Assistive Image Annotation Systems with Deep Learning and Natural Language Capabilities: A Review

Moseli Mots'oehli

Department of Information and Computer Science University of Hawai'i at Manoa Honolulu, USA moselim@hawaii.edu

Abstract-While supervised learning has achieved significant success in computer vision tasks, acquiring high-quality annotated data remains a bottleneck. This paper explores both scholarly and non-scholarly works in AI-assistive deep learning image annotation systems that provide textual suggestions, captions, or descriptions of the input image to the annotator. This potentially results in higher annotation efficiency and quality. Our exploration covers annotation for a range of computer vision tasks including image classification, object detection, regression, instance, semantic segmentation, and pose estimation. We review various datasets and how they contribute to the training and evaluation of AI-assistive annotation systems. We also examine methods leveraging neuro-symbolic learning, deep active learning, and self-supervised learning algorithms that enable semantic image understanding and generate free-text output. These include tasks such as image captioning, visual question answering, as well as multi-modal reasoning. Despite the promising potential, there is limited publicly available work on AI-assistive image annotation with textual output capabilities. We conclude by suggesting future research directions to advance this field, emphasizing the need for more publicly accessible datasets and collaborative efforts between academia and industry.

Keywords—AI-assisted Annotation, Computer Vision, Textual Hint Generation, Visual Question Answering, Image Captioning, Multi-modal Learning

I. INTRODUCTION

Deep Learning (DL) models have seen considerable success in Computer Vision (CV) tasks such as image classification, instance segmentation, Visual Question Answering (VQA), pose estimation, action recognition, and more [1]–[3]. A large proportion of this success can be attributed to the availability of large collections of annotated training data [4], advances in Graphics Processing Unit (GPU) technology, and advances in DL model architectures [5]–[7]. The time and financial costs associated with acquiring high-quality human annotations for large image datasets can be significant due to the dataset's scale and the need for expert annotation. Such is the case for most medical imaging datasets [8], or the cost of acquisitions of the images, as is the case with satellite imagery [9]. These challenges pose some problems in developing real-world computer vision applications that rely on DL models,

thus methods are needed to reduce the error rate, time, and financial cost inherent in acquiring training annotations for DL models.

One solution to these costs is to minimize the need for large volumes of annotated data. While techniques like selfsupervised pre-training and transfer learning can help, many applications still require labeled data, especially for unique or specialized datasets. Training models on domain-specific labeled data or pre-training on related datasets can capture domain-specific nuances. and improving generalization [10], [11]. With the advances in Self-Supervised Learning (SSL), Active Learning (AL), Few Shot Learning (FSL), and Multimodal Learning (MML), several Artificial Intelligence (AI) assistive annotation systems have been introduced to speed up and augment the manual annotation process, reduce the cost of acquiring annotations, and improve the quality of the annotations by eliminating or flagging annotation errors as they occur, or by guiding the annotator using a combination of large language models and image understanding. These systems enhance annotation efficiency and accuracy, enabling faster DL model development and deployment in real-world applications.

To this end, this paper explores the literature on AI-assistive image annotation systems for DL models in CV tasks, with a specific interest in systems with DL or neuro-symbolic generated textual hints, descriptions, or reasoning. Textual guidance in image annotation helps annotators understand the underlying model's reasoning and focus on aspects that require human expertise, improving the overall annotation efficiency. We show an example of an AI-assistive annotation system architecture with text suggestions in Figure 1. We compare and contrast techniques, highlighting methods from SSL, AL, FSL, and MML. Our analysis focuses on how each system addresses the challenges of annotation costs, speed, accuracy, and clarity to the annotator. We discuss evaluation metrics and benchmarks for annotation system performance, application areas, challenges, benefits, and real-world impact. We conclude this review by summarizing the state of AI-assistive image annotation systems with natural language capabilities and, explore potential avenues for future research in this area.

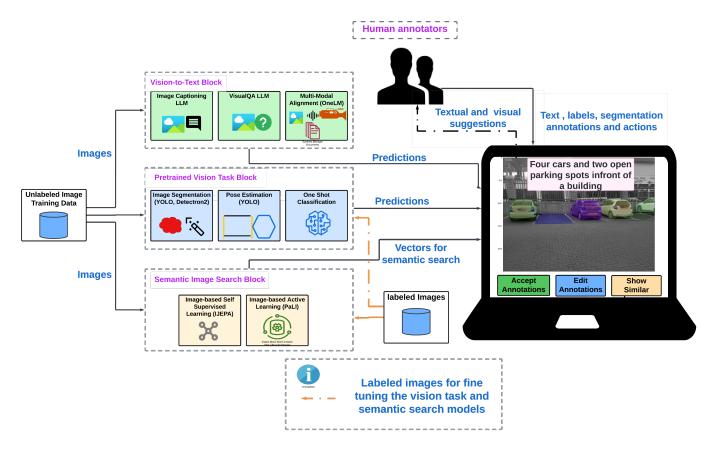


Fig. 1. An overview of an AI-assisted image annotation system. The system begins with unlabeled image training data which is processed through various blocks. The **Vision-to-Text Block** utilizes image captioning, VQA, and multi-modal alignment to provide predictions. The **Pretrained Vision Task Block** handles image segmentation, pose estimation, and one-shot classification to generate vision task predictions. The **Semantic Image Search Block** uses self-supervised learning and active learning to assist the annotator in semantic search. Human annotators receive textual and visual suggestions to annotate the images, which are then used to fine-tune the vision task and semantic search models. The final interface allows annotators to accept, edit, or show similar annotations.

