Web-based Image Annotation

Sejal Nimkar, Student, MMCOE,Pune,India, sejalnimkar2020.comp@mmcoe.edu.in

Shreya Mhetre, Student, MMCOE,Pune,India, shreyamhetre2020.comp@mmcoe.edu.in

Simantini Rembhotkar , Student, MMCOE,Pune,India,

simantinirembhotkar2020.comp@mmcoe.edu.in

Nikita Patil, Student, MMCOE, Pune, India, nikitapatil2020.comp@mmcoe.edu.in

Abstract Grounding DINO and Segment Anything Model a.k.a SAM algorithms, a recent innovation from cutting-edge research, has garnered considerable attention. Trained on extensive datasets comprising diverse annotations, this tool demonstrates an exceptional ability to accurately annotate a variety of objects within images. In the original work, the tool's efficacy was assessed through intricate zero-shot transfer tasks, showcasing its versatility. In the wake of this, numerous studies have sought to explore the tool's performance across diverse scenarios, emphasizing its proficiency in object recognition and segmentation. Furthermore, the tool has become a cornerstone in several projects, seamlessly integrating with other models such as Grounding DINO, Stable Diffusion, ChatGPT, among others. The proliferation of papers and projects revolving around this tool necessitates a comprehensive survey. In response, this work embarks on the first extensive review of the Automated Annotation Tool using Grounding DINO and Sam. As an ongoing project, we commit to regular updates to keep abreast of the evolving landscape. We welcome contributions from the research community to enrich our survey with new findings related to this innovative tool.

Keywords —SAM (Segment Anything Model), DINO, Image Segmentation, Lightweight Model Architectures, Zero-shot Detection, Instance Segmentation, Grounding DINO, Image Processing Efficiency

I. INTRODUCTION

In the past few years, the rapid expansion of digital multimedia content and the growing dependence on machine learning and computer vision applications have highlighted the crucial need for precise and streamlined image and video annotation. Annotation, which involves labeling images and videos, plays a pivotal role in establishing the foundation for building resilient computer vision systems and training effective machine learning models.

The pivotal role of annotation, however, is not without its challenges, as manual annotation processes often prove to be time-consuming and resource-intensive, presenting a bottleneck in the advancement of these technologies.

In response to the pressing need for a streamlined and user-friendly approach to annotation, our research embarks on the development of a comprehensive web-based Image and Video Annotation Platform. The motivation behind this endeavor stems from a commitment to addressing the practical challenges faced by developers, researchers, and practitioners engaged in computer vision projects. We aim to provide a platform that not only simplifies the annotation process but also caters to the diverse needs of users by offering a range of annotation tools and options.

II. PROBLEM DEFINITION

While the motivation for this project is clear, a closer examination of the existing landscape reveals significant challenges. Adapting pre-trained models, such as ResNet50, for evolving image annotation projects poses a notable hurdle, often requiring time-consuming retraining from scratch. The need for abundant labeled data exacerbates this challenge, emphasizing a significant gap in existing methodologies. In order to grasp the present state of cutting-edge techniques and their constraints, a comprehensive review of the literature was undertaken. This survey unveiled both progress and deficiencies in recent research endeavors. Notable among these are approaches like "Automated Image Annotation With Novel Features Based on Deep ResNet50-SLT" and "DeepAIA: An Automatic Image Annotation Model Based on Generative Adversarial Networks and Transfer Learning," which showcase innovative features and combinations of techniques but also acknowledge challenges in adaptation and dataset size.

As we delve into the research gap and identified challenges, our exploration of innovative algorithms like SAM (Segment Anything Model) and Grounding DINO will pave the way for addressing these limitations, contributing to the advancement of image and video annotation methodologies.

III. RELATED WORK

There have been much work proposed and acquired for images, videos and audio annotation till date. Some of them has been used for different use cases or applications. For

 DOI : 10.35291/2454-9150.2024.0084

example, the paper [2] presents an innovative semantic image annotation framework designed for generating natural language descriptions within a controlled setting. For object detection in sports images, a hybrid technique combines deep learning models with aligned annotations from ontology classes, employing a residual network and a feature pyramid network. By combining contextual information gathered from a knowledge base, this distinctive methodology advances further probabilistic identification of images. It employs a multi-tier system which takes a bottom-up approach, including an AI layer that uses RetinaNet with ResNet-50 and FPN to recognize objects in photos which is used to identify generic objects and their coordinates. The merging layer compares coordinates and probabilities to merge findings from two models, allowing for a dual-perspective examination of sports ontology annotation. One point of view makes use of generic objects for specialized annotation, whereas the other makes use of specialized classes. The mapping layer examines recognized items and ontology findings to generate accurate image annotations. For smooth integration, a RESTful service maintains input/output parameters. This multi-tiered technique provides a comprehensive framework for semantic image annotation.

