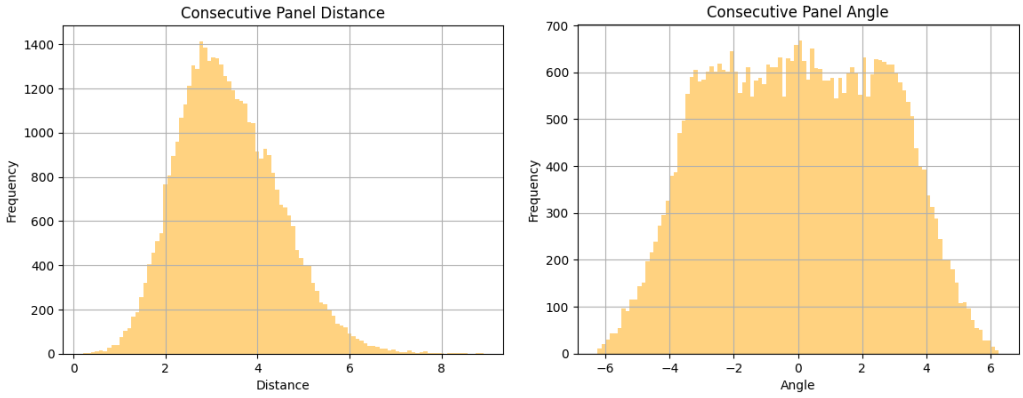
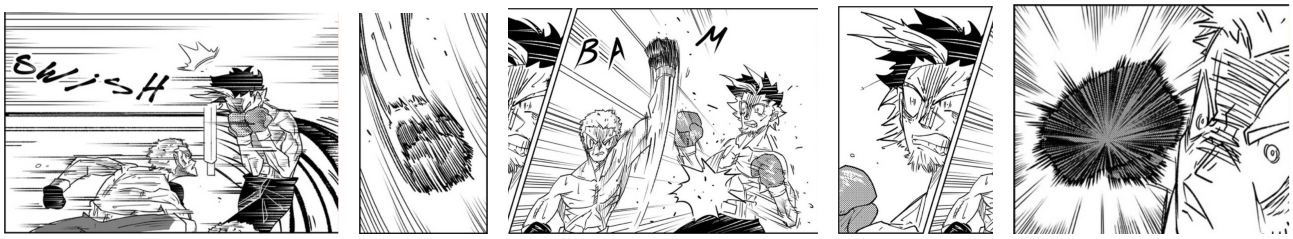


Figure 5.1: Histogram of the Euclidean distance between two consecutive panels and the angle between two consecutive pairs of panels in the features space of ASTERX of the entire TINTIN corpus. The angle is measured as the difference of the angles of two consecutive pairs of panels in radians.

Consecutive panel continuity in *Lucky Luke - Daisy Town* (Belgian Western, 1983).



Consecutive panel continuity in *Broken Fists* (Iraqi Action, 2018).

Figure 5.2: Two examples of consecutive panel continuity. According to our analysis, consecutive panels in Western comics are conceptually closer compared to consecutive panels in Action comics. Furthermore, in comics of Central Asian origin conceptual gaps between consecutive panels are larger compared to comics of European origin.

of average distance include comics in the "Romance" and "Western" genres. Comics in the "Romance" genre exhibit relatively large distances between panels, while comics in the "Western" genre show relatively small distances between panels. Similarly, comics from "Central Asia" form the upper bound within the region categorisation, while comics from "Oceania" form the lower bound. Finally, comics in the "Manga" style have the highest average distances, while those in the "Classic" style have the lowest. When examining variability (i.e., the range of the box plots), we find that comics within the "Children's" and "Autobiography" genres display low variability in panel distances, whereas "Comedy" and "Action Adventure" genres exhibit high variability. Variability across different regions and styles is similar. Where the consecutive panel distance seems to carry some descriptive value between different categories, the consecutive panel pair angle difference is very uniform across different categories, as demonstrated by figure 5.4. Across all attributes and all categories