II. FOUNDATIONS OF AI-ASSISTED ANNOTATION

In this section, we briefly outline DL as it relates to computer vision tasks as well as the role of annotation in Supervised Learning (SL). We briefly provide an overview of image captioning, VQA, MML, GPT style models, and Neuro-Symbolic Learning (NeSyL). These methods share a common theme: their ability to relate images to textual descriptions. Finally, we discuss closely related surveys to this work to provide a broader context.

A. Deep Learning for Computer Vision Tasks

Just as humans and most animals use their eyes to perceive the world for navigation and interacting with objects around them, CV is a subfield of computer science focused on creating hardware and software to assist computers in visual perception and understanding [12]. We are interested in machines with visual understanding because they can then be programmed to take actions in response to specific visual information. Many classical methods [13], [14] relied on human-engineered extraction of important features. Most real-world applications of image-based Machine Learning (ML) are trained using SL, an approach to learning that, unlike unsupervised learning

[15], requires both the input images and the target annotations. The different CV tasks get their names from the type of their target annotations as described in Section II-B.

Convolutional Neural Networks (CNNs) [16] have been the dominant DL model architecture for learning effective image representations. They accomplish this by emphasizing the spatial relationships between neighboring pixels. By stacking convolutional layers, more complex patterns can be learned. More recently, the Vision Transformer (ViT) [7] architecture uses image patches and avoids convolutional layers entirely, relying solely on multi-head self-attention to process visual data. ViTs have many training parameters, requiring substantial amounts of accurately labeled data to train without the risk of overfitting. The need for large volumes of accurately labeled data poses challenges that necessitate manual annotation, which is time-consuming, expensive, and sometimes leads to annotation errors due to factors such as fatigue, lack of focus, lack of adequate training, or just irreversible accidental mouse clicks by the annotators [17]. These errors can corrupt the training and test datasets, affecting both the training process and the reliability of the test performance evaluations.





Image belongs to Cat class

Image belongs to Dog class

Fig. 2. [Source: Image Classification]: An image depicting image classification with 2 classes, a cat and a dog. The predictions are in the form of probabilities that are then mapped to the class labels based on the highest probability.

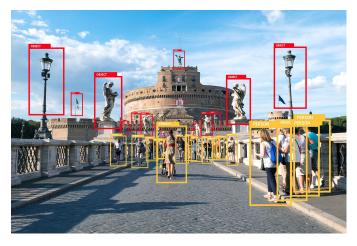


Fig. 3. [Source: Object Detection] results on a city street showing multiple instances of humans, statues, and lights. The bounding boxes for each object are regression prediction outputs for the rectangular coordinates around each region of interest

B. Image Annotation

While every CV task involves a unique annotation process with its own challenges, we limit our focus to annotation for object detection, image classification, regression, instance segmentation, and pose estimation. We outline the annotation process in each task, factors that influence the difficulty, duration, and cost of annotation, as well as potential types of errors. Figures 2, 3, 4, and 5, display annotated images representing four of the five CV tasks (excluding regression).

Image Classification: Most image classification labels consist of one or more class labels for an entire image based on its perceived content 2. Labeling for classification can be negatively affected by subtle perceived differences between classes, requiring annotators to have expert-level training to



Fig. 4. [Source: Instance segmentation] results of multiple cars detected and segmented at different distances from the viewpoint. Segmentation models typically provide bounding boxes and class labels for each detected object. Predictions usually consist of K binary masks of size $n \times m$, outlining the pixel locations for all the K detected objects in $n \times m$ image.



Fig. 5. [Source: [18]Example Pose Estimation] annotations with the head, neck, shoulders, elbows, hips, knees, and ankles as key-points

accurately discriminate between them. A large number of classes can also negatively affect annotation accuracy. Labeling is generally faster and less expensive compared to other CV tasks. The most common error in classification labeling is misclassification, even in high-quality datasets such as ImageNet [19]. Regression: In regression, annotators assign real continuous values to images, representing measures like length, height, or depth of the object of interest (OI). The annotation cost and complexity in regression are often dictated by the difficulty inherent in measuring the quantity of the OI, and this is normally done on the actual object and not the image. For example in fish stock estimation, the annotators are fishermen on a fishing vessel who measure the size of a live fish using a measuring board before capturing a picture of the fish and recording its species [20]. Annotation errors include inaccurate measurements, and typographical errors such as recording 100cm when the actual measurement is 10cm. Object Detection: In object detection, the goal is to identify the presence of particular objects within an image 3. During the annotation process, annotators create rectangular bounding boxes around objects of interest, and may also assign a class to each object for classification purposes. Annotation for object detection can be challenging due to factors such

as very small object sizes, occlusion by other objects that are not relevant to the task, and overlapping OIs. The assignment of bounding boxes can take longer and be more expensive than annotation for classification or regression as there is a need for tight bounding boxes, that consist of four data points per object. Typical errors include inaccurate bounding boxes, mislabeled object classes, and missing bounding boxes around some objects of interest.

