ARCHICLIP

Enhanced Contrastive Language–Image Pre-training Model With Architectural Prior Knowledge

SHENGTAO XIA[1], YIMING CHENG[2] and RUNJIA TIAN[3]
[1] *Architectural Association School of Architecture*
[2] *Xi'an Jiaotong-Liverpool University*
[3] *Harvard University*
[1] *shengtao.xia@aaschool.ac.uk , 0009-0007-3106-521X*
[2] *y1mingc201@gmail.com, 0009-0008-6101-8710*
[3] *runjiatian@gmail.com, 0000-0002-5983-9754*

Abstract. In the rapidly evolving field of Generative AI, architects and designers increasingly rely on generative models for their workflows. While previous efforts focused on functional or building performance aspects, designers often prioritize novelty in architectural design, necessitating machines to evaluate abstract qualities. This article aims to enhance architectural style classification using CLIP, a Contrastive Language–Image Pre-training method. The proposed workflow involves fine-tuning the CLIP model on a dataset of over 1 million architecture-specific image-text pairs. The dataset includes project descriptions and tags, aiming at capturing spatial quality. Fine-tuned CLIP models outperform pre-trained ones in architecture-specific tasks, showcasing potential applications in training diffusion models, guiding generative models, and developing specialized search engines for architecture. Although the dataset awaits human designer review, this research offers a promising avenue for advancing generative tools in architectural design.

Keywords. Machine Learning, Generative Design, Contrastive Language-Image Pre-Training, Artificial Intelligence.

1. Introduction

With the fast development of Generative AI, generative tools are taking up an increasing share of day-to-day workflows of architects and designers. Usually, generative models use algorithms to evaluate the differences between generated outputs and desired ground truth, using this feedback to improve the generation quality. Many attempts have been made to assess architecture with algorithms, but most focus on functional or performance aspects, such as structural stabilities. For example, Zheng et al. utilises an iterative machine learning algorithm to enhance the topological design exploration of compression-only shell structures, considering structural performance

ACCELERATED DESIGN, Proceedings of the 29th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2024, Volume 1, 69-78.

and construction limitations (Zheng et al., 2020).

However, designers often value spatial quality more than other aspects of generative tools in the context of architecture design. This leads to the demand for algorithms to assess more conceptual and abstract beyond pragmatic qualities in architectural design, such as spatial qualities. In the realm of architecture, a profound understanding of spatial quality is essential, directly shaping individuals' perception, experiences, and interactions. Space, extending beyond mere functional dimensions, acts as a canvas for emotions, culture, and social elements. Thoughtful spatial design guides individuals through movement, crafting unique genius loci Spatial cognition plays a central role in architectural design, influencing how individuals comprehend and interact with their surroundings. By exploring how people perceive and engage with space, AI-generated tools have the potential to produce outcomes that are both aesthetically meaningful and practical.

2. Related Work

2.1. RULE-BASED ASSESSMENT

Previous work focused on CNN/DPM finesse to achieve the purpose (Xu et al., 2014).

2.2. MACHINE LEARNING BASED ASSESSMENT

As an alternative, several state-of-the-art generative models, such as Stable Diffusions, a Latent Diffusion model that generates images from text inputs, use CLIP, a Contrastive Language–Image Pre-training method proposed by OpenAI, as a text embedding to condition the generation (Radford et al., 2021). CLIP can evaluate the semantic similarity of given text and image pairs and perform classification tasks in a zero-shot manner. Zero-shot learning offers a promising avenue to broaden the scope of cognitive capabilities in the design process, transcending limitations tied to specific application-related problems (Larochelle et al., 2008).

However, the original CLIP model performs poorly on architecture-specific classification tasks due to its dataset's lack of prior knowledge and source of truth, causing latent diffusion models trained with the default CLIP pipeline to underperform in generation tasks in the architecture domain. Improving the prior architectural knowledge in these pre-trained models is the key to enhancing the generation quality of latent diffusion models conditioned on these CLIP models. A specific model requires task-specific modifications and training from scratch for downstream tasks; fine-tuning would be a more effective way (Howard & Ruder, 2018). The finalization could be used to classify architectural components, conduct assessments throughout the process, and even directly influence design decisions.