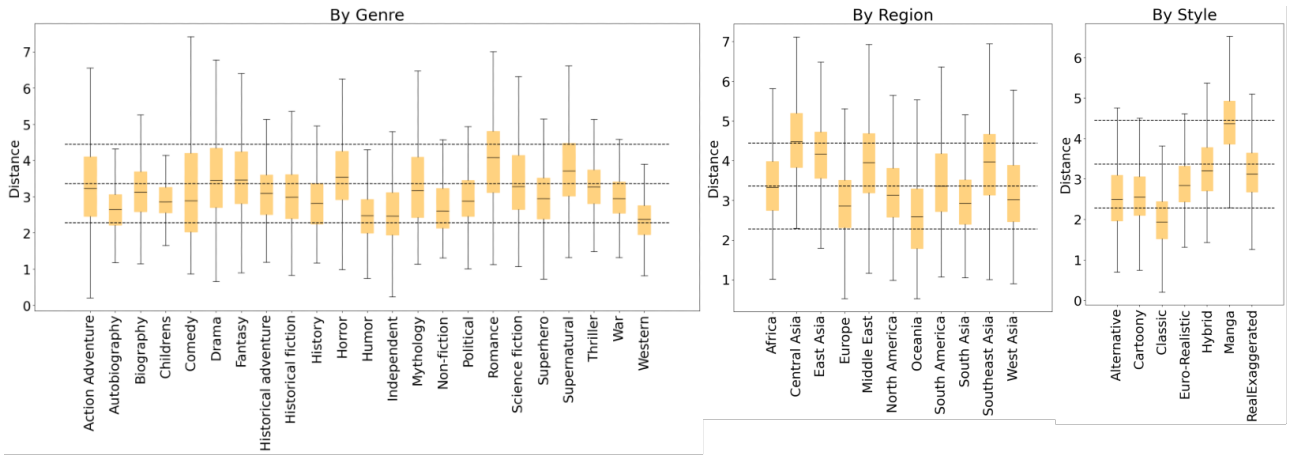


Figure 5.3: Distribution of Euclidean distance between two consecutive panels in the features space of ASTERX within genres, regions of origin and styles of the TINTIN corpus.

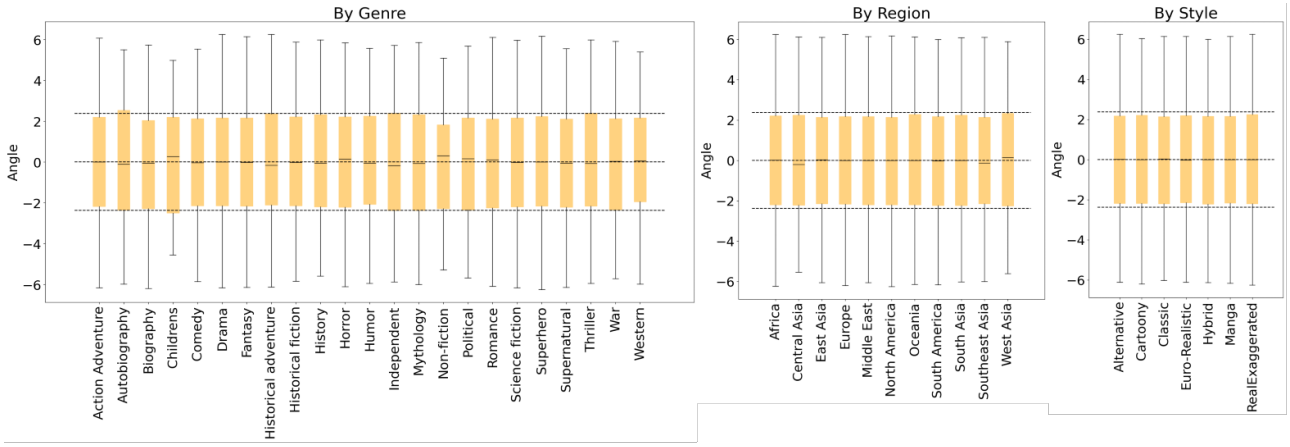


Figure 5.4: Distribution of difference of angles in radians between two consecutive pairs of panels in the features space of ASTERX within genres, regions of origin and styles of the TINTIN corpus.

within these attributes, the angle distribution is highly nondescript, indicating that this facet does not carry any significant distinguishing value with respect to these categories.

