Multimodal Retrieval and Product Structure Proposal from 2D Technical Drawings

UNIVERSITY OF TURKU
Department of Computing
Master of Science (Tech) Thesis
Robotics and Autonomous Systems
July 2026
Syed Mehdi

Supervisors:
Adjunct Prof. Hashem Haghbayan
Prof. Juha Plosila

UNIVERSITY OF TURKU
Department of Computing

SYED MEHDI: Multimodal Retrieval and Product Structure Proposal from 2D Technical Drawings

Master of Science (Tech) Thesis, 55 p.
Robotics and Autonomous Systems
July 2026

Manufacturers manually produce Enterprise Resource Planning (ERP) product structures, bill of materials (BOM) and work phase plans for every new two-dimensional technical drawing. Although they usually have an archive of past drawings with linked ERP data records, it is impractical for the engineers to find similar drawings and data within it, especially if the archive is expansive. This thesis develops and evaluates a proof-of-concept computer aided process planning system that leverages this archive. Given a new drawing, it retrieves similar past drawings, and proposes a preliminary bill of materials and work-phase plan. The system employs a two-layer pipeline architecture. The first layer performs format-specific data extraction to produce a format-neutral representation. The second layer embeds this representation using a Multi-Layer Perceptron on the textual data and a frozen DINOv2 encoder on the visual data. These embeddings drive the retrieval, classification and proposal tasks. The system is evaluated on a 1829 PDF corpus with linked ERP data. The results are distinct and divided by task. ERP prediction is primarily driven by text. Work-phase routing is predicted with multi-label classification over a 21-class production line taxonomy, and scores 0.82 macro-F_1, 0.92 micro-F_1 with 0.60 exact set match rate while staying vendor neutral. The long-tailed BOM record is addressed using a classifier for recurring components and a retrieval pool for the tail. Visual channel proves stronger for retrieving similar drawings that are human relevant. Measured against human judgement, a frozen mean-pool encoder attains a top 5 hit-rate of 0.84, but it does not improve ERP prediction. Visual similarity does not seem to suggest shared product structures, and multimodal fusion yields no performance increase. The thesis provides a vendor-agnostic, retrieve-and-propose framework and its prototype, together with a realistic discussion of its capabilities on real industrial data.

Keywords: engineering drawings, content-based retrieval, bill of materials, computer aided process planning, multi-label classification, representation learning, vendor-agnostic, DINOv2

Acknowledgements

I would like to thank my supervisor Hashem Haghbayan, for his support throughout this thesis. I am also grateful to Stera Technologies for providing the dataset and domain expertise that made this work possible, and to my family and friends for their encouragement.

Abbreviations and Notation

BOM	Bill of Materials
ERP	Enterprise Resource Planning
CBR	Case-Based Reasoning
CAPP	Computer-Aided Process Planning
GD&T	Geometric Dimensioning and Tolerancing
OCR	Optical Character Recognition
CNN	Convolutional Neural Network
GNN	Graph Neural Network
ViT	Vision Transformer
SAM	Segment Anything Model
TF-IDF	Term Frequency-Inverse Document Frequency
MLP	Multilayer Perceptron
SSL	Self-Supervised Learning
MRR	Mean Reciprocal Rank
mAP	Mean Average Precision
CAD	Computer-Aided Design
DXF / STEP	CAD file formats
F_1 **Score**	Harmonic mean of precision and recall
Macro-F_1	Mean of the per-class F_1 scores with every class weighted equally
Micro-F_1	Pooled F_1 across all classes, dominated by the frequent classes
Exact-match	Drawings with predicted label set matching the ground truth
Set Distance $\pm k$	Predictions differing from the truth by at most k labels.
P@k	Precision at rank k
Recall@k	Recall at rank k
Success@k	Fraction of queries with a relevant item in the top k (hit-rate)
Jaccard	Set overlap (intersection over union)

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

For every 2D technical drawing processed and manufactured by the industrial partner of this thesis, Stera Technologies, it must manually produce the bill of materials (BOM), listing the components the part is built from, and enterprise resource planning (ERP) data such as a work phase plan, sequence of production operations (welding, painting, bending) that the product passes through. Presently both these jobs are handled from the ground up by engineers who interpret the drawings and recall how comparable parts were previously manufactured completely dependent on their experience. The company holds a large archive of past drawings together with BOMs and work phase plans created for them. In principle every new part must have a precedent to learn from, but the archive is not usable due to its sheer scale. Finding genuinely comparable drawings among many thousands through a manual search is impractical.

This is a classic problem where variant process planning could be realized through case-based reasoning. We must retrieve similar past drawing - ERP/BOM pairs and adapt their solutions to the current drawing [1], [2]. The goal is to build a system that when given a new drawing, surfaces its closest precedents from the archive and uses them to propose preliminary BOM and work phase data therefore turning a heavily manual process into a review and adjust task. The scope of this thesis is a proof-of-concept prototype of this system evaluated on a restricted two vendor dataset, not a complete production-ready

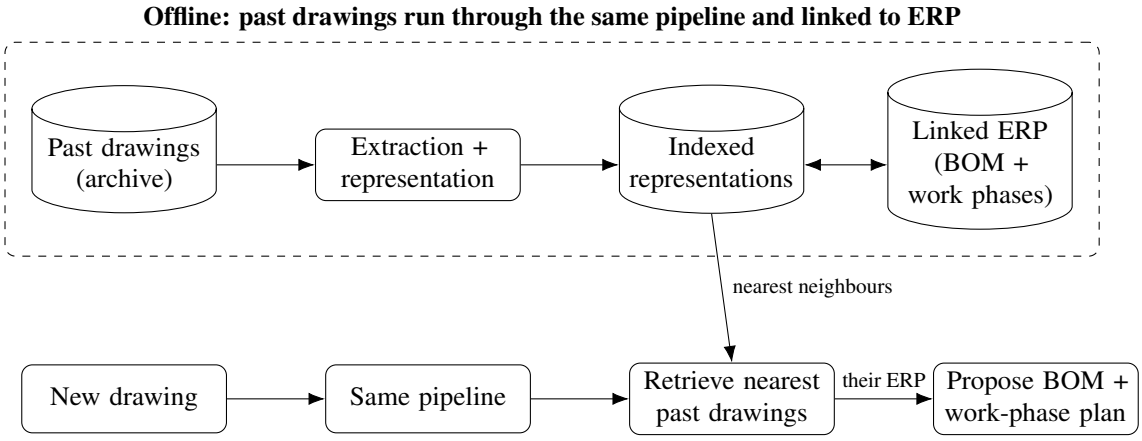


Figure 1.1: Retrieve-and-propose workflow.

system. We aim to establish what signals are extractable and what methods make ERP retrieval and generation feasible.

1.2 Problem statement and Scope

The delivered dataset contains only PDF drawings (no DXF or STEP), so the implementation is focused on PDF inputs. However, the architecture is layered so other formats could be added later without a redesign (Chapter 5). Two major goals are pursued.

Goal 1: propose ERP product structure Given a new drawing, propose its work phase plan and BOM. These two behave differently and are handled and addressed differently. The work-phase plan is small, has a closed taxonomy and is predicted by a trained classifier. The BOM signal in our dataset has long tailed distribution. Most components occur on a single drawing, so it is handled by classification for recurring components and retrieval of a candidate pool from similar drawings for unique components.

Goal 2: retrieve similar drawings Given a drawing, retrieve archived drawings of parts that are similar visually and in manufacturing. This channel is primarily dominated by visual data. As we discuss in Chapter 8, we succeed at finding similar drawings,

but these drawings do not reliably improve Goal 1 predictions according to our measurements. Looking similar does not guarantee sharing product structure.

Features that signal relevance are intrinsic to the drawing (title block data, notes, geometry, material, process specifications) so the approach is vendor agnostic. It works for any vendor without exploiting vendor specific cosmetics as a shortcut. The framing of the scope discussed above is the result of refining the original project proposal based on the nature of the available dataset, and the success and failure of experiments, specific adaptations and their reasons are discussed in Chapter 4.

Research Questions

The thesis tackles three research questions.

RQ1 Which intrinsic modality (visual or textual), and in what form, best retrieves manufacturing similar drawings? How do we quantify similarity when part type and ERP labels prove to be a poor proxy?

RQ2 Can we propose a drawing's product structure, work phase data and BOM purely through learned classification when the label space allows and by retrieval from archival data otherwise? Does multimodal fusion of text and visual channels improve performance over best single channel, and is the result vendor agnostic?

RQ3 How much data can we extract directly from vector-native PDF drawings before we require image-based fallback for our extraction pipeline?

1.3 Contributions

This thesis makes the following contributions:

1. A vendor agnostic retrieval and proposal framework for 2D technical drawings that separates format specific extraction from format independent representation, delivered as a proof-of-concept prototype.

2. A two-tier ERP proposal pipeline that classifies where the label space allows and retrieves for the unique non-repetitive data, together with its evaluation. Work-phase routing is predicted with multi-label classification over the customer's 21-class production-line taxonomy, evaluated on a held-out test set from an iterative stratified split. We reach a macro-F_1 of 0.82 (F_1 score averaged equally over all production lines), a micro-F_1 of 0.92 (F_1 score pooled over all label decisions, dominated by the common lines), and an exact-match rate of 0.60 . Recurring BOM components (that appear in multiple drawings) are predicted via a classifier (micro-F_1 0.84), while single-occurrence components are handled by a retrieval pool of suggestions.

3. A visual retrieval pipeline for drawings, evaluated against human judgement using a blind, de-identified pooling protocol. After segmentation into individual views, a frozen DINOv2 encoder with mean-pooled per-view embeddings is the strongest retriever tested (Success@5 0.84, P@1 0.69), and backbone adaptation is on par with it.

4. A demonstration of how the prediction target definition matters more than the choice of learning model. Predicting the customer's clean 21-class production-line taxonomy, instead of the raw free-text phase descriptions recorded in the ERP, raised work-phase exact-match from 0.36 to 0.60 with an unchanged learning model.

2 Background

This chapter establishes concepts that the rest of the thesis relies on. It covers what 2D technical drawings convey, the organization of enterprise resource planning (ERP) product structure (bill of materials and work phase data), the current state of engineering drawing digitization, representation learning and case-based reasoning which combine to form our system. Specific methods and respective limitations are deferred to Chapter 3.

2.1 Technical drawings as manufacturing documents

A 2D technical drawing conveys the geometry and manufacturing requirements of a part through standardized conventions. A part is illustrated in multiple orthographic views (front, top, side). It may be further supplemented by isometric views, section views that expose internal structures, and detail views that enlarge small features. Titleblocks and notes record metadata, part description, material, scale, tolerance classes, surface treatment and weight. The drawings are also annotated with dimensions, and symbols for welds, geometric dimensioning and tolerances (GD&T) and surface finish. Conventions codified in ISO and DIN standards are also present, but due to the lack of institutional accessibility of these standards for work, the vocabulary used to interpret the drawings is instead grounded in peer-reviewed manufacturing references that reproduce the taxonomies. [3]–[5]

A property worth noting for dissection is that information is heterogeneous, discrete texts exist in title blocks, notes are free-form and free-floating, dense dimension values,

graphical symbols and the geometric views of the part all coexist on a single page and are not equally recoverable.

2.2 Product structures: bill of materials and work phases

The prediction targets for this thesis come from the partner's ERP system and are described in the accompanying documentation. Bill of Materials is provided in its original ERP form, where each row provides a parent part, a component, a quantity, and the work phase at which the component is consumed. Critically to note, the BOM does not contain the full structure of multi-level products in one place. A top-level item is linked only to its immediate components and sub-assemblies, and there is no direct reference from a deep component back to the final top-level item. Most of the drawings in the corpus are top-level assembly items and their children may not always be present. This structural nature of the corpus makes generative BOM prediction infeasible and calls for our retrieval-based solution in Chapter 8.

Likewise, a work-phase plan is also provided at item-level. It is an ordered set of operations, tied to their respective work centres (for example welding consistently gets referred to work centre 250) with setup and runtimes. The work centres then consolidate into higher level production-line categories like bending, welding, laser etc. The production-line taxonomy is internal to the company and we use this clean taxonomy as our work-phase classification target. As we discuss in Chapter 8, using clean taxonomy instead of open phase description texts proves decisive in our results. Similar to BOM data, the work-phases carry no hierarchy between items.

2.3 Engineering drawing digitization

Automatic interpretation of engineering drawings is an open and established field.