3. Methodology

Therefore, we propose a workflow to compile architecture-specific data and leverage it for fine-tuning the CLIP model while preserving its zero-shot capabilities.

CLIP is a multi-modal model based on contrastive learning. Unlike some contrastive learning methods in computer vision, CLIP's training data consists of text-image pairs,

where each pair consists of an image and its corresponding text description. The aim is for the model, through contrastive learning, to learn the matching relationship between these text-image pairs. CLIP comprises two models: Text Encoder and Image Encoder. The Text Encoder is utilized to extract text features and can employ the text transformer model commonly used in Natural Language Processing (NLP), while the Image Encoder is employed to extract image features and can use common Convolutional Neural Network (CNN) models or vision transformers.In this experiment, both the Text Encoder and Image Encoder used are vanilla models. A custom dataset was employed to train and fine-tune CLIP models using various prompt methods. Multiple checkpoints were fine-tuned using OpenCLIP (Cherti et al., 2023), an open-source implementation of the OpenAI CLIP model. Some checkpoints are focused solely on project tags, while others include artistic descriptions, capturing the essence and atmosphere of the creative endeavour. The impact of fine-tuning was assessed by conducting inference on image classification tasks and image-text-pair correlation tasks, comparing the results of our fine-tuned checkpoints with the original pre-trained checkpoints from OpenCLIP.

3.1. DATASET

We construct a comprehensive image-text dataset comprising over 1 million images paired with corresponding textual descriptions. This dataset is curated by sourcing image and text pairs from publicly accessible internet resources. The images exclusively consist of photographs of recently completed projects, while the text descriptions encompass project-related details. These descriptions often convey the essence of projects in an abstract or artistic manner. Additionally, the information includes project tags, specifying components present in the images, such as building materials or architectural elements.

3.2. PROMPT

As CLIP evaluates the similarity between a sentence and an image, the training of the CLIP model with a classification dataset often involves prompt engineering (Radford et al., 2021). Typically, tags or classes are formed into sentences using determiners.

In our work, we employ a diverse array of tags to describe images from various perspectives. These include:

- MediaTags: Type of graphic media, e.g., 'photograph', 'architectural drawing'.
- UtilityTags: Describing spatial utilities of the building, e.g., 'bathroom', 'courtyard'.
- ProjectTags: Indicating the program of the project, e.g., 'residential', 'public', 'office'.
- ElementTags: Detailing architectural elements shown in the image, e.g., 'stairs', 'arch', 'handrail'.

Notably, the images we collected may not encompass all tag building types. This aspect introduces certain limitations to our dataset, which we will discuss further in Sections 5.1 and 5.3.

We developed a procedure to convert these mixed tags into sentences, prioritizing them based on their significance in the image, thereby setting up image-text pairs.

For an image encompassing all four tag types, the format is as follows:

"This is a {MediaTag} of {UtilityTag} in {ProjectTag}, showing {ElementTag}."

In cases where an image lacks certain tag types, for instance, only project and element tags, the format adapts:

"This is {ProjectTag}, including details of {ElementTag}."

Additionally, if project information collected includes the design studio and description, these elements are also incorporated:

"{TagPromptedSegment}, designed by {DesignStudio}, with the following description: '{Description}'."

A few fully prompted text-image pairs are displayed below in Figure 1.

