**哈尔滨工业大学**

**硕士学位论文开题报告**

**题 目：A study of cross-modal single-target tracking aided by natural language description**

**院 （系） 计算机科学与技术学院**

**学 科 /专 业 计算机科学与技术**

**导 师 何震宇**

**研 究 生 MONKAM KATCHA EDDY NATHAN**

**学 号 23SF51004**

**开题报告日期 2024年9月11日**

**深圳校区教务部 制**

**Table of contents**

[1. Source of the topic and the purpose and significance of the research 2](#_Toc236835824)

[1.1. Source of the topic 3](#_Toc1825419699)

[1.2. Research Background and Significance 3](#_Toc1494834519)

[2. Research status and analysis at home and abroad 5](#_Toc1861192185)

[2.1. Current research status at home and abroad 5](#_Toc641194506)

[2.2. Domestic and foreign literature review and analysis 6](#_Toc256024327)

[3. Main research content and research program 8](#_Toc1713434176)

[3.1. Research content 9](#_Toc1798213868)

[3.2. Research programs 10](#_Toc149156340)

[4. Expected objectives and research achievements 11](#_Toc33668216)

[5. Completed research work and schedule 12](#_Toc1625845252)

[5.1. Studies completed 13](#_Toc759392543)

[5.2. Schedule 13](#_Toc1630356349)

[6.Available research conditions, requirements, and funding 14](#_Toc121845794)

[6.1. Laboratory conditions and financial security 15](#_Toc916425096)

[6.2. Requirements and funding 15](#_Toc193385564)

[7. Anticipated difficulties and solutions 15](#_Toc713789855)

[7.1. Anticipated difficulties 16](#_Toc1813532431)

[7.2. Anticipated solutions 16](#_Toc1734992844)

[References 16](#_Toc253741994)

# 1. Source of the topic and the purpose and significance of the research

## 1.1. Source of the topic

My passion with the junction of computer vision and natural language processing drives my interest in cross-modal single-target tracking. The possibility of integrating these two disciplines to produce more intelligent and flexible systems has long captivated me. Applications in domains such robotics, surveillance, and augmented reality abound from the ability to track items in real-world surroundings utilizing both visual and textual information

Working on past research on computer vision and natural language processing, I discovered a consistent difficulty to properly combine textual and visual input. Models seemed to me to frequently rely mostly on visual information, ignoring the useful contextual data given by textual information. This realization piqued my interest and drove me to look at the fundamental causes of these phenomena and possible remedies.

The mounting corpus of research on cross-modal tracking and the growing interest in merging visual and textual information for different uses inspired me even more. Examining several scholarly publications and conference presentations in this field, I identified both major prospects and difficulties. The possibility of enhancing the integration of visual and textual information by means of foundation models and attention mechanisms especially caught my interest.

Many publications piqued my increasing curiosity in cross-modally single-target tracking particularly one titled Joint Visual Grounding and Tracking with Natural Language Specification by Zhou et al.1 Those papers have given insightful analysis and inspiration for this work, therefore inspiring the investigation of creative methods for cross-modal single-target tracking.

## 1.2. Research Background and Significance

Single-target tracking is a fundamental difficulty in computer vision that has a rich legacy spanning many decades. Tracking systems matched items between consecutive frames in a picture series using simple correlation methods. Although they provide a basis for tracking, these early techniques could not manage the complexity of real-world situations. With the development of computer vision technology, more complex methods like particle filters and Kalman filters appeared. These techniques presented probabilistic models able to control non-linear dynamics and uncertainty, thereby improving the tracking accuracy and robustness of objects.

