The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-G-2025
ISPRS Geospatial Week 2025 “Photogrammetry & Remote Sensing for a Better Tomorrow…”, 6–11 April 2025, Dubai, UAE

A Novel Correspondence Model for Linking Objects and Texts in Construction Plans

Shuwei Hong, Steven Landgraf, Markus Hillemann, Markus Ulrich

Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Germany
(shuwei.hong, steven.landgraf, markus.hillemann, markus.ulrich)@kit.edu

Keywords: Object Detection, Text Recognition, Multi-modal Analysis, Construction Plans, Correspondence Model

Abstract

Construction plans integrate visual and textual information that is essential for construction projects. However, the huge diversity of formats of these plans poses challenges for automated analysis. This paper presents a novel correspondence model that links objects and texts in construction plans, providing a unified approach to interpreting various formats, such as scanned blueprints, CAD drawings, and digital construction documents. Leveraging deep-learning-based object detection and text recognition techniques, our model establishes semantic correspondences between visual and textual elements. We integrate CLIP-based models with ViT-based encoders as part of our approach to enhance feature extraction and correspondence learning. By employing a threshold-based determination, our model effectively resolves cases where a single text passage may describe multiple objects or where a single object is referenced by multiple pieces of text. This capability enables the model to establish robust correspondences between objects and texts, laying a strong foundation for subsequent semantic understanding and information extraction. We evaluate its effectiveness on labeled datasets and demonstrate that our model achieves high precision, recall, F1-score, and accuracy. Hence, we provide a feasible approach to establishing object-text correspondences in construction plan analysis. The results suggest its potential to serve as a foundation for further exploration in the automated analysis of technical drawings, particularly in the context of quality assurance and construction project planning.

1. Introduction

Construction plans serve as fundamental blueprints in the architecture, engineering, and construction (AEC) industry, facilitating effective project design, coordination, and execution (Eastman et al., 2011). These plans incorporate both intricate visual object representations and textual annotations, detailing critical structural components, dimensions, material specifications, and construction guidelines. Despite their essential role, the digitization and standardization of construction plans have not kept pace with advancements in automated document analysis, leading to their frequent availability solely as raster graphics. This limitation presents substantial challenges for automated interpretation, as the information must be extracted, structured, and semantically linked for downstream applications (Zhang et al., 2024).

The extraction of visual information, such as object layouts and structural elements, from images is typically performed using advanced object detection and segmentation methods (Jamieson et al., 2024). Similarly, textual annotations, including labels, descriptions, and measurements, are commonly retrieved through Optical Character Recognition (OCR) techniques (Impedovo et al., 1991). However, these modalities are often treated independently, neglecting the complex spatial and semantic relationships that inherently exist in construction plans. Establishing accurate associations between textual elements and corresponding visual objects is crucial for ensuring data integrity, reducing errors in automated workflows, and supporting applications such as automated quality assurance (ScienceNet, 2021) and real-time project monitoring (Zhang et al., 2024).

Analyzing highly detailed and information-dense construction plans introduces further challenges. Overlapping elements, spatially close but unrelated annotations, and inconsistencies in textual descriptions contribute to ambiguities in interpretation. Moreover, variations in formatting, font styles, and handwriting in manually annotated plans further complicate automated processing (Zhang et al., 2024). Consequently, knowledge about the correspondences between visual objects and textual annotations is particularly beneficial for the automated analysis of technical drawings. In addition to handling challenging scenarios — such as overlapping elements, ambiguous annotations, and inconsistencies — this knowledge also facilitates new tasks, including detecting missing textual descriptions or objects, as well as verifying the completeness and functionality of construction plans.

To tackle these challenges, this paper presents a tailored multi-modal framework that establishes robust correspondences between objects and texts in construction plans. Instead of addressing object detection or text recognition tasks directly, our approach leverages their results as multi-modal input for a novel correspondence model and assumes that both tasks have already been accurately performed. Previous work on visual-semantic correspondence (Karpathy and Fei-Fei, 2015, Chen et al., 2020b) has demonstrated the effectiveness of shared embedding spaces for linking textual and visual information. Additionally, top-down and bottom-up attention mechanisms (Anderson et al., 2018) have been widely applied to enhance semantic and spatial reasoning in multi-modal tasks (Anderson et al., 2018, Cheng et al., 2020, Ghosh et al., 2019). However, these approaches primarily focus on generic image-text pairs and do not address the specific challenges present in construction document analysis, such as identifying missing textual descriptions or handling complex spatial correspondences between objects and labels. To bridge this gap, we propose a contrastive learning framework specifically designed for technical drawings, integrating both semantic and spatial correspondences. By considering domain-specific challenges such as structured annotations, varying text orientations, and object