5.1.2 Time: Historical and Length

The TINTIN corpus also offers insight into the year of publication of a comic and the length of a comic in the number of pages. Figure 5.5 illustrates the distribution of consecutive panel distances of ASTERX features across different decades of comic publication and various comic lengths. One notable observation is that both the average and variability of consecutive panel distances have increased over the decades. This trend suggests that comics published in more recent decades tend to exhibit greater diversity in panel transitions. Regarding comic length, the average and variability of the consecutive panel distance metric remain stable for comics up to about 30 pages. However, for comics longer than 30 pages, the average consecutive panel distance increases, indicating that longer comics may incorporate more varied and expansive panel transitions.

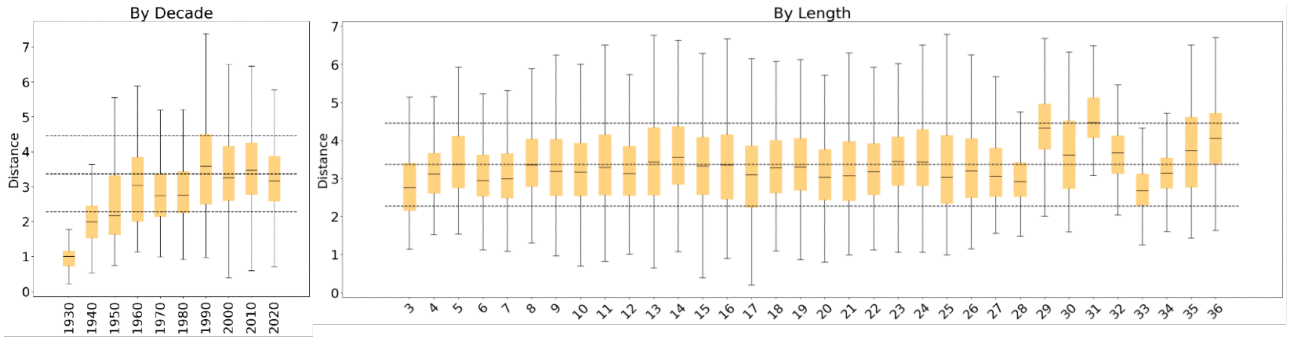


Figure 5.5: Distribution of Euclidean distance between two consecutive panels in the features space of ASTERX within different decades and comic length (in number of pages) of the TINTIN corpus.

5.1.3 Discussion

Using the ASTERX feature space, we demonstrated how comics vary in terms of consecutive panel continuity for attributes such as genre, region of origin, style, decade of publication, and comic length. While the analysis of the cosine angles between consecutive panels revealed no patterns, the analysis of the Euclidean distance provided valuable insights into differences.

Most notably, Autobiography, Children's, and Western comic genres showcase relatively small consecutive panel distances on average, while Romance and Supernatural genres exhibit larger distances. In terms of region of origin, Central Asia and Southeast Asia have large consecutive panel distances on average, whereas comics from Europe and Oceania showcase small distances. Finally, comics in the Manga style exhibit large distances, while those in the Classic style show small distances. Regarding time-based aspects, we see an increase in the distance over decades, reaching a maximum around 1980. This can be explained by the prevalence of Classics from the 1950s or earlier, while many Manga were published after 1980. Comic length does not seem to carry much significance with respect to panel distance in the ASTERX feature space.

Our findings within this analysis align with existing research and attempt to quantify continuity aspects of comics [1, 25, 56] concerning cultural aspects. Furthermore, our findings within the continual aspects of comics also align with leading theories in the sequential processing of visual narratives from a psychological perspective [34, 5, 63, 38, 47].

