Visual prompt tuning and ensemble undersampling for one-shot vehicle classification

Jan Erik van Woerden, Gertjan J. Burghouts, Sabina B. van Rooij, Frank Ruis, Judith Dijk, Hugo J. Kuijf
TNO, the Hague, the Netherlands

ABSTRACT

Vision-language foundation models for image classification, such as CLIP, suffer from a poor performance when applied to images of objects dissimilar to the training data. A relevant example of such a mismatch can be observed when classifying military vehicles. In this work, we investigate techniques to extend the capabilities of CLIP for this application. Our contribution is twofold: (a) we study various techniques to extend CLIP with knowledge on military vehicles and (b) we propose a two-stage approach to classify novel vehicles based on only one example image.

Our dataset consists of 13 military vehicle classes, with 50 images per class. Various techniques to extend CLIP with knowledge on military vehicles were studied, including: context optimization (CoOp), vision-language prompting (VLP), and visual prompt tuning (VPT); of which VPT was selected. Next, we studied one-shot learning approaches to have the extended CLIP classify novel vehicle classes based on only one image. The resulting two-stage ensemble approach was used in a number of leave-one-group-out experiments to demonstrate performance.

Results show that, by default, CLIP has a zero-shot classification performance of 48% for military vehicles. This can be improved to >80% by fine-tuning with example data, at the cost of losing the ability to classify novel (previously unseen) military vehicle types. A naive one-shot approach results in a classification performance of 19%, whereas our proposed one-shot approach achieves 70% for novel military vehicle classes.

In conclusion, our proposed two-stage approach can extend CLIP for military vehicle classification. In the first stage, CLIP is provided with knowledge on military vehicles using domain adaptation with VPT. In the second stage, this knowledge can be leveraged for previously unseen military vehicle classes in a one-shot setting.

Keywords: machine learning, classification, artificial intelligence, military vehicles, CLIP, visual prompt tuning, vision-language foundation models, EDF FaRADAI

1. INTRODUCTION

Vision-language models (VLM) are a type of artificial intelligence technology that fuses both vision and natural language foundation models. A popular example of such a vision-language foundation model is Contrastive Language-Image Pre-Training (CLIP), which has been trained on a wide variety of images and corresponding text available on the internet [1]. CLIP is known to achieve good zero-shot performance, that is: when prompted with text, it can identify concepts or objects in images that were not part of the training data.

Unfortunately, the zero-shot performance of CLIP deteriorates for data that is out-of-distribution and not similar to its training data. Handwritten digit recognition is such an example provided by the original authors [1]. Our empirical research shows that a similar mismatch, and consequently poor zero-shot performance, can be observed when classifying military vehicles.

Various approaches exist to extend the capabilities of CLIP and thereby improve its performance on datasets where it has poor zero-shot performance. Notable examples are linear probing [1], context optimization (CoOp) [2], visual prompt tuning (VPT) [3], and vision-language pre-training (or prompting) (VLP) [4]. Linear probing only employs the image encoder from CLIP, on top of which a naive linear classifier is added. CoOp maintains the original classification framework, freezes all parameters, and optimizes the textual context that is represented as input tokens. A commonly used initial context is the phrase: "This is an image of {class}", which is then further optimized. VPT also freezes the entire

Artificial Intelligence for Security and Defence Applications II, edited by Henri Bouma, Radhakrishna Prabhu, Yitzhak Yitzhaky, Hugo J. Kuijf, Proc. of SPIE Vol. 13206, 132060G · © 2024 SPIE
0277-786X · doi: 10.1117/12.3029517

model and introduces additional learnable tokens as input to each transformer layer of the image encoder, resulting in an efficient and competitive alternative to full finetuning. VLP extends the Visual Prompt Tuning to the textual encoder as well.

Besides providing CLIP with training images to extend its capabilities, its language component and the used prompts can also be altered to increase performance. This is known as prompt engineering and is common practice in vision-language models. When prompting the text encoder of a VLM, such as CLIP, there is a distinction between "hard prompts" (manually crafted prompts) and "soft prompts" (prompts that dynamically searched for) [5]. We opt for a hybrid approach, by asking ChatGPT to provide a suitable prompt for military vehicle classes [6].