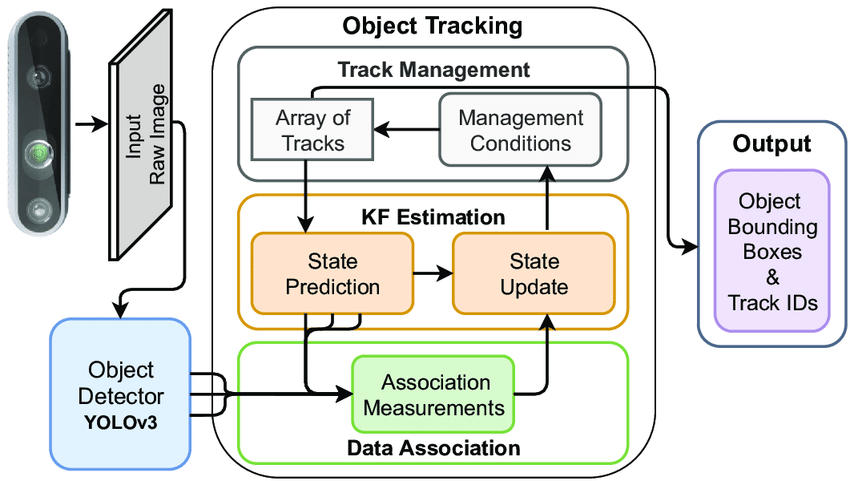
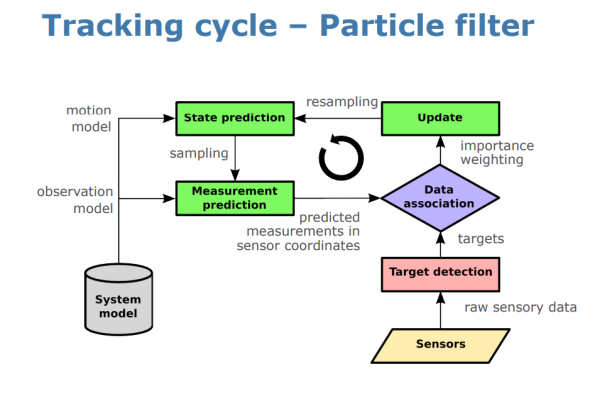
 

Fig1. Tracking with Kalman filter Fig2. Tracking with particle filter

Deep learning's arrival in recent years has fundamentally changed the field of single-target tracking. With their strong tools for extracting discriminative features from visual data, convolutional neural networks (CNNs) have grown to be pillars of this change. Tracking accuracy and resilience have also made great progress due to CNNs. Apart from CNNs, recurrent neural networks (RNNs) have been applied to simulate the temporal dynamics of object motion, hence enhancing the capacity to monitor objects across challenging environments including those involving fast motion and different environmental conditions.

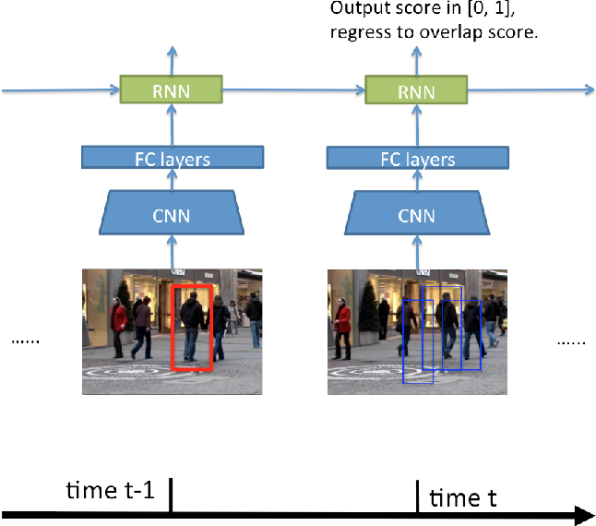


Fig3. Tracking with Recurrent neural networks (RNNs)

Although there have been some advancements, traditional tracking techniques—including those using deep learning—face ongoing difficulties. Still major challenges including occlusions, scale changes, and complicated backgrounds are recurrent. In response, scientists have started looking at using several modalities—especially visual and textual information—to meet these difficulties. By utilizing the complementing qualities of both textual and visual data, the integration of the two modalities may improve tracking systems. While textual descriptions provide contextual and semantic information that may help with object classification and enhance tracking accuracy, visual data provides specific spatial information.

Large-scale databases labeled with both kinds of data (visual and textual) have made cross-modal tracking—which mixes visual and textual information—more and more relevant. This mix of modalities has the potential to revolutionize several areas including augmented reality, robotics, and surveillance. Combining textual and visual data, for instance, might make it simpler for the surveillance system to keep an eye on people in places that are difficult to reach. By enabling a robot to follow both textual and visual input, such integration has the potential to improve a robot's understanding and interaction with its surroundings in robotics. Likewise with augmented reality, integrating various modalities can result in more contextually aware and intuitive applications.

The relevance of this study reaches both theoretical and practical domains. It should theoretically help us to better grasp how to combine textual and visual data for challenging projects. This includes investigating cutting-edge approaches for balancing and aligning various types of data, collecting semantic links, and developing advanced fusion procedures. The study's practical goal is to develop tracking systems that are more advanced and useful in a wider range of circumstances. By addressing significant problems including temporal misalignment and modality imbalance, our work may result in more robust and adaptable tracking systems that perform well in a variety of dynamic environments.