5.2 ELRIC Studies - Character Instance Variation

Following the analysis performed using ASTERX features, we conduct a similar analysis using ELRIC features at the character level. Specifically, we calculate the intra-cluster distance of ELRIC feature vectors within character clusters in the TINTIN corpus. These statistics provide insights into how varied different instances of the same characters are for various descriptive attributes such as genre and style. Figure 5.6 illustrates the distribution of all intra-cluster distances across the entire TINTIN corpus. Similar to the ASTERX distance distribution, this distribution roughly follows a gamma distribution, with a tail extending

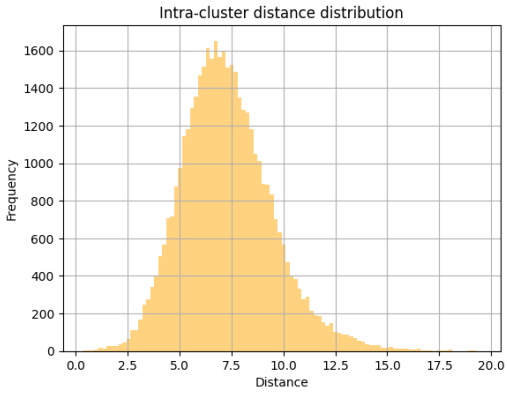
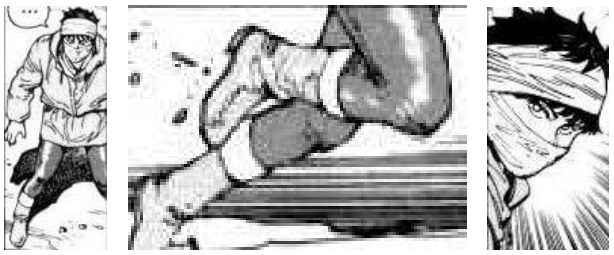


Figure 5.6: Distribution of intra-cluster distance of character-level features of ELRIC over the entire TINTIN corpus.



Character variation in *Akira* (Japanese Manga, 1987).

Character variation in *Nell'Impero degli Incas* (Italian Classic, 1936).

Figure 5.7: Two examples of character instance variation in two different regions of origin, styles and decades. In each case, the leftmost example represents the closest instance to the character cluster mean in the ELRIC feature space, and the other two instances represent the two least similar instances in the ELRIC feature space.

towards longer distances.

5.2.1 Genre, Region of origin and Style

Figure 5.8 showcases the intra-cluster distance distribution per category for the genre, global region of origin, and style attributes within the TINTIN corpus. Within the genre attribute, notable observations include the "Biography" and "Thriller" genres, which exhibit relatively high average intra-cluster distances, and the "Children's" and "Western" genres, which show relatively low average intra-cluster distances. In terms of variability, the "Biography" genre stands out with high variability, whereas the "Children's" genre shows low variability.

For the global region of origin attribute, the lowest average intra-cluster distance is observed in "Oceania," while the highest is recorded in "Central Asia". The average intra-cluster distance is very similar between all regions. Regarding the style attribute, the "Classic" style exhibits the lowest intra-cluster distances, whereas the "Manga" and "Real Exaggerated" styles show the highest intra-cluster distances. Across both global regions of origin and style attributes, the variability in intra-cluster distances is similar.

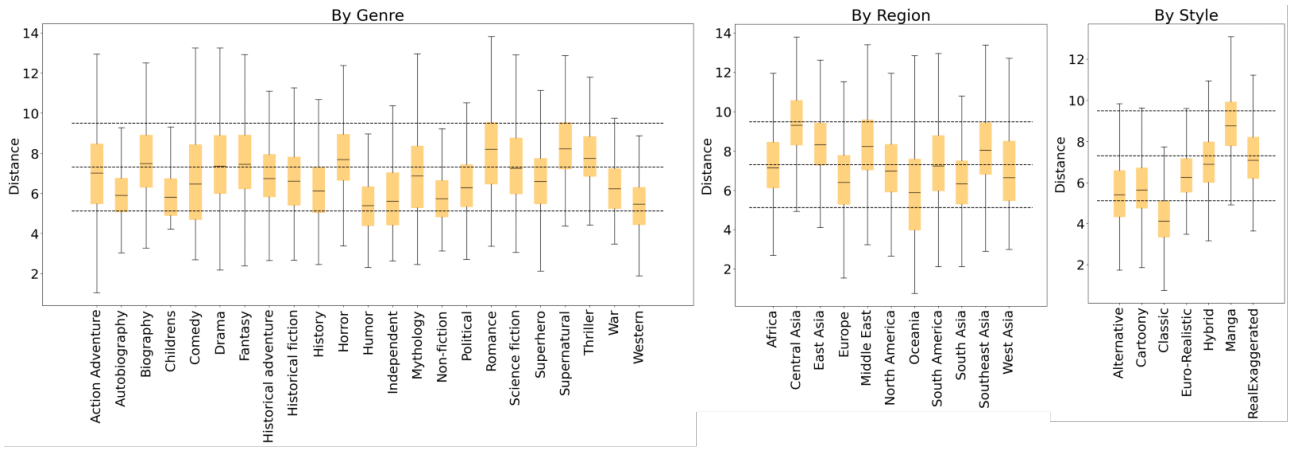


Figure 5.8: Distribution of Euclidean distance of ELRIC feature vectors within character clusters for different genres, regions of origin and styles of the TINTIN corpus.