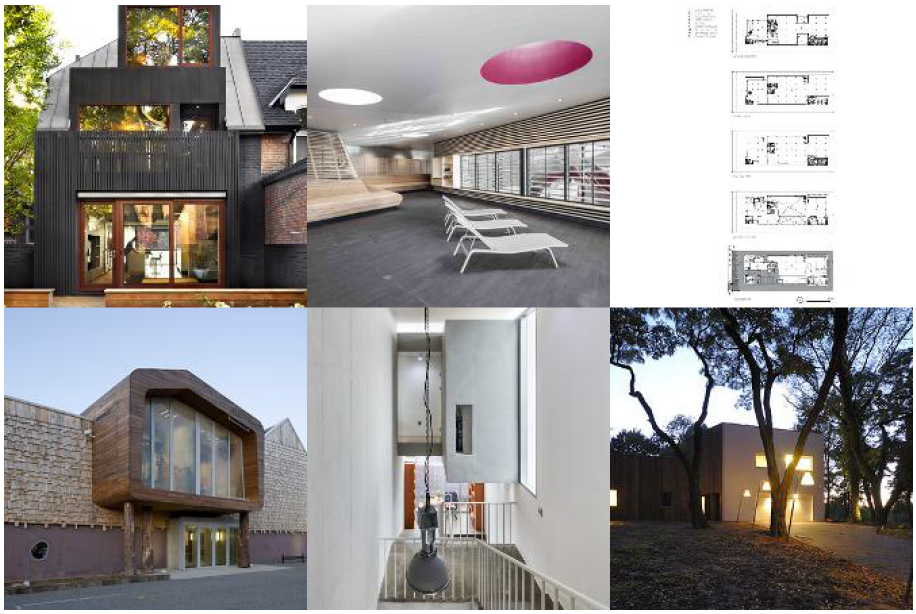

Figure 1. 6 images extracted from training dataset with their corresponding text listed below (from left to right, top to bottom). Images were upscaled from 224x244

- this is a garden in a house, include details of doors, facades, lighting, designed by plus tongtong, that has the following descriptions: "+tongtong Transforms Traditional Toronto House into Tasteful Modern Home that Honors East-end Neighborhood"
- this is an image of an architecture project, designed by auer weber, that has the following descriptions: "Sourcane is a new sports, leisure and wellness swimming hall in Douai, in the North of France. Due to its location, the aquatic centre will be an essential part of the future eco-quarter of Le Raquet, a new city district, at whose

heart will be a landscaped park. The new swimming hall lies at its northeastern end, at the interface between the artificial landscape of the park and the urban structure of the city. The project is oriented towards the future tramway station and a central urban square."

- this is an architectural drawing of a factory, designed by dp architects, that has the following descriptions: "The building housing the new headquarters of Sunray Woodcraft Construction is one of the first to be completed as part of the newly positioned International Furniture Hub in Sungei Kadut, Singapore. It presents an opportunity to look afresh at the light industrial factory type, stacking production processes in order to optimise working conditions."
- this is a museum, include details of facades, doors, designed by guinee et potin architects, that has the following descriptions: "French architectural photographer Stphane Chalmeau shared with us the Rennes Metropole Museum by french architects Guine et Potin."
- this is a kitchen in an apartment, include details of stairs, facades, handrails, lighting, designed by leau design, that has the following descriptions: "The apartment estate market which seemed not to be withered is cooling now. As the imprudent house-poor, who borrows the money to buy a house even if the interest rate is low, disappears, the fever of becoming to farming for a moment with the longing for the life in a country house is passing like a wind. And as changing the interest to the investment of the profitable real estate, the interest about a leasing profitable building of neighbourhood living facilities with an integration of a habitation house is increasing."
- this is a house, include details of facades, designed by neostudio architekci, that has the following descriptions: "This project is located on a picturesque plot that originally was a home for seed drying installation of Agricultural University - and with its magnificent Acacia trees plantation and natural splendor was a design challenge for us."

3.3. GENERAL PURPOSE DATASETS AND PRE-TRAINED MODEL SELECTION

We developed two distinct types of image-text paired datasets: the first comprising solely prompted tags, and the second encompassing both the prompted tags and project descriptions.

