

Zero-Shot Multimodal Deep Learning Models for Military Vehicle Detection – An Analysis

Philipp J. Rösch, Fabian Deuser, Konrad Habel and Norbert Oswald

University of the Bundeswehr Munich
Werner-Heisenberg-Weg 39
85577 Neubiberg
GERMANY

philipp.roesch@unibw.de, fabian.deuser@unibw.de, konrad.habel@unibw.de, norbert.oswald@unibw.de

ABSTRACT

Cognitive superiority using artificial intelligence aims to extract relevant information from a huge amount of data to create military and non-military situational awareness. Reliable and timely interpretations of visual information are contributing factors to gain such superiority. With the rise of large-scale, multimodal deep learning models like Contrastive Language-Image Pre-training (CLIP), a promising type of neural network is emerging to perform such visual recognition tasks. This kind of network is able to extract knowledge from visual input by applying Optical Character Recognition (OCR), facial recognition, or object classification at once and without being explicitly fine-tuned. This zero-shot capability of CLIP is enabled by the choice of specific text prompts targeting the searched object within an image.

In this paper, we investigate how CLIP can be used to identify vehicles in the military domain and use lessons learned from the Ukraine-Russia war. For analysis, a new dataset was created containing images with military and civilian vehicles, but also images without vehicles. First, we search for appropriate queries to leverage single search results and then ensemble multiple prompts. Second, we explore whether this approach can be used to identify military vehicles from video streams based on surveillance cameras and smartphones. We show on our image dataset that with thoughtful prompt engineering, the CLIP model is able to identify military vehicles with high precision and recall. The performance of the video dataset depends on object size and video quality. With this approach, allies, as well as hostile parties, can systematically analyze large amounts of video and image data, without time-consuming data collection and training.

1.0 INTRODUCTION

In recent years, deep neural networks, such as AlexNet [1] and ResNet [2], have achieved outstanding performance on image classification benchmarks like ImageNet [3]. Image classification is mostly concerned with the presence of a specific object in an image. For each data sample, an image and its label is available. These samples are used to update the model parameters. This type of learning paradigm is called “supervised learning.”

Deep neural networks enable wide-ranging applications for robotics, self-driving cars, traffic monitoring, and pedestrian recognition. However, a great amount of data has to be collected and manually annotated for each new use case. This is not only expensive and time consuming, but also problematic if the data is classified and may not be passed to third parties. Furthermore, in supervised deep learning, the networks are typically restricted to predict a certain number of classes. For example, the ImageNet classification dataset has 1,000 classes and the MS COCO [4] dataset has only 80 classes. To learn new concepts not embodied in these datasets, additional annotations have to be added.

One possible solution for these issues is zero-shot learning. Larochelle et al. [5] defined zero-shot learning as a training process where only descriptions of the novel classes are present during the training. This is typically achieved by grouping images and related texts as pairs. Neural networks are then used to find common

representations between these modalities [6], [7], [8], [9]. This joint learning embeds information from the text into the representation of the image, and vice versa. In traditional supervised classification, the output of the neural network is fixed and only probabilities according to the learned classes are predicted. In contrast, zero-shot learning uses flexible textual descriptions of an arbitrary number of classes that minimize the distance between text and image representation. These textual descriptions, also known as text prompts, must be carefully chosen during inference to represent a specific class. This type of research, called “prompt engineering,” is currently a very active field of research [10], [11].

Modern zero-shot approaches, such as CLIP [10] or ALIGN [11], use large-scale datasets to cover as many concepts as possible. In doing so, they show wide-ranging capabilities on many datasets, which implies a high degree of generalization. In both approaches, an image encoder and a text encoder are trained together to create vector embeddings that minimize the distance between the text and the image representation.

In this work, we analyze the capabilities on the basis of military-related data without additional training on them. To evaluate these skills, we collect two datasets:

- An image dataset with military vehicles, civilian vehicles, and without vehicles to find appropriate prompts for each class. This image dataset is used for additional analysis to understand which text prompt syntax is important for powerful zero-shot classification. See details in Section 2.4.
- A video dataset to examine the found prompts, for different in-the-wild settings, such as dash cameras, traffic surveillance, and mobile devices in context of the Ukraine war. See Section 2.5.

The settings in our experiments refer to the present situation in Ukraine. Here, surveillance and public webcams were used to show the invasion of Russian tanks and logistics into Ukrainian territory in the media. However, cognition of Ukrainian or Western logistics would also be possible. An enemy actor can use open-source data to obtain situational awareness of movement of troops, logistics, or humanitarian aid. To hinder this evaluation, Germany’s public traffic surveillance cameras have been switched off [12]. Our results show that a malicious actor can use CLIP to screen masses of traffic surveillance data and videos from public sources for reconnaissance purposes. For this type of solution, neither time and labor-intensive data collection and labeling is needed, nor do computationally expensive models need to be trained.