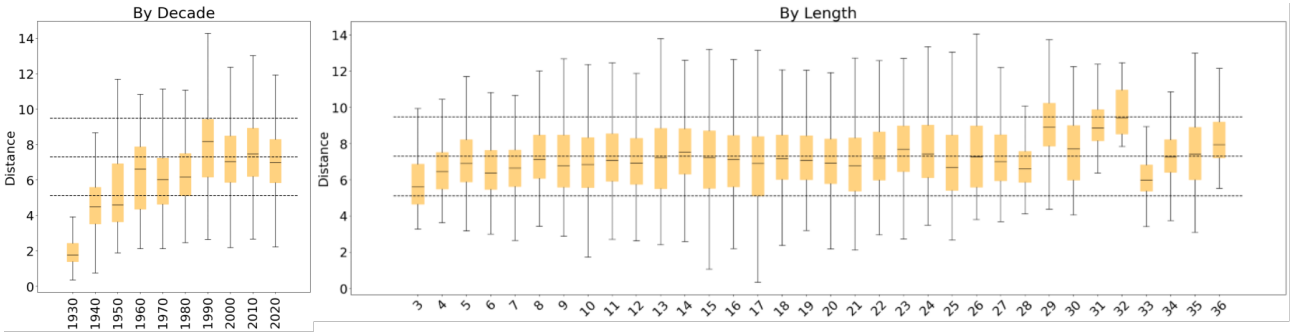


Figure 5.9: Distribution of Euclidean distance of ELRIC feature vectors within character clusters for different decades and comic lengths of the TINTIN corpus.

5.2.2 Time: Historical and Length

A similar analysis is applied to the time-based aspects of the TINTIN corpus using ELRIC character-level features. The results, displayed in Figure 5.9, mirror those obtained with ASTERX features. Specifically, we observe that the average intra-cluster distance has increased over the decades, suggesting that character representations become more diverse in more recent publications. Additionally, the average intra-cluster distance remains steady for comics up to about 30 pages, after which it increases, indicating that longer comics exhibit greater variability in character features.

5.2.3 Discussion

The ELRIC feature space analysis reveals significant variations in character instance features across different attributes, paralleling the ASTERX panel continuity findings.

Genres such as Autobiography, Children's, and Western exhibit smaller average intra-cluster distances, indicating more consistent character representations within these categories. In contrast, genres like Romance and Supernatural show larger distances, reflecting greater variability in character design and attributes. This implies that Romance and Supernatural genres tend to have a broader spectrum of character depictions. From a regional perspective, comics originating from Central Asia, Southeast Asia, and the Middle East demonstrate larger intra-cluster distances. This suggests that character designs in these regions are more varied, potentially reflecting the diverse cultural influences and storytelling techniques prevalent in these areas. Conversely, comics from Oceania and Europe show smaller distances, indicating a more uniform approach to character representation.

The analysis of style attributes reveals that the Manga style has the highest average intra-cluster distances, indicating significant character variability. This is consistent with related research into diversity and creativity in Manga character design and framing [45, 48], often spanning a wider range of visual elements. On the other hand, the Classic style exhibits the smallest distances, reflecting more consistent and perhaps traditional approaches to character representation - also in line with related work [40].