For the task of military vehicle classification, training data is often scarce and this hampers re-training or fine-tuning of classification models. Although a larger collection of training images with military vehicles might be available, e.g. for use in extending CLIP, specific examples of exact military vehicle types may be limited to only one training image. Simply training or fine-tuning CLIP with all available images results in a class imbalance problem, where the majority class dominates the learning process and consequently skews the classification results. A strategy to circumvent this issue is the design of an undersampled ensemble, which has proven effective in other domains [7] [8] [9]. Multiple weak classifiers are trained on a balanced selection of all training data, which are then combined in an ensemble to create the final classifier.

In this work, we extend the capabilities of CLIP with limited training data and investigate a one-shot approach to recognize a previously unseen type of military vehicles. Our contribution is twofold: (a) various techniques will be evaluated to extend the capabilities of CLIP with knowledge on military vehicles both in the visual domain (CoOP, VLP, VPT) and language domain (ChatGPT), and (b) a two-stage approach based on an undersampled ensemble will be used to classify novel military vehicle types based on one training image.

2. MATERIALS AND METHODS

Data

For this work, two datasets were used: one was collected from public sources and contained videos of 13 different military vehicles and the other was recorded during the European Defence Agency (EDA) CAT B project Detect Before Launch (DEBELA) [10].

The first dataset, was compiled from public sources and consisted of 13 different classes of military vehicles: Boxer, BRDM-2, BTR-80, Fennek, Fuchs, Leopard, M109, M1A2, MSTA, Patria, Panzerhaubitze 2000, T90, and a military truck. Some example images are shown in Figure 1. For each class, approximately 50 images were collected; of which randomly 16 were placed in a training set and the remainder in a separate test set. In the one-shot experiments, only one image from the training set will be used.

The second dataset was recorded during the DEBELA-trial at WTD-52, Oberjettenberg, Germany; in June 2023. At this trial, four vehicle types were present: a BMW car, a Mercedes van, an electric UTV (utility terrain vehicle), and a STANDCAM (Standard Decoy for Camouflage Materials). Examples are shown in Figure 2. For each class, between 50 and 100 images were taken from the video recordings; of which randomly 16 were placed in a training set and the remainder in a separate test set. In the one-shot experiments, only one image will be used.

Methods

A two-stage approach was implemented to first extend CLIP with knowledge on military vehicles and, second, to classify novel military vehicle types based on only one training image. For the first stage of extending the capabilities of CLIP, various strategies both in the visual and language domains were investigated: linear probing, CoOp, VLP, and VPT. Additionally, the effect of improved prompts obtained from ChatGPT was investigated. Based on these results, the best performing method to extend CLIP was selected. In the second stage, the extended CLIP was fine-tuned in an undersampled ensemble strategy to obtain improved one-shot performance.

Extending CLIP

Figure 3 shows the default, zero-shot classification setup of CLIP, where N encoded class descriptions are compared to the encoded image using a cosine similarity. The ViT-B/16 variant of CLIP was used for all experiments, which combines a BERT-variant text encoder and a ViT image encoder. CLIP was extended using various strategies: linear probing, CoOp, VLP, and VPT.

Figure 1: Example images of various military vehicles obtained from public sources. Top row: a Panzerhaubitze 2000 (left image credit: John van den Boogaart, defensiefotografie.nl; right image credit: Netherlands Ministry of Defence); bottom left: a Boxer (image credit: Martin Bos, defensiefotografie.nl); bottom right: a Fennek.

Figure 2: Examples of the different vehicle types present at the DEBELA trial. Left to right: a BMW car, an electric UTV, a STANDCAM, and a Mercedes van.

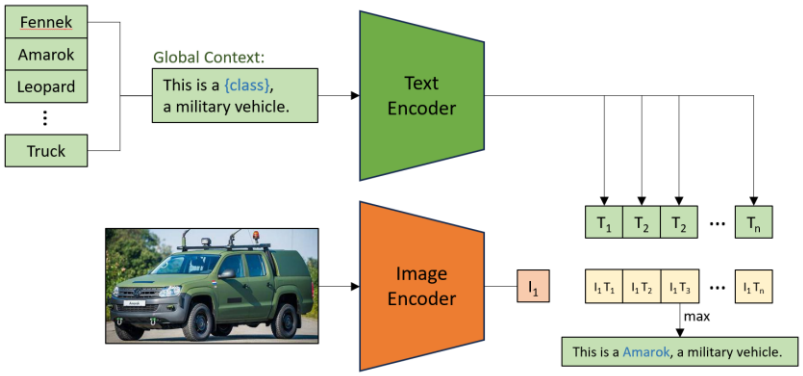

Figure 3: Classifying images with CLIP. On top, multiple class descriptions are inserted into the global context and fed through the text encoders. On bottom, the input image is fed through the image encoder. The encoded texts and image are compared with a cosine similarity and the best match is selected as the final class.