Instance Segmentation: In segmentation annotation, humans manually label each pixel in an image according to the object it belongs to. The goal is to precisely outline the boundaries of each object, enclosing it as tightly as possible. The complexity of this task varies widely based on factors such as the shape, size, and quantity of objects within the image, as well as the extent to which these OIs are obscured by background objects. Due to the increased complexity, Segmentation annotation typically takes longer and is more costly. The most common errors are missing segmentation masks for some OIs, inaccurate segmentation boundaries, or mistakenly segmenting background objects as OIs. **Pose Estimation:** The common goal in pose estimation is detecting the position and orientation of a person or OI by identifying key points or joints 5. An example annotation involves marking and connecting the dots representing a person's head, neck, shoulders, arms, waist, knees, and feet in motion [18]. Annotation in pose estimation is negatively affected by object occlusion, the complexity of the person's pose, image and OI size, unexpected orientational changes as well as difficulties due to a bad camera viewpoint [21]. The annotation process is not as complex and costly as instance segmentation annotation, but can often be more challenging than classification and regression. Common annotation errors include missing or inaccurate key points and joint markers, as well as misaligned key point markers, for instance, placing one eye key point marker slightly higher on the face than the other.

C. Textual Description of Images

Since this work focuses on AI assistive annotation systems with natural language hint generation or reasoning, it's important to establish existing methods for learning imageto-text mappings. DL and NeSyL offer various approaches. Two prominent DL image-to-text tasks are image captioning and VQA. In **Image Captioning**, a CNN is typically used as an encoder, extracting a compressed representation of the image. This is then fed to a Recurrent Neural Network (RNN), typically a Long Short-Term Memory (LSTM) network [22], which decodes the image representation and generates a textual description. These networks are trained jointly in an encoder-decoder fashion. VQA systems build upon image captioning in that the encoder receives both the image and the question text. To achieve this, the encoder combines a CNN for image representation learning and an LSTM for learning the hidden representation of the input text. The decoder, also an LSTM-like network, generates the response one token/word at a time. However, with the introduction of the ViTs and standard NLP transformer, both image and text encoders, along with the text decoder, can be built using a single architecture for image captioning and VQA [23].

Multi-modal Learning (MML) is a DL framework that bridges the gap between image and text representation learning by processing images and the accompanying text data jointly [24]. By doing so, the model can leverage the strengths of each data modality in such a manner that the image tokens provide context for the question, and the question textual tokens guide the model to attend to specific tokens of the image in producing the response or caption. Training a joint image and text transformer architecture reduces the complexities that come with setting up and training a mix of CNN and LSTM encoders and decoders. Being able to analyze and visualize the attention maps of the Multi-modal learner can also aid in the interpretability of the question in a VQA setting or understanding of the generated textual descriptions with respect to the visual attention maps [25], Current state-ofthe-art MML models can handle images, audio, text, video as well as point cloud inputs [26], [27]. This can be useful in developing AI assistive image annotation systems with language generation capabilities.

NeSyL image-to-text systems [28] combine neural image representation learning with high-level symbolic representations to model image content. This is achieved by representing the objects, actions, attributes, and abstract concepts contained in an image as nodes of a knowledge graph, and the edges representing the relationships between them. For instance, an image of a boy wearing red shoes kicking a ball would have a corresponding knowledge graph with nodes "boy", "shoes", "ball", "wearing", "kicking", "red", and edges explaining the semantic, logical and spacial relations between nodes. For instance, a semantic relationship "is a" can connect two nodes "red" and "color" or "boy" and "person". This involves utilizing images, question text, and a knowledge graph as inputs to generate the expected output text. This is achieved by adjusting the DL model weights and the importance weights within the knowledge graph. Since NeSyL relies on logical rules and established domain knowledge, it minimizes the likelihood of producing text that contradicts common sense or fundamental physics laws. More recent works in NeSyL with image-to-text outputs, (Captioning, VOA, reasoning), that are potentially applicable to AI assistive image annotation with textual hints include [29], [30].

D. Related Surveys

To highlight the significance of this survey and distinguish it from prior works, we present a summary of topics covered in related surveys and identify areas that have not been addressed as they relate to this work. The most relevant survey, by Tousch et al. [31], explores prior research on semantic hierarchies for image annotation. Given that this work predates the breakthroughs in NLP and vision transformers and effective transfer learning from large language models (both appearing later in the decade), it's understandable that the authors focus on literature using structured vocabularies to

describe image content for automatic annotation. Structured hierarchies of the semantics of an image are used to construct a semantic network, similar to a knowledge graph. This approach models the task of describing image contents in a NeSyL fashion. The survey is limited to annotation for CV tasks such as classification or image captioning where one or more words from a fixed structured vocabulary suffice for a valid annotation. This is not the case for image annotation with textual cues for tasks such as pose estimation, instance segmentation, and regression. In [32], Sager et al. provide a comprehensive survey of image annotation for computer vision applications. By grouping annotation software into manual, semi-automated, and fully automated, show different methods used, their strength, and weaknesses. They highlight the use of clustering and transfer learning in semi-automated and fully automated annotation. However, their review does not cover the use of NLP models to generate free text and hints for AI assistive annotation that we cover in this work.

The authors of the survey [33] follow a similar approach to [32], but focus only on annotation software for medical imaging, specifically the graphical user interface (GUI), and component tools of the software meant to make annotation easier. Still closely related to [32], the authors of [34] focus only on Automatic Image Annotation (AIA). Their survey groups methods into five broad categories: generative, nearest neighbors, discriminative, tag completion, and DL-based methods, based on how the annotations are automatically generated from the images. They compare the five categories of AIA based on computational complexity, time, and annotation accuracy. Similarly, this work does not address the combined use of image and text-based DL models for assistive annotation through hint generation or text descriptions. Notable surveys on Human-inthe-Loop (HITL) and human-computer joint exploration are [35] and [36]. The first survey [35] reviews HITL in ML, emphasizing humans as domain experts throughout the data pipeline, from collection and annotation to model training and deployment. Our survey, however, focuses on image and textbased DL models as primary agents in annotating data for CV tasks, unlike [35], where humans play the central role. The second survey [36] explores literature related to multimedia tagging. The covered methods address assistive and automated assignment, recommendation, and organization of keywords to multimedia files for internet retrieval. However, these methods are limited to keyword selection from a dictionary and do not include free text hint generation or description in natural language leveraging DL models. Additionally, they do not cover assistive annotation in other CV tasks such as instance segmentation, pose estimation, and VQA beyond keyword tagging.