Examining the time-based aspects, the data indicates an increase in character representation variability over the decades, with modern comics displaying more diverse designs. This trend aligns with broader changes in the comic industry over the decades described

within the data [41, 55]. Furthermore, longer comics also exhibit greater character variability, suggesting that extended narratives provide more opportunities for allowing readers to make inferences based on a foundation laid in earlier pages of comics. This data aligns with leading theories within the domain inference theory and visual narratives [72, 71, 16]

Chapter 6. Conclusion

This thesis explores the complex domain of multi-level representation learning in comics, introducing ASTERX and ELRIC, two novelm methods in this domain. ASTERX is designed to capture sequential representations within comic panels through masked language modelling for visual language. ELRIC aims to normalise this contextual information, increasing performance in panel region-specific tasks. Our research demonstrates the efficacy of the ASTERX model in capturing the continuity and progression of comic narratives through sequential panel representations. ASTERX's ability to maintain contextual relationships between panels is validated through various experiments and evaluation tasks. ELRIC significantly outperforms baseline methods, in character matching and emotion classification tasks. The incorporation of context-independent features allows ELRIC to achieve higher performance in various metrics. In a broader scope, we showcase how domains which are not as data rich can benefit from tailored representation learning setups, such as the sequential setup for ASTERX and the context-normalised setup for ELRIC. The resulting feature spaces are shown to be a strong basis for cultural analysis of visual narratives. The results described in our cultural analyses align with existing theories within visual linguistics and psychology, but additionally providing a set of metrics that enable quantification of these theories from a novel perspective.

6.1 Limitations

The main limitation of our introduced methods is the lack of understanding of page layout. Certain aspects of visual narrative are encoded within the layout of a comic page. Sometimes panels within a comic page can even be read in multiple directions. Aspects such as narrative continuity on the level of page layouts are not entirely modelled in our approach, but often form a basis for both cultural analysis [58, 41, 62] and inferential analysis [13, 55, 4].

In addition, the multi-modality of comics and visual narratives as a whole are left unconsidered in this thesis. Where comics consist of both textual and visual features, our approaches are solely optimised for the visual aspects. The interplay between these multiple modalities are shown to convey [60, 65, 19] many aspects that are not captured by just one of the two modalities.

Bibliography

[1] Hans-Christian Christiansen Anne Magnussen. *Comics and Culture: Analytical and Theoretical Approaches to Comics*. 2000.

[2] Fred Atilla, Bien Klomberg, Bruno Cardoso, and Neil Cohn. Background check: cross-cultural differences in the spatial context of comic scenes. *Multimodal Communication*, 12:179–189, 2023.

[3] Olivier Augereau, Motoi Iwata, and Koichi Kise. A survey of comics research in computer science. *Journal of imaging*, 4:87, 2018.

[4] Irmak Hacımusaoğlu Bien Klomberg and Neil Cohn. Running through the who, where, and when: A cross-cultural analysis of situational changes in comics. *Discourse Processes*, 59(9):669–684, 2022.

[5] Marie-Thérèse Bornens. Problems brought about by "reading" a sequence of pictures. *Journal of Experimental Child Psychology*, 49:189–226, 1990.

[6] C. Brienza and P. Johnston. *Cultures of Comics Work*. 2016.

[7] Bruno Cardoso and Neil Cohn. The multimodal annotation software tool (MAST). pages 6822–6828, 2022.

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. pages 9630–9640, 2021.

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020.

[10] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning, 2022.

[11] Hillary Chute. Ragtime, kavalier amp; clay , and the framing of comics. *MFS Modern Fiction Studies*, 54(2):268–301, 2008.

[12] Neil Cohn. A different kind of cultural frame: An analysis of panels in american comics and japanese manga. *Image and Narrative*, 12(1):120–134, 2011.

[13] Neil Cohn. Visual narrative structure. *Cognitive science*, 37 3:413–52, 2013.

[14] Neil Cohn. In defense of a "grammar" in the visual language of comics. *Journal of Pragmatics*, 127:1–19, 2018.

[15] Neil Cohn. Being explicit about the implicit: inference generating techniques in visual narrative. *Language and Cognition*, 11(1):66–97, 2019.

[16] Neil Cohn. Your brain on comics: A cognitive model of visual narrative comprehension. *Topics in cognitive science*, 12(1):352–386, 2020.

[17] Neil Cohn. A starring role for inference in the neurocognition of visual narratives. *Cognitive Research: Principles and Implications*, 6(1), 2021.