Recent surveys by [6], [7] describe the field largely as recognition and transcription

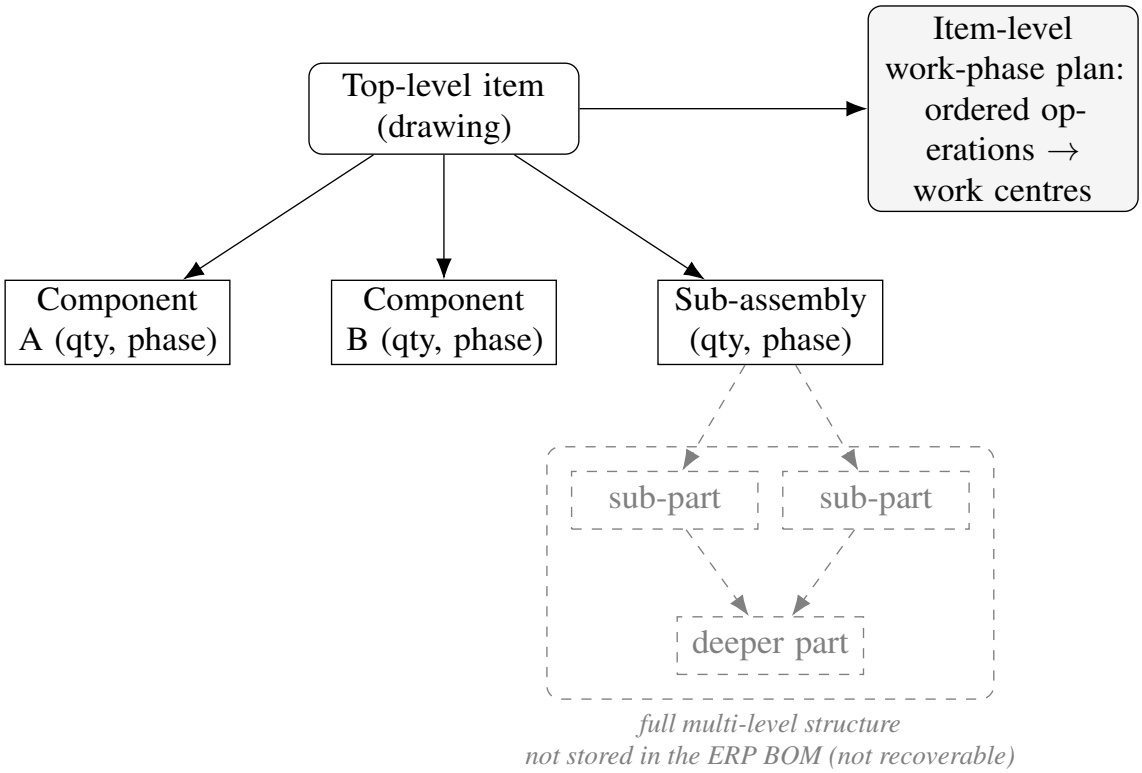


Figure 2.1: ERP product structure.

of text, symbols, dimensions and geometry on a sheet into a structured representation. All efforts largely and consistently encounter the same obstacles, contents where the text, symbology and dimensions are spatially and semantically entangled, scarcity of annotated training data and apparent class imbalances such as the under-representation of mechanical/production drawings relative to piping and instrumentation schematics. The surveys note that all solutions lie on a spectrum from bottom-up (assembling primitive structures into meaning) to top-down (Imposing an expected structure given the top-level data). This observation guides our extraction pipeline design in Chapter 5. We emphasize on one key distinction, currently the field of engineering drawing digitization evaluates how well a drawing is read (converted to meaningful information), not how usefully it can be matched to precedents and other similar drawings for downstream manufacturing decision-making, and that is the gap addressed by the thesis.

2.4 Representation learning and content-based retrieval

The ability to retrieve similar items by content relies on representation learning, being able to map each item to a vector, such that their proximity in vector space reflects similarity, so retrieval is essentially reduced to a nearest neighbor search. In the context of this thesis, a drawing's representation means fixed-form numerical encoding of its intrinsic signals, textual or visual, over which all retrieval and prediction operate. Representation learning is the act of training the encoder that produces it. In the visual domain, vision backbones such as DINOv2 [8] encode images into general-purpose features. In the textual domain, sentence encoders like Sentence-BERT [9] can produce embeddings that can approximate semantic similarity through cosine similarity. These models can be used either frozen (as a fixed feature extractor based on pretrained data) or adapted (further trained on domain specific data) and a central empirical question that we examine in Chapter 8, is whether adaptation is worth the cost for our use-case.

The way our thesis uses the representation created by models can be best explained as case-based reasoning (CBR), a learning friendly formula where we retrieve similar past cases, then reuse and revise their solutions [1]. CBR has seen broad industrial adoption across manufacturing and its value lies in the explainability of its decisions grounded in concrete precedents [2]. The central theory is directly relevant to our thesis, that overall performance of our solution will be dominated by the retrieval quality and how we define our case-similarity function, and that is why making proper representations is the heart of this thesis. The learned distance between the drawings (their representations) is our case-similarity function, and the contribution we aim to make is to define and evaluate this distance against actual ERP payoff instead of a nominal label of similarity.

3 Literature Review

This section will position the thesis against five prior tracks of work. (i) surveys on engineering drawing digitization (ii) domain specific prior work on mechanical drawings (iii) content based retrieval of technical drawings (iv) visual and textual representation learning (v) the case-based reasoning paradigm. As per observation, the drawing digitization field currently evaluates extraction i.e. how accurately text, symbols or dimensions are read, whereas our thesis, evaluates retrieval and prediction against downstream ERP target. To our understanding and reading, no prior paper on drawings reports on this combination.

3.1 Surveys and contextualization framing

Two recent surveys map the engineering drawing digitization landscape, Jamieson et al. [7] review deep learning methods applied for digitization of complex documents and engineering drawings, and Moreno-García et al. [6] navigates newer trends in the area. Both converge on the same problems. Drawing content is a tangled mixture of texts, symbols and dimensions. Annotated data is scarce, and class imbalances are severe. Both consider the task to comprise of recognition and extraction where the output is a structured transcription of the sheet. Our goal is not to judge the quality of extraction, rather the quality of downstream manufacturing ERP payoff. And that is where we deviate, we use the extracted signal not as the end itself but a query to be used against historical archives. The surveys caution that extraction on realistic data is usually imperfect. Haar et al. [10] report recognition extracting nearly 70% of information from production drawings and

Strien et al. [11] show that OCR error can harm downstream tasks unevenly and even over-aggressive cleaning can destroy the signal quality. We understand the risks as stated by both, and keep text cleaning light and to also judge it, by its effect on the downstream metrics, rather than extraction coverage and quality.

3.2 Prior domain work on mechanical drawings

The closest analogue to our current task is Xie et al. [12] where graph neural networks classify (single label) manufacturing method based on an engineering drawing (turning, sheet metal bending and others). It establishes that we can learn manufacturing routing labels from drawing structures, but it is a single-label classifier over one method, whereas our work-phase classification is a multi-label question needing a set of production lines per drawing. Van Daele et al. [13] neurally learn symbolic parsers that read tabular content of technical drawings (like title blocks), support the view of hybrid extraction pipelines but stopping at extraction. Classical feature recognition techniques like You and Yang [14] assume single parts represented in all orthographic views which is often violated by our corpus, that is composed dominantly of top-level assemblies (Chapter 4). This mismatch itself tells us why an assembly level retrieval framing is required. Component level problems have been targeted with graph matching and graph convolutional neural networks [15], [16], and some recent work also applies vision-language models for multi-view interpretation [17], however both operate below whole-drawing levels and therefore may not be directly applicable to a real manufacturing pipeline.

3.3 Content based retrieval of technical drawings

Baba et al. [18] build a similarity driven partial image retrieval system for engineering drawings, and another paper by the group technology tradition pursues encoding parts for similarity-based reuse [19]. More recently, Quan et al. [20] present a self-supervised con-

trastive (Contrastive SSL) graph neural network for mechanical CAD drawings, motivated by the expensive nature of manual labelling of similar parts for machine learning. We share this motivation and it has guided our model choice and use of human relevance judgement over nominal labels for some results. However, they [20] work with parametric 3D CAD data not 2D drawings. We repeatedly notice the common trend that there is an evaluation gap, retrieval quality is measured against hand-curated similarity and never downstream product structure targets. The present thesis positions itself to use a combination of both.

3.4 Representation learning for drawings and line art

Our visual channel builds on the self-supervised vision backbone DINOv2 [8], that produces strong general-purpose features without labels. Literature finds that frozen DINO backbone is weak on line art and adaptation is the key. Saavedra et al. [21] adapts a DINOv2 backbone for self-supervised sketch based image retrieval, along with its companion study [22]. We revisit these claims in Chapter 8, and since automatic and ERP proxies proved unreliable in our setting, contrasting against human judgement as our metric. If we consider the state of the art for retrieval, the strongest results come not from DINO but rather convolutional encoders with margin based metric losses and multi-view aggregation. Higuchi and Yanai [23] retrieve patent figures with a pipeline of this format, Su et al. [24] establish multi-view convolutional aggregation, and Deng et al. [25] provide the additive angular margin loss widely used in this family of work. DeepPatent line [26], [27] provides large scale drawing retrieval benchmarks. The contrastive objectives underpinning our adaptation experiments, supervised contrastive learning [28] and SimCLR [29], are standard, and the transfer-learning literature [30] tells us that out-of-domain backbones may not transfer, motivating a measure and then adapt philosophy. It is to be noted that a substantial subset of representation learning methods, operate on 3D point clouds and graphs, dynamic graphs CNNs [31] and differentiable graph pooling [32], but they demand

3D geometry or point cloud data which a flat 2D PDF drawing simply cannot provide, so they are not directly applicable for our setting and are mentioned only to demarcate the border.

3.5 Textual and OCR work

Title block metadata extraction has been performed with frame detection with successive per-cell Optical Character Recognition (OCR) and keyword dictionary lookup [33], and by Ondrejcek et al. [34] earlier. Both evaluate on per-field accuracy. Scheibel et al. [35] extract dimension requirements directly from vector PDFs using bounding box clustering and no training, the closest single methodology to our vector-native data setting, although the scope is tied to dimensions rather than semantic data. François et al. [36] show that clustering OCR tags by edit distance lifts recognition without any lexicon. This normaliser is applicable on any extractor. As established by Strien et al. [11], OCR noise harms downstream task performance and overcleaning is also harmful. This is in line with our finding (Chapter 8) that text representation barely changes downstream performance of representation learning tasks, however it may be useful for other purposes.

3.6 Text similarity representations

Our research also questions if a dedicated text similarity model can rerank visual candidates. Sentence BERT [9] learns dense sentence embeddings through Siamese objectives such that cosine distances can approximate semantic similarity between embeddings. SimCSE [37] obtains competitive sentence embeddings without labels through contrastive dropout objectives. It is also established that lexical and dense representations of text are complementary. Lexical matching excels at precise identifiers (standard codes, material grades) and dense representations can capture paraphrasing so a hybrid of both beats either in terms of similarity retrieval. We apply unsupervised SimCSE and lexical matching to

our extracted textual data in Chapter 8, but neither improves the similarity by reranking of drawings on our small corpus.

3.7 Symbolic detection

Manufacturing symbols, welding callouts, geometrics dimensioning and tolerances, surface finish markers carry process information and are graphical rather than textual. Recent detectors target these symbols. Reddy et al. [38] compares modern object detectors for GD&T symbols, Bickel et al. [39] detect symbols in mechanical drawings with synthetic training data, and Fan et al. [40] perform panoptic symbol detection in CAD drawings. This corresponds with our findings that there is no textual shortcut for these symbols, and a dedicated visual extractor is required for them. Additionally, OCR is unable to pick these up as it is designed to work with standard ASCI or adjacent characters. We, therefore, need to scope the symbolic channel out of the current thesis, treating these as a known visual gap until we can train a visual pipeline to detect them specifically.

3.8 View segmentation and BOM from drawing

Our pipeline is designed to isolate individual drawing views before embedding. For segmentation we are using Segment Anything Model (SAM) [41] as the basis of our vendor agnostic view filter. The requirement of per-view embedding over whole-page embedding will be discussed later (Chapter 6). For creation of bill of materials from drawings, [42] is our closest precedent, which generates a BOM from drawings using a MASK R-CNN detector. The work is entirely vision based, single domain on construction formwork and reports on detection mean average precision. Our BOM construction is framed as retrieval and classification against ERP ground truth data, therefore we define different set-overlap metrics, as mAP does not measure the product structure overlap we are interested in.

3.9 Case-based reasoning paradigm

This process is simple, retrieve similar past cases, then reuse and revise their solutions. Zhou et al. [1] makes a concrete case-based reasoning mechanism for manufacturing process planning, local similarity computed per factor global similarity for retrieval, and adapting the nearest past plan. This is exactly the kind of function our representation should learn. Khosravani and Nasiri [2] review the use of case-based reasoning across an entire manufacturing process and emphasize its explainability and industrial uptake. Zhao and Melkote [43] show that material and part-type or quality are reliable indicators of process capability. These exist as clean and readable text in our textual title block data and gives some explanations to why textual channel massively dominated work phase predictions. The taxonomy we have based our vocabulary matching on is taken from references [3]–[5] as we did not have direct access to the underlying ISO/DIN standards.

3.10 Positioning

After reviewing our literature, our thesis is framed into a specific and unfilled position: (i) the digitization field measures extraction quality per field, while we evaluate retrieval quality of downstream ERP target; (ii) the nearest task analogues are either single-label (manufacturing method classification) or vision based single-domain (BOM generation), while we predict multi element product structures, through classification where there is sufficient data and with retrieval for long tailed data in a vendor agnostic manner; (iii) the vision similarity is tested through human judgement instead of automatic proxies; (iv) pretrained text similarity models exist but are untested on technical drawing text, we apply them and report negative results; (v) finally, our design uses already established CAPP/CBR paradigm, with novelty residing in our formulation and evaluation.

Table 3.1: Closest prior work.

Work	Task / approach	Gap vs. this thesis
Xie et al. [12]	Manufacturing method classification (GNN)	Single-label. We predict a multi-label set
Chowdhury and Moon [42]	BOM from drawings (Mask R-CNN, mAP)	Vision-only and single-domain. We use ERP set
Quan et al. [20]	Self-supervised contrastive CAD retrieval	3D parametric CAD, not 2D-drawing text/image
Kashevnik et al. [33]	Title-block extraction (per-field)	Extraction, not downstream ERP retrieval
This thesis	Vendor agnostic retrieve-and-propose	Retrieval judged by ERP payoff, classification where possible

4 Dataset and problem

This chapter details the characteristics of the data the system is built on and evaluated against, the separation of training signals from prediction targets and how the project objectives departed from the original proposal and why. We detail all the data findings that direct every design choice that followed.