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-G-2025-597-2025

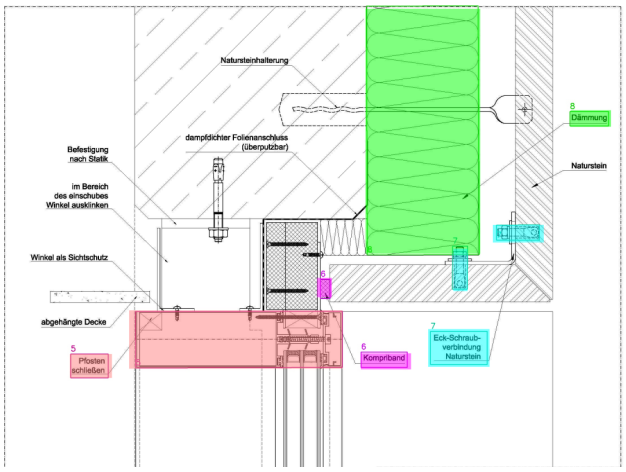

Figure 1. Example image of the dataset that was utilized to evaluate our correspondence model. Corresponding objects and texts are illustrated in the same color.

co-occurrence patterns, our method ensures robust correspondences between visual objects and textual annotations, thereby improving the accuracy of object-text linking in construction plans.

We evaluate our novel correspondence model on a labeled dataset of construction plans, specifically curated for this task and labeled by domain experts. The dataset consists of construction plans annotated with object categories, their locations, as well as textual content and corresponding positions. Additionally, the correspondences between objects and their associated texts are labeled, by assigning them the same group ID. Figure 1 presents an example from our dataset, illustrating the annotated construction plans with object bounding boxes and text bounding boxes. This dataset enables the testing of upstream tasks such as object detection and text recognition, as well as downstream tasks like establishing correspondences in this particular domain. Experimental results from two Vision-Transformer (ViT)-based models (Dosovitskiy, 2020) confirm the feasibility of our approach, showing that it achieves satisfactory performance in object-text correspondence tasks within construction plan analysis.

2. Related Work

2.1 Object Detection and Text Recognition

Object detection and text recognition are fundamental components in the analysis of technical drawings, such as constructions plans. Modern object detection models, such as YOLO (Redmon et al., 2016), have demonstrated high efficiency in identifying and localizing objects in images, while Optical Character Recognition (OCR) methods (Smith, 2007) enable the extraction of textual information from scanned images and documents (Appalaraju et al., 2021). These techniques have been applied to the analysis of technical drawings (Nguyen et al., 2021, Huang et al., 2019, Rezvanifar et al., 2020), where identifying object categories and extracting textual content are essential for various downstream applications.

However, most existing object detection models are primarily optimized for real-world images, making their direct application to technical drawings challenging. Unlike real-world images, construction plans contain domain-specific textual annotations, geometric dependencies, and hierarchical relationships. While these factors introduce unique challenges, they can often be addressed through careful fine-tuning (Rothmeier et al., 2024) and domain adaptation techniques (Sarkar and Stricker, 2019). Although conventional OCR techniques (Smith, 2007) are primarily designed for printed and handwritten text, they often struggle with the specific characteristics of construction plans, such as non-standard fonts, rotated text orientations, and domain-specific notation systems. These challenges necessitate domain-specific adaptations in both object detection and text recognition pipelines to ensure accurate interpretation.

2.2 Challenges in Document Parsing

Parsing technical drawings requires advanced methods that integrate textual and graphical elements. Traditional approaches, such as heuristic segmentation (Moreno-García et al., 2017) and object-text recognition (Nguyen et al., 2021), treat object detection and text extraction as separate tasks, limiting their ability to capture contextual correspondences.

A major challenge in construction plans is the precise spatial alignment of objects. Unlike standard documents with predictable layouts, technical drawings vary in format, scale, and object positioning. Overlapping symbols, multi-line annotations, and mixed structured and unstructured content further complicate parsing (Zhang et al., 2024).

Recent advancements, such as LayoutLM (Xu et al., 2020a), integrate textual and spatial embeddings to enhance document understanding, while LayoutLMv2 (Xu et al., 2020b) further incorporates visual features for structured document processing. StructText (Li et al., 2021) learns hierarchical text-layout representations for improved information extraction. However, these models are primarily designed for structured documents like forms and receipts, where text follows a fixed layout, rather than for establishing object-text correspondence in construction plans with unknown and diverse layouts. Their effectiveness on technical drawings remains limited due to the domain-specific symbols and intricate spatial correspondences. Consequently, construction and engineering workflows require specialized models that can accurately capture the interplay between texts and geometric objects.

2.3 Multimodal Learning for Drawings

Recent advancements in multimodal learning have introduced new possibilities for the simultaneous analysis of visual and textual information in technical drawings. Contrastive learning techniques (Oord et al., 2018, Chen et al., 2020a) have been employed to align representations from different modalities, facilitating improved cross-modal retrieval and interpretation. Similarly, transformer-based architectures (Devlin, 2018, Li et al., 2022) have demonstrated strong capabilities in structured document understanding by leveraging self-attention mechanisms to model contextual dependencies. However, while these approaches (Oord et al., 2018, Chen et al., 2020a, Devlin, 2018, Li et al., 2022) have proven useful for general document processing, they do not address the specific challenges posed by engineering and construction plans, where there is no standard template for spatial arrangement, making interpretation dependent on complex correspondences between objects and textual annotations.