[18] Neil Cohn, Jessika Axnér, Michaela Diercks, Rebecca Yeh, and Kaitlin Pederson. The cultural pages of comics: cross-cultural variation in page layouts. *Journal of Graphic Novels and Comics*, 10(1):67–86, 2019.

[19] Neil Cohn, Ryan Taylor, and Kaitlin Pederson. A picture is worth more words over time: Multimodality and narrative structure across eight decades of american superhero comics. *Multimodal Communication*, 6(1):19–37, 2017.

[20] Neil Cohn, Amaro Taylor-Weiner, and Suzanne Grossman. Framing attention in japanese and american comics: Cross-cultural differences in attentional structure. *Frontiers in Psychology*, 3, 2012.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. pages 248–255, 2009.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

[24] David Dubray and Jochen Laubrock. Deep cnn-based speech balloon detection and segmentation for comic books. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1237–1243, 2019.

[25] Randy Duncan and Matthew J Smith. *The power of comics*. 2009.

[26] Arpita Dutta and Samit Biswas. Cnn based extraction of panels/characters from bengali comic book page images. volume 1, pages 38–43, 2019.

[27] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023.

[28] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. 2020.

[29] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. 2021.

[30] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. 2019.

[31] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, abs/1512.03385:770–778, 2015.

[32] Zheqi He, Yafeng Zhou, Yongtao Wang, Siwei Wang, Xiaoqing Lu, Zhi Tang, and Ling Cai. An end-to-end quadrilateral regression network for comic panel extraction. In *Proceedings of the 26th ACM International Conference on Multimedia*, page 887–895, 2018.

[33] Anh Khoi Ngo Ho, Jean-Christophe Burie, and Jean-Marc Ogier. Panel and speech balloon extraction from comic books. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 424–428, 2012.

[34] John P. Hutson, Joseph P. Magliano, and Lester C. Loschky. Understanding moment-to-moment processing of visual narratives. *Cognitive Science*, 42(8):2999–3033, Nov. 2018.

[35] Bien Klomberg Irmak Hacımusaoğlu and Neil Cohn. Navigating meaning in the spatial layouts of comics: A cross-cultural corpus analysis. *Visual Cognition*, 31(2):126–137, 2023.

[36] Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan L. Boyd-Graber, Hal Daumé III, and Larry S. Davis. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. 2016.

[37] Brendan Jou and Shih-Fu Chang. Deep cross residual learning for multitask visual recognition. *Proceedings of the 24th ACM international conference on Multimedia*, 2016.

[38] Panayiota Kendeou, Catherine Bohn-Gettler, Mary Jane White, and Paul Van Den Broek. Children's inference generation across different media. *Journal of Research in Reading*, 31(3):259–272, Aug. 2008.

[39] Bien Klomberg and Neil Cohn. Picture perfect peaks: comprehension of inferential techniques in visual narratives. *Language and Cognition*, 14(4):596–621, 2022.

[40] Martha Kuhlman. *12. Design in Comics: Panels and Pages*, pages 172–192. 2020.

[41] Ralph LaRossa, Charles Jaret, Malati Gadgil, and G Robert Wynn. The changing culture of fatherhood in comic-strip families: A six-decade analysis. *Journal of Marriage and Family*, 62(2):375–387, 2000.

[42] Jochen Laubrock and Alexander Dunst. Computational approaches to comics analysis. *Topics in Cognitive Science*, 12(1):274–310, Nov. 2019.

[43] Janghyeon Lee, Donggyu Joo, H. Hong, and Junmo Kim. Residual continual learning. pages 4553–4560, 2020.

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[45] Dana Elizabeth Lubow. Framing the superheroine: Form and character in contemporary comics and manga. 2019.

[46] Javier Lucas, Antonio Javier Gallego, Jorge Calvo-Zaragoza, and Juan Carlos Martinez-Sevilla. Automatic detection of comic characters: An analysis of model robustness across domains. In *Document Analysis and Recognition – ICDAR 2023 Workshops*, pages 151–162, 2023.

[47] Joseph P. Magliano, Adam M. Larson, Karyn Higgs, and Lester C. Loschky. The relative roles of visuospatial and linguistic working memory systems in generating inferences during visual narrative comprehension. *Memory amp; Cognition*, 44(2):207–219, Oct. 2015.