4.1 The corpus

The partner provided a proof-of-concept dataset of 1829 PDF technical drawings drawn from two customers. Company A parts are primarily sheet metal, and Company B parts are thicker, welded structures. In our processing the data split of usable drawings out of the total 1829 is 857 Company A and 970 Company B drawings, 1827 that yield extractable text. The drawings are vector-native PDFs rather than scans, and 52.1% are multi-page (typically one or two pages, up to eighteen). However, being a vector does not guarantee extractable text; about 7-9% of drawings carry nearly no extractable text layer because it is path rendered as vector outlines (predominantly in the Company B half of the corpus). Each drawing is joined with ERP records through a drawing identifier embedded in the item description. The join is nearly complete 1828 of 1829 drawings (99.95%) resolve to at least one ERP item, with a single genuine orphan absent from the ERP. As the dataset was assembled from top-level items, the corpus is 94.2% top-level assemblies by construction (1723/1829 items that appear in the bill of materials as a parent and are never consumed as a component), and only 11.6% of drawings appearing as a sub-component of something

Table 4.1: Corpus statistics.

Quantity	Value
PDF drawings (text-extractable)	1829 (1827)
Vendor split (Company A / Company B)	857 / 970
Multi-page drawings	52.1 %
Linked to ≥ 1 ERP item	1828/1829 (99.95 %)
Top-level assemblies (by construction)	1723/1829 (94.2 %)
Distinct BOM components	3401
Single-use components	67 % (79 % by BOM row)

larger. At the item level only 31% of all referenced ERP items are top-level while 68% are external components that are never drawn. This is the structural reason a generative bill of materials is infeasible on this data.

4.2 ERP data

The drawing dataset is accompanied by four CSV exports, item basic data, bill of materials, work phases, and work-centre basic data. The bill of materials is in native ERP form, each row containing a direct parent to component link, and therefore does not contain a full multi-level structure. The BOM file contains 3401 distinct components and is disproportionately long tailed. 67% of distinct sub-components occur on a single drawing only (79% if counted by BOM rows). The median drawing links to a small number of sub-components with a few having several dozen.

The work-phase target is the customer's production-line taxonomy. Out of all the production lines defined, 26 are present in our drawings and 21 occur frequently enough to train and evaluate (≥ 20 examples each). The remaining lines are either never used on this corpus or are too rare to learn. These 21 lines, bending, blasting, inspection, laser, machining, packing, painting, programming, threading, welding, and so on, including automated (robotics) production lines indicated by ".A" suffixes are the multi-label target of the work-phase classifier (Chapter 8).

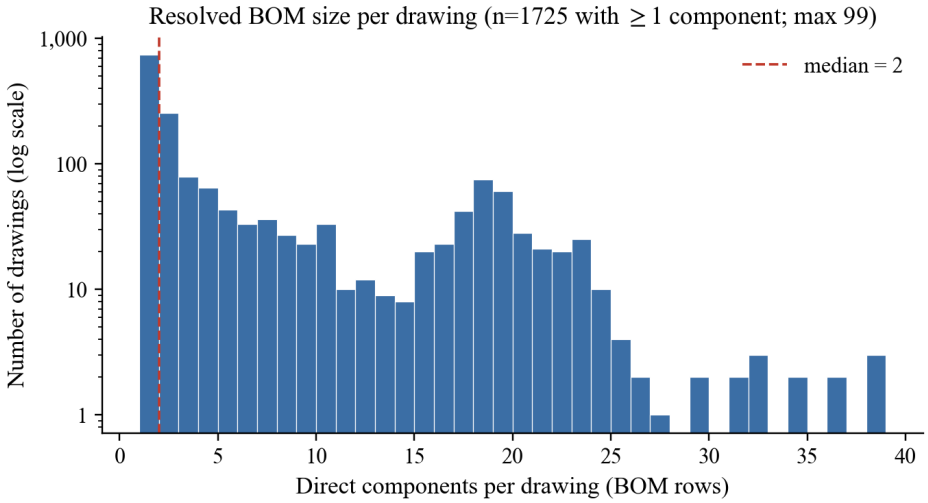


Figure 4.1: Distribution of resolved BOM size per drawing.

4.3 Signal-Target distinction

To avoid information leakage a contract is needed to govern design decisions, so the input signal and prediction targets do not mix. The input signal is everything we extract from the drawing itself such as title block fields, notes, dimensions, geometry, and the embedded BOM table when one is printed on the sheet, and it is available at inference for any drawing from any vendor. The target is the ERP data: the work-phase plan (pure prediction target) and the BOM (a retrieval-prediction target, since most components are non-recurring). The target is available only during training and must be predicted after training is complete.

To clarify one key point, the BOM table printed on a drawing is an input signal not a target, it is intrinsic to the sheet, legitimately readable, and it does not join with ERP BOM rows, so using it as input is not the same as leaking the ERP target.

4.4 Data findings and design consequences

The objectives stated here are a result of refining the original proposal around the nature of the data provided and as experiments succeeded or failed. Several properties of the data fix the shape of the system and account for how those objectives have been adapted away from

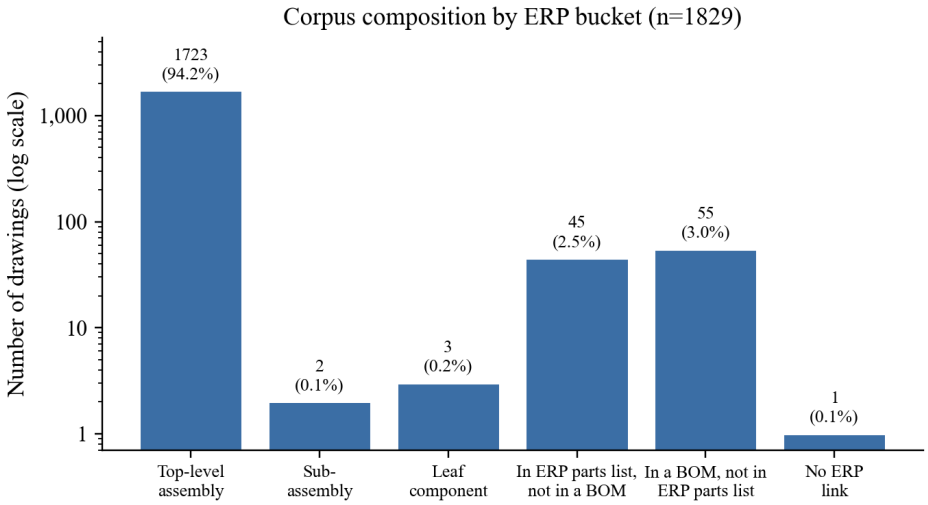


Figure 4.2: Corpus composition.

the original proposal. Table 4.2 pairs each finding with the design decision it forces, and each is traceable to a result reported later. These findings fix the per-view visual channel, decide the classify or retrieve, provide a vendor agnostic shape to the system (Chapter 5) while also setting the evaluation expectations (Chapter 7).

Table 4.2: Data-level findings and the design decisions they force, including the departures from the original proposal.

Finding / data property	Design consequence (and departure from the proposal)
The BOM is long tailed: 67 % of components are single-use, and the corpus is predominantly top-level assemblies whose children are never drawn	Generative or pure-aggregation BOM is infeasible. The proposal's retrieval-aggregation becomes a *classifier* for recurring components plus a *retrieval pool* for the long tail.
Work-phase content is near-ubiquitous across drawings.	Copying phases from neighbours is uninformative. The proposal's retrieval-aggregation becomes *direct classification* over the production-line taxonomy.

continued on next page

Table 4.2 (continued)

Finding / data property	Design consequence (and departure from the proposal)
Whole-page visual embeddings cluster drawings by vendor cosmetics, not part content (Chapter 6).	The visual channel must operate *per-view*. This channel, parked in the proposal, is resumed as the core of the similarity goal.
There is no reason a Company A part has a Company B analogue.	Cross-vendor retrieval, a proposal goal, is ill-posed and is replaced by a *vendor agnostic* constraint.
Process symbology is graphical, not textual: welding, GD&T and surface finish leave essentially no text.	The symbolic processes are a visual-only gap, scoped out. Only cheap textual signals (tolerance class, paint code) are kept.
Many drawings are short and information poor.	Work-phase accuracy is bounded below independently of the model (Chapter 8).
Text representation is performance neutral for the prediction models (Chapter 8).	The proposal's "hybrid representation as the performance lever" is dropped. Structured text is retained for human-readable output, not model accuracy.

5 Methodology

This chapter describes the system design. The two layer pipeline, per format extraction, textual and visual extraction channels and the downstream classification, retrieval and similarity mechanisms. Vendor neutrality is treated as an explicit design choice rather than an afterthought.

5.1 Two-layer pipeline architecture

The pipeline is divided into two distinct layers so later work can generalize beyond the single delivered format (PDF). Layer A comprises of input specific extractors that convert *input file* data into an intermediate format. For PDF drawings, these include vector rendering, vector to text extraction, OCR fallback for text extraction and view segmentation. Layer B uses different models (conventional, neural, graphical etc.) to convert the visual and textual data in the intermediate format into representations over which all the downstream prediction and retrieval operations can be performed. The input format agnostic intermediate, consists of sets of text units, region-tagged textual data (a title block, cell tables, body notes, BOM tables), and per-view images. Even for different input file types, like DXF or STEP, only a separate Layer A is required to produce the same intermediate, that will still be routed through the same Layer B and multi-format support could be achieved. The implementation of multi-format support is deferred to later work, not an abandoned objective, and we will keep the current thesis implementation focused on vector PDFs.

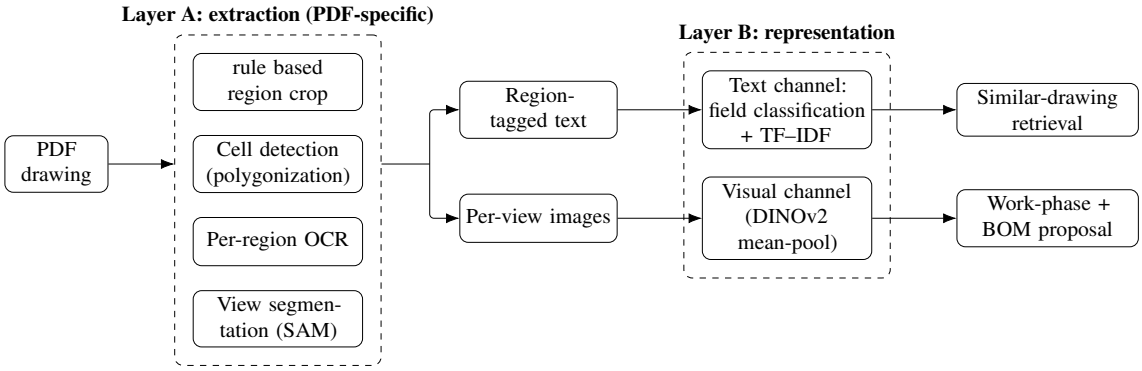


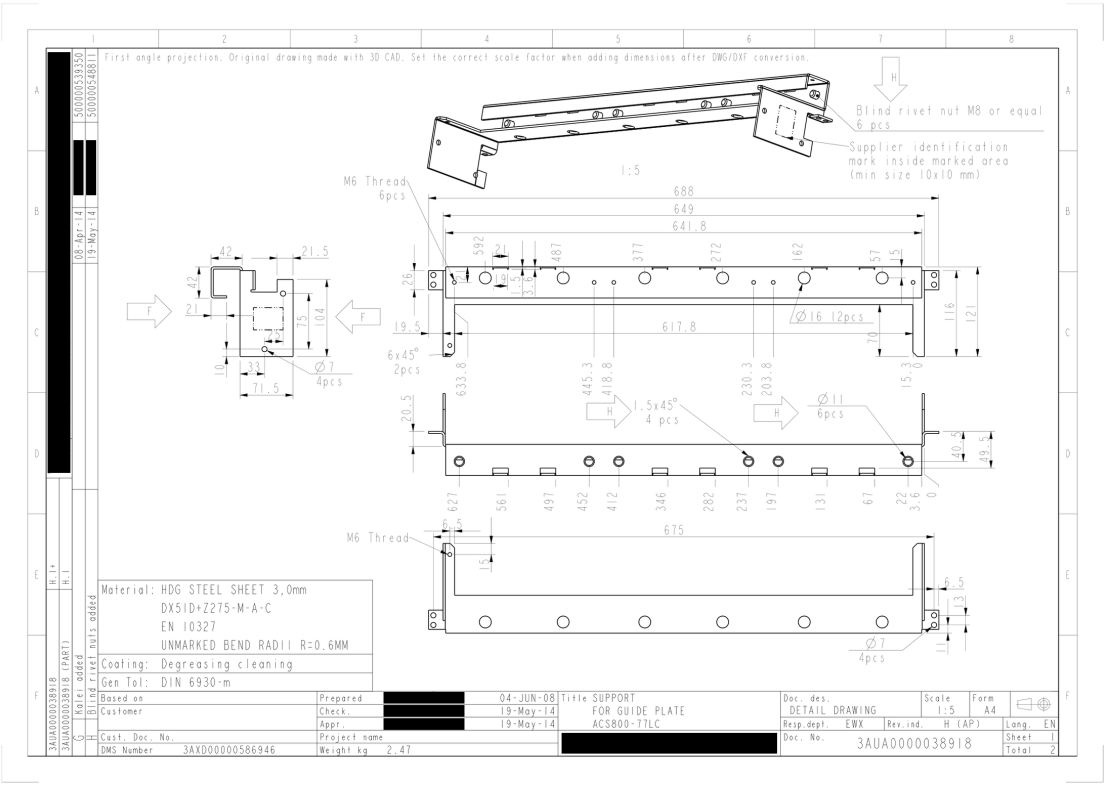
Figure 5.1: Two-layer pipeline architecture. Layer A (PDF-specific extraction) emits a format agnostic intermediate, region-tagged text and per-view images, which Layer B represents and the retrieval and ERP heads consume.