Existing multimodal frameworks primarily rely on pre-trained visual and textual encoders optimized for natural images and general text understanding. These models struggle with the unique characteristics of construction plans, where textual elements often describe geometric objects, and their meaning is highly dependent on spatial context. Some prior works (Sun et al., 2021, Lin and Hu, 2022) have explored multimodal analysis in structured documents, but they typically focus on either object detection or text recognition, treating them as independent tasks. This results in a critical limitation: while previous approaches can identify objects and extract text, they lack a mechanism to establish semantic correspondences between these elements, which is essential for understanding construction plans at a deeper level.

To bridge this gap, we introduce a correspondence model that explicitly links detected objects with related textual annotations. Instead of treating text and object recognition as separate outputs, our approach integrates their spatial and semantic relationships into a structured representation. By leveraging contrastive learning and domain-specific adaptations, our model learns to correspond objects with their respective descriptions, supporting tasks such as automated quality assurance of technical drawings and digital construction management. This goes beyond mere detection, allowing us to interpret construction plans in a way that aligns with real-world engineering workflows, such as verifying plan completeness, identifying thermal bridges, and supporting BIM processes. Our method has the potential to significantly increase precision in the inspection of technical drawings and contributes to automation in the AEC industry.

3. Methodology

The proposed framework is illustrated in Figure 2. This section provides an overview of the methodology, detailing the input preprocessing, the dataset structure for correspondence learning, the correspondence modeling, and the correspondence score computation.

3.1 Input Preprocessing

The input to our correspondence model comprises Oriented Bounding Boxes (OBBs) for detected objects and textual elements, extracted text content, and the full construction plan image. Also as shown in Figure 2, the preprocessing stage is divided into two primary components, which serve as preliminary tasks rather than integral parts of the correspondence model itself:

Object Detection: To identify relevant construction elements and text regions, we employ YOLOv8 (Redmon et al., 2016). We choose version 8 because, unlike newer versions, this version is also able to predict the orientations of the bounding boxes. The orientation is important because construction plans often contain objects that are not aligned horizontally (e.g., a tilted screw in an assembly drawing). The object detection model is fine-tuned on a domain-specific dataset to enhance recognition accuracy for construction-related objects. The output consists of OBBs containing positional and category attributes of detected elements. Notably, objects classified as ”text” are further processed in the text recognition stage.

Text Recognition: The primary goal of this stage is to extract textual content from the detected text regions and transcribe it into machine-readable text. For this purpose, we utilize Tesseract OCR (Smith, 2007).

Once the objects and texts are extracted, the detected OBBs are highlighted in the original image, which then serves as *overlay image*. An example of a processed overlay image is shown in Fig. 3, where one object-text correspondence is visually encoded. This visualization ensures that the correspondence model captures a comprehensive representation of the construction plan, preserving global spatial structures and contextual correspondences.

3.2 Dataset Structure for Correspondence Learning

To facilitate effective model training, we construct a structured ground-truth dataset for correspondence learning. This dataset comprises the original construction plans, detected object categories along with their OBBs, recognized text content with their corresponding OBBs, and the existence of relationships between them. For each construction plan, we generate the following data representations:

- **Object regions:** Cropped image patches corresponding to the OBBs of detected objects, where each patch is assigned a category label.
- **Text regions:** Cropped image patches corresponding to the OBBs of detected text regions, where each patch is mapped to the recognized textual content.
- **Overlay images:** Visual overlays where each detected object and text region is assigned a unique overlay image. Each overlay highlights the corresponding OBB region, ensuring that objects and their associated texts are distinctly represented while maintaining their spatial relationships.
- **Object-text correspondences:** Binary labels indicating for each object-text pair whether they correspond (1) or not (0).

3.3 Correspondence Modeling

The core of our methodology is the establishment of semantic relationships between detected objects and texts to find corresponding pairs. The correspondence model consists of the following key stages:

Image Feature Extraction: We employ a ViT-based (Dosovitskiy, 2020) image feature extractor to extract image features from three different inputs: object regions, text regions, and the overlay images highlighting object-text relations. The combined image feature vector f_c is calculated by the weighted sum

$$f_c = \alpha f_1 + \beta f_2 + \gamma f_3, \tag{1}$$

where all feature vectors f_1, f_2, and f_3 are normalized to have unit norm (e.g., $\|f_i\| = 1$, for $i \in \{1, 2, 3\}$). Specifically, f_1 represents the feature of the object region, f_2 corresponds to the feature of the text region, f_3 denotes the feature of the overlay image, and α, β, $\gamma \in [0, 1]$ denote the weighting factors. Initially, the weights are set to be equally distributed as $\alpha = 0.333, \beta = 0.333, \gamma = 1 - \alpha - \beta = 0.333$.