III. TYPES OF ASSISTIVE DEEP LEARNING ANNOTATION SYSTEMS

In this section, we explore the literature on the broad systems for assistive image annotation with the help of DL, highlighting their key characteristics, limitations, and strengths. These systems leverage DL models to help human annotators in the image annotation process by generating textual hints, tags, descriptions, or logical steps. We cover Deep Active learning-based methods, self-supervised, Semi-supervised learning-based annotation systems, and Human-in-the-loop annotation platforms. We focus on the feature extraction process from input images and how it translates into textual guidance for the CV annotation task at hand.

A. Deep Active Learning-based Systems

Deep Active Learning (DAL) seeks to train the bestperforming model with as little annotated data as possible, by iteratively, and strategically selecting the most informative samples based on a DL image model, for annotation by a human annotator [37], [38]. In this setting, the assistive part of the annotation process is in the form of a DL model selecting the candidate images for annotation, as opposed to selecting potentially redundant samples that do not improve the quality of information in the annotated data. Most DAL methods rely on the extracted image features [39], [40], prediction probability [41], or training dynamics [42], [43] for sample selection. Despite the extensive literature on both CV, and NLP tasks, systems leveraging AI-assisted image annotation methods that leverage DAL, and generate textual outputs remain scarce. This is likely due to limited research in this area, or the high monetization potential of such systems, resulting in proprietary industry efforts and breakthroughs remaining unpublished in scholarly articles.

Focusing on commercial products, [44] offers automated annotations based on DAL for the following CV tasks: singlelabel image classification, semantic segmentation, and object detection. They also handle single-label text classification. They however do not have any cross-modal, VGA, or image captioning capabilities. Roboflow [45], similarly covers only a few CV tasks by leveraging large pre-trained vision models but also falls short when it comes to image-to-free-text image descriptions for annotation assistance. Labelbox [46], offers AI-Assisted auto annotation as a service covering more input modalities such as images, text, video, Geospatial data, audio, and multiple document formats. They handle most CV tasks: classification, segmentation, cloud point prediction, and object detection, as well as text-to-text annotations for NLP-type tasks. However the system is not cross-modal, it neither offers assistive annotation features for image-to-text tasks such as image captioning, and VQA, nor does it provide text hints or suggestions for image inputs that could lead to higherquality annotations. HumanSignal [47] has increased annotator efficiency by a factor of 1.2 on the number of annotations per oracle in medical imaging through DAL-based AI-assistive annotation. Similar to [46], HumanSignal handles most input data modalities and provides assistive annotations for most CV tasks, but like SageMaker [44], HumanSignal only offers single-label image classification suggestions during the DAL cycle for assistive annotation.

HumanSignal also provides a quasi-VQA/captioning search functionality on their platform that allows the user to search the unlabeled image dataset based on predefined queries such as "find similar". While users might interpret these predefined text-based searches as a sign of natural language comprehension, the interface button most likely triggers a pre-programmed function that analyzes uploaded images and selects similar ones based on the DAL model's learned visual features. Neither the DAL implementations nor underlying vision models (CNN, ViT), for these systems are disclosed. However, assuming the class of suitable DAL algorithms at the scale these companies operate at dictates training and inference efficiency in the DAL setting as well as the interactive nature of suggestive annotation, one can rely on existing literature around the complexity and performance balance of DAL algorithms for a reasonable approximation of the underlying methods. Telus International, CloudFactory, Encord, Datagym and Scale AI [48]-[52] respectively cover more varied CV tasks over and above the previously stated methods at different levels of data specificity for DAL. These include keypoint pose estimations, video action recognition, autonomous driving road signs, and vehicle image mapping. However, again these systems assume adequate knowledge about the annotation task by the annotator, hence the automated assistive annotation is visual in nature, and formatted for the output of the CV task, not textual assistive hints that are capable of instilling new knowledge. While most of these tools are primarily based on DAL, they likely use pre-trained CNNs, ViTs, and rely on SSL pre-training on the input images. In the next Section, we look at methods based on SSL as well as weak supervision for AI assistive annotation with NLP capabilities.

B. Self-supervised, Semi-supervised Learning-based Systems

Self-supervised learning(SSL) and semi-supervised assistive image-to-text annotation systems seek to use the majority of the unlabeled data properties in both images and the accompanying text to learn meaningful image and text representations. These are normally learned by training separate Vision and Text representation models for the different input tokens in each modality. For example, in visual SSL we have generative models [53], joint embedding models [54] that predict pixel-level information, as well as joint-predictive embeddingstyle models [55], that are trained to predict an intermediate representation of image patches. This forces a fundamental semantic understanding of the image contents and avoids wasted computational resources on approximating pixel-level details. In text-based foundation models, the current standard is to fine-tune an auto-regressive Large Language Model (LLM), pre-trained on next-word prediction [56], and some level of reinforcement learning with human feedback [57]. Fine-tuning on a downstream task such as VOA, classification, and captioning is normally simpler for languages with large datasets, and problematic for Low-Resource-Languages [58], [59].