Each dataset underwent a split, allocating 60% for the training dataset and 40% for use as a validation/benchmark dataset. The 60% designated for training was then combined with two extensive image-text-pair datasets: LAION 400M (Schuhmann et al., 2021) and CC3M (Sharma et al., 2018). These datasets are widely recognized for their broad range of general content and are typically utilized in CLIP training.

Owing to limitations in computational resources, we incorporated only a small portion of the LAION 400M dataset and one-fifth of the CC3M dataset. This selective integration, in combination with our training dataset, was employed to fine-tune our model.

For the pre-training model, we utilized ViT-B/16 LAION400M (Release Pretrained Weights · mlfoundations/open_clip. (n.d.). GitHub. 2023), a model previously trained by OpenCLIP (Cherti et al., 2023).

Model	Process	Archiclip Dataset (ours)				Laion-400m		Cc3m	Training dataset size
		tag and descriptions		only tag					
		60%	40%	60%	40%	0.15 %	100 %	18.1 %	
ArchiCLIP-tagdcsp (ours)	finetune	~600k				~600k		~600k	~1.8m
	validation/ benchmark		~400k						
ArchiCLIP-tagonly (ours)	finetune			~600k		~600k		~600k	~1.8m
	validation/ benchmark				~400k				
ViT-B/16 LAION400M (open-clip)	training (open-clip)						~413m		~413m
	benchmark		~400k		~400k				

Table 1. Table showing the datasets we used and how they were merged before proceed with training

4. Experiments

We fine-tuned the ViT-B/16 LAION400M model on our two merged datasets for 8 epochs. The initial learning rate was set to 1e-5 with a learning rate scheduler featuring cosine decay to optimize the gradient descent.

The finetuning process was carried out on an Nvidia A100 80GB graphic card with an batch size of 600 to maximize the utilisation of GPU memory.

4.1. TRAINING

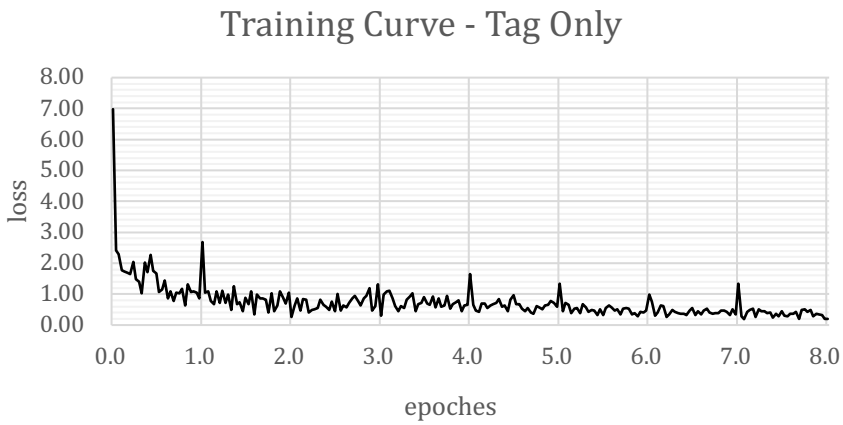

Figure 2. The training curve of tag only dataset showing CLIP loss against epoches

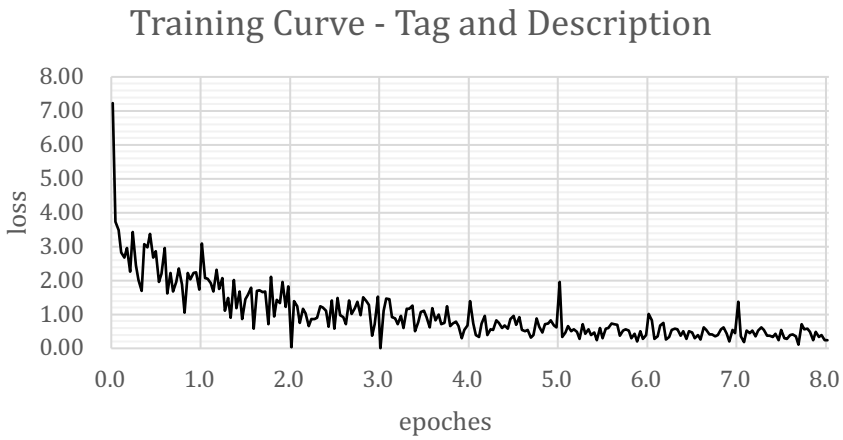

Figure 3. The training curve of tag and description dataset showing CLIP loss against epoches