[3] The effectiveness of the framework's object detection and prediction capabilities was evaluated using two diverse datasets. The first dataset consisted of individual instances with everyday scenes featuring common objects, while the second custom dataset focused on sports images gathered from the web. To enhance model understanding of object interactions and relationships within images, we meticulously annotated objects, attributes, and relationships.

Notably, the performance of object detectors varies based on where they are applied in an image. Challenges arise when the object doesn't completely fill the box boundaries, leading to the inclusion of various samples from other parts of the image. This complexity makes it challenging for our framework to accurately detect the color of the object. During analysis, an invalid coefficient was identified when examining correlations between variable pairs, particularly involving the generic object variable.

The annotation results obtained from both pre-merging and post-merging stages included correct sentences with all the expected objects. Surprisingly, even with two generic objects detected and predicted by the pre-trained model, and two false objects predicted by the custom-trained model, the expected objects were found in the annotation results with 100.0 percent accuracy.

To address the semantic gap between computer-generated characteristics and human interpretation, we propose the integration of an Automatic Image Annotation (AIA) system. This system aims to enhance the interpretability of our framework's outputs, facilitating a more seamless understanding of the visual content by bridging the gap between computational features and human perception.

[4] The AIA framework ResNet50-SLT combines ResNet50, slantlet transform, word2vec, and principal component analysis with t-distributed stochastic neighbor embedding. This integration allows raw image pixels to be converted into meaningful semantic concepts, enhancing information retrieval. For sentence generation based on feature vectors, the AIA system use seq2seq. When tested on popular datasets (Flickr8k, Corel-5k, ESP-Game), the proposed framework outperforms existing approaches in terms of flexibility, accuracy, and computing cost. The study's contributions are in overcoming temporal challenges during large dataset training, giving a valuable tool for efficiently choosing and extracting picture features, and finally permitting exact image annotation and retrieval from enormous databases.

[5] In the realm of video scene text detection, a critical challenge arises from the dearth of annotated scene text video datasets, which are often limited in scale and lack instances that truly challenge detection algorithms. To surmount this hurdle, an innovative tracking-based semi-automatic labeling strategy is introduced in this research. The proposed methodology involves initial manual labeling for the first frame, followed by automatic tracking for subsequent frames, thereby mitigating the resource-intensive nature of manual annotation. Furthermore, to address the deficiency in existing datasets, a novel low-quality scene text video dataset named Text-RBL is presented. This dataset encompasses raw, blurry, and low-resolution videos, all labeled using the devised semi-automatic labeling strategy. To gauge the efficacy of Text-RBL, a baseline model is proposed, integrating a text detector and tracker for video scene text detection. An additional contribution is a failure detection scheme designed to handle intricate scenes. The experimental results underscore the value of Text-RBL, showcasing that the incorporation of low-quality labeled videos enhances the performance of the text detector, particularly in challenging low-quality scenes.

[7] This paper introduces an inventive approach to tackle the challenge of annotating concepts in videos and images using a deep convolutional neural network (DCNN) architecture. The method capitalizes on the relationships between concepts at two distinct levels to enhance the annotation process.

At the first level, drawing inspiration from multi-task learning, a novel strategy is presented for learning concept-specific representations. These representations are crafted to be sparse, linear combinations of latent concept representations. The encouragement of sharing these latent concept representations serves to exploit the implicit relationships between the target concepts.

Taking the approach further, at the second level, insights from structured output learning are incorporated. During training, a new cost term is introduced that explicitly models

 DOI : 10.35291/2454-9150.2024.0084

the correlations between concepts. This explicit modeling allows for the capture of the structure within the output space, specifically the relationships among concept labels.

Both of these levels are seamlessly integrated into a unified DCNN architecture, utilizing standard convolutional layers. The entire network is trained end-to-end using standard back-propagation, providing a cohesive solution to the problem at hand.