In AI-assistive image annotation using SSL and textual hints, the goal is to learn a good image, textual, or multi-modal representation from high-quality datasets. This is achieved by predicting parts of the input or an intermediate representation of the image, requiring only a few annotated image-text pairs to learn the mapping from image-to-text embeddings for downstream tasks. The state-of-the-art methods in this space utilize a combination of image and text SSL, before employing a fine-tuning cross-modal block that learns the image-to-text alignment [27]. As opposed to methods using DAL, SSL methods are normally used in the pre-training phase for representation learning. This means SSL primes suggestive DL annotation. Examples of industry-level AIassisted annotation products based on SSL for automated image captioning and VQA include [60]-[62], all with a partial image-to-text component. Label Studio [63] on the other hand focuses on LLM integration for assistive annotation for text-based inputs. They also have the same image captioning and single-target classification suggestions as the following solutions [60]–[62]. All that said, none of these methods produce free-text-type output suggestions for image annotation based on the underlying SSL model. Scholarly contributions in this area are limited but include works such as [64], [65], [61], and [66]. Nevertheless, these are mostly not production-level systems for image annotation, but rather experimental works on assistive annotating based on SSL. Again we see a shortage in descriptive textual interactive AIassisted annotation methods, mainly for CV tasks such as pose estimation [67] and instance segmentation.

C. Neuro-Symbolic Learning systems

Assistive NeSyL for image annotation is a relatively unexplored area in both scholarly publications and industry applications with limited literature [29], [30]. This may be in part due to the current overwhelming success of DL models claiming a big share of the AI research funding and talent. It would appear NeSyL is going through it's own winter like the DL winter of the 80s. However, there are some notable NeSyL-based methods for suggestive annotation, such as those presented in [30], [68]. While NeSyL adds to the reliability and factual nature of AI systems, its strict and limited data introduces a higher annotation quality requirement in the form of domain expertise and formalism. Again, we see a deficit in free-text image annotation hints/suggestions based on NeSyL assistive annotation.

D. Human-In-The-Loop and Crowd Sourcing

Human-In-The-Loop (HITL) image annotation systems improve AI-based auto annotation by involving a human expert annotator in the annotation process. By combining an AI-based vision model with interactive human annotation, HITL image annotation can be more efficient than manual annotation, as humans can be trained to annotate images quickly and accurately, while at the same time allowing for scalability through crowd-sourcing. The underlying AI model can be either based on SSL or a DAL model. Existing HITL and crowd-sourcing annotation systems with AI capabilities include Amazon Mechanical Turk [69] and CloudFactory [70]. They both focus on making the annotation task affordable and fast by using multiple annotators from many places across the world and using suggestive AI models in image annotation.

CrowdFlower [71] focuses on providing high-quality image data annotations, metadata creation, as well as providing real-time transcriptions. The implementation details of the underlying AI models for all these commercial products are not disclosed. However, upon exploring the platforms, it's evident that none of these platforms offer assistive AI annotation tools for generating free-text outputs from images. In the next Section, we cover common evaluation metrics and benchmarks for assessing the effectiveness of assistive annotation systems.

IV. EVALUATION METRICS AND BENCHMARKS

By using assistive annotation or automatic annotation tools, the goal is typically to improve the annotation speed, quality of the annotations, and usability of the tool. Evaluating such systems has been performed using metrics such as classification accuracy [72], [73], F-1 Score [74], [75], Intersection over Union [72], [76], average annotation time [73], [76], as well as metrics such as Cohen's Kappa that measure the level of agreement between two or more annotators [74]. During the evaluation, it is common to measure the perceived benefits of AI-assistive annotation for expert and non-expert annotators in the domain of the images. It has been shown that the largest improvement in annotation efficiency and accuracy due to AIassistive annotation is seen in non-expert annotators [72], [73], as experts normally perform near or above the underlying DL model used in suggesting annotations. For scholarly research and model development, the following datasets in Table I are typically used for training and evaluating DL models capable of contributing towards assistive annotation as well as image understanding tasks. We include the dataset name, size, number of images, and the CV tasks it is capable of being used in towards AI assistive annotation.

V. USE CASES

AI-assistive annotation, based on underlying AI models, finds most applications in the medical imaging and biological fields due to high annotation costs, the need for expert annotators, and its potential benefits in various tasks such as instance segmentation, object detection, pose estimation, and counting [33], [85]. AI annotation also plays a role in autonomous driving, where cars equipped with sensors and high-quality cameras stream training data for crowd-sourced and AI-based annotation [86]. Similar patterns are observed in these areas of application. While the literature is promising in assistive annotation, image-to-textual hint suggestions remain overlooked despite the potential benefits for the annotator and performance on a predefined downstream task.