Thus, the system maintains two degrees of freedom, allowing adjustments to α and β, while γ is constrained by their sum to ensure proper normalization. A detailed analysis of weight optimization and its impact on model performance will be discussed in Section 5.3, where we provide a performance map

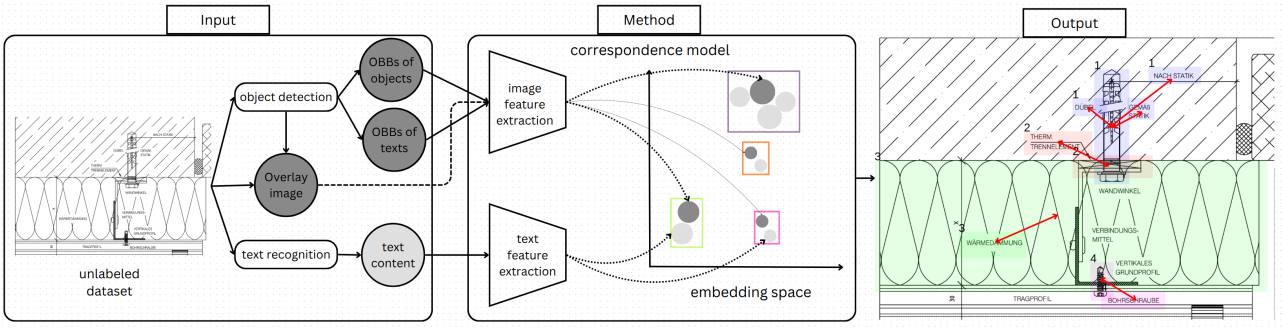

Figure 2. Overview of the proposed correspondence model for linking objects and texts in construction plans.

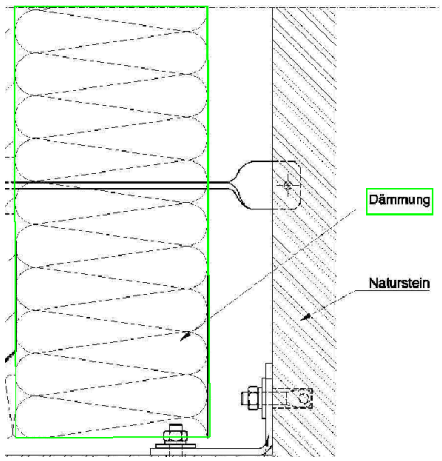

Figure 3. Overlay image (cropped) with one positive object-text pair.

visualization illustrating how different weight configurations influence key evaluation metrics.

Text Feature Extraction: The textual content extracted from the OCR module (Smith, 2007) is processed using a ViT-based (Dosovitskiy, 2020) text feature extractor. Through this process, the text is transformed into feature representations, denoted as f_t, embedding the extracted content into the same shared space as the image features to facilitate effective correspondence learning.

To ensure feature consistency, we normalize the combined image feature vector and the text feature vector:

$$\hat{f}_c = \frac{f_c}{\|f_c\|_2}, \quad \hat{f}_t = \frac{f_t}{\|f_t\|_2} \tag{2}$$

Correspondence Modeling and Loss Function: To establish the correspondence between image and text features, we employ contrastive learning. Both image and text features are projected into a common embedding space through separate linear projection layers. The model is trained to minimize distances between positive and maximize distances between negative object-text pairs.

The correspondence between image and text features is measured using cosine similarity

$$s_{\text{object-text}} = \frac{\cos(\hat{f}_c, \quad \hat{f}_t) + 1}{2}, \tag{3}$$

which rescales the similarity score to $[0, 1]$ to be compatible with the loss function.

We utilize the Binary Cross-Entropy Loss

$$\mathcal{L}_{\text{object-text},b} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log s_i + (1 - y_i)\log(1 - s_i)\right], \tag{4}$$

where s_i is the predicted probability output by the model, i.e., $s_i = s_{\text{object-text},i}$, and y_i is the ground-truth label indicating whether the object-text pair corresponds. N represents the total number of object-text pairs in a single batch.

Given a total of B batches in an epoch, the final loss for the epoch is computed as

$$\mathcal{L}_{\text{total}} = \frac{1}{B}\sum_{b=1}^{B}\mathcal{L}_{\text{object-text},b}. \tag{5}$$

This loss function enforces a strong correspondence between positive object-text pairs while ensuring clear separation of negative ones.

3.4 Correspondence Score Computation

As described before, once object detection and text recognition have been performed, the model extracts feature representations for both objects (f_c) and text (f_t). It then computes a similarity between these representations to obtain the correspondence score. If this score exceeds a predefined threshold τ, the associated object and text are considered semantically related. By using threshold-based determination, our model resolves cases where multiple objects and texts need to be linked, ensuring accurate semantic correspondences for downstream information extraction.

4. Experimental Setup

4.1 Dataset

Our dataset contains 10,187 potential object-text correspondences, among which 215 are positive pairs, i.e., actual correspondences. The correspondences are distributed across 30 high-resolution construction plan images. The dataset also includes complicated cases where a single object corresponds to multiple texts. These annotations ensure high-quality data for object-text correspondence learning.

The dataset is split based on the number of construction plan images, with 80% used for training and 20% for testing, maintaining a representative balance of the visual appearances of

Table 1. Statistics of the test dataset, including the number of positive and negative object-text pairs, object counts, text box counts, and the total potential correspondences.