2.0 MODEL AND DATASETS

In the following, we describe the used zero-shot model and our metrics for experiments and discussion. Moreover, we give a description of the two datasets used in our analysis.

2.1 CLIP Model

One of the best zero-shot models currently available is CLIP. Radford et al. [10] developed a completely new way to learn as many concepts as possible using a simple contrastive pre-training objective. CLIP is pre-trained on 400 million image-text pairs. However, the dataset is not publicly available and therefore no details about the training data are known. The images are embedded by an image encoder and the texts by a separate text encoder. The objective is to reduce the distance of the embeddings using a symmetric cross-entropy loss, depicted in Figure 1 (left). The cosine similarity is used as the distance metric. Based on this simple pre-training objective, CLIP learns general concepts without supervised annotations and thus enables strong zero-shot capabilities. Both ResNet [2] with various improvements [13], [14] and Vision Transformer [15] are used as image encoders, and the Transformer architecture [16] is used for the text embeddings. Radford et al. provides nine different configurations of their CLIP model. For our analyses, we use ViT-B/16, a mid-sized model with 86.2 million and 37.8 million parameters for the image and text encoder, respectively. To prevent overfitting, several data augmentations are normally used, but these can be neglected due to the size of the pre-training dataset, and only simple cropping takes place. The pre-training dataset is not public, therefore the amount of

military-related data during training is not known. During inference, the searched classes are encoded using different prompts (T_1, \dots, T_N) and then the class is determined based on the distance between the text vectors and the image vector (I_1), as shown in Figure 1.

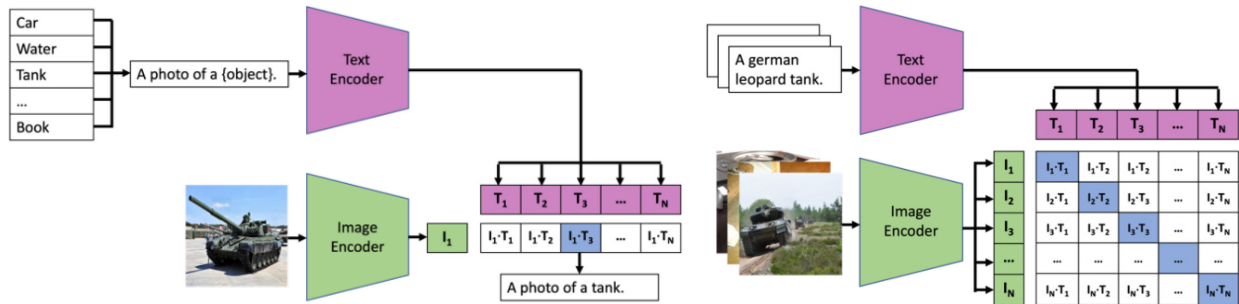


Figure 1: Pre-training concept of CLIP (left) and during inference (right). Tank images from Refs. [24], [25].

2.2 Prompt Engineering

For our analysis, we do not train the model explained in Section 2.1 on any additional data, and only use it for inference. We use our image dataset to design appropriate text prompts that are as general as possible, in order to specify a certain class. Radford et al. [10] have shown in their evaluation of the CLIP approach that the accuracy of the models increases when a whole sentence is formed. For example, to detect a tank, it is more effective to use the text prompt “this is an image of a tank, a type of vehicle” instead of just using the word “tank”. This also avoids inaccuracies. For example, in ImageNet, there are the classes “construction cranes” and “animal cranes.” If we search for “crane” only, the result is ambiguous. A simple specification to get the desired class would be “a picture of the crane, a type of bird.” Our contribution is to engineer prompts for the desired classes in a military context in order to obtain the most accurate results possible.

2.3 Evaluation Metrics

During the analysis of the datasets, we report precision, recall, and the F1 score for the classification based on the text prompts. Due to the imbalance of the dataset, we mainly highlight the F1 score, which is the harmonic mean of precision and recall.

For the extended evaluation of our results, we use the gScoreCAM [17]. The gScoreCAM algorithm is based on the idea of GradCAM [18] to provide a visual explanation of important regions within an image based on gradients. Regions that appear important to CLIP based on a caption are highlighted with heat maps.