VI. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This review has highlighted the scarcity of AI-assistive annotation systems that provide non-expert annotators with textual annotation suggestions based on image understanding and the specific CV task. We attribute the lack of progress in this area to the decades-long dependence on using different model architectures for images and text, hindering progress in cross-modal representation alignment. With advances in

transformer-based representation learning [87], hardware acceleration, and breakthroughs in DL training, we foresee significant progress in multi-modal learning, enabling more effective and explainable AI-assistive or automated annotation. A prime example of an assistive and interactive vision-based annotation system is Meta's segment Anything model (SAM) [77]. SAM is capable of performing multiple CV tasks based on both visual and text-based prompts. Based on this review we see a need for developing an annotation system capable of free-text hints and suggestions for CV tasks. A viable approach would be a combination of efficient self-supervised image and language pre-training with multi-modal alignment methods and advances in NeSyL for factual grounding, as shown in Figure 1. Additionally, promising recent image-totext retrieval methods that may be applicable to AI-assistive annotation are presented in the following survey [88]. The use of text-based annotation hints/suggestions not only lowers the requirements for expert-level human annotation but also improves the annotation speed, model interpretability, and accuracy for CV tasks. Lastly and more ambitiously, it seems obvious to us that prioritizing research around multi-modal methods capable of setting their own goals to optimize can solve most cross-domain problems. Efforts are also required to ensure such models are easily understandable, safe for use, and morally and ethically aligned with human interests, considering diverse cultures and beliefs

VII. ACKNOWLEDGMENTS

A big thank you to Khotso Ramoreboli and Itumeleng Tlali for the insightful discussions during the development of this paper.

REFERENCES

- H. Thisanke, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, and D. Herath, "Semantic segmentation using vision transformers: A survey," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106669, 2023.
- [2] Y. Srivastava, V. Murali, S. R. Dubey, and S. Mukherjee, "Visual question answering using deep learning: A survey and performance analysis," in *Computer Vision and Image Processing*, (Singapore), pp. 75–86, Springer Singapore, 2021.
- [3] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," ACM Comput. Surv., vol. 56, aug 2023.
- [4] A. Krizhevsky, I. Sutskever, and G.Hinton, "Imagenet classification with deep convolutional neural networks," *Neural Information Processing* Systems, vol. 25, 01 2012.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- [6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [7] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, "An image is worth 16x16 words: Transformers for image recognition at scale," in 9th International Conference on Learning Representations, ICLR 2021, 2021.
- [8] P. Bangert, H. Moon, J. O. Woo, S. Didari, and H. Hao, "Active learning performance in labeling radiology images is 90
- [9] J. Kremer, K. Stensbo-Smidt, F. Gieseke, K. Pedersen, and C. Igel, "Big universe, big data: Machine learning and image analysis for astronomy," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 16–22, 2017.

Dataset	Year	Size	Detection	Segmentation	Panoptic Segmentation	Pose Estimation	Captioning	VQA
SA-1B Dataset [77]	2023	11M	✓	✓	✓	×	✓	√
VizWiz [78]	2018	20,523	×	×	×	×	✓	\checkmark
CityScapes [79]	2016	25,000	×	\checkmark	✓	×	×	×
COCO-QA [80]	2014	123,287	×	×	×	×	\checkmark	\checkmark
MSCOCO [81]	2014	328,000	\checkmark	\checkmark	✓	\checkmark	\checkmark	\checkmark
MPII Human Pose [82]	2014	25,000	×	×	×	\checkmark	\checkmark	×
Flickr30k [83]	2014	31,000	✓	×	×	×	✓	×
PASCAL VOC [84]	2012	11,530	✓		×	×	×	×

Summary of common datasets: their respective capabilities for detection, instance segmentation, panoptic segmentation, pose estimation, image captioning, and visual question answering.

- [10] S. R. Chitnis, S. Liu, T. Dash, T. T. Verlekar, A. Di Ieva, S. Berkovsky, L. Vig, and A. Srinivasan, "Domain-specific pre-training improves confidence in whole slide image classification," in 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1–4, 2023.
- [11] T. Kataria, B. Knudsen, and S. Elhabian, "To pretrain or not to pretrain? a case study of domain-specific pretraining for semantic segmentation in histopathology," in *Medical Image Learning with Limited and Noisy Data*, pp. 246–256, Springer Nature Switzerland, 2023.
- [12] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [13] E. Karami, M. Shehata, and A. Smith, "Image identification using sift algorithm: performance analysis against different image deformations," arXiv preprint arXiv:1710.02728, 2017.
- [14] O. Lorente, I. Riera, and A. Rana, "Image classification with classic and deep learning techniques," arXiv preprint arXiv:2105.04895, 2021.
- [15] H. Barlow, "Unsupervised Learning," Neural Computation, vol. 1, pp. 295–311, 09 1989.
- [16] Y. LeCun and Y. Bengio, Convolutional Networks for Images, Speech, and Time Series, p. 255–258. Cambridge, MA, USA: MIT Press, 1998.
- [17] C. G. Northcutt, A. Athalye, and J. W. Mueller, "Pervasive label errors in test sets destabilize machine learning benchmarks," in ArXiv, 2021.
- [18] A. Singh, S. Agarwal, P. Nagrath, A. Saxena, and N. Thakur, "Human pose estimation using convolutional neural networks," in 2019 Amity International Conference on Artificial Intelligence (AICAI), pp. 946– 952, 2019.
- [19] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?," in *International Conference on Machine Learning*, 2019.
- [20] M. Mots'oehli, A. Nikolaev, W. IGede, J. Lynham, P. Mous, and P. Sadowski, "Fishnet: Deep neural networks for low-cost fish stock estimation," arXiv preprint arXiv:2403.10916, 2024.
- [21] C. Redondo-Cabrera, R. J. López-Sastre, Y. Xiang, T. Tuytelaars, and S. Savarese, "Pose estimation errors, the ultimate diagnosis," in *Computer Vision – ECCV 2016* (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 118–134, Springer International Publishing, 2016.
- [22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," in *Neural Computation*, vol. 9, pp. 1735–1780, 1997.
- [23] X. Yang, C. Gao, H. Zhang, and J. Cai, "Auto-parsing network for image captioning and visual question answering," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2177–2187, 2021
- [24] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12113–12132, 2023.
- [25] N. Das, A. Joshi, P. Yenigalla, and G. Agrwal, "Maps: Multimodal attention for product similarity," in 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2988–2996, 2022.
- [26] R. Zhang, J. Han, C. Liu, P. Gao, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, and Y. Qiao, "Llama-adapter: Efficient fine-tuning of language models with zero-init attention," arXiv preprint arXiv:2303.16199, 2023.
- [27] J. Han, K. Gong, Y. Zhang, J. Wang, K. Zhang, D. Lin, Y. Qiao, P. Gao, and X. Yue, "Onellm: One framework to align all modalities with language," ArXiv, vol. abs/2312.03700, 2023.
- [28] X. Wang, L. Ma, Y. Fu, and X. Xue, "Neural symbolic representation learning for image captioning," in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, ICMR '21, (New York, NY, USA), p. 312–321, Association for Computing Machinery, 2021.