Though much progress has been made, there are still some significant obstacles in the sector. One of the primary issues is the imbalance between the visual and textual modalities, wherein existing models occasionally prioritize visual input above textual information. Taking up this task is necessary to develop a more balanced tracking system that utilizes both modalities. Additionally, improved contextual knowledge is necessary to capture the semantic connections between textual and visual data, which will aid the system in accurately tracking objects. Furthermore, enhanced contextual awareness is required to extract the semantic relationships between textual and visual information, which will let the system track things more precisely. Enhancing fusion strategies—which seek to develop techniques that optimize the complementary qualities of textual and visual knowledge—is another crucial area of focus. In order to effectively synchronize several modalities, the system must manage temporal dynamics in order to control the misalignment between visual and textual input.

# 2. Research status and analysis at home and abroad

## 2.1. Current research status at home and abroad

In computer vision, cross-modal single-target tracking—that is, combining visual and textual information for tracking a particular object—has shown great promise. Using the complementing qualities of several modalities, this method seeks to surpass the constraints of conventional single-modal trackers. Driven by developments in deep learning and natural language processing, this discipline has seen major progress recently.

Research in this area is expanding in China using creative approaches combining textual and visual data. Researchers are focusing on capsule networks to improve tracking robustness by capturing complex interactions between visual and textual inputs2. Another important method that is currently being explored is contrastive learning, which brings textual and visual features into a shared representation space thereby enhancing tracking accuracy15. With initiatives and efforts to optimize algorithms that balance computational efficiency and tracking accuracy, real-time processing of natural language descriptions is another main topic of attention as well9. Setting standards for further developments, researchers are also developing new benchmarks to evaluate tracking systems depending on criteria including accuracy and adaptability2. Notwithstanding these advances, problems still exist in balancing computing loads and improving model generalization across diverse environment.

The research is similarly advanced internationally, and different research institutes all around are making progress. To combine visual and textual input for enhanced tracking performance, researchers are using sophisticated neural network designs such Transformers and Siamese networks511. The alienation of features between modalities via contrastive learning and attention mechanisms is leading to better tracking accuracy results68. Furthermore, real-time systems that effectively process natural language inputs while preserving high performance are under focus7. With measures created to assess tracking systems on flexibility, accuracy, and adaptability11, comprehensive benchmarking—a vital component of international research—is underlined. Difficulties including computing efficiency and model generalization are being tackled, much as in research in China reflecting a worldwide attempt to improve cross-modal tracking technology.

## 2.2. Domestic and foreign literature review and analysis

One interesting path for improving the performance of single target tracking systems is the merging of visual and textual information. Researchers have aimed to create more strong and accurate tracking algorithms that can manage challenging situations and generalize to unexplored domains by using the complimentary characteristics of various modalities.

The construction of unified models that combine visual and textual modalities is one well-known topic of study. Using their unified vision-language tracking model—which stresses multi-modal alignment—Zhang et al. investigate this2. By generating a shared representation space that lowers modality differences, their method efficiently synchronizes visual and textual elements, hence improving the accuracy and resilience of tracking systems. Comparably, Zhou et al. suggest a joint visual grounding and tracking system with natural language requirements1. This one model grounds items within the framework given by textual descriptions in addition to tracking them. Leveraging the extra information given by natural language, the integration of grounding and tracking into a single model enables a more contextually aware and exact tracking mechanism.

The application of cutting-edge neural network topologies and methodologies makes still another important addition to the discipline. Ma et al. introduces capsule-based object tracking with natural language specifications3. Detailed natural language descriptions help to provide a more sophisticated and precise tracking system by which to record and depict complicated interactions between visual and textual inputs. Ma et al. investigates furthermore how contrastive learning may unite vision-language tracking with visual data3. Their work shows how contrastive learning may match textual and visual elements, therefore improving the accuracy of the system in object tracking depending on natural language inputs. This method emphasizes how well contrasting learning might enhance cross-modal tracking performance.

Practical uses of cross-modal tracking systems depend on real-time processing capability. Using natural language descriptions, Feng et al. create models for real-time visual object tracking9. Their work highlights how quickly and precisely one may parse textual descriptions in real time. Applications needing quick reactions and changes depending on natural language inputs depend on this real-time processing capacity. Using Siamese networks to address natural language descriptions, Feng et al. also propose the Siamese Natural Language Tracker10. This model shows how well Siamese trackers combined with natural language improve tracking accuracy and resilience.