Experiments conducted on four extensive video and image datasets demonstrate the effectiveness of the proposed DCNN. Significant improvements in concept annotation accuracy are observed compared to state-of-the-art methods in the field. This underscores the potential of the approach in advancing the state of the art in video/image concept annotation.[7]In the realm of computer vision and deep learning for autonomous driving applications, the establishment of accurate ground-truth annotations is a fundamental prerequisite. While existing public datasets predominantly capture urban driving scenarios, there exists a notable underrepresentation of countryside roads and nighttime driving conditions. Addressing this gap, this paper introduces a novel semi-automated method for bounding box annotation tailored specifically for nighttime driving videos. The proposed three-step approach encompasses (a) the generation of trajectory proposals using a tracking-by-detection method, (b) the extension and verification of object trajectories through single-object tracking, and (c) the formulation of an efficient pipeline for semi-automatic annotation of object bounding boxes in video sequences. To assess the effectiveness of this methodology, the CVL dataset, which focuses on nighttime driving conditions in European countryside roads, serves as the experimental benchmark. The results highlight notable improvements achieved at each processing step, revealing a substantial 23 percent increase in recall, while maintaining almost constant precision when compared to the initial tracking-by-detection approach. This innovative approach contributes to the domain of semi-automatic video annotation, particularly in the context of autonomous vehicles navigating nighttime environments, and showcases promising advancements in visual object tracking.

IV. PROPOSED SYSTEM ARCHITECTURE

Meta published a paper [1] introducing SAM, or Segmentation with Prompt-based Attention Model, which is a pioneering image segmentation architecture designed for efficiency and flexibility. The model features an image encoder that generates a one-time embedding for the input image, capturing its contextual information. To enable real-time adaptation, a lightweight encoder is incorporated, converting any prompt into an embedding vector on-the-fly. This allows dynamic adjustment of the model's focus based on user-provided prompts.

The core of SAM lies in its lightweight decoder, where the image and prompt embeddings are synergistically fused to predict accurate segmentation masks. The prompt-based attention mechanism enhances the model's ability to respond to specific user instructions, making it versatile across diverse segmentation tasks. This innovative design not only ensures computational efficiency but also facilitates seamless integration into various applications where real-time, prompt-guided image segmentation is essential. SAM represents a significant advancement in the intersection of image processing, prompt-based techniques, and lightweight model architectures.

[10] DINO, a novel self-supervised system developed by Facebook AI, demonstrates an impressive ability to extract meaningful representations from unlabeled data. The key innovation of DINO lies in treating self-supervision as a unique form of self-distillation, eliminating the need for any labeled data in the learning process.

In the DINO framework, a student network is trained by aligning its output with that of a teacher network across various perspectives of the same image. During training, images undergo a specific cropping process, incorporating both local views (small crops covering less than 50 percent of the image) and global views (large crops covering more than 50 percent of the image) for the student model. Notably, the teacher model is exposed solely to global views.

The teacher, functioning as a momentum teacher, adjusts its weights based on an exponentially weighted average of the student model. The training objective is for the teacher model to predict high-level features, and the student model aims to replicate this prediction using cross-entropy to align the two distributions.

In simpler terms, the student is essentially learning to imitate the teacher's output. The unique aspect of self-distillation in DINO is that the student model learns from its own predictions, eliminating the need for external labels. This distinctive self-supervised learning approach makes DINO stand out as a method that doesn't rely on manual data annotation.

This innovative self-supervised learning technique opens up new possibilities for representation learning without the need for labeled datasets, showcasing the potential of DINO in advancing the field.

Grounding DINO [9] and the Segment Anything Model (SAM) stand out as cutting-edge solutions that significantly enhance the efficiency of image processing. Grounding DINO, in particular, excels in zero-shot detection, showcasing its ability to identify any object within an image without the need for prior training on specific object classes. This unique feature sets Grounding DINO apart by providing a versatile and rapid solution for object detection.

On the other hand, SAM complements the capabilities of Grounding DINO by transforming the generated bounding boxes into precise instance segmentation masks(refer to Fig. 2). SAM's strength lies in its capacity to take the identified object boundaries and produce detailed masks that precisely

 DOI : 10.35291/2454-9150.2024.0084

outline each object instance within the image. This step is crucial for applications requiring a more granular understanding of the scene.

Together, Grounding DINO and SAM form a powerful duo, streamlining the image processing pipeline. Grounding DINO's proficiency in zero-shot detection and SAM's ability to convert bounding boxes into segmentation masks create a seamless workflow that can be invaluable for tasks such as object recognition and scene understanding. These models collectively contribute to the advancement of image analysis, offering efficiency and accuracy in the extraction of valuable information from visual data.