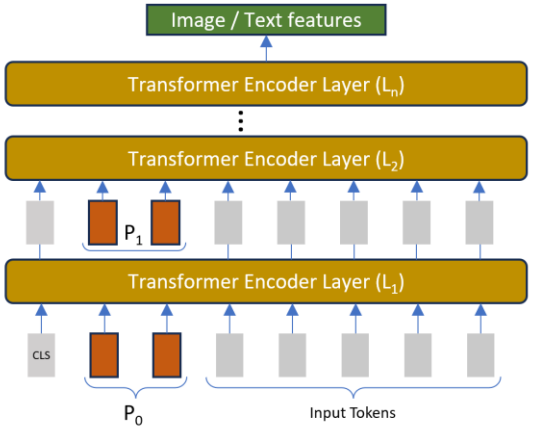

Figure 4: For both VPT and VLP, all layers of the CLIP model are frozen. New randomly initialized tokens (P) are added to the image encoder (VPT) or both the text and image encoders (VLP). Only these new tokens P are trainable.

For linear probing, we only use the image encoder of CLIP and add a linear projection layer. The method proposed by [1] was used to train the linear probe model.

For CoOp, the method proposed in the original paper was used [2]. All weights were frozen and only the initialized global context tokens were trainable. The global context was initialized with the prompt: “this is a {class}, a military vehicle.”

For VPT, both the global context and the weights were frozen and only the image encoder was trainable. For each transformer layer in the image encoder, five tokens were added that were randomly initialized. This is how it is described in the original paper on VPT [3], where each trainable token is initialized after the CLS token and before the input tokens. This is summarized in Figure 4.

VLP extends this approach to both the text and image encoders. It additionally makes the global context trainable.

Training

For training we use the following setup for all models. When training with two stages, we train the first stage with 150 epochs and the second stage with 100 epochs. In a one-stage training, we train similarly to the first stage of the two-stage setup. Each stage is trained using SGD as optimizer, using 10 epochs of warmup, and a cosine LR scheduler that starts at 0.01. A batch size of 32 is used. As augmentations we use random cropping and flipping. All augmented images are resized to 224x224 before being fed into the CLIP architecture.

During inference, logits per class are created based on the cosine similarity between the image and the text (see Figure X). Instead of applying a contrastive loss as in CLIP pre-training [1], the optimization is done using a standard Cross Entropy on the cosine similarity logits.

Improved prompts

The prompt "a photo of a [X]" was used by default, where [X] was replaced with one of the vehicle class names listed in the Data section (e.g. "a photo of a M1A2"). To provide more context, the phrase "a military vehicle" was added to the prompts (e.g. "a photo of a M1A2, a military vehicle"). Finally, ChatGPT was asked to provide more elaborate prompts suitable for CLIP and these were added to the default prompt (e.g. "a photo of a M1A2, American main battle tank with a 120mm gun. Crew of 4. Size: 9x4x2m, a military vehicle").

One-shot capabilities

To assess its one-shot capabilities, the extended CLIP model was trained with a single image of a previously unseen type of military vehicle. To address the class imbalance between the first stage (16 images per class to extend CLIP) and the second stage (one-shot), an undersampled ensemble strategy was used. Here, the model was fine-tuned using one image from each of the existing classes and the one image of the new class. This approach was repeated sixteen times, each time using a different image from the existing classes and the same image of the new class. The sixteen resulting models were combined in an ensemble. The average of the different predicted cosine similarities was used to determine the final prediction.

Experiments

First, the performance of different strategies to extend the capabilities of CLIP will be compared on the dataset of military vehicles, including a baseline experiment that shows CLIP's zero-shot performance. All methods to extend CLIP will be provided with 16 images per class (the training set) and evaluated on the test set. Recall (mean ± sd) will be used as the evaluation metric on classification performance. The best performing strategy will be selected for further experiments.

Second, the effect of improved prompts will be assessed on the test set. This will be evaluated using a confusion matrix, macro-recall, and macro-F1. The best performing strategy for prompts will be selected for further experiments.