5.2 Extraction layer (A)

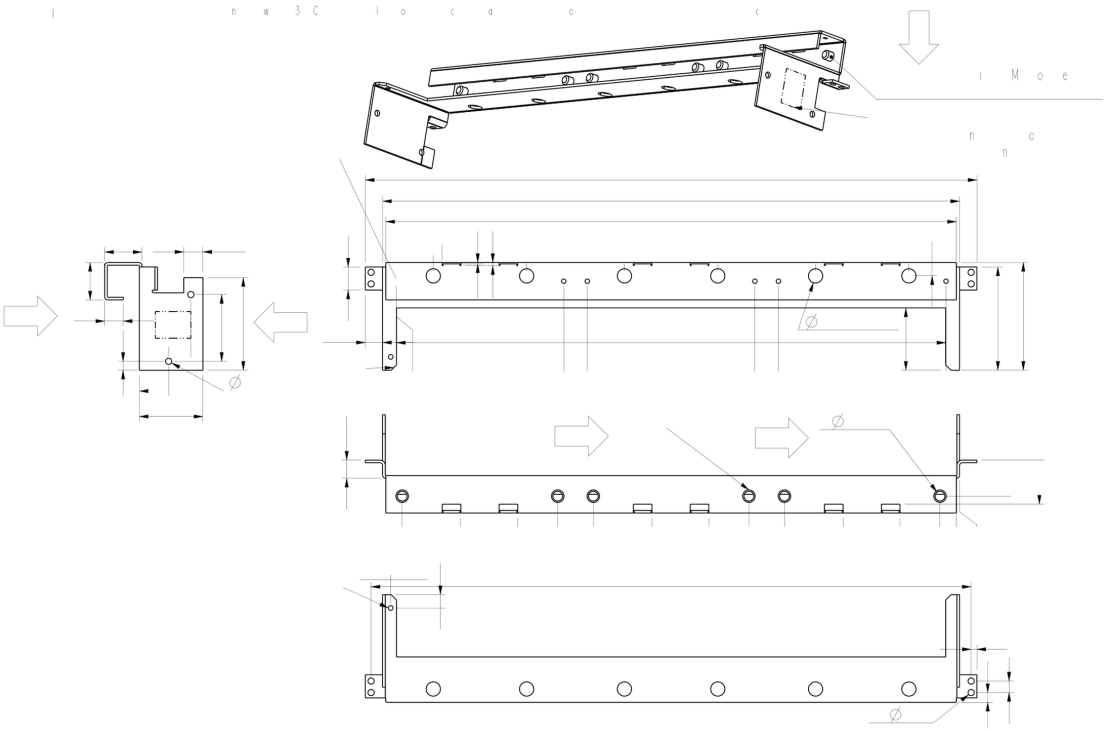
Extraction takes place in a number of steps that are listed as follows:

Region detection: Both visual cleanup and textual subdivision require region segmentation. For our vector data, axis-aligned line segments that usually make up the drawing's frame and title block grid are picked up, merged into closed polygons using Shapely's polygonize function. Edge-connected polygons make up the title block and cells, while the remainder are part of the drawing views. Hence, we obtain a cell map of the technical drawing.

Vector-native re-rendering: Rather than using a rasterized page, geometry is drawn directly from the PDF vector paths (pymupdf Shape API) to produce clean per-sheet images. The advantage lies in how any vector (for example text) can be individually removed without painting white bounding boxes, erasing adjacent data. To produce a cleaner input for visual extraction, the title block and table cells, and the data they enclose is removed and the result is cropped in with some padding, allowing the visual extractor to see part geometries rather than annotations. By operating in vector domain, we also avoid the resolution loss of rasterization. The extent of information we can extract vector natively before we need an image-based fallback is RQ3 and

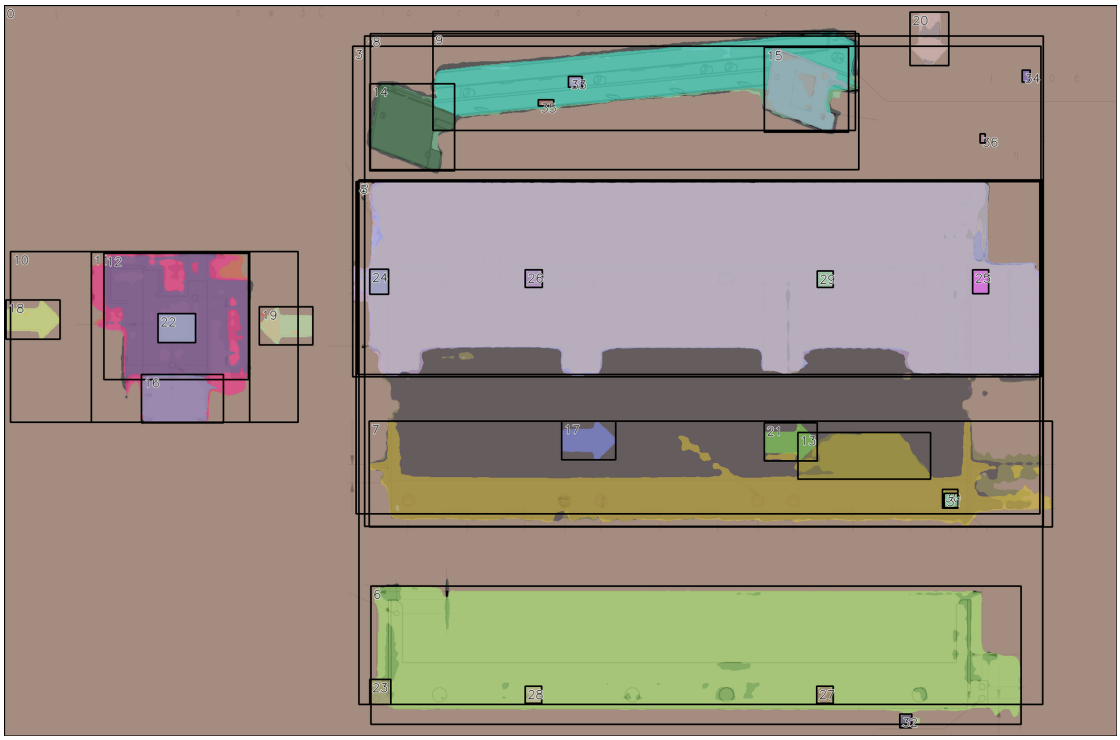


(a) Rendered page.

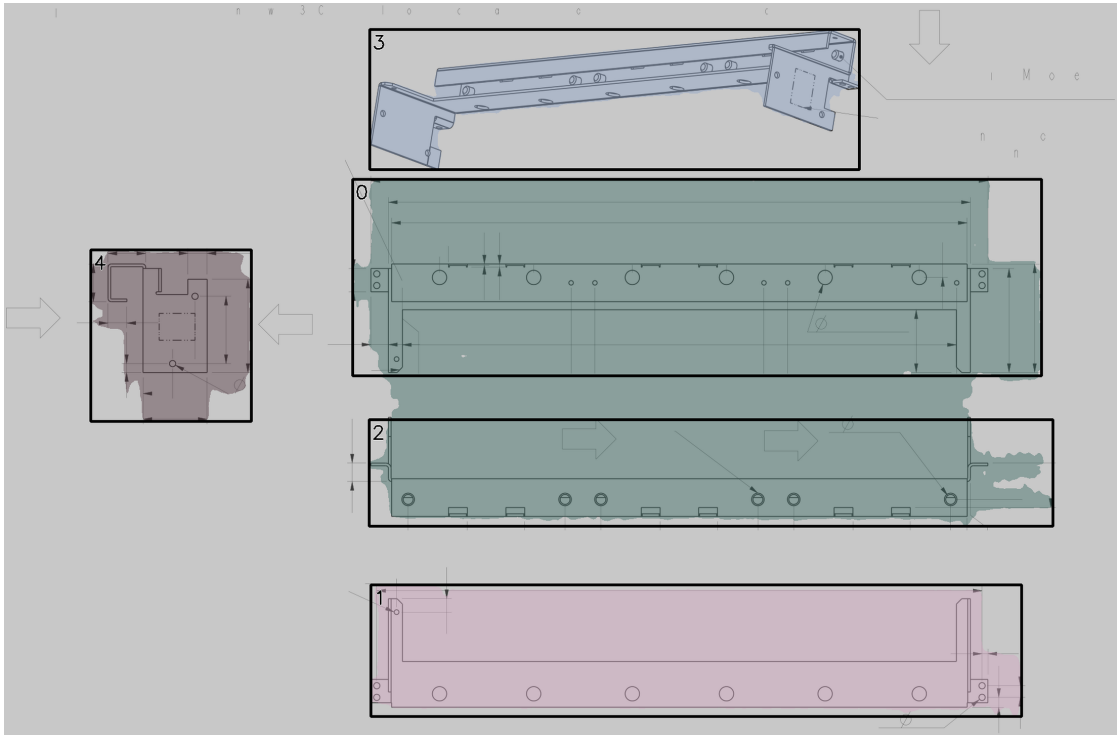


(b) Vector region crop with text and title block removed.

Figure 5.2: Vector-native view segmentation on an example sheet-metal drawing.



(c) Segment Anything Model (SAM) candidate masks.



(d) Views kept by the filter.

Figure 5.2: Vector-native view segmentation *(continued)*.

is discussed in Chapter 6.

Text extraction and OCR fallback: Text is read per region. We read inside the region of each detected cell individually. If the PDF text layer is available in the drawing, it is used directly, however if the text is path rendered (drawn as vectors without recoverable characters), the cell region is passed as an image for Optical Character Recognition OCR with PaddleOCR. Body notes are texts lying outside the cells, when present they are recovered by grouping the grid cells into columns and parsing header and rows, similar extraction logic and fallbacks as cell-bound text applies. OCR engine was selected through a three-way comparison against Tesseract and EasyOCR.

View segmentation: Each sheet of a drawing generally carries several distinct views. The visual channel in Layer B embeds each view separately rather than whole-page embeddings (Chapter 6). Candidate regions are produced using Meta's Segment Anything Model (SAM) [41] in "everything" mode. These are then reduced to genuine views by a simple filter consisting of removing background regions hugging the border, splitting merged regions at ink gaps and text blocks (floating notes) rejection. The filter was developed and validated on 10 representative drawings and then applied corpus-wide (Chapter 8).

5.3 Textual representation (Layer B)

Extraction yields region-tagged raw text. A separate classification and pattern logic orders it into title block or table cells, body notes and BOM tables. The raw text is then parsed into typed fields, material, part type, tolerance class, labelled keyword rules. Legal and template texts repeated by vendors are detected as high frequency boilerplate noise and removed. BOM table text is classified separately. Whatever remains after classification passes is labelled and stored together as unclassified blocks. Every text unit thus falls into

Table 5.1: Textual channel: the raw title-block text of the drawing in Figure 5.2(a) is typed into structured fields by the labelled classification rules. Repetitive legal and template text is shelved as boilerplate.

Extracted field	Value
Part type	support
Material	DX51D+Z275-M-A-C (hot-dip galvanized steel sheet, $3.0\,\mathrm{mm}$)
Coating	degreasing cleaning
Tolerance class	DIN 6930-m
Scale	1:5
Weight	$2.47\,\mathrm{kg}$
Standards	EN 10327, DIN 6930

either classified fields, unclassified blocks, embedded BOM table, or discarded chrome. Cleaning is intentionally kept light and only unambiguously boilerplate text gets removed because over cleaning may destroy a weak signal [11]. A prediction head is the task-specific predictor mapping a drawing's representation to one ERP output. For handling by prediction head, the text is tokenized at different word and character levels, into a TF-IDF vector. Chapter 8 provides a key empirical result, that once the prediction head is trained and fixed, the choice of text representation (raw, lightly cleaned, fully classified, lexical or dense) has little effect on the neural model performance. The prediction channel itself does not depend on representational sophistication. However, this does not make structured representations redundant because classified representation data (material, part type, tolerance, coating, and internal BOMs) are valuable human readable output, that can serve as output product features and explainers for the classification and retrieval guided proposals. As the extracted BOMs do not cleanly join to the ERP data, they are legitimate training inputs, but not a retrieval target for the ERP BOM prediction.

5.4 Visual channel (Layer B)

Segmented views from the previous layer are preprocessed as line art (ink binarized, dilated proportionally, square padded, and resized to 224×224 through ImageNet normalization)

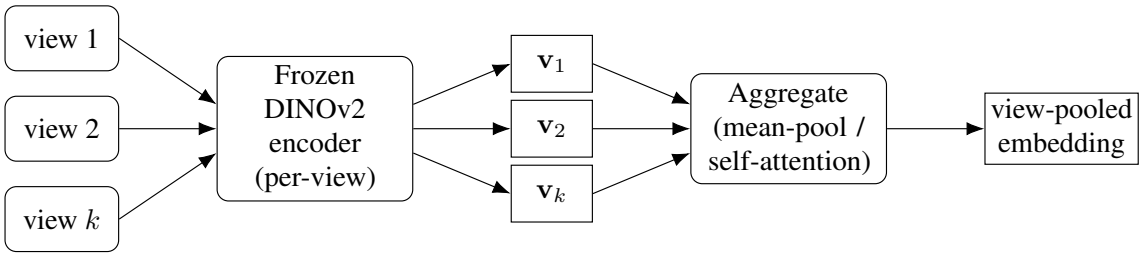


Figure 5.3: Visual channel: per-view embedding and aggregation.

and embedded using a frozen DINOv2 encoder [8] to produce a 384-dimensional view vector. Each drawing can be visually represented as an aggregation of its view vectors into a single view-pooled embedding (distinct from whole-page embeddings). The default aggregator used is mean pooling. A trainable cross-view self-attention option is also present. As per Chapter 8, frozen mean-pool DINOv2 is the strongest single retrieved and training the aggregator using a nominal label like part type degrades it. Whether DINOv2 backbone should be adapted to the domain is considered in the same chapter. The proof-of-concept uses a frozen backbone due to lower cost and simplicity, and the thesis reports on par similarity retrieval.