This Model shall be used for image annotation in the proposed system

V. METHODOLOGY

A. Training Procedure

The methodology for training SAM (Segmentation Annotation Model) adopts a dual approach to ensure proficiency in generating accurate segmentation masks for diverse prompts. This training procedure is meticulously designed to optimize SAM's segmentation capabilities.

The Pretraining Task serves as the foundational phase, where SAM acquires essential knowledge. This initial step establishes a base for subsequent learning, providing the model with fundamental insights crucial for effective image annotation. The knowledge gained during pretraining forms the bedrock upon which SAM's interactive learning process is built.

The Interactive Data Collection phase is a pivotal element of SAM's training. This iterative process involves annotators interacting with SAM in real time, guiding the model to refine its capabilities according to the specific requirements of image annotation tasks. SAM's adaptability is honed through this interactive learning, ensuring it aligns more closely with intended annotation objectives.

Real-time functionality is paramount for practical utility. The emphasis on achieving Real-time Functionality underscores SAM's commitment to responsiveness. This feature enables seamless and interactive annotation, even when executed on a CPU within a web browser. SAM's ability to deliver results in real-time enhances its practical applicability in various contexts, where dynamic and interactive annotation is essential.

B. Performance Evaluation

Segmentation Accuracy, Efficiency, and Generalization: SAM's performance is comprehensively evaluated in terms of accuracy, efficiency, and generalization across diverse tasks and domains. This rigorous assessment ensures SAM's effectiveness in various scenarios, providing insights into its segmentation capabilities in real-world applications.

In handling ambiguity, SAM's capabilities are scrutinized to assess its Handling Ambiguity. This involves evaluating SAM's capacity to generate multiple valid masks in situations where prompts may be ambiguous. SAM's versatility shines as it demonstrates its ability to produce accurate masks even in uncertain scenarios, showcasing its robustness.

Examining Real-time Segmentation is crucial to understanding SAM's efficiency in dynamic scenarios. SAM's capacity to segment objects in real-time after pre-computing image embeddings is assessed. This evaluation highlights SAM's responsiveness and efficiency, confirming its suitability for applications requiring immediate and dynamic segmentation tasks.

C. Bias Analysis

To ensure fairness and equity in SAM's performance, a thorough Demographic Bias Analysis is conducted. This involves scrutinizing potential biases across perceived gender presentation, perceived skin tone, and perceived age range. The objective is to identify and rectify any biases, making SAM's performance consistent and unbiased across different demographic groups.

D. Application Scenarios:

Exploring potential applications of SAM in various domains is crucial. SAM's versatility is highlighted as it is envisioned to play a pivotal role in Domain-specific Applications. From AR/VR to content creation, scientific research, and general AI systems, SAM's applicability is discussed. Specific scenarios, such as identifying objects via AR glasses, assisting farmers, and aiding biologists in their research, are explored to showcase SAM's diverse range of applications.

E. Algorithm:

SAM (Segment Anything Model), DINO (Differentiable Neural Optimizer), and Zero-Shot Detection

Segment Anything Model (SAM): SAM, as a foundational model for image segmentation, employs prompting techniques for versatile annotation. The model's architecture comprises key components: Image Encoder, Lightweight Encoder, and Lightweight Decoder. SAM's Training Approach is detailed, emphasizing the significance of the pretraining task, interactive data collection, and real-time functionality. This comprehensive strategy optimizes SAM's segmentation capabilities across diverse annotation requirements.

Grounding DINO (Differentiable Neural Optimizer): Grounding DINO enhances SAM's object recognition and segmentation capabilities. The Applications section illustrates the seamless integration of Grounding DINO with SAM, contributing to the overall improvement of annotation processes. The unique contributions of Grounding DINO to SAM's performance are expounded, showcasing the synergy between these advanced algorithms.

Zero-Shot Detection: Zero-shot detection is a pivotal aspect of SAM's capabilities. This technique allows SAM to recognize and annotate objects even in scenarios where it has not encountered them during training. The essence of Zero-

 DOI : 10.35291/2454-9150.2024.0084

Shot Detection lies in SAM's ability to generalize across new tasks and domains without the need for specific training on those tasks. SAM's zero-shot capabilities empower it to handle diverse and evolving annotation projects, making it a versatile tool for real-world applications.