Major areas of study include the creation of thorough benchmarks and the generalizing tracking system capability to unexplored fields. With natural language, Wang et al. concentrate on accuracy and adaptability in object tracking11; they provide methods and benchmarks evaluating tracking systems depending on performance criteria. Their efforts emphasize the need of assessing tracking systems in several contexts to guarantee their efficiency and flexibility. Yu et al. introduces natural language representation6, so helping to generalize many objects tracking. Their method shows how natural language could increase the adaptability of tracking systems outside of the training data by aiming to extend tracking capabilities to new and diverse surroundings.

Another key component of cross-modal monitoring is the relationship between visuals and text. The CiteTracker approach closes the distance between textual and visual data by stressing the efficient matching of textual descriptions with visual characteristics, hence improving tracking performance4. This method emphasizes the importance of exact correlation in order to enhance tracking results. Furthermore, highlighting the progress in including contextual information into tracking systems are the investigations on tracking by natural language specification and long short-term context decoupling. These methods stress the need of contextual knowledge on tracking accuracy and offer a more complete way to combine natural language with visual tracking.

In the realm of cross-modal single-target tracking, numerous typical constraints still exist although there were some advances achieved by different researchers. The mismatch between processing textual and visual input is one of the main difficulties here. Many models—including those put out by Zhang et al. and Zhou et al. —have trouble making sure visual and textual modalities have equal weight during processing12. This sometimes leads to situations whereby the model performs poorly because one modality is overused over the other. The computational expense connected with real-time processing is still another important restriction. Although models like the ones created by Feng et al. show the possibility for real-time processing, their usefulness in contexts with limited resources may be limited by their often high computational resources demand9. For researchers, lowering the computing burden while keeping monitoring precision remains difficult. Moreover, generalization to unexplored fields is still a difficult question. Although Yu et al. and Wang et al. have made progress in enhancing the adaptability of tracking systems to different sectors, more strong models that can manage several environments without appreciable performance loss is still much needed611. These models show the need of improved generalizing procedures since they commonly suffer with invisible items or descriptions. Finally, a difficult component of cross-modal tracking still is the relationship between textual descriptions and visual elements. Although CiteTracker and other related models have advanced in this field, many systems still find it difficult to properly map textual elements to their matching visual counterparts4. When dealing with vague or incomplete textual descriptions, where the lack of precision can provide erroneous tracking results, this problem is more noticeable.

# 3. Main research content and research program

## 3.1. Research content

The main goal of this thesis is to investigate how the integration of visual and textual information might enhance tracking performance, therefore strengthening the field of cross-modal single-target tracking. How can the imbalance between visual and textual modalities be resolved in cross-modally tracking systems is a fundamental concern driving this work. Most current models show visual data as the main focus, with textual data serving just a secondary importance. The aim is to build a model that equally uses both modalities therefore enhancing general tracking performance. This work will thus look at various important approaches for combining textual and visual data. Particularly the instances including occlusion, scale fluctuation, and complicated or crowded backgrounds—that are typically difficult for tracking systems—will get special attention. We investigate methods such transformers, attention processes, or graph-based models to find the best way to combine several data sources.

Temporal mismatch between visual and textual input presents is another major obstacle in cross-modal tracking. Whereas verbal descriptions are stationary or fluctuate at different intervals, visual data is often recorded constantly in the form of video frames. This work will investigate techniques to reduce this temporal misalignment therefore enabling more accurate and synchronized tracking. The project intends to generate a tracking system able to operate more efficiently in real-world applications by creating methods to better synchronize the time between visual and textual inputs.

Apart from temporal misalignment, the thesis will investigate how a cross-modal tracking system may generalize over several domains and surroundings devoid of significant retraining. Domain adaption methods will be examined to guarantee the system's performance under new conditions as well as unseen ones. Real-world applications depend on this capacity to generalize across several fields since conditions can vary greatly between one situation to another. The study will also consider how contextual knowledge could help to increase tracking accuracy. It will especially look at how the system might record interactions between objects mentioned in the book and those found in the visual scene. Graph neural networks and other models able to represent these interactions will be assessed to improve the contextual awareness of the system.

The approach of this thesis will include experimental and theoretical elements. A thorough analysis of the literature will be the first stage to pinpoint the shortcomings of current techniques and direct the design of a new architecture. With special focus on including advanced fusion techniques, such attention mechanisms, the new architecture will strive to balance the integration of visual and textual information. Experiments using publicly available datasets including visual data matched with textual descriptions will be carried out to test this architecture. Custom datasets will also be produced to replicate more difficult tracking situations such as cluttered backdrops, scale variance, and occlusion.