[48] Dean Mervyn and Eryc Eryc. Analysis of character design in gacha games through manga matrix theory. In *CoMBInES-Conference on Management, Business, Innovation, Education and Social Sciences*, volume 4, pages 79–89, 2024.

[49] Irmak Hacımusaoğlu Neil Cohn and Bien Klomberg. The framing of subjectivity: Point-of-view in a cross-cultural analysis of comics. *Journal of Graphic Novels and Comics*, 14(3):336–350, 2023.

[50] Nhu-Van Nguyen, Christophe Rigaud, and Jean-Christophe Burie. Comic characters detection using deep learning. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 03, pages 41–46, 2017.

[51] Xuan-Sonand Rigaud Christopheand Jiang Liliand Burie Jean-Christophe Nguyen, Nhu-Vanand Vu. Icdar 2021 competition on multimodal emotion recognition on comics scenes. In *Document Analysis and Recognition – ICDAR 2021*, pages 767–782, 2021.

[52] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[53] Xufang Pang, Ying Cao, Rynson W.H. Lau, and Antoni B. Chan. A robust panel extraction method for manga. In *Proceedings of the 22nd ACM International Conference on Multimedia*, page 1125–1128, 2014.

[54] G. Parikesit. Quantitative image analysis of tintin comics. *Int. J. Arts Technol.*, 10:231–240, 2017.

[55] Kaitlin Pederson and Neil Cohn. The changing pages of comics: Page layouts across eight decades of american superhero comics. *Studies in Comics*, 7(1):7–28, 2016.

[56] Henry John Pratt. Narrative in Comics. *The Journal of Aesthetics and Art Criticism*, 67(1):107–117, 02 2009.

[57] Christophe Rigaud, Jean-Christophe Burie, and Jean-Marc Ogier. Text-independent speech balloon segmentation for comics and manga. In *Graphic Recognition. Current Trends and Challenges*, pages 133–147, 2017.

[58] Anne Magnussen Rikke Platz Cortsen, Erin La Cour. *Comics and Power: Representing and Questioning Culture, Subjects, and Communities*. 2015.

[59] Burieand Arnaud Revel Ruddy, Théodoseand Jean-Christophe. Automated emotion recognition through graphical cues on comics at character scale. In *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, pages 61–75, 2023.

[60] Adam Schwartz and Eliane Rubinstein-Ávila. Understanding the manga hype: Uncovering the multimodality of comic-book literacies. *Journal of Adolescent & Adult Literacy*, 50(1):40–49, 2006.

[61] Rishabh Sharma and Vinay Kukreja. Image segmentation, classification and recognition methods for comics: A decade systematic literature review. *Engineering Applications of Artificial Intelligence*, 131:107715, 2024.

[62] A. Silbermann. *Comics and Visual Culture: Research Studies from ten Countries*. 2010.

[63] E.J. Stainbrook. *Reading Comics: A Theoretical Analysis of Textuality and Discourse in the Comics Medium*. 2003.

[64] Weihan Sun, Jean-Christophe Burie, Jean-Marc Ogier, and Koichi Kise. Specific comic character detection using local feature matching. In *2013 12th International Conference on Document Analysis and Recognition*, pages 275–279, 2013.

[65] Miloš Tasić and Dušan Stamenković. The interplay of words and images in expressing multimodal metaphors in comics. *Procedia - Social and Behavioral Sciences*, 212:117–122, 2015.

[66] Barış Batuhan Topal, Deniz Yuret, and Tevfik Metin Sezgin. Domain-adaptive self-supervised pre-training for face body detection in drawings, 2023.

[67] Gido M. van de Ven, Nicholas Soures, and Dhireesha Kudithipudi. Continual learning and catastrophic forgetting, 2024.

[68] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[70] Haoqing Wang, Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhi-Hong Deng, and Kai Han. Masked image modeling with local multi-scale reconstruction, 2023.

[71] Janina Wildfeuer. The inferential semantics of comics: Panels and their meanings. *Poetics Today*, 40(2):215–234, 2019.

[72] Francisco Yus. Inferring from comics: A multi-stage account. *Quaderns de Filologia. Estudis de Comunicacio*, 3:223–249, 2008.