5.5 Similarity retrieval

When provided a query drawing, similar drawings are retrieved against the visual representation of the query drawing, using nearest neighbor search in the visual representation vector space. A mean pooling aggregator is used to surface neighbors, so the output similarity retrieved drawing set is the top-k results ($k = 5$) produced. Textual channel is used to re-rank the k retrieved drawings using cosine similarity on the TF-IDF of the drawings. Combining the two channels as such is needed to answer RQ2.

5.6 ERP prediction heads

Work phase plan is predicted using multi-label classification over the (realistically trainable) 21-class production-line taxonomy. Text representation is provided to a Batch-Normalized MLP (Multi-Layer Perceptron), trained with multi-label targets, using an iterative stratified split. Predicting clean taxonomy over raw text phase descriptions is the dominant design decision here.

Bill of Materials BOM is addressed through two modes. Frequent non-unique components are learnable by a multi-label classifier to predict directly. Unique single or low occurrence components are not predicted, instead a candidate pool is retrieved from the most text-similar neighbors and proposed for the engineer to choose from. The evaluation is only done on non-unique data subset, as metrics are only meaningful here, and it is an indicator how accurately components could be predicted provided we have trainable large amounts of component occurrences (a larger dataset).

5.7 Design constraint of vendor neutrality

Similarity and prediction rely on signals intrinsic to the drawing and never on vendor identity and vendor specific data. Vendor is never used as a text feature, the shared prediction target of work phases is common to both vendors and the view filtering is vendor neutral by design. Because parts made by both vendors are highly distinguishable from their drawings, without such a constraint, a model could score well by learning vendor cosmetics instead of manufacturing semantics. How well our constraint holds is measured in Chapter 8.

5.8 Multimodal fusion

When both channels are applicable for solution, they are combined at scoring stage. Similarity and prediction scores are normalized and merged while reciprocal rank fusion is used for union in a retrieval case. We treat fusion as an ablation with its impact needing measurement, not an automatic upgrade. Our findings (Chapter 8) report that fusion does not reliably improve product structure prediction over the best single modality results for our problem.

6 Foundational Experiments

Before processing and representation pipelines were built at scale, a set of foundational experiments were performed on curated subsets of our dataset, and they established the current shape of the pipeline. There are several negative results, and they rule out certain design directions and justify the current processing structure.

6.1 Deficiency of whole-page embeddings

The most significant negative result rules out the most obvious baseline for our pipeline, embedding the whole drawing with a pretrained vision encoder and retrieving nearest neighbor to judge similarity. Drawings were rendered at three resolutions (224p, 448p, 896p) and embedded with two raw pretrained backbones, ResNet-50 and DINOv2 [8] without adaptations. Clustering quality was measured using silhouette scores for part-type and vendor. All six results generated negative scores on part-type and uniformly positive scores on vendor that climbed at higher resolution (+0.075 to +0.120 for ResNet-50). This implied that page level embeddings were more likely to cluster drawings by vendor than by part type with increasing resolution. Cropping away title blocks did not rescue the results, and further revealed that vendor cosmetics existed across the drawing, in the dimensioning and line weights to name a few.

A second experiment was performed using a lightly supervised finetune like that of Van Daele et al. [13] where we retrained only the final layer. The model once again failed to make part-type structures converge. We concluded that page level rendering

Table 6.1: Whole-page cluster quality (*silhouette score (higher is better), negative means worse than random*). Grouping by part-type is negative for every combination, while grouping by vendor is positive and strengthens with resolution for ResNet-50. The encoders lock on vendor cosmetics, not part geometry.

Backbone	Grouping	224 px	448 px	896 px
ResNet-50	Part-type	-0.049	-0.074	-0.099
ResNet-50	Vendor	0.075	0.095	0.120
DINOv2-small	Part-type	-0.075	-0.076	-0.091
DINOv2-small	Vendor	0.085	0.087	0.076

was insufficient at all resolutions, with or without supervision and the visual channel must operate on more granular data like individual views. The underlying reasoning is clear, drawings carry two types of similarity, drawing-style similarity (similar sheet layout, title block design, body note locations, dimensioning) and manufacturing-style similarity (similar looking parts, potentially made the same way) and the visual encoding channel must therefore see only the part geometries to avoid fitting to the wrong similarity.

6.2 Symbolic channel

If we audit the textual data for how and where symbolic data appears textually in the corpus, we find that welding callouts (ISO 2553) are extractable in 0% of the drawings, surface finish indications on roughly 4%, tolerancing glyphs are caught in 1%, and region-tagged dimension and tolerance data is not as useful as initially assumed. Most of these processes are encoded graphically and textual shortcuts are rare if present aside from the usually mentioned ISO/DIN standards. Recovery would require a separate dedicated visual symbol only detector, which is solvable but another layer of complexity as Hu et al. [44] have developed. Cheaper process relevant signals are present textually, such as tolerance classes (ISO 2768) is present in 70% of the Company A and 5% of the Company B corpus, and paint colours (RAL codes) on 3% Company A and 59% Company B. We make a scoping decision based on these findings. Graphical symbol classes exist but visual detection is not

Table 6.2: Textual footprint of process symbology (corpus-wide).

Signal	Textual presence
Welding callouts (ISO 2553)	$0.0\,\%$
GD&T glyphs	$1.0\,\%$
Surface-finish	$4.1\,\%$
Tolerance class (ISO 2768)	Company A $70\,\%$ / Company B $5\,\%$
Paint colour (RAL)	Company A $3\,\%$ / Company B $59\,\%$

cheap and left for future work, and cheap in-text process signals are folded in along with the textual channel.

6.3 Standards vocabulary for text normalization

To convert textual data from different drawings, vendors and sources into a common form before passing it on for representation, we rely on a vocabulary of manufacturing terms (materials, processes, tolerances, and surface classes) to normalize different forms into the same tokens. The ISO and DIN standards that define these were not institutionally accessible, so the vocabulary is grounded in manufacturing reference papers that reproduce those taxonomies [3], [4]. Every entry carries source tags and sub-codes that could not be verified against cited sources were removed from memory. It is important to note that the vocabulary is only used for tagging and classification and not as a knowledge store, or an allow list. Its value lies in interpretability and human-readable output only, as findings (Chapter 8) show it is performance neutral for prediction models.

7 Evaluation Metrics

This chapter details how we will evaluate the system created. The use of data-splits and leakage controls, metrics to quantify the results of prediction and retrieval tasks, diagnosing vendor agnosticism. The thesis learns the hard way that the choice of target metrics must fit the structure of the data to be meaningful.

7.1 Splits and leakage control

For all learning tasks, the results are tested using hold-out test sets produced using iterative multi-label stratification [45] that balances rare label combination and stratifies to ensure variety of drawings in both training and test splits based on ERP data. The ratios of the Train/Validate/Test split are 70/15/15 and we used a fixed seed to ensure repeatable and comparable results between models. Each drawing is a single sample so no drawing appears in more than one set, and the splits are computed once and reused across all models for a given task, preventing leakage between sets through repeated re-splitting. Work phase classification splits 1762 drawings carrying usable phase labels into 1234/264/264, BOM classifier applies the same split to 1761 drawings with usable BOMs, and all visual/multimodal experiments run on a shared single split of 1218/261/261. All these are written to disk and reused when needed. For vendor agnosticism test, the testing set is further partitioned by vendor.

Signal-Target contract mentioned in Section 4.3 is the other leakage control mechanism. ERP data never enters the inference phase for classification tasks, embedded BOMs are

Table 7.1: Train/Validate/Test split sizes (iterative multi-label stratification, 70/15/15, seed 42).

Task	Train	Val	Test
Work-phase ($n = 1762$)	1234	264	264
BOM classifier ($n = 1761$)	1233	264	264
Visual / multimodal	1218	261	261

used as a legitimate input signal as they do not meaningfully join to ERP but provides a negligible but present F_1 lift of +0.012 (0.341 to 0.353) on ERP-BOM retrieval metrics.

7.2 Metrics for prediction

Work-phase plans and recurring BOM components are multi-label classification tasks. They are evaluated using conventional multi-label metrics. Macro-F_1 is the mean of the per-class F_1 scores,where every class is weighted equally and rare classes are not masked. Micro-F_1 pools the true positives, false positives and false negatives across all classes into a single count and then computes F_1, therefore it is dominated by the frequent classes and reflects overall decision accuracy. Exact-match is the fraction of predictions where all predicted labels are correct. Set distance $\pm k$ measures predictions differing from the truth by at most k labels. Finally, Jaccard overlap is the intersection over union of the predicted and true label sets. Each metric provides uniquely important insights. Micro-F_1 shows if the model gets common labels right, macro-F_1 exposes the performance on the rare labels, and exact-match provides the strict output performance without errors.

In the case of BOM data which is long tailed, exact and set scores are bound not by the model but by the data itself. A component seen once cannot be predicted from a precedent. In that case we need to only report the performance on the trainable recurring subset of the data where the score may reflect the model's performance and treat the remaining single-use components as a suggestive proposal that cannot be defined by metrics.

7.3 Metrics for retrieval and human-judged protocols

Similarity retrieval (RQ1) cannot be scored against nominal field labels, because the nearest thing to a nominal proxy, the part type label, only agrees with perceived similarity 0.44 to 0.60 of the time. It is better to therefore evaluate visual retrieval by a human relevance judgement us a pooled and blinded protocol like traditional information retrieval evaluation. For a query drawing a set of k neighbors from competing methods are pooled together, they are de-identified (lettered and method and score removed), and then shuffled. A human then marks which of the candidates look genuinely similar to the query and unjudged candidates are treated as wrongly matched. Queries are excluded from their prediction candidate lists (leave-one-out).

Based on the human results we report hit-rate Success@5 (queries with at least one similar drawing in the top 5), Recall@5 (fraction of the similar set in the top 5), Precision P@1 (similarity success of the top matched drawing), and mean average precision (mAP). We do not produce Precision@5 because precision is inherently unfair for this retrieval as the dataset might not have more than a few similar drawings capping precision at 0.2-0.4 regardless of model quality, and therefore, we lead with Success@5 and Recall@5. Comparisons are made between methods on the same pooled judgement: frozen versus adapted encoders, mean-pooled versus self-attention, so the conclusion rests on performance judged by humans and not an absolute threshold.

7.4 Diagnosis on vendor Neutrality

We evaluate vendor agnosticism (RQ2) in two parts: (i) we measure how well a classifier can be trained to predict a vendor given a feature channel. We find vendor is near perfectly predictable from text embedding (0.998) and also holds strongly for visual embeddings (0.88), confirming that a model could exploit vendor cosmetics if allowed to. (ii) as we try to exclude vendor from the features and labels by construction, we measure whether

task performance holds per vendor. A task is vendor neutral if both vendors are served comparably, and not if a strong result is being carried because of aggregating around a strong vendor. This reporting is important because it distinguishes whether our result is genuinely transferable or simply averages around the vendor's data.

7.5 Evaluation discipline and target choice

Choosing a target or metric that fits the data being predicted or retrieved is load bearing for our evaluation success. Choosing incorrectly results in silent inflation or deflation of results, hindering progress. Predicting clean customer production-line taxonomy versus free/text phase descriptions resulted in an increase in exact-match rates from roughly 0.36 to 0.60 with the models unchanged. Nominal Similarity proxy of part type being used to judge visual similarity proved to be very weak and degraded the retriever when used for training or tuning . In each case correction only came from measuring against the right target, with the right metrics, which is why evaluation discipline is treated as a result not a footnote.

8 Results

This chapter reports our quantitative and qualitative findings, organized around the two goals and three research questions. Goal 1 (ERP product structure proposal for new drawings) is addressed in Section 8.2 (work phase plan) and Section 8.3 (bill of materials), followed by Goal 2 (similarity drawing retrieval) in Section 8.4. Critical questions are answered: text versus visual similarity (Section 8.5), vendor neutrality (Section 8.6) and multimodal fusion (Section 8.7). Methodological findings in Section 8.8 report that target definition and correct evaluation were more dominant levers in prediction performance than model sophistication.

8.1 Ground truth

ERP to drawing join resolves for 99.95% of drawings (1828/1829) that link to at least one ERP item. A singular genuine orphan absent from the ERP dataset. As Stera selected the dataset from top-level items, the corpus is 94.2% top-level assemblies (1723/1829). These are items that appear in the BOM as a parent and are never used as a component. Only 11.6% of drawn items appear as a sub-component of a larger assembly. If we consider all items referenced in the ERP just 31% are top-level while 68% are external components that we do not have drawings for. For this structural reason a generative bill of materials is infeasible on this data. The multi-level structure is not directly recoverable, by Stera's own documentation, and it motivates the retrieval-and-propose framing of Goal 1's BOM branch.

How much product structure content do drawings share usually? Two drawings chosen randomly share at least one BOM component 11.7% of the times, and at least one work-phase 63.8% of the time. This is why work phase routing is treated as a direct classification task rather than by copy from neighbors. By contrast the BOM content sharing (11.7%) is sparse, idiosyncratic and genuinely discriminative as a product structure signal. It is the harder target, handled by a two tier scheme.

8.2 Goal 1: work phase plans (RQ2)

The clearest output signals were obtained on work phase routing plans. The target is the customer's production-line taxonomy, a documented Stera field, not an author-defined set of labels. Of the work centres reachable from our drawings, 26 production lines occur and 21 are trainable with ≥ 20 examples each (BEND, BLAST, COLLECT, DEBURR, DRILL, GRIND, INSPECT, LASER, MACHINING, PACK, PAINT, PEMM, PLASMA, PROGRAM, THREAD, TPP, WASH, WELD, WORK, the two automated CABLE.A, and BEND.A). The task is multi-label and evaluated with macro-F_1, micro-F_1 and exact-set and set distance $\pm k$ matching on an iterative multi-label stratified split.