This study's data collecting stage will make use of several large-scale datasets including well-known benchmarks such as LaSOT, TrackingNet, and VidText. These data will serve for assessing the proposed tracking system. Apart from the already available data, this study might create unique datasets to evaluate the performance of the system in challenging real-world scenarios including metropolitan surroundings with several moving objects and partial occlusions. Furthermore, real-world data will be gathered to evaluate the system's performance in handling pragmatic applications whereby pre-defined tasks or user interactions yield textual descriptions.

I will evaluate my system using standard tracking metrics like Intersection over Union (IoU), tracking accuracy (TA), and precision-recall curves. For cross-modal evaluation, I will use metrics that assess how well the model aligns textual and visual data, such as cross-modal retrieval accuracy. I will also benchmark my model against state-of-the-art tracking systems that either use only visual data or incorporate other modalities.

Techniques of analysis will combine qualitative and quantitative assessments to assess the performance of the suggested model. Multi-modal alignment scores will be used to evaluate the quality of the modality fusion; visual and textual cues will be displayed using attention maps to help one understand how the model interprets them. To find the significance of particular model components—such as fusion layers and attention mechanisms—in supporting overall performance, ablation tests will be carried out. These investigations will assist in determining which elements of the model are most essential for success.

At last, the proposed tracking system will be evaluated in relation to other cross-modal based tracking systems. Especially in challenging tracking situations, this comparison study will show how the innovative approaches either enhance or complement current systems. By means of this extensive research approach, the thesis seeks to challenge the limits of cross-modally single-target tracking and support developments in the balance between the visual and textual information.

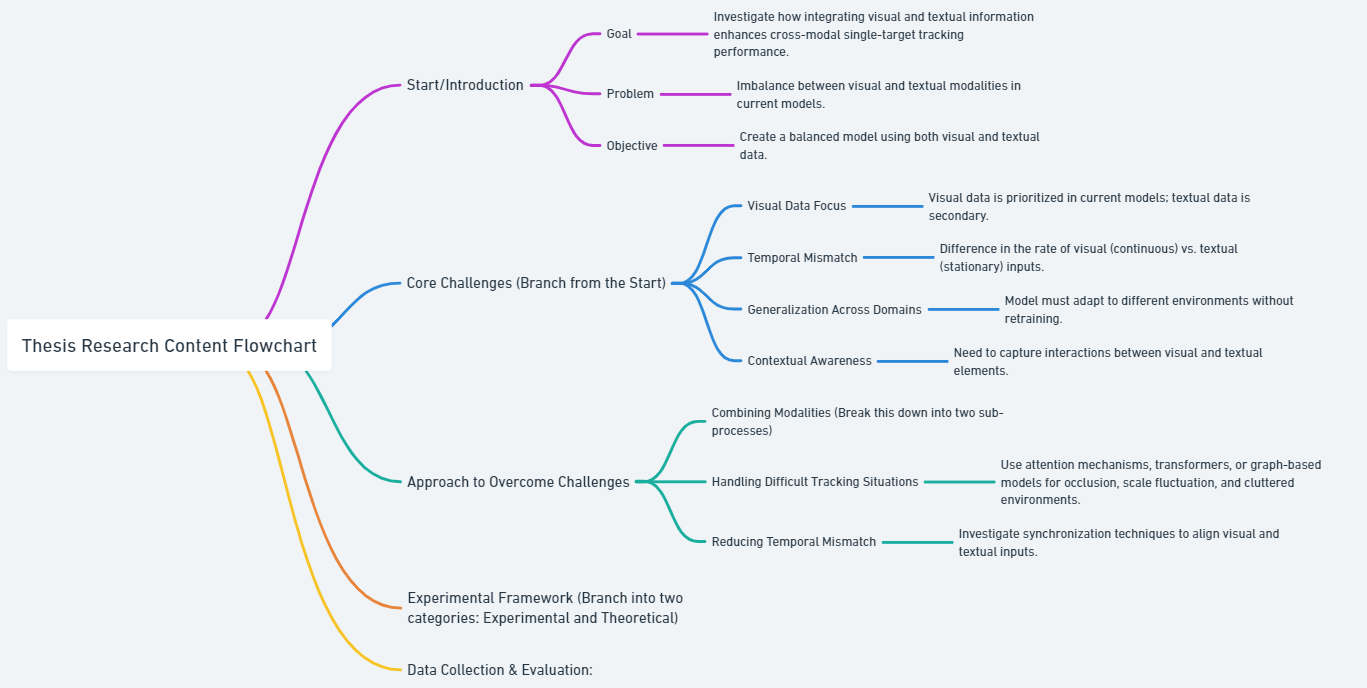


Fig4. Research content

## 3.2. Research programs

While allowing for iterative development and testing of the suggested model, the research program for this thesis is set to follow a disciplined schedule guaranteeing consistent progress and timely completion. The program consists in different phases with particular goals and benchmarks.

