Final Assignment: Data Annotation

1. Overview

For your final assignment, you will work in a group of **2-4 students** to complete a project involving **data annotation**. Your group will be responsible for choosing a specific topic or field, collecting and annotating data related to that field, and evaluating the utility of the dataset for model training and research.

2. Assignment Tasks

1) Topic Selection:

- Your group should choose a **specific field or topic** for which you will create and annotate a dataset. This could be from any area you are interested (e.g., object detection, facial recognition, entity extraction, etc.).

2) Dataset Annotation:

- Once you have chosen the field, your group should collect data resources (either by capturing your own or sourcing them from public datasets), and annotate them with labels. These labels could include object boundaries, categories, keypoints, or any other relevant information, depending on the type of dataset.
- The annotation should be done using an appropriate tool

(e.g., LabelImg for object detection, Labelbox for image segmentation).

- It’s essential to maintain **annotation quality** by ensuring consistency and accuracy. If possible, use multiple annotators and apply quality control measures (e.g., double-checking annotations).
- Annotate **at least 500 images/sentences/audios**, ensuring a **reasonable balance of labels** or categories based on the dataset's scope.

3) Dataset Evaluation:

- After annotating your dataset, evaluate it by applying at least one **model** from the field you chose (e.g., a pre-trained object detection or segmentation model).
- Use standard **evaluation metrics** such as **precision**, **recall**, **mean Average Precision (mAP)**, or **IoU (Intersection over Union)** depending on the task.
- Provide an analysis of how well the model performs with your dataset, identifying any potential challenges, such as data imbalance, labeling errors, or difficult cases.

4) Essay Writing:

- After completing the annotation and evaluation process, each group must write an **essay of approximately 2000**

words that describes the entire process. The general writing requirements are listed below.

- The essay should be well-organized and follow academic writing standards. Use figures, tables, and references as needed to support your points.
- Use the IEEE citation style for any references, tools, or datasets you mention.

3. Submission Requirements

1) Dataset:

- Submit your **annotated dataset** along with any associated metadata or documentation that explains the structure of your dataset. If applicable, include scripts or tools used for annotation.
- Ensure the dataset is in a format that can be easily accessed (e.g., .xml, .json, .csv for labels).

2) Essay:

- Submit your **2000-word essay** in IEEE format (see Appendix B) that covers all sections outlined above.

3) Deadline:

- Both the dataset and the essay must be submitted by May 30, 23:59.

4. Grading Criteria

Your final grade will be based on the following:

- **Dataset Quality (40%):** The completeness, accuracy, and usability of the annotated dataset. Ensure that your dataset is well-structured and ready for machine learning model training.
- **Evaluation and Analysis (30%)**: The thoroughness and depth of your model evaluation and analysis. This includes the use of appropriate metrics, a clear evaluation of model performance, and addressing dataset challenges.
- **Essay Quality (20%):** The clarity, organization, and academic rigor of your essay. Ensure your writing is clear, concise, and well-supported by data and references.
- **Collaboration (10%):** Your group's ability to collaborate effectively and divide responsibilities. All group members should contribute equally to the project.

5. Important Notes

- **Collaboration**: Effective teamwork is essential. Make sure each group member has a clear role in

the project (e.g., data collection, annotation, model evaluation, writing the essay).

- **Ethics**: Ensure that your dataset does not infringe on any copyrights, and if using public datasets, properly acknowledge the original creators.
- **Reproducibility**: Make sure that your dataset and results are reproducible. Document the process clearly so that others can replicate your work.

6. General Writing Requirements

a) Length:

Target word count: **2000 words**, with a tolerance of **±10%** (1800-2200 words).

b) Format:

IEEE Format: Follow the IEEE referencing and formatting guidelines. This includes:

- **Title**: Descriptive, concise, and informative.
- **Abstract**: 150-250 words.
- **Sections and subsections:** Use clear, numbered section titles (e.g., 1. Introduction, 2. Related Work).
- **References**: Use IEEE citation style for references and citations (in-text citations in square brackets, e.g., [1], [2]). Ensure that your **reference list** is comprehensive, covering

all datasets, tools, and related works you mention.

- **Figures/Tables**: If you include figures or tables (e.g., sample annotations, model results), ensure they are well-labeled and referenced correctly in the text. Always provide **captions** for figures and tables.
- **Font**: Use the predefined font in Appendix B.

c) Clarity and Focus:

- Write in clear, concise academic English. Avoid overly complex sentences and jargon unless necessary (usually about 20-30 words in one sentence, no more than 40 words).
- Maintain a **formal** and **objective** tone throughout the essay. Use third-person passive voice where appropriate (e.g., "The dataset was annotated by a team of experts").
- Stay focused on the topic of **labeled datasets in the field you chosen**. Ensure that each section contributes directly to your main objective.

d) Structure:

Each section should be written logically, with smooth transitions between ideas. See Appendix A for the detailed content outline for each section’s specific points.