Plan	Positive Pairs	Negative Pairs	Object Counts	Text Box Counts	Potential Correspondences
Plan 1	5	235	20	12	240
Plan 2	1	230	33	7	231
Plan 3	2	163	15	11	165
Plan 4	7	175	14	13	182
Plan 5	7	185	12	16	192
Plan 6	5	145	10	15	150
Overall	27	1133	104	74	1160

the construction plans (i.e., formatting, line styles, font types, etc.). The data representation and preprocessing steps follow the structure outlined in Section 3.2.

Table 1 summarizes the key statistics of the test dataset, including the number of positive and negative pairs, objects, and text boxes for object-text correspondence detection. The number of all potential correspondences is computed as the product of the object counts and the text box counts.

4.2 Training Configuration

Model Architecture. For our correspondence model, we use a ViT-L/14 (Dosovitskiy, 2020) for both the image encoder and the text encoder. Since Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) has demonstrated strong performance in aligning visual and textual representations by training on large-scale image-text pairs, we initialize all models with weights from CLIP (Radford et al., 2021). CLIP (Radford et al., 2021) employs contrastive learning, where images and text are mapped into a shared latent space, encouraging semantically related pairs to be close while pushing unrelated pairs apart. This enables our model to effectively learn the correspondence between textual and visual information.

Training Hyperparameters. For all experiments, we use the following hyperparameters. The batch size is set to 8, and the learning rate is initialized at 1×10^{-6}. The model is trained for a total of 300 epochs. The optimization process utilizes the AdamW optimizer (Loshchilov and Hutter, 2019) with a weight decay of 0.001, ensuring effective weight regularization and preventing overfitting. Furthermore, we employ early stopping based on the performance on the test dataset.

Balancing positive and negative object-text pairs. As shown in Table 1, the number of negative object-text pairs is significantly higher than that of positive pairs within the same construction plan. To address this imbalance, we ensure that in each epoch, all positive pairs are included in the training, along with an equal number of randomly sampled negative pairs, in order to maintain a 1:1 ratio during training.

Hardware and Acceleration: The training process is fully optimized for GPU acceleration, using an NVIDIA A100-PCIE-40GB GPU with CUDA enabled.

4.3 Evaluation Metrics

To assess the effectiveness of the proposed framework in establishing correct object-text correspondences, a comprehensive set of evaluation metrics is employed. The formal definitions and computation of these metrics follow standard formulations in classification tasks (Manning et al., 2008):

- **Precision** quantifies the proportion of correctly established object-text correspondences among all pairs predicted as valid.

Algorithm 1: Evaluation Logic

Input : A dataset from Section 3.2, i.e., each potential correspondence contains a [Binary-label]
Output: TP, FN, FP, and TN
foreach Correpondence $\in$ dataset **do**
  label $\leftarrow$ [Binary-label];
  $S \leftarrow$ model.predict(Correpondence);
  if label $= 1$ **then**
    if $S \geq \tau$ **then**
      TP $\leftarrow$ TP $+ 1$;
    else
      FN $\leftarrow$ FN $+ 1$;
  else if label $= 0$ **then**
    if $S \geq \tau$ **then**
      FP $\leftarrow$ FP $+ 1$;
    else
      TN $\leftarrow$ TN $+ 1$;

- **Recall** measures the model's ability to correctly retrieve all true object-text correspondences from the dataset.
- **F1-Score** is the harmonic mean of Precision and Recall, providing a balanced measure of classification performance. This metric is particularly useful in scenarios where an optimal trade-off between Precision and Recall is required.
- **Accuracy** measures the overall proportion of correctly classified object-text pairs, encompassing both valid and invalid correspondences.

Evaluation Procedure. Algorithm 1 outlines the evaluation logic used to compute classification performance metrics, specifically the counts of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). TP represents the number of correctly established object-text correspondences, while FP denotes incorrectly established correspondences. Similarly, FN corresponds to actual correspondences that the model failed to establish, whereas TN represent the number of correctly identified non-correspondences. The process relies on a scoring mechanism to determine the validity of object-text correspondences in a given construction plan. For each object-text pair, the model assigns a confidence score S, which is then compared against a predefined threshold τ. Based on this comparison and the ground truth label, the pair is classified into one of the four categories, providing the foundation for assessing the model's overall performance in terms of precision, recall, and other relevant metrics.

5. Results

This section presents the evaluation results of our proposed framework on the test dataset. First, we describe the quantitative results that focus on the four evaluation metrics: Precision, Recall, F1-Score, and Accuracy, as well as qualitative results of our model with optimal parameter settings. Thereafter, we describe various ablation studies that justify the model's architecture and parameter choices.

5.1 Overall Performance

The performance of our method across all test images is summarized in the right column of Table 2. Almost all correspondences predicted by our method are indeed valid correspondences, but not all correspondences are found, which is reflected in the high accuracy and slightly lower recall. This can be

Table 2. Performance of our method with optimal parameter settings. The right column shows the results of our method, whereas the left column shows the considerably worse results obtained with a smaller encoder (cf. Section 5.2).