In practical terms, zero-shot detection means that SAM can accurately segment objects in images or videos, even if it has never been explicitly trained on those specific objects. This is achieved through SAM's promptable design, allowing it to generate valid segmentation masks for any prompt, whether it be foreground/background points, bounding boxes, or freeform text. SAM's proficiency in zero-shot detection significantly expands its applicability, making it adept at handling unforeseen annotation tasks without the need for extensive retraining. This aspect of SAM's algorithmic prowess contributes to its effectiveness in addressing the challenges posed by evolving image annotation projects.

VI. CONCLUSION

Conclusively, the Automated Annotation Tool employing Grounding DINO and Sam algorithm stands as a pivotal advancement in the field of automated image annotation. Through its adeptness in accurately annotating diverse objects in images, it has proven to be a versatile and powerful tool. The extensive evaluation in zero-shot transfer tasks and its seamless integration with various models underscore its adaptability and potential for widespread applicability. As our survey reveals, the tool has not only sparked numerous studies exploring its performance across different scenarios but has also become an integral component in diverse projects. The ongoing commitment to regular updates ensures that our comprehensive survey remains a valuable resource in navigating the ever-evolving landscape of this innovative tool. We encourage researchers and practitioners to engage with this survey and contribute new insights, fostering a collaborative and informed community dedicated to advancing automated image annotation technology.

ACKNOWLEDGMENT

This research addresses a problem suggested by Equations work (https://eqw.ai/)

REFERENCES

[1] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment Anything. arXiv preprint arXiv:2304.02643.

[2] Sezen, A., Turhan, C., & Sengul, G. (2021). A Hybrid Approach for Semantic Image Annotation. IEEE Access, 9, 131977-131994.

[3] Liang, H., Wang, F., Bing, L., Yu, D., & Wang, J. (2021, October). Automatic annotation algorithm based on sliding window moment feature matching. In 2021 International Conference on Computer Engineering and Application (ICCEA) (pp. 216-219). IEEE.

[4] Adnan, M. M., Rahim, M. S. M., Khan, A. R., Alkhayyat, A., Alamri, F. S., Saba, T., & Bahaj, S. A. (2023). Automated Image Annotation With Novel Features Based on Deep ResNet50-SLT. IEEE Access, 11, 40258-40277.

[5] Huang, R., Zheng, F., & Huang, W. (2021). Multilabel Remote Sensing Image Annotation With Multiscale Attention and Label Correlation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 6951-6961.

[6] Lin, J., Yu, T., & Wang, Z. J. (2022). Rethinking Crowdsourcing Annotation: Partial Annotation With Salient Labels for Multilabel Aerial Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-12.

[7] Markatopoulou, F., Mezaris, V., & Patras, I. (2019). Implicit and Explicit Concept Relations in Deep Neural Networks for Multi-Label Video/Image Annotation. IEEE Transactions on Circuits and Systems for Video Technology, 29(6), 1631-1644.

[8] Adnan, M. M., Rahim, M. S. M., Khan, A. R., Saba, T., Fati, S. M., & Bahaj, S. A. (2022). An Improved Automatic Image Annotation Approach Using Convolutional Neural Network-Slantlet Transform. IEEE Access, 10, 7520-7532.

[9] Zhu, J., Jiang, X., Jia, Z., Xu, S., & Cao, S. (2021). Tracking Based Semi-Automatic Annotation for Scene Text Videos. IEEE Access, 9, 46325-46338.

[10] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., ... & Zhang, L. (2023). Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv preprint arXiv:2303.05499.

[11] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. arXiv preprint arXiv:2104.14294.

[12] Schörkhuber, D., Groh, F., & Gelautz, M. (2021, September). Bounding Box Propagation for Semi-automatic Video Annotation of Nighttime Driving Scenes. In 2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA) (pp. 131-137). IEEE.

[13] Le, T. N., Akihiro, S., Ono, S., & Kawasaki, H. (2020, March). Toward Interactive Self-Annotation For Video Object Bounding Box: Recurrent Self-Learning And Hierarchical Annotation Based Framework. In 2020 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 3220-3229). IEEE.

[14] Hosseinnia, M., & Behrad, A. (2023, February). 3D Image Annotation using Deep Learning and View-based Image Features. In 2023 6th International Conference on Pattern Recognition and Image Analysis (IPRIA) (pp. 1-6). IEEE.

 DOI : 10.35291/2454-9150.2024.0084