I am now finishing the first step of the investigation. Aiming at mapping out current methods in cross-modal tracking, it entails an extensive literature review. Emphasizing the integration of visual and textual data, this paper addresses important issues including temporal misalignment and modality mismatch. The aim is to get a comprehensive knowledge of present approaches and their constraints, therefore guiding the development of the suggested study.

The second step, following the literature study, will concentrate on data preparation and cross-modal tracker basic architectural design. This phase will consist in locating and preparing pertinent datasets. The aim is to create a first model architecture balancing feature extraction from both visual and textual modalities. Early and late fusion as well as other fusion methods will be investigated in design and tested on a preliminary dataset.

The emphasis in the third phase will turn to using sophisticated multi-modal fusion techniques to improve the integration of textual and visual data. This phase will entail the creation of dynamic fusion methods that change depending on the importance of every modality in several frames. Furthermore, included will be temporal alignment methods to solve visual-textual temporal misalignment. Using graph-based models or scene parsing approaches to increase the model's capacity to grasp relationships between things specified in text and those present in the visual scene will also be a top focus on the investigation of contextual comprehension.

Training the model and assessing its performance under several conditions will take front stage in the fourth phase. Standard cross-modal benchmarks will be used for training the model with special focus on handling occlusions, scale fluctuations, and crowded settings. Furthermore, the model will be tested using tailored datasets meant to add complexity, including metropolitan settings including several moving objects. Additionally included in this phase will be ablation studies to evaluate the contribution of every element to the general model performance.

The fifth phase will consist in a comparison of the suggested model with current techniques. This study will apply consistent rating criteria. Various tracking situations will allow the model to combine visual and textual data using visualization methods. Performance of the model will be assessed by means of an extensive qualitative and quantitative study.

Writing and finishing the thesis will be the main emphasis of the last section of the research program. Drafting and editing several chapters—including the Introduction, Literary Review, Methodology, Results, and Discussion—will be part of this step. Graphs and visuals will be ready to succinctly show the performance and results of the model. The thesis will be turned in for review; comments will guide changes to it. The turning in the final thesis will signal the end of this phase.

# 4. Expected objectives and research achievements

Focusing on the efficient integration of visual and textual information, this thesis is ready to offer major contributions to the discipline of cross-modal single-target tracking. The main goal is to create fresh approaches that solve current problems and improve monitoring quality. Correcting the disparity between visual and textual modes is one of the main goals. Many times, current models give visual data top priority, which underuses textual data. This study will aim to provide a better-balanced strategy so that both modalities equally support tracking accuracy.

Still another important goal is to enhance the methods for combining textual and visual data. To improve the complementarity of visual and textual cues, the work will investigate cutting-edge fusion methods including transformer models and attention mechanisms. This seeks to solve challenging tracking situations including occlusion, scale variation, and crowded backdrops where current approaches sometimes find difficulty.

The thesis will also stress reducing temporal mismatch between textual and visual data. Maintaining tracking accuracy over time depends on proper synchronizing of several senses. The study will look at methods to match visual frames with textual descriptions so that the tracking system may more strongly manage dynamic and changing environments.

The expected results of this work include the creation of a strong cross-modal tracking model surpassing present state-of- the-art techniques. This model is supposed to show better adaptability and accuracy over a spectrum of demanding surroundings. Furthermore, the studies will provide insightful analysis of the integration of multi-modal data, thereby expanding theoretical knowledge and useful applications in fields such surveillance, robotics, and augmented reality.