Metric	ViT-B/32	ViT-L/14
Precision (%)	23.5	86.3
Recall (%)	50.1	79.2
F1-Score (%)	32.3	82.6
Accuracy (%)	91.8	99.2
Parameters (M)	151.28	427.62
Inference Time (ms)	27.4	45.2

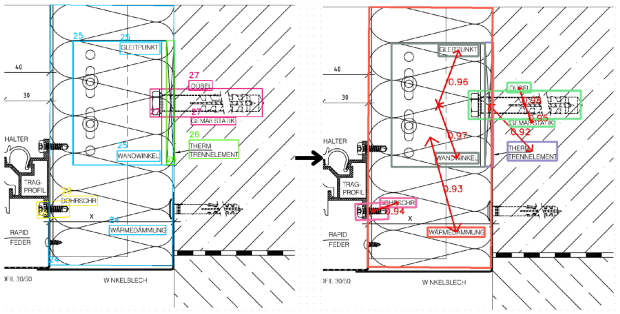

Figure 4. Ground truth correspondences (left) and predicted correspondences (right) for a successful example (cropped). The predicted correspondences match the ground truth exactly, demonstrating the model’s reliability when trained with sufficient plans with a similar visual appearance.

attributed to the relatively high number of negative pairs compared to positive pairs, which are inherent to the task. The reliability in distinguishing negative pairs from actual correspondences underscores the robustness of our method.

Figures 4 and 5 illustrate the qualitative results. In each figure, the left side presents the ground truth annotations for the correspondences between objects and text in the construction plans, while the right side shows the predicted correspondences generated by our model. If the confidence score between an object and a text exceeds the predefined threshold τ (in this case, $\tau = 0.88$), the model establishes a correspondence, represented by a red arrow linking the object and text. Additionally, the confidence score is displayed on the arrow.

Figure 4 shows a case where the predicted correspondences match perfectly with the ground truth. This highlights the robustness of the model in scenarios where there is no knowledge gap, that is, when sufficient training data of the given planning style is available.

In contrast, Figure 5 demonstrates a limitation of the model. When faced with construction plans with a divergent appearance compared to the majority of training data, the model struggles to establish the correct correspondences. This case emphasizes the importance of diverse and representative training data to improve generalization.

5.2 Impact of the Encoder Size

In a first ablation study, we test the impact of the encoder size. For this, we use ViT-B/32 (Dosovitskiy, 2020), which is a smaller encoder compared to ViT-L/14 with a patch size of 32 to assess efficiency. The results are shown in the left column of Table 2. The performance of ViT-B/32 is significantly lower than that of ViT-L/14 in all relevant metrics even though the

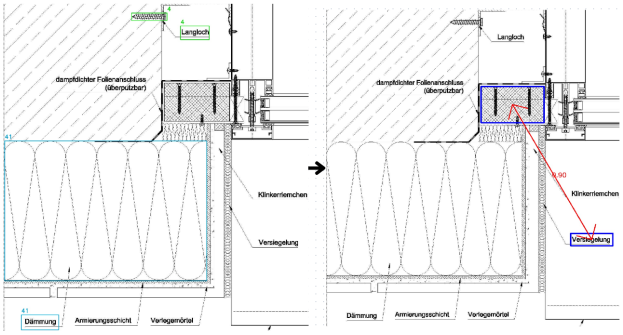

Figure 5. Ground truth correspondences (left) and predicted correspondences (right) for an unsuccessful example (cropped). The model struggles to recognize the correct correspondences due to insufficient construction plans with this visual appearance.

model has fewer parameters and a shorter inference time. This highlights the limitations of smaller encoders in complex classification tasks. Due to its lower F1-Score, ViT-B/32 is not considered further in our analysis.

5.3 Impact of the Weight Parameters α, β, and γ on Model Performance

To further analyze the models’ behavior, we investigate the impact of the weight parameters α, β, and $\gamma = 1 - \alpha - \beta$, which control the contribution of different features in the correspondence model. Their values significantly affect the model’s ability to correctly associate textual and object elements. To analyze this effect, we systematically vary these parameters and evaluate their impact on F1-Score.

For this analysis, we fix the threshold τ at its default value of 0.8 and explore different weight combinations. Since it is infeasible to exhaustively evaluate all possible parameter combinations, our analysis is based on a representative subset of configurations rather than an absolute global optimum. From the resulting performance map in Figure 6, we observe that the best-performing region is characterized by relatively high values of α and lower values of β and γ, with γ being noticeably larger than β. This suggests that object features have a significant impact on performance, while text box features contribute relatively less, additionally, the overlay image feature also plays an important role, providing essential global contextual information.

The best of the tested parameter combinations is $\alpha = 0.62, \beta = 0.05, \gamma = 1 - \alpha - \beta = 0.33$. This parameter combination is used in Section 5.4 to analyze the impact of the threshold τ.

5.4 Impact of Threshold τ on Evaluation Metrics

After determining the optimal weight parameters, we now analyze how different values of the threshold τ affect Precision, Recall, F1-Score, and Accuracy. In this experiment, we fix the weight parameters at the values from Section 5.3.

The results are presented in Figure 7. It illustrates how Precision, Recall, F1-Score and Accuracy change as we adjust the threshold τ. Lower thresholds generally lead to higher recall and lower precision due to an increased number of positive predictions. Conversely, higher thresholds improve precision by reducing false positives but may also slightly impact recall.