2.4 Military Image Dataset (Mid)

For our prompt engineering analysis, we create a new dataset, called mid, which contains military and civilian vehicles, but also images without vehicles. The images are retrieved from different sources. Military vehicles are web scraped from Google Search, and the civilian vehicles come from both a Kaggle dataset [19] and Google Search. The non-vehicle images are extracted from the COCO 2017 validation subset [4], which is a web-scraped dataset as well. Here, care was taken to remove all motor vehicles to obtain a vehicle-free class.

At the most detailed level, the dataset distinguishes between five classes: “no vehicle,” “civilian car,” “civilian truck,” “military truck,” and “Armored Fighting Vehicle (AFV)/tank.” For our analysis, an aggregated level is used. Here, there are 3,041 images of military vehicles, 977 civilian vehicles, and 1,475 non-vehicle images. Images have different resolutions and quality, yet if there is a vehicle in an image, it is usually in the center and fills much of the area. See Figure 2 for some examples.

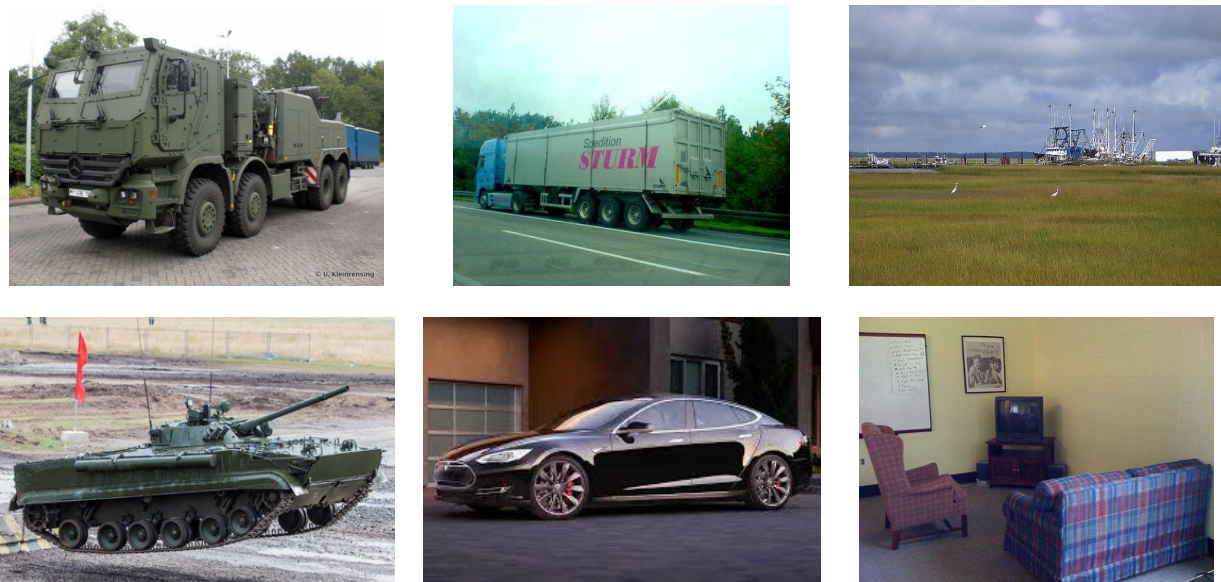


Figure 2: Examples from the mid dataset. Military (Left), Civilian Vehicles (Center), and Images Without Vehicles (Right).

2.5 Military Video Image Dataset (mvid)

The military video image dataset (mvid) consists of images extracted from 10 different videos that all originate from YouTube. A more detailed overview with titles and links is provided in Appendix 1. There are five videos with military content and five with civilian content. The first five are labeled with the classes “military vehicle” and “no vehicle.” The remaining videos have the labels “civilian vehicle” and “no vehicle.” For both domains, there are three dash-camera-like videos (filmed from the vehicle) and two surveillance-camera-like videos that are directed to a street. The image quality of these images is significantly worse than in the mid dataset. Vehicles are not necessarily in the center of the image and are often small. Video streams are usually heavily pixelated, and edges are not very clear to detect. Moreover, surveillance cameras have a different angle to images from mid. In this real-world scenario, multiple objects can appear on each video frame. In our case, there are often multiple objects of the same type, but different classes are never mixed.

3.0 PROMPT ENGINEERING FOR ZERO-SHOT CLASSIFICATION

3.1 Prompt Engineering Analysis

CLIP was pre-trained to identify whether images and text descriptions match. This can be used for multiclass zero-shot classification, where features from the CLIP image encoder and the CLIP language encoder are used, and their similarity is calculated. Yet a broad analysis of the military domain is missing. In this experiment, we use the mid dataset with the three classes: “military vehicle,” “civilian vehicle,” and “no vehicle,” and evaluate diverse prompt setups.