Figure 2 and Figure 3 demonstrate the training curve during fine-tuning. Model already started showing traces of overfitting as the training curve flatten around 6th epoch. We believe that this is partially due to the size of our dataset, further discussed in 5.1

4.2. EVALUATION

Due to our dataset not being a single-label classification dataset, as mentioned earlier in section 3.2, the common benchmarks for CLIP models, which typically evaluate zero-shot classification tasks, cannot be applied to ArchiCLIP.

Therefore, we developed our own benchmark. For each image, this benchmark calculates a CLIP correlation against all possible tags. The tags with the highest

correlation scores are selected for comparison with the ground truth in the dataset. The number of tags chosen matches the count of tags in the ground truth. For each correct tag identified, the model earns a score equal to the reciprocal of the number of ground truth tags in that corresponding tag category.

$$S_{tagtype} = score\ of\ model\ achieved\ in\ evaluating\ \{tagtype\}\ tags$$

$$N_{tagtype} = counts\ of\ images\ that\ have\ at\ least\ one\ \{tagtype\}\ tags$$

$$n_{i,tagtype} = counts\ of\ \{tagtype\}\ tags\ belong\ to\ image\ \{i\}$$

$$c_{i,tagtype} = correct\ \{tagtype\}\ predictions\ belong\ to\ image\ \{i\}$$

$$S_{tagtype} = \frac{1}{N_{tagtype}} \sum_{i=1}^{N_{tagtype}} \frac{c_{i,tagtype}}{n_{i,tagtype}}$$

Checkpoint	epoch	MediaTag	UtilityTag	ProjectTag	ElementTag
ViT-B-16 (open-clip)	n/a	0.620	0.450	0.245	0.467
ArchiCLIP description (ours)	5	0.752	0.698	0.135	0.644
	6	0.510	0.677	0.090	0.559
	7	0.508	0.640	0.141	0.627
	8	0.610	0.679	0.116	0.575
ArchiCLIP tagonly (ours)	5	0.575	0.763	0.204	0.690
	6	0.563	0.762	0.240	0.707
	7	0.573	0.742	0.231	0.735
	8	0.541	0.783	0.275	0.707

5. Limitations and Future Improvements

When evaluating our finetuned model, we identified several limitations affecting its performance.

5.1. DATASET

As mentioned in section 3.2, not all images in our dataset contain all four types of tags. Upon further manual examination, we found that in many cases, the tags sourced from the internet are a subset of what is depicted in the images. For instance, an image may feature a 'staircase' that is not reflected in the tags. This discrepancy likely arises from sourcing images from the internet, where the information on webpages may not fully capture all elements in the image.

This inconsistency impacts the model's performance. The loss function of the CLIP model is defined as the difference between the cosine similarity matrix and the identity matrix (Radford et al., 2021). Text that fails to fully describe the image contributes to loss, even if the model's prediction is accurate, thus making the dataset less reliable.

Moreover, most CLIP models are trained on much larger datasets with significant computational power. Studies have shown that the performance of CLIP models generally improves with larger datasets (Gadre et al., 2023). For comparison, the pre-trained model we used for fine-tuning, ViT-B-16 LAION400M, was initially trained on the Laion400M dataset, comprising approximately 413 million data points (Cherti et al., 2023). Our dataset, combined with a portion of Laion400M and CC3M, only reached 1.8 million data points, less than 0.5% of the initial pre-trained model's size. This substantial difference contributes to the performance loss.