A word + character TF-IDF representation learned with a Batch Normalized residual-MLP scores macro-F_1 0.82, micro-F_1 0.92, exact-match 0.60 (Table 8.1). The result is robust to the text representation so raw, lightly cleaned and fully classified inputs score within run-to-run variance of each other.The architecture choice (resMLP, self-attention, gatedMLP) moves exact-match by only a few percentage points. The more important factors are target definition for label outputs and the training regime. Most production lines lie above the macro-F_1 of 0.82, and the average is pulled down by the two rarest lines, BEND.A and DRILL ($n = 7$ each).

Table 8.1: Work-phase routing over the 21 production-line classes (word+char TF-IDF classifier).

Text input	Model	Macro-F_1	Micro-F_1	Exact	± 1
classified (page 1)	BatchNorm-ResMLP	0.824	0.916	0.602	0.750
classified (page 1)	Self-attention	0.804	0.914	0.598	0.758
raw (page 1)	Self-attention	0.807	0.912	0.595	0.746
raw (page 1)	Gated MLP	0.812	0.912	0.583	0.742

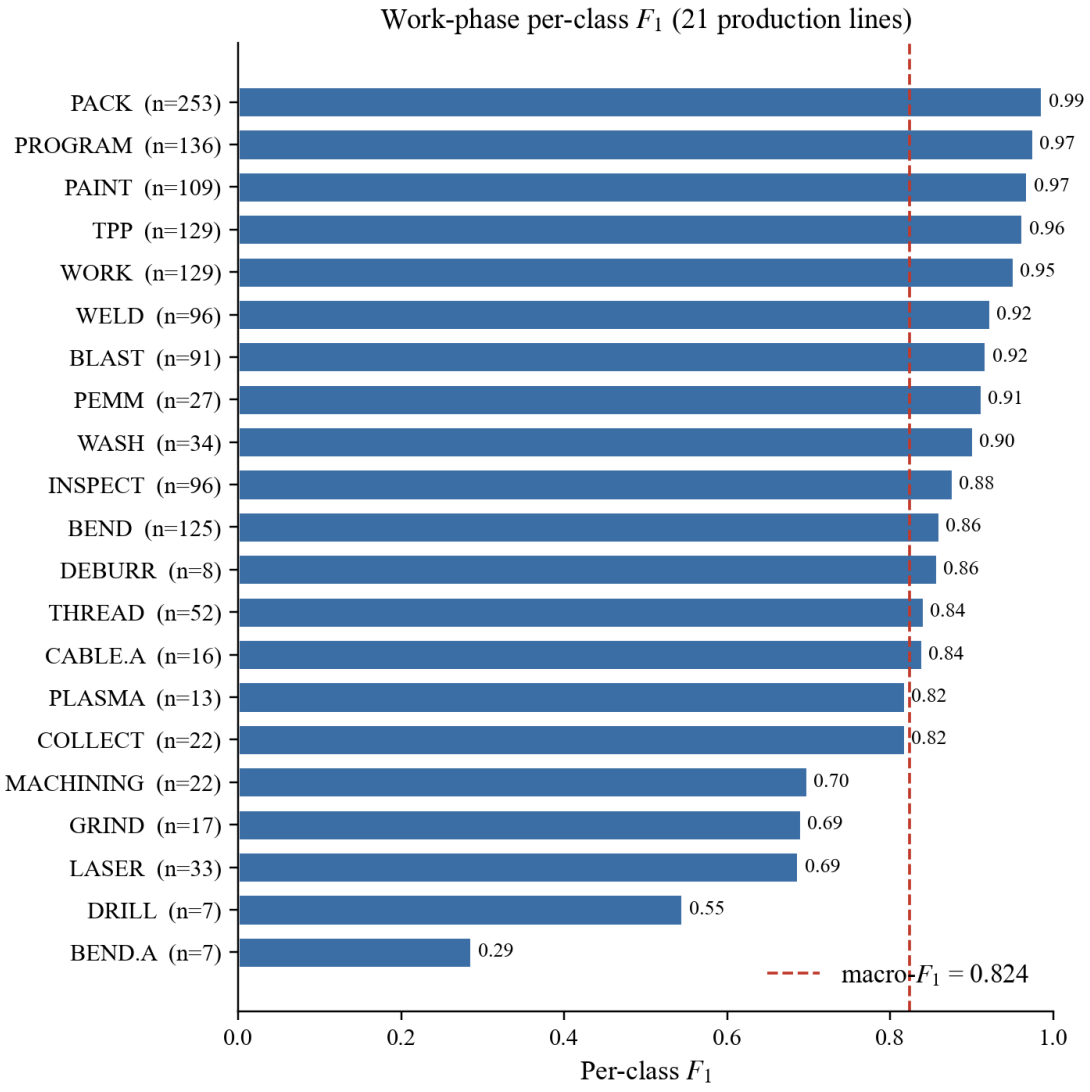


Figure 8.1: Per-class F_1 for the 21 work-phase production lines.

Short drawings are the information bottleneck. Accuracy variation correlated with the amount of text extracted from a drawing. If the test set is binned by token count, the

shortest quartile scores 0.470 exact while the longest scores 0.803 (overall 0.610). Text length to accuracy correlation is +0.119 and it is not a label count artifact. Average label count does not vary among the quartiles. The bottleneck is the lack of information and not lack of training. Lightly worded drawings simply carry less information to predict routing. Addition of the visual channel does not rescue these drawings (Section 8.7).

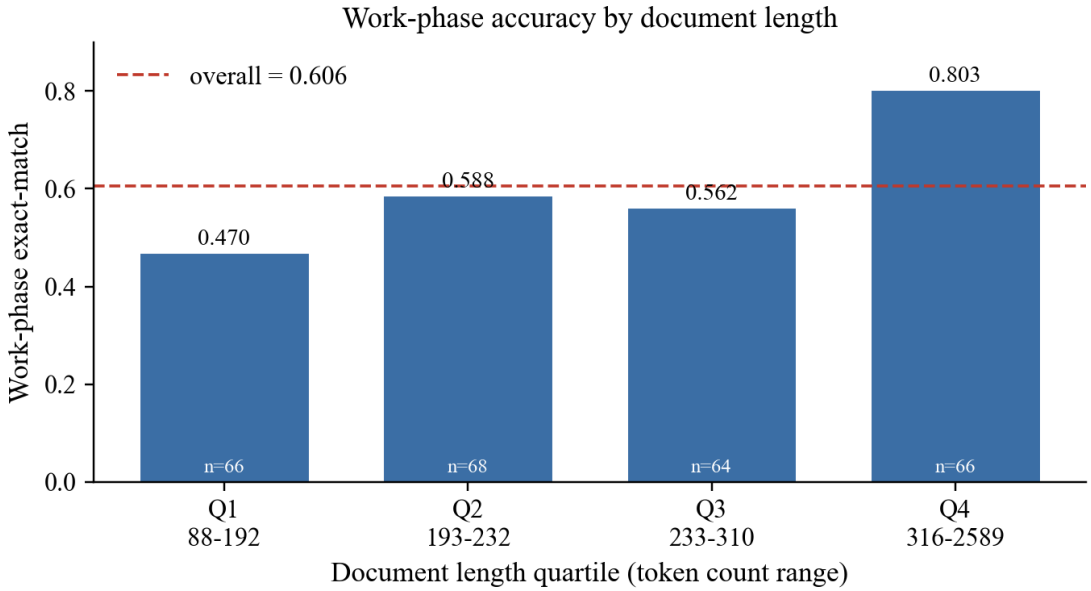


Figure 8.2: Work-phase accuracy by document length.

8.3 Goal 1: Bill of Materials (RQ2)

BOM data is long tailed, of the 3401 distinct ERP components, 67% are unique (one drawing only), that number rises to 79% if we count by BOM rows. Components seen only once are definitively unlearnable, therefore this goal is addressed in two tiers. A classifier for the small recurring subset (the head), and a retrieval pool suggestion to address the tail of the dataset.

(i) Head: direct classification. A multi-label classifier was trained on components above a frequency threshold (similar to work-phase). The results produced, tabulated in Table 8.2, reach micro-F_1 0.84 at freq $\geq$ 20, 0.85 at freq $\geq$ 10 while macro-F_1 is 0.76 and

exact-match 0.38 . Direct classification far exceeds retrieval-aggregation on the recurring set, but the gain is capped by the size of the dataset and hence the number of components frequent enough to learn. Critically, it was found that BOM classification is not vendor agnostic (Section 8.6).

Table 8.2: BOM as a classifier over recurring components.

Frequency	Components	Macro-F_1	Micro-F_1	Exact
≥ 20	57	0.761	0.838	0.386
≥ 10	98	0.761	0.851	0.382
≥ 5	254	0.775	0.848	0.372
≥ 2	1130	0.270	0.726	0.236

(ii) Tail: retrieval pool. To account for unique BOM components, the system proposes a candidate component pool from the most text-similar drawings. It is purely suggestive in nature and for an engineer to review. Two parameters define the pool. k refers to the number of nearest drawings retrieved, and vote tells us how many drawings must repeat a BOM component before it is proposed into the pool (vote=1 is essentially a union of BOM components of all k proposed drawings).

A unique component appears with only one drawing and can never be recovered from a neighbour and its recall is not measurable. We therefore evaluate the retrieval pool on the recurring components. We achieved a recall of 0.736 with k=10 and vote=2 and 0.908 with k=5 and vote=1, with precision remaining around 0.37. Details are present in Table 8.3. With a looser pool more true components can be captured at the cost of more incorrect ones. This suits why we chose recall to represent its results, where a human can then prune through the result for real components.

Table 8.3: BOM tail retrieval evaluated over non-unique components. k is the number of neighbours retrieved and *vote* the number of them a component must appear in to be proposed.

k	Vote	Recall	Precision	Micro-F_1
5	1	0.908	0.377	0.533
10	2	0.736	0.370	0.492

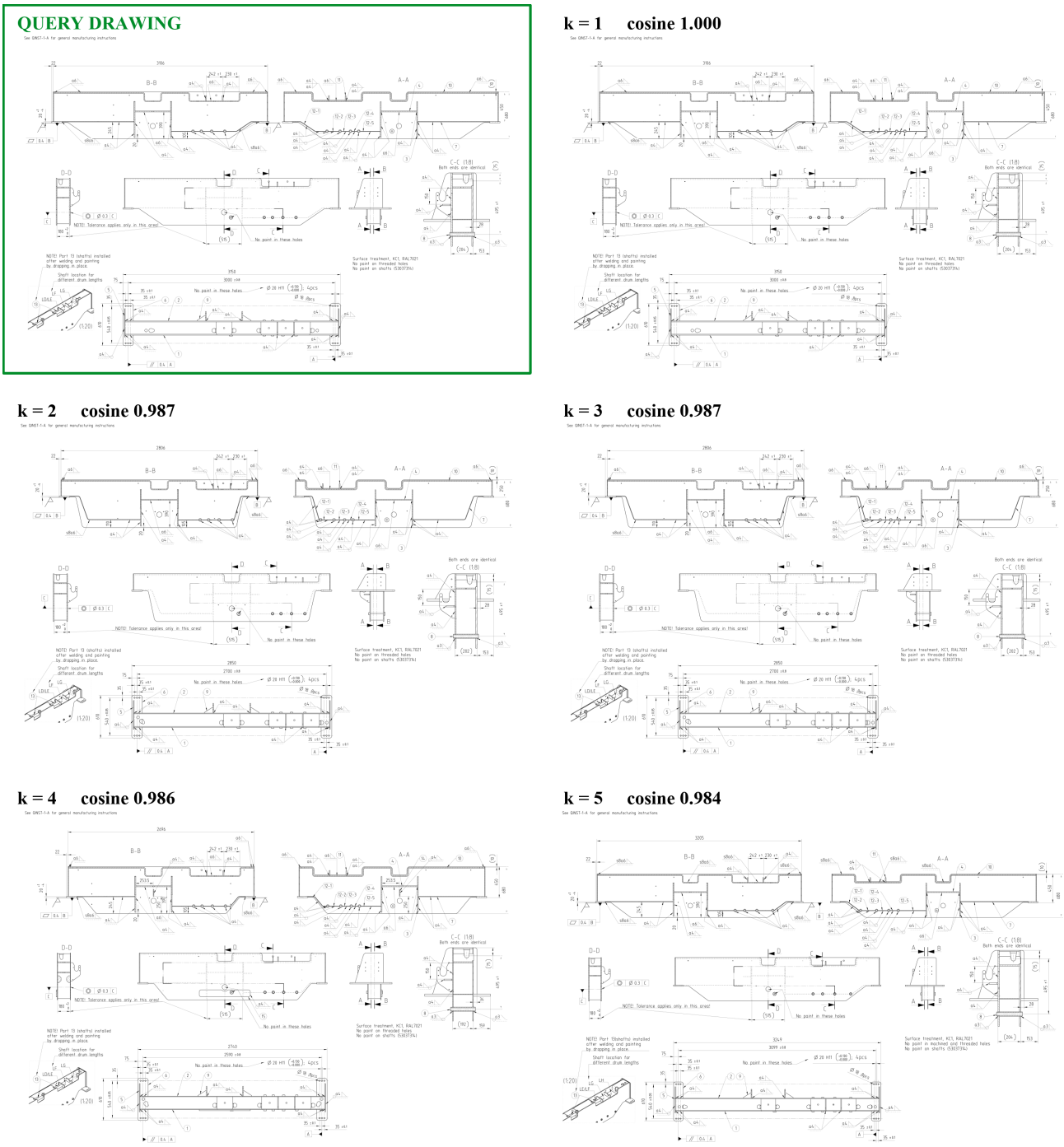
The evaluation results of the head classifier and tail pool are not directly comparable. Macro-F_1 (the per-class average) of the classifier over the recurring-component space is held down by the lower frequency components, and micro-F_1 stays high because it is dominated by the frequent components. On the other hand, the retrieval pool results measure the overlap between a drawing and a broad candidate data pool. The two tiers measure different quantities and should be read on their own terms and not scored against each other.