We create different text prompts for each class and evaluate their performance for the positive class “military.” For each of the three classes, we start with a simple single word prompt (“word” in Table 1). Therefore, we use the terms “military vehicle,” “civilian vehicle,” and “everyday object,” respectively. Radford et al. [10] claim that it is often useful to phrase a whole sentence and add an auxiliary clause to the text description (“sentence”). Hence, we create the sentence template “an image of a/an {object}” for each class. Moreover, we add the clause “; a type of vehicle” to the end of the sentence for both vehicle classes. Lastly, we want to analyze how an ensemble of multiple prompts performs (“ensemble”). For this analysis, we create five

different prompts for each class and take the mean of the textual embeddings. In this setting, we also apply the sentence template with the auxiliary clause from before. The list of all prompts for the ensemble can be found in Table 2. Results are displayed in Table 1.

Table 1: F1 Score, precision, and recall metric for classifying the “military vehicle” class in mid.

Prompt	F1 score	Precision	Recall
word	0.9125	0.8585	0.9736
sentence	0.9424	0.8934	0.997
ensemble	0.9878	0.9888	0.9868

It can be observed that with increasing complexity of the prompts, the F1 score improves. With the very simple word prompt, recall is already very good, but the downside is that 15% of non-military vehicles were detected as “military.” However, the performance is already surprisingly good for a not-fine-tuned model. The introduction of a sentence template in combination with the auxiliary clause proves to be helpful for high precision and very good recall. The F1 score reaches 0.94 for this setting. The best results are achieved if several prompts are combined. With five prompts per class, we reach a 0.9895 F1 score. Both precision and recall show very good results.

In this experiment, we illustrate that CLIP is able to classify military vehicles correctly from non-military vehicles (i.e., it has powerful zero-shot capability in the military domain). Good prompt engineering is important to receive competitive results. Also, simple single word prompts can already be sufficient.

Table 2: Prompts used for the ensemble version.

Military Vehicle	Civilian Vehicle	No Vehicle
Military vehicle, a type of vehicle	Vehicle, a type of vehicle	Everyday object
Military truck or military tank, a type of vehicle	Civilian vehicle, a type of vehicle	Standard object
Military lorry or panzer, a type of vehicle	Car or truck, a type of vehicle	Empty street without cars
Armored fighting vehicles, a type of vehicle	Normal vehicle, a type of vehicle	Random object
Military transporter or military tank, a type of vehicle	Civilian car or civilian truck, a type of vehicle	Empty street

3.2 Color Analysis

It might be possible that the color of objects plays an important role for classifying specific objects. In our dataset, civilian vehicles have many different colors, whereas military vehicles are usually brown, green, or sand colored. To prove or disprove the hypothesis, the previous experiment is repeated, but on grayscale images. If the hypothesis that CLIP is merely a color detector proves to be true, the performance should drop sharply.

However, the experiment with grayscale images shows that performance differences are minimal. This is indicated in Table 3. Therefore, we can conclude that CLIP in this setting does not rely on colors for military vehicle detection.

Table 3: Prompt engineering analysis on grayscale images.

Prompt	F1 score	Precision	Recall
word	0.9152	0.8624	0.9750
sentence	0.9422	0.8941	0.9957
ensemble	0.9825	0.9832	0.9819

4.0 ZERO-SHOT EVALUATION ON VIDEO STREAMS

In this experiment, we want to apply the results of the prompt engineering to appropriate video streams. We analyze the classification performance per video on images extracted every three seconds. As described in Section 2.3, we inspect 10 videos, five from the military and five from the civilian domain. For each domain, we specify the target class as a positive class and hence report F1 score for “military vehicle” and “civilian vehicle,” respectively.

The performance highly depends on the video stream analyzed and is displayed in Table 4. The F1 score on mvid ranges from 0.381 to 1.0. When we take a look at the five military videos, the worst result within the military domain is for **mil_border**, with an F1 score of 0.6471. During an in-depth analysis of all video frames, we discover that the object of interest is often at the very right corner of the video stream, and hence gets removed when CLIP’s image transform is applied. Using a transform without cropping (only quadratic resize), the F1 score rises to 0.8718, which is a reasonably good result. The **mil_cctv** video comes with poor resolution, which is due to the fact that the screen of a surveillance camera was filmed with a smartphone. Despite the very poor picture quality, the result of 0.7778 F1 score is good. When analyzing the **mil_tank** video, the F1 score achieves a score of 0.7473. The model has degrading performance if a tank is in the background and hence does not fill a significant area of the video frame. The performance can be increased by applying the focus area on the foreground. The **mil_sun** video also has a low resolution and is highly pixelated. The objects of interest are in the center of the image and thus in the recognizable area. This leads to a high F1 score of 0.8986. The **mil_sumy** video shows a street scene filmed from a car. Due to the format of the news provider, the region of interest also remains visible after the left/right areas were removed by CLIP’s pre-processing. In this video stream, all videos were correctly classified. To sum things up, the object of interest needs to cover a significant part of the image area. Neither camouflage nor blur due to moving objects appear to have a significant impact on the performance.