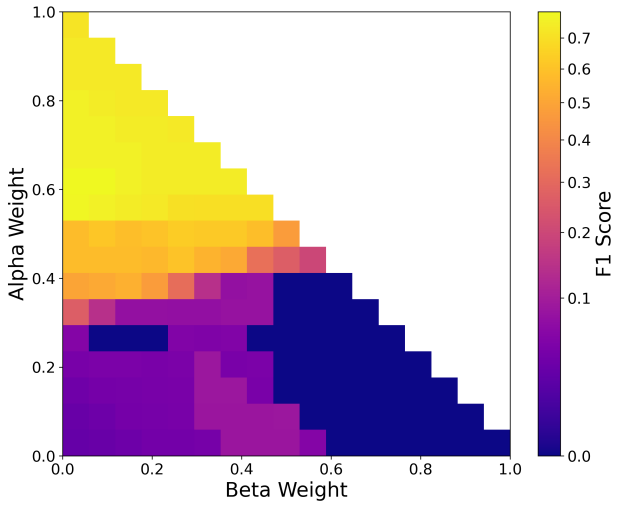

Figure 6. Effect of weight parameters (α, β) on F1-Score. Lighter regions indicate better performance.

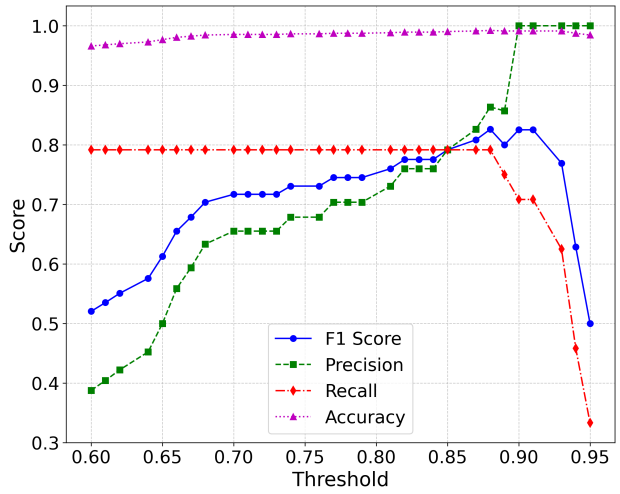

Figure 7. Effect of different threshold values τ on F1-Score, Precision, Recall and Accuracy.

However, since we only consider threshold values starting from 0.6, recall remains relatively stable across this range, with noticeable changes only for values above 0.88. This trade-off must be carefully adjusted based on the application's requirements.

6. Discussion

Our method enables the identification of corresponding objects and texts in highly complex construction plans using a simple threshold-based determination system, a task that has not yet been addressed by previous work. By fine-tuning a CLIP-based ViT within a contrastive learning framework, we achieve robust performance with a balanced trade-off between precision and recall. Notably, the larger encoder ViT-L/14 offers significantly better performance than the smaller encoder ViT-B/32, which is particularly interesting given that we are training on a very limited number of training samples. A possible explanation for the superior performance of ViT-L/14 is that its more potent feature representations, derived from CLIP pre-training, enable more effective generalization – even with limited fine-tuning data. These results underscore the benefits of using advanced models like ViT-L/14, pre-trained on vast amounts of data, for real-world applications such as parsing complex construction documents. The ability to achieve strong performance with limited training data further demonstrates the practical viability of our approach, especially in domains where annotated data is scarce.

Limitations. While our approach offers clear advantages, it is important to address potential limitations and explore opportunities for improvement. One key aspect is dataset composition. As shown in Table 1, the test dataset includes three images that have the same visual appearance regarding formatting, linestyles, and font types. Although unbiased, this relatively uniform structure may lead to an overestimation of model performance. To enhance robustness, future evaluations should incorporate more diverse datasets with variations in text styles, annotation conventions, and document quality as well as real-world scanned documents that often contain noise, distortions, and inconsistencies in text placement. Next to the expansion of the dataset with real-world samples and synthetic augmentations, leveraging domain adaptation techniques can improve generalizability and ensure reliable performance across various document conditions.

7. Conclusion

This paper introduces a completely novel approach capable of linking objects and texts in construction plans. By exploiting CLIP (Radford et al., 2021), a Vision Transformer pre-trained to connect text and images, we establish semantic correspondences between visual and textual elements. Our correspondence model, fine-tuned within a contrastive learning framework, effectively resolves even complicated cases where a single text corresponds to multiple objects or vice versa, employing a simple yet highly effective threshold-based decision system.

Experimental results demonstrate robust performance in terms of Precision, Recall, F1-Score, and Accuracy, reinforcing the feasibility of deep learning for construction plan analysis. Based on a limited yet representative test dataset, our correspondence model with a ViT-L/14 encoder achieves an impressive F1-Score of 82.6% and a remarkable Accuracy of 99.2%. By capturing spatial and semantic relationships, our method reduces reliance on explicit rule-based parsing, paving the way for more automated workflows in architecture and engineering.