Third, different strategies will be evaluated to study the one-shot capabilities of CLIP, i.e.: classifying a previously unseen type of military vehicle based on one training image. A number of approaches will be evaluated, including baseline (zero-shot) approaches, one-stage and two-stage approaches:

- Baseline approaches
    1. Zero-shot performance of CLIP, without any retraining
    2. Zero-shot performance of CLIP, with retraining on 12 of the 13 classes
- One-stage one-shot approaches
    3. Retraining CLIP with the 16 training images for each of 12 classes, and 1 training image of the 13th class added 16 times (to counter the imbalance)
    4. Retraining CLIP with 1 training image for each of all 13 classes
    5. Retraining CLIP in an undersampled ensemble. A total of 16 models will be included in the ensemble, each trained with 1 selected training image for each of 12 classes and the same image of the 13th class.
- Two-stage one-shot approaches
    6. Retraining CLIP on 12 of the 13 classes (approach 2), followed by retraining CLIP with 1 training image for each of all 13 classes (approach 4)
    7. Retraining CLIP on 12 of the 13 classes (approach 2), followed by retraining that model 16 times in an undersampled ensemble (approach 5).
    8. The same as approach 7, but when it predicts a class that is not the novel class, the model of approach 2 is used instead.

For all approaches, the reported metric is the recall (mean ± sd) for classification. This will be reported separately for the baseline classes (the 12 classes for which 16 training images were available) and the novel class (the 13th class for which 1 training image was available). All results are averaged over a leave-one-class-out cross-validation.

Finally, the experiments will be repeated on the dataset acquired at the DEBELA-trial.

3. RESULTS

The results obtained on the dataset of 13 different classes of military vehicles obtained from public sources are provided first. Zero-shot performance of CLIP achieves a recall of 21% when evaluated on the test set. This can be improved with linear probing (57%), CoOp (64%), VPT (76%), and VLP (75%). VLP is a combination of CoOp and VPT, but achieves only 75% recall; possibly because the larger number of trainable parameters (text and image) degrades performance. Therefore, VPT was selected for the remainder of the experiments.

The improved prompts obtained from ChatGPT are:

Boxer: German 8x8 armored vehicle for transport and combat. Crew of 3-8. Size: 8x2x2m.
BRDM: Russian 4x4 armored reconnaissance vehicle. Crew of 3. Size: 6x2x2m.
BTR: Russian 8x8 armored vehicle for transport and combat. Crew of 3-10. Size: 8x3x2m.
Fennek: German 4x4 armored reconnaissance vehicle for recon. Crew of 3. Size: 6x2x2m.
Fuchs: German 6x6 armored vehicle for transport and recon. Crew of 9. Size: 6x2x2m.
Leopard: German main battle tank with a 120mm gun. Crew of 4. Size: 10x4x3m.
M109: American self-propelled howitzer with 155mm gun. Crew of 6. Size: 10x3x3m.
M1A2: American main battle tank with a 120mm gun. Crew of 4. Size: 9x4x2m.
MSTA: Russian self-propelled howitzer with 152mm gun. Crew of 5. Size: 12x3x3m.
Patria: Finnish 8x8 armored vehicle for transport and combat. Crew of 3-9. Size: 8x3x2m.
PzH 2000: German self-propelled howitzer with 155mm gun. Crew of 5. Size: 11x4x3m.
T90: Russian main battle tank with a 125mm gun. Crew of 3. Size: 10x4x2m.
Truck: American 6x6 vehicle for transport and logistics. Crew of 2-6. Size: 7x2x2m.

Results of the improved prompts are shown in Figure 5 and Table 1. The confusion matrix suggests that, initially, the model is able to roughly distinguish between American (M109) and Russian (BRT-80) military vehicle types. This suggests that these classes may be overrepresented in the original text encoder of CLIP. After providing the refined prompts, model performance increases from 24% (baseline) to 48% (ChatGPT descriptions).