Table 4: Performance metrics for classifying the “military vehicle” or “civilian vehicle” respectively for each video in the mvid dataset.

Name	F1 score	Precision	Recall
mil_sumy	1.0000	1.0000	1.0000
mil_border	0.6471	0.9167	0.5000
mil_sun	0.8986	0.8378	0.9688
mil_tank	0.7473	1.0000	0.5965
mil_cctv	0.7778	1.0000	0.6364
civ_red	0.3810	0.4000	0.3636
civ_seoulday	0.5625	0.9000	0.4091
civ_seoulnight	0.6000	1.0000	0.4286
civ_street	0.7857	1.0000	0.6471
civ_ohiotraffic	0.8764	0.7800	1.0000

When inspecting the civilian videos, the **civ_red** stream shows difficulties and only reaches an F1 score of 0.381. However, on closer inspection, it can be noticed that vehicles are often on the edge and not in the detectable area of CLIP. The results for **civ_seoulday** and **civ_seoulnight** are intermediate (0.5625 and 0.6000). Both datasets are challenging because vehicles are often present at a far distance and thus occupy only a small visual area. A focus area in the foreground would be helpful for better results. **civ_street** is a typical street surveillance camera with a good resolution. It shows a very high precision, but the recall is intermediate (0.6471). The best results are achieved for the **civ_ohiotraffic** video. This video provides grayscale images only. Vehicles are far away but without noisy objects. Due to the angle of the camera, vehicles are captured from the front and the side, which increases the area of the object and shows the typical vehicle shapes. The F1 score is at 0.8764.

5.0 DISCUSSION AND LIMITATIONS

CLIP is a strong tool for image and video analysis. During our evaluation, we became aware of several issues to consider when using it in a production environment.

5.1 Limitations on Video Streams

We show that CLIP also works with low-resolution videos. However, the performance is not as high as in the previous analysis. It is best to have the object of interest in the center and foreground. For videos, it is recommended that a quadratic region of interest where vehicles usually fill a larger area is specified. Results for civilian vehicles are slightly worse. This might be due to the fact that during pre-training, civilian vehicles are not “that special” in image data and often appear in the surroundings of another main object. This is usually not the case with military content. In addition, CLIP does not regard temporal dependency. Since we have achieved our results only on single extracted frames, a post-processing extension would be necessary in practice. For this, the predictions on several frames would be used to obtain robust results.

5.2 Data Pre-Processing

During the CLIP pre-training, the only data augmentation used is a center quadratic crop on the images. During the evaluation on ImageNet and other datasets, this was not changed because the searched object appears mostly in the center of an image. Our analysis in Table 5 shows that the different proportions in only resized images result in minimal performance losses. As an alternative in other real-life settings such as phone camera videos or drone footage in aspect ratios other than square, square overlapping patches can also be entered into CLIP to further minimize this effect.

Table 5: Comparison of image pre-processing steps on mid dataset. Original setup with center-crop and version with resize-only.

Prompt	F1 score		Precision		Recall	
	Original	Resize	Original	Resize	Original	Resize
word	0.9125	0.9124	0.8585	0.8526	0.9736	0.9812
sentence	0.9424	0.939	0.8934	0.8876	0.9970	0.9967
ensemble	0.9878	0.9880	0.9888	0.9878	0.9868	0.9881

5.3 Typographic Attacks

During the experiments with various webcam videos, the Optical Character Recognition (OCR) capability from CLIP became apparent [10], [20]. However, CLIP can be easily tricked because of this capability [21]. An example of this can be seen in Figure 3. The dog in the center of the image is misrecognized as soon as a text is added that is very similar to a queried prompt. This is usually rare with traffic cameras, but of course, limits the transferability to social media videos where text or watermarks are often faded in. A possible pre-processing with the help of a text spotting component [22], [23] can help to increase the quality of the predictions in this setting. The exact position of the detected text can be used to perform targeted blur operations similar to Figure 3.