# 5. Completed research work and schedule

## 5.1. Studies completed

Although this suggestion presents the future direction of research, in the field of cross-modal single-target tracking I have done preliminary studies using a comprehensive literature analysis and theoretical investigation. These initiatives have given the research a strong conceptual basis and helped to pinpoint important problems and possible areas of innovation. Reviewing and evaluating current approaches has helped me to better understand the present situation of cross-modal tracking, especially the problems with the imbalance between visual and textual data, constraints in fusion procedures, and temporal misalignment across several modalities. Though not yet any tests have been carried out, this thorough review of the literature has greatly helped to shape the suggested methodology and improve the research questions. These early studies will direct the stages of practical experimentation and model creation, thereby guaranteeing a clear and orderly path as I enter the next phases of the research.

## 5.2. Schedule

Over a six-month period, the research proposal for this thesis is organized with each phase concentrating on important project components to guarantee consistent development and timely completion. With special attention to visual and textual data integration techniques, the first month's emphasis will be on doing an extensive literature analysis to map out the present status of cross-modal single-target tracking. Key difficulties include modality imbalance, and temporal misalignment will be emphasized in this review, therefore guiding the construction of a concise problem statement. A thorough literature evaluation will be turned in by the end of the month, pointing up areas of interest for this kind of study as well as gaps.

Data collecting and the first model architectural design will occupy the second month. We will find and get ready for processing datasets. At the same time, a first model will be developed with an eye on dual-stream networks handling separate textual and visual input. At this point the deliverable will be a draft of the first model design and a paper on data preparation.

Third month model development will start with the building of a dual-stream network meant to independently process textual and visual data. Integration of attention mechanisms will help to enable early-stage feature fusion amongst several modalities. A progress report with first test findings on small-scale datasets will be presented before the end of the month. Development will keep during next month, honing the model and adding more sophisticated fusion methods including cross-attention or transformer models. Furthermore, investigated will be temporal alignment techniques to improve the synchronizing of visual and textual data streams. By the end of the fourth month, small dataset test performance and revised model design will be turned in.

Model testing and evaluation will take front stage in the fifth month. Larger datasets will be used for training the model so that it may maximize performance in challenging situations. Performance of the model will be assessed using important measures. The evaluation results will be collected into a report stressing the strengths and opportunities for additional improvement of the model.

The last month will be devoted to a cross-modally single-target tracking comparative study between the created model and current ones. This study will enable a comparison of the performance of the suggested model with known methods. Simultaneously, the thesis will start to be written, concentrating on crafting the sections on Introduction, Methodology, and first Results. By the end of this term, the first thesis draft and a thorough study will be turned in for review.

# 6.Available research conditions, requirements, and funding

## 6.1. Laboratory conditions and financial security

My laboratory will be the scene of the research. It will offer the best setting for doing experiments. Regarding tools, strong computers are easily reachable. Especially in the model training and inference phases, these systems are indispensable for managing the computing requirements of cross-modal tracking.

## 6.2. Requirements and funding

Regarding software, the project will make use of well-supported deep learning frameworks as TensorFlow and PyTorch, which are fresh from the research environment. Libraries like as OpenCV are accessible for image and video processing chores; natural language processing (NLP) will be handled using tools such NLTK and spaCy. Furthermore, data visualization instruments such as Matplotlib and Seaborn will be employed to show the experimental outcomes clearly and powerfully.

Regarding financial stability, the project does not now need for more money. All required tools and programs are now available. External funding is not mandated.

# 7. Anticipated difficulties and solutions

## 7.1. Anticipated difficulties

This research project is expected to face challenges in managing the complexities of cross-modal data fusion. Combining textual and visual data into a single model presents a number of challenges, particularly when dealing with the inherent imbalance between the two modalities. Textual data is frequently discontinuous and sparse, whereas visual data is typically high-dimensional and continuous. Because of the differences in data structure, it may be difficult to efficiently synchronize two types of information for tracking. The timing mismatch between the textual and visual inputs adds to the difficulties. For example, written data, such as natural language descriptions, may only be updated periodically, whereas visual data from video feeds is continuously updated. Addressing these challenges is critical to ensuring the tracking system's capacity to maintain precision and robustness in a variety of environments.

Ensuring that the system can generalize to new domains and circumstances effectively is another important issue. In controlled or pre-trained environments, cross-modal models often perform well; however, they may not perform as well when applied to real-world circumstances when the data distributions differ significantly from the training set. This problem, referred to as "domain adaptation," could reduce the model's performance in novel or unexpected contexts, which is crucial for real-world uses like robots or surveillance.