5.2. BENCH MARK AND EVALUATION

CLIP models are generally adept at evaluating correlations between objects in an image and descriptive text. However, their ability to assess the correlation between an image and abstract descriptions remains undetermined. To our knowledge, no existing research focuses on evaluating these correlations in the field of architecture. This lack of precedent makes it challenging to gauge our model's success in this area.

5.3. AREAS OF IMPROVEMENTS

We believe that with targeted improvements, the model's performance can be further enhanced.

- Manual Labeling: With sufficient architects involved, we could more accurately relabel our dataset, thereby improving its accuracy and optimizing tag comprehensiveness and consistency. This addresses the limitation mentioned in 5.1
- Human Benchmark: Engaging a sufficient number of architects, we could establish a human benchmark for assessing the correlation between images and abstract descriptions. This would provide deeper insights into standard performance for such image-text tasks, addressing the limitation mentioned in 5.2
- Classification Tasks: One of the primary research areas in computer vision is image classification. The CLIP model, a sub-field of this area, has made significant strides in processing natural language and in zero-shot learning. Nonetheless, due to the unique format of its training dataset, certain methods that improve image classification accuracy are not applicable for enhancing the accuracy of CLIP's image-text pair matching tasks. By reformating our dataset, we might transform the training dataset into multiple image classification datasets and apply existing methods (Wortsman et al., 2022) from related fields to optimize the model's performance

6. Conclusion

Compared to the pre-trained checkpoints, our fine-tuned model performs better in architecture-specific image classification tasks and demonstrates some capabilities in relating abstract and artistic descriptions with architectural photographs. In a practical context, the post-optimized CLIP model can find application in various ways. For example, it can train diffusion models, serve as a guiding discriminator for other generative models, and even be harnessed to develop a specialised semantic-based

searching engine explicitly tailored for architectural purposes. It is foreseeable that ArchiCLIP will play a critical role in AI-generated process. Limited by our resources, our dataset has yet to undergo review nor filtering by human designers, leaving room for potential improvements in dataset qualities.

References

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., ... & Jitsev, J. (2023). Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2818-2829).

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint *arXiv preprint arXiv:1801.06146*.

Larochelle, H., Erhan, D., & Bengio, Y. (2008, July). Zero-data learning of new tasks. In *AAAI* (Vol. 1, No. 2, p. 3).

Phelan, N., Davis, D., & Anderson, C. (2017, May). Evaluating architectural layouts with neural networks. In *Proceedings of the Symposium on Simulation for Architecture and Urban Design* (pp. 1-7).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

Xu, Z., Tao, D., Zhang, Y., Wu, J., & Tsoi, A. C. (2014). Architectural style classification using multinomial latent logistic regression. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13* (pp. 600-615). Springer International Publishing.

Zheng, H., Moosavi, V., & Akbarzadeh, M. (2020). Machine learning assisted evaluations in structural design and construction. *Automation in Construction*, 119, 103346.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., & Komatsuzaki, A. (2021). LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *ArXiv:2111.02114* [Cs]. https://arxiv.org/abs/2111.02114

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). https://doi.org/10.18653/v1/p18-1238

Release Pretrained Weights · mlfoundations/open_clip. (n.d.). GitHub. Retrieved September 2023, from https://github.com/mlfoundations/open_clip/releases/tag/v0.2-weights

Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., Orgad, E., Entezari, R., Daras, G., Pratt, S., Ramanujan, V., Bitton, Y., Marathe, K., Mussmann, S., Vencu, R., & Cherti, M. (2023, October 20). DataComp: In search of the next generation of multimodal datasets. *ArXiv.org.* https://doi.org/10.48550/arXiv.2304.14108

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Gontijo-Lopes, R., Hajishirzi, H., Farhadi, A., Namkoong, H., & Schmidt, L. (2022). Robust fine-tuning of zero-shot models. *ArXiv:2109.01903* [Cs]. https://arxiv.org/abs/2109.01903