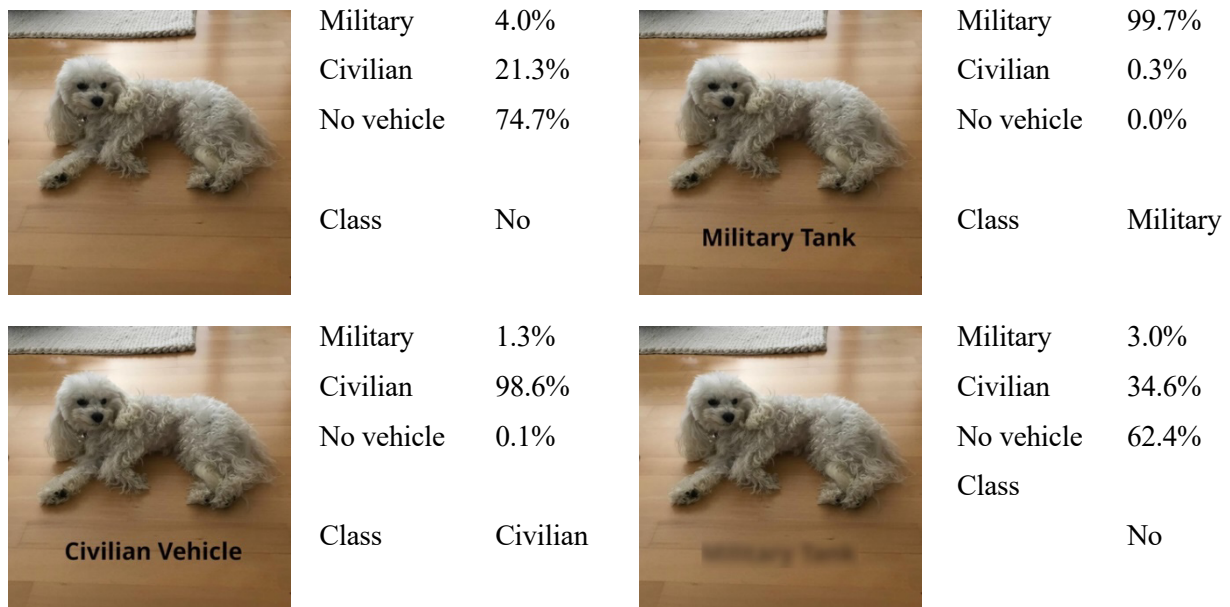


Figure 3: Typographic attack on a non-vehicle image.

According to the predicted probabilities, the presence of text in an image dramatically changes the prediction. Especially for the non-civilian/military class, a prompt targeting dogs was added. CLIP also shows a preference for the text within the image instead of the actual object. This can also be observed in the gScoreCAM visualization shown in Figure 4, where the correct areas in the image are detected depending on the prompt.

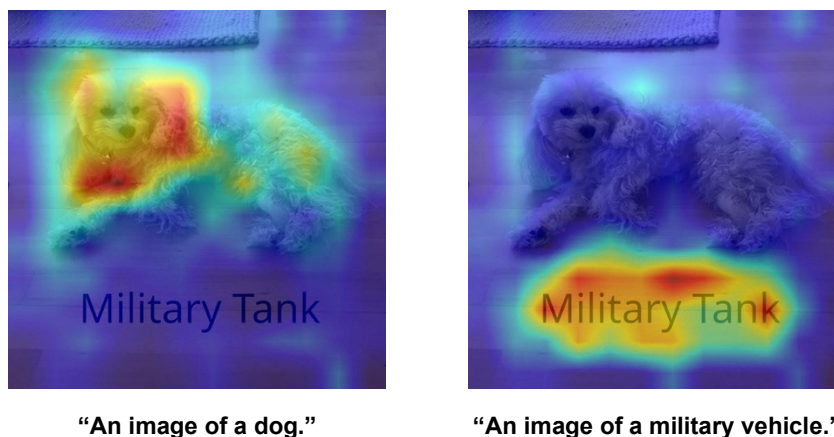


Figure 4: Heatmap using gScoreCAM for two different prompts.

An example of such an attack is shown in Figure 5. A civilian truck with a neutral tarp may have sprayed writing on it to cause an automated detector based on CLIP to false alarm. Furthermore, the depiction of military vehicles on the tarpaulin may also lead to false prediction.