Future Work. Looking ahead, our findings underscore the broader potential of deep learning-based approaches for automated construction plan analysis. A promising direction for future research lies in the development of an end-to-end framework that integrates object detection, text recognition and correspondence establishment. This would not only streamline the processing pipeline but also enhance efficiency and scalability. By sharing feature representations across tasks, such a framework could improve generalizability while reducing computational overhead.

Acknowledgments

The authors sincerely appreciate Michael Werkmann's assistance in annotating the dataset, which contributed to the quality of our experimental evaluation.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., 2018. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6077–6086.

Appalaraju, S., Jasani, B., Kota, B. U., Xie, Y., Manmatha, R., 2021. Docformer: End-to-end transformer for document understanding. *Proceedings of the IEEE/CVF international conference on computer vision*, 993–1003.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations. *International conference on machine learning*, PMLR, 1597–1607.

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J., 2020b. Uniter: Learning universal image-text representations. *Computer Vision – ECCV 2020*, Springer International Publishing, Cham, 104–120.

Cheng, L., Wei, W., Mao, X., Liu, Y., Miao, C., 2020. Stack-VS: Stacked Visual-Semantic Attention for Image Caption Generation. *IEEE Access*, 8, 154953-154965.

Devlin, J., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eastman, C., Teicholz, P., Sacks, R., Liston, K., 2011. *BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers, and Contractors*. John Wiley & Sons.

Ghosh, S., Burachas, G., Ray, A., Ziskind, A., 2019. Generating natural language explanations for visual question answering using scene graphs and visual attention.

Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C. V., 2019. ICDAR2019 competition on scanned receipt ocr and information extraction. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, IEEE.

Impedovo, S., Ottaviano, L., Occhinegro, S., 1991. OPTICAL CHARACTER RECOGNITION — A SURVEY. *International Journal of Pattern Recognition and Artificial Intelligence*, 05(01n02), 1-24. https://doi.org/10.1142/S0218001491000041.

Jamieson, L., Moreno-Garcia, C. F., Elyan, E., 2024. Towards fully automated processing and analysis of construction diagrams: AI-powered symbol detection. *International Journal on Document Analysis and Recognition (IJDAR)*. https://doi.org/10.1007/s10032-024-00492-9.

Karpathy, A., Fei-Fei, L., 2015. Deep visual-semantic alignments for generating image descriptions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3128–3137.

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F., 2022. Trocr: Transformer-based optical character recognition with pre-trained models.

Li, Y., Qian, Y., Yu, Y., Qin, X., Zhang, C., Liu, Y., Yao, K., Han, J., Liu, J., Ding, E., 2021. Structext: Structured text understanding with multi-modal transformers. *Proceedings of the 29th ACM International Conference on Multimedia*, 1912–1920.

Lin, R., Hu, H., 2022. Multimodal contrastive learning via uni-modal coding and cross-modal prediction for multimodal sentiment analysis. *arXiv preprint arXiv:2210.14556*.

Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization. *International Conference on Learning Represention (ICLR)*.

Manning, C. D., Raghavan, P., Schütze, H., 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Moreno-García, C. F., Elyan, E., Jayne, C., 2017. Heuristics-based detection to improve text/graphics segmentation in complex engineering drawings. *Proceedings of the 18th International Conference on Engineering Applications of Neural Networks (EANN 2017)*, Springer International Publishing, 87–98.

Nguyen, M. T., Pham, V. L., Nguyen, C. C., Nguyen, V. V., 2021. Object detection and text recognition in large-scale technical drawings. *arXiv preprint arXiv:2107.12345*.

Oord, A., Li, Y., Vinyals, O., 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint*, arXiv:1807.03748.

Radford, A., Kim, J. W., Hallacy, C., et al., 2021. Learning transferable visual models from natural language supervision. *Proceedings of the International Conference on Machine Learning (ICML)*, 8748–8763.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788.

Rezvanifar, A., Cote, M., Albu, A. B., 2020. Symbol spotting on digital architectural floor plans using a deep learning-based framework. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 568–569.

Rothmeier, T., Huber, W., Knoll, A. C., 2024. Time to shine: Fine-tuning object detection models with synthetic adverse weather images. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4435–4444.

Sarkar, K., Stricker, D., 2019. Simple domain adaptation for cad based object recognition. *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM,*, INSTICC, SciTePress, 429–437.

ScienceNet, 2021. A review of automated quality control methods based on image processing. *Science Net News*. Accessed: 2024-12-01.

Smith, R., 2007. An overview of the Tesseract OCR engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2, 629–633.

Sun, H., Kuang, Z., Yue, X., Lin, C., Zhang, W., 2021. Spatial dual-modality graph reasoning for key information extraction. *arXiv preprint arXiv:2103.14470*.

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., 2020a. Layoutlm: Pre-training of text and layout for document image understanding. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACM, 1192–1200.

Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W. et al., 2020b. Layoutlmv2: Multimodal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

Zhang, Q., Huang, V. S.-J., Wang, B., Zhang, J., Wang, Z., Liang, H., Wang, S., Lin, M., He, C., Zhang, W., 2024. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *arXiv preprint arXiv:2410.21169*.