These results show that the ceiling in terms of BOM prediction is the existence of unique and low frequency data, with availability of more data theoretically increasing prediction performance further.

8.4 Goal 2: retrieval of visually similar drawings (RQ1)

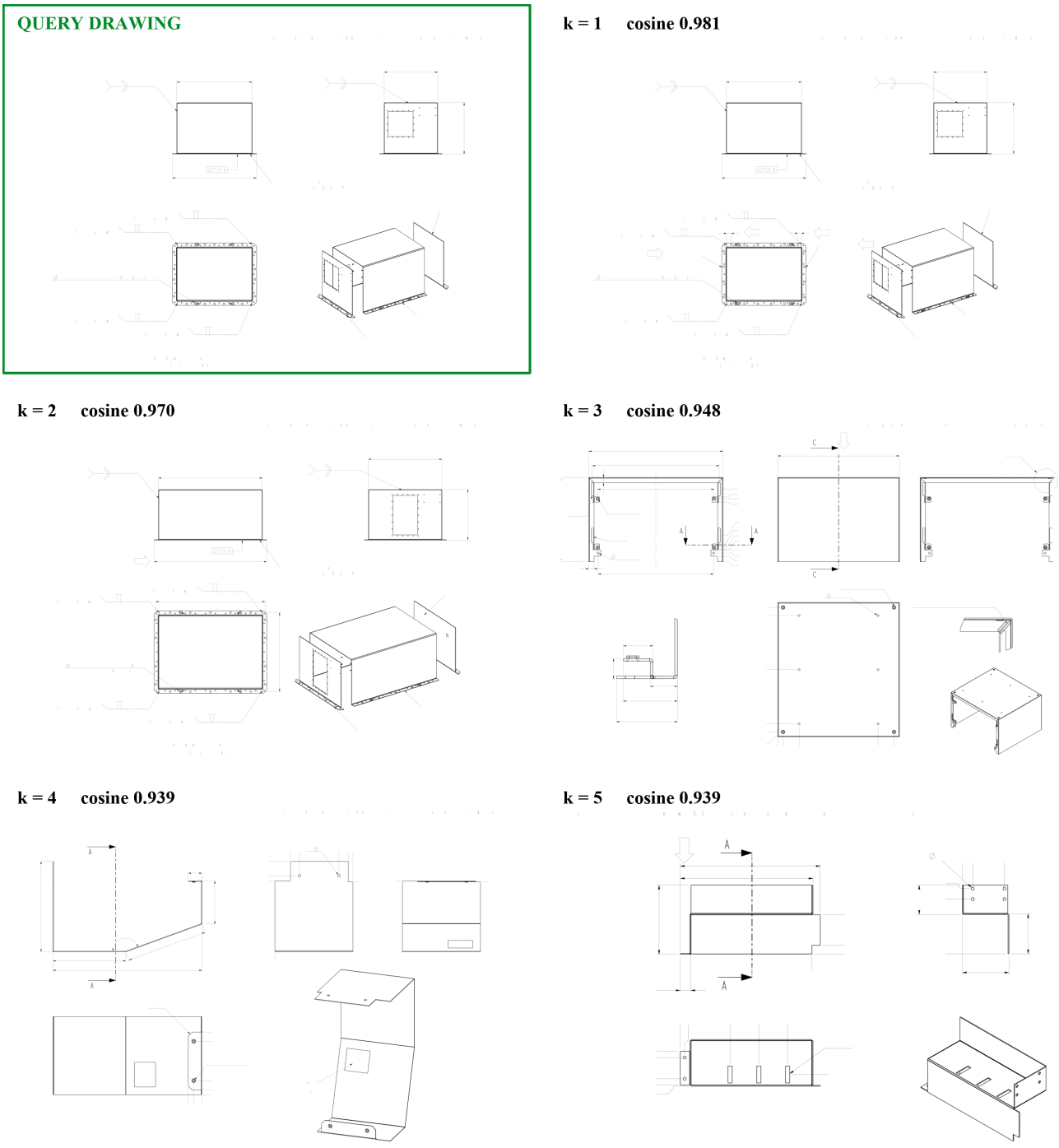
As stated, the similarity pipeline primarily uses visual features. Each drawing is segmented into views, and the views are embedded with a frozen DINOv2 encoder, with the resultant vectors aggregated. We have established that nominal labels are unreliable as a similarity goal, for example part type agreed with human-perceived similarity only 0.44 to 0.60 of the time. Therefore, it is better to judge the retrieval quality against a human relevance assessment on a blind, de-identified pool of all the top k drawings suggested by competing approaches, and scoring with Success@5, Recall@5, mAP and P@1 metrics.

Results show that simple frozen DINOv2 backbone with mean pooling is the best



(a) Query: support beam.

Figure 8.3: Visual retrieval examples: the query drawing (green) and its five nearest neighbors by mean-pool DINOv2 cosine similarity, ranked $k = 1$ (most similar) to $k = 5$.



(b) Query: cover.

Figure 8.3: Visual retrieval examples *(continued)*.

Table 8.4: Human-judged visual retrieval ($n = 32$ blind queries).

Aggregator	Success@5	Recall@5	mAP@5	P@1
Frozen mean-pool	0.84	0.80	0.68	0.69
Part-type self-attention	0.53	0.48	0.45	0.47

retriever. On 32 human-judged queries, we recorded a Success@5 of 0.84, beating the part type trained self-attention pooling alternative on all metrics (Table 8.4). Training the encoder to recognize part type actively degraded similarity retrieval.

Backbone Adaptations: whether adapting the DINOv2 backbone in domain instead of training or changing the aggregator helps is another question that needs answering. A three model blind comparison was made with the frozen backbone against unsupervised same-drawing adaptation and supervised part-type adaptation (all with mean-pool aggregation) on a 20 drawing pool. No clear winner was found (Table 8.5). However, we do note that in this case part type adaptation neither harms nor lifts the human-perceived similarity in contrast with the part-type self-attention aggregator harming the results. This is an interesting finding and the two different uses of part type to adapt our model should not be conflated. Because adaptation is on par at best, our prototype keeps the simple frozen encoder.

Table 8.5: Backbone adaptation, human-judged ($n = 20$). No clear winner.

Backbone	Success@1	Success@5	Recall@5
Frozen	0.700	0.850	0.620
Unsupervised-adapted	0.650	0.800	0.667
Part-type-adapted	0.650	0.850	0.673

The problem that bounds Goal 2's value for Goal 1 is that visual similarity does not imply shared product structures. Two drawings that look alike might not share the same BOM codes. As Section 8.7 finds that fusion with visual channel adds no reliable lift to our ERP predictions. A visual neighbor is a good place to start looking but not always a

structural match.

8.5 Text versus visual for similarity

We tested if textual information was useful for visual similarity ranking by reranking the candidate pool using several text models and rescoring the reranked list against human-judged results (Table 8.6). No text model was able to improve the top of the ranking compared to the visual mean-pool results (Success@1 0.688). TF-IDF, zero-shot sentence encoder, or unsupervised SimCSE model trained on the drawing text all failed.

Table 8.6: Text reranking of the fixed visual pool ($n = 31$).

Ranker	Success@1	Recall@5	mAP
Visual mean-pool	**0.710**	0.803	**0.739**
TF–IDF (cleaned text)	0.548	0.791	0.691
Sentence encoder (zero-shot)	0.613	0.805	0.696
SimCSE (unsupervised, adapted)	0.516	0.687	0.620
Fusion (visual + TF–IDF)	0.645	**0.855**	0.757

8.6 Vendor agnosticism (RQ2)

The system needs to work for all future data without exploiting vendor specific tricks to "hack" its way. The vendors in our case are very distinguishable from each other. A classifier can predict the vendor from raw text almost perfectly (0.998) and from view-pooled embeddings about 0.88 of the time. The shortcut is strong and all effort has been made to exclude it by design. The real test for vendor neutrality is whether the task performance of our model holds in each vendor subset (Table 8.7).

Work phase routing is vendor agnostic and both Company A and Company B data

Table 8.7: Per-vendor task performance (text classifier, vendor-split test). Work-phase is agnostic, BOM is carried by Company B.

Task	Vendor	Exact	Micro-F_1	Jaccard
Work-phase (21 class)	Company A	0.583	0.928	0.876
	Company B	0.607	0.902	0.854
BOM (freq ≥ 10)	Company A	0.318	0.464	0.351
	Company B	0.462	0.925	0.793

produce similar numbers. However, if we look at BOM prediction for non-unique high frequency components, the strong aggregate score is being carried by Company B data (micro-F_1=0.925) while it struggles on Company A (0.464). Upon further investigation, it can be explained by the nature of components linked to each company in the Stera dataset. Because Company B assemblies are mostly welded structures that use a lot of standard components, its BOMs are recurring and learnable compared to one-off BOMs from Company A's sheet metal parts. BOM results should therefore be reported per vendor and not as an aggregate.

8.7 Multimodal fusion for ERP prediction (RQ2)

Fusion of text and visual channel does not improve ERP prediction tasks. On work phase plan prediction tasks, textual channel dominates the visual channel and fusion at score level yields performance within run-to-run variance of textual channel alone. Visual channel is also unable to rescue short documents with low word count that bound the overall accuracy (Section 8.2). The conclusion is clear, visual channel is the right tool to find similar looking drawings but provides no dependable lift to Goal 1 (product structure proposal). And therefore, multimodal fusion does not bring more performance to the table.

8.8 Evaluation discipline (a methodological finding)

An important methodological result of the project is the value of correct target definition. Predicting a clean label taxonomy (the 21 production lines) rather than noisy ERP text field raised work phase exact-match from 0.36 to 0.60 with the models unchanged. Target definition, and not architecture, was the driving force. On the evaluation side: a nominal proxy (part type) that looked like "manufacturing similarity" agreed with human-judged similarity only 0.44 to 0.60 of the time and, when used as a training signal, degraded the very retriever it was meant to improve. Each correction came only from measuring against the right target with the right metric an important evaluation discipline lesson.

9 Discussion

In this chapter we discuss the results and the central claims of the thesis together. We examine how the results took shape the way they did and the limitations.

9.1 Different modalities lead different tasks

The clearest pattern we can notice in the result is how the two different modalities of extracted information are suited to different tasks. Textual channel dominates ERP predictions. It carries material, part type, tolerance, surface finish and other information that determine work-phase plans and recurring BOM components independently of how the text is represented. Visual signals conversely, are stronger at finding drawings that look like the query, and here text is weaker and as reranking experiments show, does not help. It reinforces that the channels do not help reinforce each other's prediction tasks. Multimodal fusion did not provide a reliable lift in ERP prediction and did not rescue short information poor documents that bound the work phase accuracy results. With no dependable gain to be found, the correct design decision is to route each task to the modality that serves it best rather than combining them indiscriminately.

9.2 Visual similarity does not imply similar structures

Contrary to the expected outcome, visual similarity does not drive ERP proposal, and copying product structures and BOM from visually similar images is not supported by the

results. Visual similarity tracks appearance, which seems to not carry any product structure information, and for that reason fails to provide a dependable lift to ERP prediction. We reach the same practical result once again, that the two goals are best served individually, where visual retrieval is used to propose visually similar precedents for the human in the loop to inspect.

9.3 Vendor neutrality is task dependent

There is nuance to the vendor agnosticism result. Work phase data was predicted purely vendor agnostically, both vendors served equally, because we had the necessary textual information for prediction commonly present in drawings from both vendors. The BOM data that showed to be vendor dependent with the aggregate scores being carried by Company B drawings is an explainable artifact of the drawing properties of the two vendors. The welded structures in Company B parts reuse more components than the sheet metal parts of Company A, with relatively sparse component reuse.

9.4 Text representation is performance neutral but still holds value

Choice of text representation barely affects neural model training, which appears to be a negative result for the classification efforts. However, what is found is that text representation is performance neutral for prediction. Rule classified data like material, part type, tolerance, coatings and surface finishes, internal BOMs are valuable human readable outputs. While a raw token stream seems to suffice for a neural classifier, a human reviewing the outputs might benefit from such data, for interpretability of results if nothing else.

9.5 Adaptation under honest judgement

In most cases during our thesis, adaptations made to the visual encoder or aggregator have led to worse results (when guided by automatic proxies) or been on par with frozen models. The frozen models seemed like the most sensible default to go for in the thesis because they are the simplest for virtually no loss of performance. However, there is a gap in this conclusion. Most of these models have millions of parameters per layer that need training, while our corpus consists of roughly 1800 drawings which is very modest for adapting something like a vision encoder which typically want large batches of data to learn. We therefore, cannot conclusively establish if "adaptation is not useful" or if "there is too little data to adapt a visual encoder in a useful manner". The literature on on-modality drawing retrieval suggests the latter is a very real possibility. Nevertheless, for data at this scale, frozen encoders seem the most effective, with adaptation being impossible. A larger archive may make adaptations a worthwhile effort (Chapter 10).

9.6 Limitations of the Data

Several results are constrained not by the availability of methods but rather structure of data. The BOM being long tailed, and the corpus being exclusively top-level assemblies with components that are never drawn, restricts us from predicting many components from precedents with an honest target only being found in the high frequency subset. Likewise, work phase data is bound by drawings that are information poor. The reported numbers are meaningful and realistic in the light of the limitations of dataset, not failures against unattainable goals.