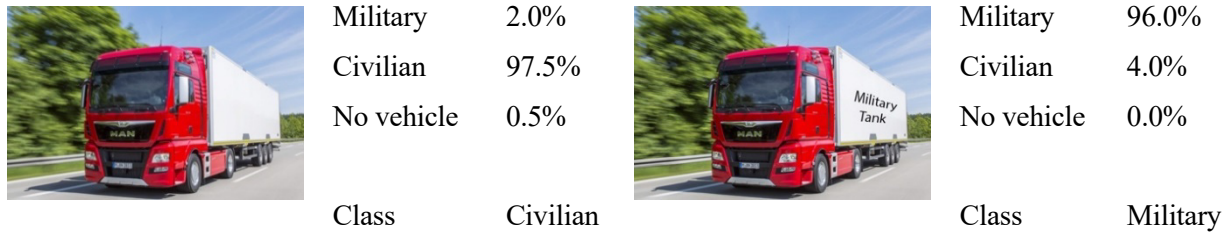


Figure 5: Example of an attack on an image with a civilian truck.

5.4 Prompt Selection and Bias

The conducted research highlights the following recommendations during the selection of suitable prompts. Based on templates like “an image of a/an {object}” the performance during inference rises as opposed to a single “{object}” prompt. Additionally specifying “an image of a/an {object}, a type of {object category}” boosts performance as well, even if keywords occur twice in the sentence.

Due to the unknown pre-training data of CLIP, some prompts can mislead the model and therefore lead to unwanted bias. For example, the word “tank” has the following meanings: a type of military vehicle or a type of container. Unfortunately, we cannot evaluate which type of tank is more common in CLIP’s pre-training data. Hence, for ambiguous words, it is useful to add an adjective or another type of description. We choose to use the term “military tank” in our analysis to reduce bias towards the term “container tank.”

6.0 CONCLUSION

In our work we have analyzed the zero-shot capabilities of CLIP in the military domain. We evaluated prompts of increasing complexity and have shown that, with an ensemble of five prompts, an almost perfect detection for military vehicles can be achieved on our mid dataset. Furthermore, the use of grayscale images only results in insignificant performance loss. There are special characteristics that must be taken into account when applying CLIP. Especially for surveillance cameras or dash cams, it is important to know that background objects are very hard to detect. Hence, it is useful to specify a region of interest where objects in the foreground usually appear. Without this process, an F1 score between 0.381 and 1.0 was achieved on our datasets. If no focus area is specified, it is important to consider that the edges of an image are cropped during pre-processing. Moreover, we have shown that due to the pre-training objective, CLIP can be fooled by written statements in the images. This is hardly a problem with surveillance cameras, but it can be a challenge when analyzing data from social media.

CLIP is known to be a powerful model for visual and textual applications. There is ongoing research investigating the capabilities of this deep learning model. Our results show that a malicious actor can use CLIP to screen masses of traffic surveillance data and videos from public sources without the need for training specifically on military vehicle datasets. Doing so publicly available open-source models can be used by reconnaissance units to analyze troop and logistic movement. From a military perspective, it is vital to have an overview over such movements. Automated analysis without – or at least with little – human control, allows hundreds or thousands of video camera streams to be analyzed and to supply command and control with the results of the analysis. The availability of this information can lead to cognitive superiority on the battleground. The approach shown here demonstrates that public models and intelligent prompt engineering without the need of time-consuming data collection and model training can be used to create such software solutions for the military reconnaissance sector.

7.0 ACKNOWLEDGEMENT

The authors gratefully acknowledge the computing time granted by the Institute for Distributed Intelligent Systems and provided on the GPU cluster, Monacum One, at the University of the Bundeswehr Munich. We also want to thank Mr. Albert Lattke, OF-1, for his contribution to data preparation.


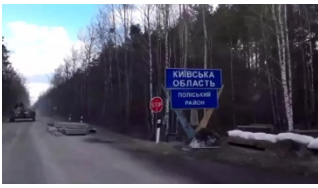


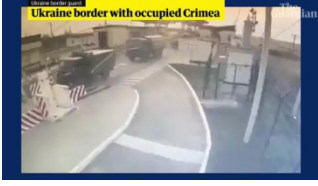


8.0 REFERENCES

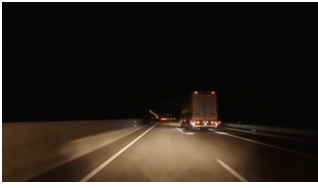
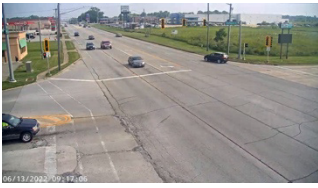