## 7.2. Anticipated solutions

Several solutions are suggested to handle these challenges. First, to increase the integration of visual and textual input, advanced fusion approaches including transformer-based models and attention processes will be applied. These methods let the model dynamically balance the value of every modality, harmonizing their contributions under challenging situations. The study will investigate synchronizing methods that match the timing of visual and textual inputs to manage the temporal misalignment, so ensuring that both modalities significantly contribute meaningfully at the appropriate times. The research will include domain adaptation methods, such transfer learning, or data augmentation, so enabling the model to better adjust to new surroundings without significant retraining generalizing across domains. These techniques seek to make the model more robust, adaptable, and relevant in several contexts.

# References

1. Zhou, L., Zhou, Z., Mao, K., & He, Z. (2023). Joint Visual Grounding and Tracking with Natural Language Specification.
2. Zhang, Y., Zhu, J., & Zhang, H. (2022). All in One: Exploring Unified Vision-Language Tracking with Multi-Modal Alignment.
3. Ma, K., Wu, D., & Zhang, H. (2021). Capsule-based Object Tracking with Natural Language Specification.
4. CiteTracker: Correlating Image and Text for Visual Tracking.
5. Li, X., Liu, J., & Wang, Z. (2020). Cross-modal Target Retrieval for Tracking by Natural Language.
6. Yu, Z., Wang, Y., & Huang, T. (2020). Generalizing Multiple Object Tracking to Unseen Domains by Introducing Natural Language Representation.
7. Du, Y., Chen, X., & Wang, L. (2020). iKUN: Speak to Trackers without Retraining.
8. Hong, Y., Li, Y., & Liu, Z. (2020). OneTracker: Unifying Visual Object Tracking with Foundation Models and Efficient Tuning.
9. Feng, J., Wang, X., & Zhang, H. (2020). Real-time Visual Object Tracking with Natural Language Description.
10. Feng, J., Wang, X., & Zhang, H. (2020). Siamese Natural Language Tracker: Tracking by Natural Language Descriptions with Siamese Trackers.
11. Wang, Y., Li, X., & Liu, J. (2020). Towards More Flexible and Accurate Object Tracking with Natural Language: Algorithms and Benchmark.
12. Zheng, Y., Li, J., & Liu, Z. (2020). Towards Unified Token Learning for Vision-Language Tracking.
13. Tracking by Natural Language Specification.
14. Tracking by Natural Language Specification with Long Short-term Context Decoupling.
15. Ma, K., Wu, D., & Zhang, H. (2020). Unifying Visual and Vision-Language Tracking via Contrastive Learning.
16. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl vain Gelly, et al. An image is worth 16x16 words: Trans formers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
17. Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition, pages 15789–15798, 2021.
18. Parminder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. Comparative analysis on cross-modal information retrieval: a review. Computer Science Review, 39:100336, 2021.
19. Chenliang Li, Ming Yan, Haiyang Xu, Fuli Luo, Wei Wang, Bin Bi, and Songfang Huang. Semvlp: Vision-language pre-training by aligning semantics at multiple levels. arXiv preprint arXiv:2103.07829, 2021.
20. Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and lan guage by cross-modal pre-training. In Proceedings of the AAAIConferenceonArtificial Intelligence, volume 34, pages 11336–11344, 2020.
21. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to end object detection with transformers. In ECCV, 2020.
22. Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In CVPR, 2020.
23. Martin Danelljan, Luc Van Gool, and Radu Timofte. Proba bilistic regression for visual tracking. In CVPR, 2020.
24. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll´ ar, and Ross Girshick. Masked autoencoders are scalable vision learners. In CVPR, 2022.
25. Mingzhe Guo, Zhipeng Zhang, Heng Fan, and Liping Jing. Divert more attention to vision-language tracking. In NeurIPS, 2022.
26. Q. Feng, V. Ablavsky, Q. Bai, G. Li, and S. Sclaroff, “Real-time visual object tracking with natural language description,” in WACV, pp. 689 698, 2020.
27. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in ECCV, pp. 213–229, 2020.
28. Z. Yang, Z. Gan, J. Wang, X. Hu, F. Ahmed, Z. Liu, Y. Lu, and L. Wang, “Unitab: Unifying text and box outputs for grounded vision-language modeling,” in ECCV (36), pp. 521–539, 2022.
29. A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, “MDETR- modulated detection for end-to-end multi-modal understand ing,” in ICCV, pp. 1760–1770, 2021.
30. R. Hu and A. Singh, “Unit: Multimodal multitask learning with a unified transformer,” in ICCV, pp. 1419–1429, 2021.
31. Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taix´e, and Bastian Leibe. Hota: A higher order metric for evaluating multi object tracking. International journal of computer vision, 129(2):548–578, 2021.