- [29] S. Amizadeh, H. Palangi, A. Polozov, Y. Huang, and K. Koishida, "Neuro-symbolic visual reasoning: Disentangling "Visual" from "Reasoning"," in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, pp. 279–290, PMLR, 13–18 Jul 2020.
- [30] T. EITER, N. HIGUERA, J. OETSCH, and M. PRITZ, "A neuro-symbolic asp pipeline for visual question answering," *Theory and Practice of Logic Programming*, vol. 22, no. 5, p. 739–754, 2022.
- [31] A.-M. Tousch, S. Herbin, and J.-Y. Audibert, "Semantic hierarchies for image annotation: A survey," *Pattern Recognition*, vol. 45, no. 1, pp. 333–345, 2012.
- [32] C. J. Christoph Sager and P. Zschech, "A survey of image labelling for computer vision applications," *Journal of Business Analytics*, vol. 4, no. 2, pp. 91–110, 2021.
- [33] M. Aljabri, M. AlAmir, M. AlGhamdi, M. Abdel-Mottaleb, and F. Collado-Mesa, "Towards a better understanding of annotation tools for medical imaging: a survey," *Multimedia Tools and Applications*, vol. 81, p. 25877–25911, jul 2022.
- [34] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Li, "A survey and analysis on automatic image annotation," *Pattern Recognition*, vol. 79, pp. 242–259, 2018.
- [35] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," *Future Generation Computer Systems*, vol. 135, pp. 364–381, 2022.
- [36] M. Wang, B. Ni, X.-S. Hua, and T.-S. Chua, "Assistive tagging: A survey of multimedia tagging with human-computer joint exploration," ACM Comput. Surv., vol. 44, sep 2012.
- [37] P. Ren, Y. Xiao, X. Chang, P. Huang, Z. Li, X. Chen, and X. Wang, "A survey of deep active learning," ACM Computing Surveys (CSUR), vol. 54, pp. 1 – 40, 2020.
- [38] M. Mots'oehli and K. Baek, "Deep active learning in the presence of label noise: A survey," arXiv preprint arXiv:2302.11075, 2023.
- [39] G. Rotman and R. Reichart, "Multi-task Active Learning for Pretrained Transformer-based Models," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 1209–1228, 11 2022.
- [40] H. Li and Z. Yin, "Attention, suggestion and annotation: a deep active learning framework for biomedical image segmentation," in Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23, pp. 3–13, Springer, 2020.
- [41] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pp. 399–407, Springer, 2017.
- [42] H. Wang, W. Huang, Z. Wu, H. Tong, A. J. Margenot, and J. He, "Deep active learning by leveraging training dynamics," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25171–25184, 2022.
- [43] J. Aklilu and S. Yeung, "Alges: active learning with gradient embeddings for semantic segmentation of laparoscopic surgical images," in *Machine Learning for Healthcare Conference*, pp. 892–911, PMLR, 2022.
- [44] Amazon, "Import annotations as pre-labels." https://docs.aws. amazon.com/sagemaker/latest/dg/sms-automated-labeling.html# samurai-automated-labeling-byom, 2024.
- [45] Roboflow, "Roboflow annotate." https://roboflow.com/annotate, 2024.
- [46] Labelbox, "Import annotations as pre-labels." https://docs.labelbox.com/ docs/model-assisted-labeling, 2024.
- [47] HumanSignal, "Labeling automation and active learning." https:// humansignal.com/, 2024.