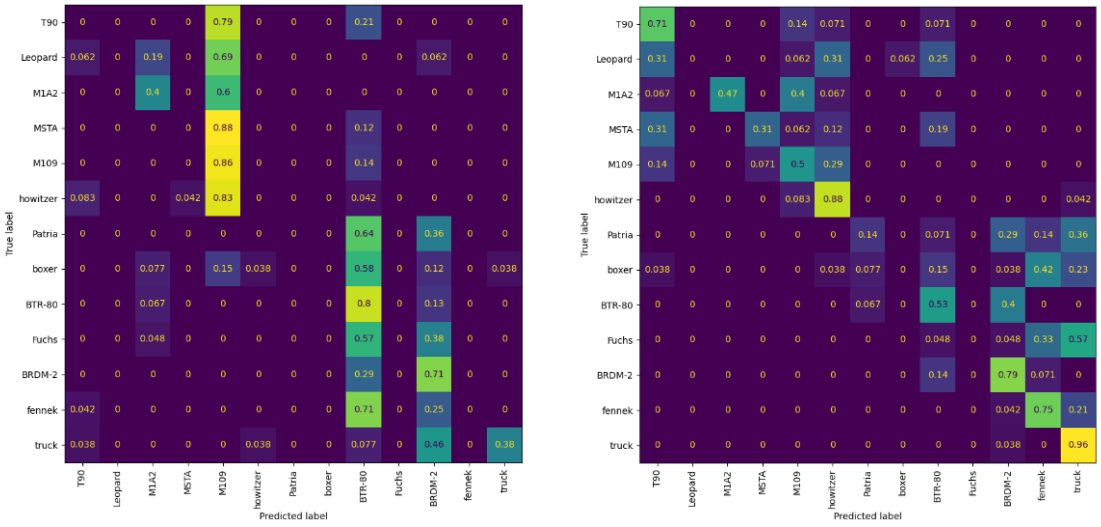

Figure 5: confusion matrices on military vehicle types before (left) and after (right) providing improved prompts generated by ChatGPT.

Table 1: results of providing improved prompts to our extended CLIP model. For example: a baseline prompt would be "a photo of a M1A2", a tuned context prompt would be "a photo of a M1A2, a military vehicle", and the ChatGPT provided description prompt would be "a photo of a M1A2, *American main battle tank with a 120mm gun. Crew of 4. Size: 9x4x2m,* a military vehicle" (ChatGPT part in *italics*).

	Macro Recall	**Macro F1**
Baseline	24.3%	13.9%
Tuned context	32.0%	28.1%
ChatGPT descriptions	47.7%	41.1%

Table 2: Results of the various experiments to assess the one-shot capabilities of our extended CLIP model. The column "novel classes" reports the recall (mean ± sd) for the full test set of the previously unseen class that is added in the one-shot setting. The column "base classes" reports the performance of the test sets on the classes with which the CLIP model was originally extended. This is repeated in a leave-one-class-out cross-validation.

Approach	**Recall (mean ± sd)**	
Baseline	**Novel classes**	**Base classes**
1) zero-shot, no retraining	47.7% ± 33.4%	47.7% ± 2.8%
2) zero-shot, with retraining	3.5% ± 11.2%	82.4% ± 2.6%
One-stage one-shot		
3) retraining, 16 images	19.2% ± 22.1%	81.5% ± 2.1%
4) retraining, 1 image	48.5% ± 17.2%	47.5% ± 4.7%
5) retraining undersampled ensemble	64.1% ± 20.3%	62.7% ± 3.0%
Two-stage one-shot		
6) (2) + (4)	56.6% ± 15.1%	65.8% ± 5.1%
7) (2) + (5)	69.6% ± 14.9%	78.3% ± 3.2%
8) (7) followed by (2) for base class	69.6% ± 14.9%	80.1% ± 2.8%

Results of the one-shot capabilities are provided in Table 2. Zero-shot, CLIP is able to achieve a classification performance of 48% for military vehicle types (approach 1; with the ChatGPT prompts). By retraining CLIP with images of military vehicles (approach 2), its performance increases to 82%, but it completely loses its zero-shot abilities and novel/unseen classes cannot be accurately classified anymore. A one-stage one-shot approach can improve the performance for novel classes when providing a single training image (approach 4), where our undersampled ensemble approach achieves good performance on novel classes (recall of 64%; approach 5), but loses some performance on the base classes (63%). Our proposed two-stage undersampled ensemble approach (approach 7) is able to further improve the one-shot performance on novel classes (recall 70%) as well as the performance on the base classes (recall of 78%). The slight drop in performance of the base classes is likely because the additional novel class increases the possibility for confusion between classes.

DEBELA-recordings

Results of the one-shot capabilities for the DEBELA-recordings are provided in Table 3. Zero-shot, CLIP achieves a classification accuracy of 80% for the four classes. Extending CLIP with images from the vehicles present in the DEBELA-recordings improves the accuracy to 93%, but thereby again losing the ability to accurately classify unseen/novel classes. These zero-shot results with retraining are not as bad as for the military vehicles (Table X), likely because the car and van have quite good performance regardless. A one-stage one-shot approach (approach 4) improves the accuracy to well over 80%, both for the unseen/novel classes and the base classes. Here, the undersampled ensemble approach (approach 5) performs best for both the novel classes (86%) and the base classes (88%). Our proposed two-stage undersampled ensemble approach (approach 7) improves this slightly with respect to the one-stage approach.