- [1] Krizhevsky, A., Sutskever, I., and Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Pereira, F., Burges, C., Bottou, L., and Weinberger, K, Eds. Vol. 25. Curran Associates, Inc., 2012. <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>
- [2] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 770-778.
- [3] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S. et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3), 2015, 211-252.
- [4] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P. et al. Microsoft COCO: Common objects in context. *European Conference on Computer Vision*, Springer, 2014, 740-755.
- [5] Larochelle, H., Erhan, D., and Bengio, Y. Zero-data learning of new tasks. 2(01), 2008, 646-651.
- [6] Xian, Y., Lampert, C.H., Schiele, B., and Akata, Z. Zero-shot learning – a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9), 2018, 2251-2265.
- [7] Schönfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., Akata, Z. Generalized zero- and few-shot learning via aligned variational autoencoders. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, 8239-8247.
- [8] Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., and Schiele, B. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 69-77.
- [9] Zhang, L., Xiang, T., and Gong, S. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 2021-2030.
- [10] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S. et al. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, PMLR, 2021. 8748-8763.
- [11] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H. et al. Scaling up visual and vision-language representation learning with noisy text supervision. *International Conference on Machine Learning*, PMLR, 2021, 4904-4916.
- [12] Krieg in Ukraine: In niedersachsen wurden autobahnwebcams abgeschaltet. [Because of Putin’s war: motorway webcam switched off in Lower Saxony]. 12 March 2022. <https://www.rnd.de/panorama/krieg-in-ukraine-in-niedersachsen-wurden-autobahnwebcams-abgeschaltet-4ODDZNGOEBADHNBA PLXFUPW46Y.html>

- [13] Zhang, R. Making convolutional networks shift-invariant again. International Conference on Machine Learning, PMLR, 2019, 7324-7334.
- [14] He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M. Bag of tricks for image classification with convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 558-567.
- [15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T. et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
- [16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N. et al. Attention is all you need. Advances in Neural Information Processing Systems, 30, 2017.
- [17] Chen, P., Li, Q., Biaz, S., Bui, T., and Nguyen, A. gscorecam: What is clip looking at? 2022.
- [18] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, 2017, 618-626.
- [19] Umcu, E. Car and truck. August 2018. <https://www.kaggle.com/datasets/enesumcu/car-and-truck>
- [20] Deuser, F., Habel, K., Rösch, P.J., and Oswald, N. Less is more: Linear layers on clip features as powerful vizwiz model. arXiv preprint arXiv:2206.05281, 2022.
- [21] Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L. et al. Multimodal neurons in artificial neural networks. Distill, 2021. <https://distill.pub/2021/multimodal-neurons>.
- [22] Zhang, X., Su, Y., Tripathi, S., and Tu, Z. Text spotting transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 9519-9528.
- [23] Wang, P., Zhang, C., Qi, F., Liu, S., Zhang, X., Lyu, P. et al. Pgnet: Real-time arbitrarily-shaped text spotting with point gathering network. Proceedings of the AAAI Conference on Artificial Intelligence 35(4), 2021, 2782-2790.
- [24] U.S. Army Europe. 2010. https://www.flickr.com/photos/usarmyeurope_images/14431618363/in/photolist-2ngmHBV-nZgPmg
- [25] Kuzmin, V. Vitaly Kuzmin Military Blog. 2015. <https://www.vitalykuzmin.net/>

Appendix 1: DETAILS OF MVID

Table A1-1: Name, title and url for videos in mvid.

Name	Title	Example	URL
mil_summy	Ukrainian citizen confronts Russian soldiers after tank runs out of fuel		https://www.youtube.com/watch?v=14gVDF2b1vA
mil_border	Russian video shows troops entering Kyiv region		https://www.youtube.com/watch?v=fWc-gtblvOU
mil_sun	Russian military convoy moves through Kharkiv region, littered with destroyed tanks		https://www.youtube.com/watch?v=BIS8XWEhZdM
mil_tank	Russian tanks drive through town in Kyiv, Ukraine		https://www.youtube.com/watch?v=F5tpehGgN_Q
mil_cctv	CCTV shows Russian tanks entering Ukraine from Belarus and Crimea		https://www.youtube.com/watch?v=wfUKVsjhclY
civ_red	Red light camera flash		https://www.youtube.com/watch?v=bYw56iTqeKU
civ_seoulday	ASMR Highway Driving in the Rain – Day to Night (No Talking, No Music) – Seoul to Daegu, Korea		https://www.youtube.com/watch?v=Dwswey-GqQc&t=6492s

Name	Title	Example	URL
civ_seoulnight	ASMR Highway Driving at Night (No Talking, No Music) – Busan to Seoul, Korea		https://www.youtube.com/watch?v=nABR88G_2cE&t=11093s
civ_street	Village of Tilton – Traffic Camera		https://www.youtube.com/watch?v=5_XSYIAfJZM
civ_ohiotraffic	Ohio Traffic Camera Captures Tornado on Video		https://www.youtube.com/watch?v=PmrSOPMkfAo