9.7 Limitations

The thesis has clear and obvious limitations. Visual channel currently embeds the primary page of each drawing, over half the corpus is multipage, and while per page visual treatment is designed, it is computationally expensive and not yet integrated, but is likely to increase certain metrics. Sub-assembly decomposition is out of scope as the ERP data does not expose multi-level structures, and the system has to propose product structures based on top-level data only. Graphical process symbols remain a visual gap that a dedicated trained encoder could help close. The evaluation rests on a 2 vendor corpus and vendor neutrality could be strengthened by introducing additional vendors. None of these limitations undermine the current results, but each marks an avenue for extension to it.

10 Conclusions and Future Work

10.1 Summary

The thesis set out to help an industrial partner reuse its archive of historical drawings and ERP records for authoring bill of materials and work phase plans for new 2D technical drawings. It pursued two goals, proposing ERP product structures for drawings and retrieving similar drawings, aiming to maintain a vendor neutral result with a proof-of-concept designed and evaluated around a 2 vendor corpus. Three research questions are tackled.

RQ1 (similarity) Frozen DINOv2 encoder with mean pooling over per-view embeddings is the strongest single retriever of potentially similar manufacturing drawings, judged against human evaluation. Nominal proxies like part-type are poor proxies for perceived similarity and re-ranking proposed visual candidates based on text similarity does not improve the top of the ranking.

RQ2 (ERP proposal) Work phase plans are well predicted by direct classification over production-line taxonomy (0.82 macro-F_1, 0.92 micro-F_1 and 0.60 exact-match). Likewise an MLP based classifier works well over recurring BOM components (0.84 micro-F_1, 0.76 macro-F_1), with retrieval pool for long tail. Multimodal fusion does not reliably improve results of best single modality. The result is only partly vendor neutral. The work phase plan is predicted equally for both vendors, being capped by low data drawings, while BOM score aggregates are carried heavily by the

vendor whose parts reuse standard components, so these results must be measured by vendor.

RQ3 (extraction) Vector-native treatment of PDFs carries extraction a long way, with OCR fallback required only for path rendered textual data in select files.

10.2 Proof-of-concept prototype

The methods used and their case based reasoning are stored in a single end-to-end prototype that, when given a new drawing, extracts its content, retrieves similar drawings and proposes a work phase plan and BOM candidate list. The retrieval of similar drawings produces the five nearest neighbors in the visual encoder's embedding space. Aside from ERP proposal, the prototype also surfaces content that was extracted and classified from the text directly like title block fields, specifications and internal BOM tables independent to the prediction tasks. The prototype is a feasibility demonstration and not a product ready system, the value lies in identifying the most reusable signals to integrate into a workflow.

10.3 Future work

Several directions would improve the work. On the visual encoder side, on par results by adapted backbones and negative results from reranking by textual similarity suggest improvements need to come from elsewhere. Implementing non-DINO encoder families being used in current on-modality drawing retrieval, that combine convolutional encoders, margin based metric losses and learned multi-view aggregation is a valid direction. On par adaptation results potentially bounded by the modest corpus size available, make a larger corpus a pre-requisite before implementing a newer architecture and open more learning payoff from adaptation efforts. The results of BOM classifier are also bound by data sparsity. With more historical data, the prediction head can be extended to predict

more components directly. Sub-assembly decomposition is out of reach based on ERP export and notes, leaving only direct parent-component relations to work on. A richer substructure for drawings could allow for proposal of more complex product structures. Graphical symbols are a visual gap that could be closed by training a dedicated symbol encoder against the symbols. Vendor neutrality could also be strengthened by additional vendors in the dataset. A large language model could be used for text normalization, however the outputs would need verification to avoid hallucinations. Together these would move the system from a proof-of-concept to a deployable multiformat framework.

References

[1] F. Zhou, Z. Jiang, H. Zhang, and Y. Wang, "A case-based reasoning method for remanufacturing process planning", *Discrete Dynamics in Nature and Society*, vol. 2014, 2014, Art. no. 168631. DOI: 10.1155/2014/168631.

[2] M. R. Khosravani and S. Nasiri, "Injection molding manufacturing process: Review of case-based reasoning applications", *Journal of Intelligent Manufacturing*, vol. 31, no. 4, pp. 847–864, 2020. DOI: 10.1007/s10845-019-01481-0.

[3] M. P. Groover, *Fundamentals of Modern Manufacturing: Materials, Processes, and Systems*, 4th ed. Hoboken, NJ: Wiley, 2010.

[4] F. Klocke, *Manufacturing Processes 1: Cutting*. Berlin, Heidelberg: Springer, 2011. DOI: 10.1007/978-3-642-11979-8.

[5] F. Klocke, *Manufacturing Processes 4: Forming*. Berlin, Heidelberg: Springer, 2013. DOI: 10.1007/978-3-642-36772-4.

[6] C. F. Moreno-García, E. Elyan, and C. Jayne, "New trends on digitisation of complex engineering drawings", *Neural Computing and Applications*, vol. 31, no. 6, pp. 1695–1712, 2019. DOI: 10.1007/s00521-018-3583-1.

[7] L. Jamieson, C. F. Moreno-García, and E. Elyan, "A review of deep learning methods for digitisation of complex documents and engineering diagrams", *Artificial Intelligence Review*, vol. 57, 2024, Art. no. 136. DOI: 10.1007/s10462-024-10779-2.

[8] M. Oquab et al., "DINOv2: Learning robust visual features without supervision", *Transactions on Machine Learning Research*, 2024. [Online]. Available: https://openreview.net/forum?id=a68SUt6zFt.

[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410.

[10] C. Haar, H. Kim, and L. Koberg, "AI-based engineering and production drawing information extraction", in *Flexible Automation and Intelligent Manufacturing: The Human-Data-Technology Nexus (FAIM 2022)*, Cham: Springer, 2023, pp. 374–382. DOI: 10.1007/978-3-031-18326-3_36.

[11] D. van Strien, K. Beelen, M. Coll Ardanuy, K. Hosseini, B. McGillivray, and G. Colavizza, "Assessing the impact of OCR quality on downstream NLP tasks", in *Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART)*, SciTePress, 2020, pp. 484–496. DOI: 10.5220/0009169004840496.

[12] L. Xie et al., "Graph neural network-enabled manufacturing method classification from engineering drawings", *Computers in Industry*, vol. 142, 2022, Art. no. 103697. DOI: 10.1016/j.compind.2022.103697.

[13] D. Van Daele, N. Decleyre, H. Dubois, and W. Meert, "An automated engineering assistant: Learning parsers for technical drawings", in *Proceedings of the AAAI Conference on Artificial Intelligence (IAAI)*, vol. 35, 2021, pp. 15 195–15 203. DOI: 10.1609/aaai.v35i17.17783.

[14] C. F. You and S. S. Yang, "Automatic feature recognition from engineering drawings", *The International Journal of Advanced Manufacturing Technology*, vol. 14, pp. 495–507, 1998. DOI: 10.1007/BF01351395.

[15] J. Pu and K. Ramani, "On visual similarity based 2D drawing retrieval", *Computer-Aided Design*, vol. 38, no. 3, pp. 249–259, 2006. DOI: 10.1016/j.cad.2005.10.009.

[16] W. Zhang et al., "Component segmentation of engineering drawings using graph convolutional networks", *Computers in Industry*, vol. 147, 2023, Art. no. 103885. DOI: 10.1016/j.compind.2023.103885.

[17] M. T. Khan, Z. Yong, L. Chen, W. Feng, N. Y. J. Tan, and S. K. Moon, "A multi-stage hybrid framework for automated interpretation of multi-view engineering drawings using vision language model", in *Proceedings of the 13th International Conference on Industrial Engineering and Applications (ICIEA 2026)*, In press, 2026. arXiv: 2510.21862 [cs.CV].

[18] T. Baba et al., "Similarity-based partial image retrieval system for engineering drawings", in *Seventh IEEE International Symposium on Multimedia (ISM'05)*, 2005, pp. 303–310. DOI: 10.1109/ISM.2005.105.

[19] E. Rica, S. Álvarez, and F. Serratosa, "Group of components detection in engineering drawings based on graph matching", *Engineering Applications of Artificial Intelligence*, vol. 104, 2021, Art. no. 104404. DOI: 10.1016/j.engappai.2021.104404.

[20] Y. Quan, H. Zhao, J. Yi, and Y. Chen, *Self-supervised graph neural network for mechanical CAD retrieval*, 2024. arXiv: 2406.08863 [cs.IR].

[21] J. M. Saavedra, C. Stears, and W. Campos, "Achieving high performance on sketch-based image retrieval without real sketches for training", *Pattern Recognition Letters*, vol. 193, pp. 94–100, 2025. DOI: 10.1016/j.patrec.2025.04.018.

[22] W. Campos, J. M. Saavedra, and C. Stears, "A study on self-supervised sketch-based image retrieval on unpaired data (S3BIR)", *Neural Computing and Applications*, vol. 37, no. 18, pp. 11 945–11 963, 2025. DOI: 10.1007/s00521-025-11142-4.

[23] K. Higuchi and K. Yanai, "Patent image retrieval using transformer-based deep metric learning", *World Patent Information*, vol. 74, 2023, Art. no. 102217. DOI: 10.1016/j.wpi.2023.102217.

[24] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition", in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 945–953. DOI: 10.1109/ICCV.2015.114.

[25] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694. DOI: 10.1109/CVPR.2019.00482.

[26] M. Kucer, D. Oyen, J. Castorena, and J. Wu, "DeepPatent: Large scale patent drawing recognition and retrieval", in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 557–566. DOI: 10.1109/WACV51458.2022.00063.

[27] K. Ajayi, X. Wei, M. Gryder, et al., "DeepPatent2: A large-scale benchmarking corpus for technical drawing understanding", *Scientific Data*, vol. 10, 2023, Art. no. 772. DOI: 10.1038/s41597-023-02653-7.

[28] P. Khosla et al., "Supervised contrastive learning", in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)*, Curran Associates Inc., 2020, 1567. arXiv: 2004.11362.

[29] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations", in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020, pp. 1597–1607. arXiv: 2002.05709.

[30] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging", in *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019)*, Curran Associates Inc., 2019, pp. 3347–3357. arXiv: 1902.07208.

[31] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds", *ACM Transactions on Graphics*, vol. 38, no. 5, 2019, Art. no. 146. DOI: 10.1145/3326362.

[32] R. Ying, J. You, C. Morris, X. Ren, W. L. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling", in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018)*, Curran Associates Inc., 2018, pp. 4805–4815. arXiv: 1806.08804.

[33] A. Kashevnik et al., "An approach to engineering drawing organization: Title block detection and processing", *IEEE Access*, vol. 11, 2023. DOI: 10.1109/ACCESS.2023.3244603.

[34] M. Ondrejcek, J. Kastner, R. Kooper, and P. Bajcsy, "Information extraction from scanned engineering drawings", National Center for Supercomputing Applications, Urbana, IL, Tech. Rep. NCSA-ISDA09-001, 2009.

[35] B. Scheibel, J. Mangler, and S. Rinderle-Ma, "Extraction of dimension requirements from engineering drawings for supporting quality control in production processes", *Computers in Industry*, vol. 129, 2021, Art. no. 103442. DOI: 10.1016/j.compind.2021.103442.

[36] M. François, V. Eglin, and M. Biou, "Text detection and post-OCR correction in engineering documents", in *Document Analysis Systems (DAS 2022)*, Cham: Springer, 2022, pp. 726–740. DOI: 10.1007/978-3-031-06555-2_49.

[37] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings", in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2021, pp. 6894–6910. DOI: 10.18653/v1/2021.emnlp-main.552.

[38] T. N. Reddy et al., "Intelligent GD&T symbol detection in mechanical drawings: A comparative study of YOLOv11, Faster R-CNN, and RetinaNet for quality assurance", *Journal of Intelligent Manufacturing*, vol. 37, pp. 2941–2959, 2026. DOI: 10.1007/s10845-025-02669-3.

[39] S. Bickel, S. Goetz, and S. Wartzack, "Symbol detection in mechanical engineering sketches: Experimental study on principle sketches with synthetic data generation and deep learning", *Applied Sciences*, vol. 14, no. 14, 2024, Art. no. 6106. DOI: 10.3390/app14146106.

[40] Z. Fan, T. Chen, P. Wang, and Z. Wang, "CADTransformer: Panoptic symbol spotting transformer for CAD drawings", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 976–10 986. DOI: 10.1109/CVPR52688.2022.01071.

[41] A. Kirillov et al., "Segment anything", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3992–4003. DOI: 10.1109/ICCV51070.2023.00371. arXiv: 2304.02643.

[42] A. M. Chowdhury and S. Moon, "Generating integrated bill of materials using mask R-CNN artificial intelligence model", *Automation in Construction*, vol. 145, 2023, Art. no. 104644. DOI: 10.1016/j.autcon.2022.104644.

[43] C. Zhao and S. N. Melkote, "Learning the manufacturing capabilities of machining and finishing processes using a deep neural network model", *Journal of Intelligent Manufacturing*, vol. 35, no. 4, pp. 1845–1865, 2024. DOI: 10.1007/s10845-023-02134-z.

[44] H. Hu, C. Zhang, and Y. Liang, "Detection of surface roughness of mechanical drawings with deep learning", *Journal of Mechanical Science and Technology*, vol. 35, pp. 5541–5549, 2021. DOI: 10.1007/s12206-021-1125-8.

[45] P. Szymański and T. Kajdanowicz, "A network perspective on stratification of multi-label data", in *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA)*, ser. Proceedings of Machine Learning Research, vol. 74, 2017, pp. 22–35. arXiv: 1704.08756.