The results from the DEBELA-recordings show a slightly different trend than the military vehicles dataset, where the performance of the novel classes (slightly) outperforms the performance of the base classes. This seems counter-intuitive, since the novel classes are only trained with a single image as compared to the base classes that are trained with sixteen images. This might be caused by having fewer classes (only four vehicle types) in total, for some of which CLIP already has good baseline performance, and having one or two outliers impacts performance more.

Table 3: Results of the various experiments on the dataset from the DEBELA-recordings. The column "novel classes" reports the recall (mean ± sd) for the full test set of the previously unseen class that is added in the one-shot setting. The column "base classes" reports the performance of the test sets on the classes with which the CLIP model was originally extended. This is repeated in a leave-one-class-out cross-validation.

Approach (DEBELA)	**Recall (mean ± sd)**	
Baseline	**Novel classes**	**Base classes**
1) zero-shot, no retraining	79.5% ± 17.7%	79.5% ± 17.7%
2) zero-shot, with retraining	32.7% ± 22.2%	92.9% ± 2.9%
One-stage one-shot		
3) retraining, 16 images	69.9% ± 21.0%	92.8% ± 3.3%
4) retraining, 1 image	83.2% ± 14.4%	84.7% ± 7.9%
5) retraining undersampled ensemble	86.1% ± 14.5%	87.8% ± 4.0%
Two-stage one-shot		
6) (2) + (4)	85.0% ± 15.3%	81.6% ± 15.3%
7) (2) + (5)	88.6% ± 12.6%	86.9% ± 5.4%
8) (7) followed by (2) for base class	88.6% ± 12.6%	89.0% ± 5.0%

4. CONCLUSION

Our proposed two-stage undersampled ensemble approach is able to achieve good performance in one-shot classification of novel military vehicles. In the first stage, CLIP is provided with general knowledge on military vehicles, for which it only has moderate zero-shot performance. We used visual prompt tuning (VPT) to provide visual features from images of military vehicles; and obtained improved textual prompts from ChatGPT with the overall objective to improve military vehicle classification with CLIP. In the second stage, an undersampled ensemble is used to counteract the common data imbalance in one-shot learning for novel classes.

ACKNOWLEDGEMENTS

This work was supported by the European Commission under the European Defence Fund project "Frugal and Robust AI for Defence Advanced Intelligence" (FaRADAI) № 101103386, and the European Defence Agency (EDA) CAT B project Detect Before Launch (DEBELA).

REFERENCES

[1] A. Radford, J. Wook Kim, C. Hallacy, A. Ramesh, G. Gob, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *International conference on machine learning*, 2021.

[2] K. Zhou, J. Yang, C. C. Loy and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision,* vol. 130, no. 9, pp. 2337-2348, 2022.

[3] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision*, 2022.

[4] Y. Xing, Q. Wu, D. Cheng, S. Zhang, G. Liang, P. Wang and Y. Zhang, "Dual Modality Prompt Tuning for Vision-Language Pre-Trained Model," 2023.

[5] J. Gu, Z. Han, S. Chen, A. Beirami, B. He, G. Zhang, R. Liao, Y. Qin, V. Tresp and P. Torr, "A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models," arXiv preprint arXiv:2307.12980, 2023.

[6] OpenAI, "ChatGPT v3.5," 2023.

[7] X.-Y. Liu, J. Wu and Z.-H. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 39, no. 2, pp. 539-550, 2009.

[8] L. Liu, X. Wu, S. Li, Y. Li, S. Tan and Y. Bai, "Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection," *BMC Medical Informatics and Decision Making,* vol. 22, no. 1, p. 82, 2022.

[9] P. Ksieniewicz, "Undersampled Majority Class Ensemble for highly imbalanced," in *Second International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 2018.

[10] C. Eisele, D. P. Seiffer, E. Sucher, L. Sjöqvist, M. Henriksson, C. Lavigne, R. Domel, P. Déliot, J. Dijk, H. Kuijf and N. Boehrer, "DEBELA – Investigations on potential Detect before Launch technologies," in *SPIE Sensors+Imaging, Electro-optical and Infrared Systems: Technology and Applications XXI*, Edinburgh, 2024.