- [48] telusInternational, "training data types." https://www.telusinternational. com/solutions/ai-data-solutions/ai-training-data/image-data?INTCMP= ti_data-annotation_link_view-all-image-data-services_card-content, 2024.
- [49] cloudfactory, "Computer vision annotation to power ai." https://info.cloudfactory.com/, 2024.
- [50] Encord, "All the tools you need to build better models, faster." https://encord.com/, 2024.
- [51] Datagym, "training data types." https://www.datagym.ai/, 2024.
- [52] S. Waterbury, "Machine learning assisted image segmentation." https://scale.com/blog/ml-image-segmentation, 2022.
- [53] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 1, pp. 857–876, 2021.
- [54] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International* conference on machine learning, pp. 1597–1607, PMLR, 2020.
- [55] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," 2023.
- [56] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings* of naacL-HLT, vol. 1, p. 2, 2019.
- [57] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., "Llama: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.
- [58] A. Ragni, K. M. Knill, S. P. Rath, and M. J. Gales, "Data augmentation for low resource languages," in *INTERSPEECH 2014: 15th annual conference of the international speech communication association*, pp. 810– 814, International Speech Communication Association (ISCA), 2014.
- [59] V. Marivate, M. Mots'Oehli, V. Wagnerinst, R. Lastrucci, and I. Dzingirai, "Puoberta: Training and evaluation of a curated language model for setswana," in *Artificial Intelligence Research* (A. Pillay, E. Jembere, and A. J. Gerber, eds.), (Cham), pp. 253–266, Springer Nature Switzerland, 2023.
- [60] S. inc, "Shade helps you to organize and understands your content." https://www.shade.inc/webtools/imageCaption, 2024.
- [61] V. AI, "Vertexai:mml customize and deploy generative models," 2024.
- [62] SceneXplain, "Scenexplain: Scenexplain lets you attach images to your prompt. explore image storytelling beyond pixels.." https://iina.ai/, 2024.
- [63] H. L. S. Jimmy Whitaker, "Interactive data labeling with llms and label studio's prompt interface." https://labelstud.io/blog/ automate-data-labeling-with-llms-and-prompt-interface/, 2024.
- [64] W. Bai, C. Chen, G. Tarroni, J. Duan, F. Guitton, S. E. Petersen, Y. Guo, P. M. Matthews, and D. Rueckert, "Self-supervised learning for cardiac mr image segmentation by anatomical position prediction," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pp. 541–549, Springer, 2019.
- [65] R. E. Ferreira, Y. J. Lee, and J. R. Dórea, "Using pseudo-labeling to improve performance of deep neural networks for animal identification," *Scientific Reports*, vol. 13, no. 1, p. 13875, 2023.
- [66] G. Naithani, J. Kivinummi, T. Virtanen, O. Tammela, M. J. Peltola, and J. M. Leppänen, "Automatic segmentation of infant cry signals using hidden markov models," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, pp. 1–14, 2018.
- [67] T. Davy Neven, "Accelerating human pose labeling: Effortless efficiency with trainyolo." https://www.trainyolo.com/blog/ accelerating-human-pose-labeling/, 2024.
- [68] J. S. V. Adnan Siddiqui, Nischcol Mishra, "A survey on automatic image annotation and retrieval," *International Journal of Computer Applications*, vol. 118, pp. 27–32, May 2015.
- [69] Amazon, "Amazon mechanical turk: Access a global, on-demand, 24x7 workforce." https://www.mturk.com/, 2024.
- [70] cloudfactory, "Accelerate the ai lifecycle with human-in-the-loop solutions." https://www.cloudfactory.com/, 2024.

- [71] CrowdFlower, "Collect, clean, and label your data at scale with crowdflower.." https://visit.figure-eight.com/People-Powered-Data-Enrichment T/, 2024.
- [72] G. Pavoni, M. Corsini, F. Ponchio, C. Edwards, N. Pedersen, S. Sandin, and P. Cignoni, "Taglab: Ai-assisted annotation for the fast and accurate semantic segmentation of coral reef orthoimages," *Journal of Field Robotics*, pp. 1–17, 2021.
 [73] M. Radeta, R. Freitas, C. Rodrigues, A. Zuniga, N. T. Nguyen, H. Flores,
- [73] M. Radeta, R. Freitas, C. Rodrigues, A. Zuniga, N. T. Nguyen, H. Flores, and P. Nurmi, "Man and the machine: Effects of ai-assisted human labeling on interactive annotation of real-time video streams," ACM Trans. Interact. Intell. Syst., vol. 14, apr 2024.
- [74] A. Hoelzemann, M. Bock, and K. V. Laerhoven, "Evaluation of videoassisted annotation of human imu data across expertise, datasets, and tools," in 2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), pp. 1–6, 2024.
- [75] S. Chatterjee, B. Mitra, and S. Chakraborty, "Amicron: Framework for generating micro-activity annotations for human activity recognition," in 2022 IEEE International Conference on Smart Computing (SMART-COMP), pp. 26–31, 2022.
- [76] S. Vinayahalingam, B. Berends, F. Baan, D. A. Moin, R. van Luijn, S. Bergé, and T. Xi, "Deep learning for automated segmentation of the temporomandibular joint," *Journal of Dentistry*, vol. 132, p. 104475, 2023.
- [77] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023.
- [78] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (Los Alamitos, CA, USA), pp. 3608–3617, IEEE Computer Society, jun 2018.
- [79] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [80] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," Advances in neural information processing systems, vol. 28, 2015.
- [81] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision ECCV 2014* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), (Cham), pp. 740–755, Springer International Publishing, 2014.
- [82] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [83] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.
- [84] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, June 2010.
- [85] I. Dimitrovski, D. Kocev, S. Loskovska, and S. Džeroski, "Hierarchical annotation of medical images," *Pattern Recognition*, vol. 44, no. 10, pp. 2436–2449, 2011. Semi-Supervised Learning for Visual Content Analysis and Understanding.
- [86] A. Swief and M. El-Habrouk, "A survey of automotive driving assistance systems technologies," in 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), pp. 1–12, IEEE, 2018.
- [87] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.
- [88] S. Wang, L. Zhu, L. Shi, H. Mo, and S. Tan, "A survey of full-cycle cross-modal retrieval: From a representation learning perspective," *Applied Sciences*, vol. 13, no. 7, 2023.