CIA: Controllable Image Augmentation Framework based on Stable Diffusion

Mohamed Benkedadra
University of Mons
Mons, Belgium
mohamed.benkedadra@umons.ac.be

Dany Rimez
UCLouvain
Louvain-La-Neuve, Belgium
dany.rimez@uclouvain.be

Tiffanie Godelaine
UCLouvain
Louvain-La-Neuve, Belgium
tiffanie.godelaine@uclouvain.be

Natarajan Chidambaram
University of Mons
Mons, Belgium
natarajan.chidambaram@umons.ac.be

Hamed Razavi Khosroshahi
Université libre de Bruxelles
Brussels, Belgium
hamed.razavi.khosroshahi@ulb.be

Horacio Tellez
Multitel
Mons, Belgium
hatellezp@gmail.com

Matei Mancas
University of Mons
Mons, Belgium
Matei.MANCAS@umons.ac.be

Benoit Macq
UCLouvain
Louvain-La-Neuve, Belgium
benoit.macq@uclouvain.be

Sidi Ahmed Mahmoudi
University of Mons
Mons, Belgium
sidi.mahmoudi@umons.ac.be

***Abstract*—Computer vision tasks such as object detection and segmentation rely on the availability of extensive, accurately annotated datasets. In this work, We present CIA, a modular pipeline, for (1) generating synthetic images for dataset augmentation using Stable Diffusion, (2) filtering out low quality samples using defined quality metrics, (3) forcing the existence of specific patterns in generated images using accurate prompting and ControlNet. In order to show how CIA can be used to search for an optimal augmentation pipeline of training data, we study human object detection in a data constrained scenario, using YOLOv8n on COCO and Flickr30k datasets. We have recorded significant improvement using CIA-generated images, approaching the performances obtained when doubling the amount of real images in the dataset. Our findings suggest that our modular framework can significantly enhance object detection systems, and make it possible for future research to be done on data-constrained scenarios. The framework is available at: github.com/multitel-ai/CIA.**

***Index Terms*—Computer Vision, Generative AI, Stable Diffusion, Object Detection**

I. Introduction

The performance of deep learning models is dependent on the quality and diversity of the dataset they were trained on.

Unfortunately, the creation of such high-quality and accurately annotated datasets is often challenged by the scarcity of data and the substantial costs associated with annotation [1], especially in specialized and evolving Computer Vision tasks. Hence, other strategies are commonly used to enhance dataset quality, like Active Learning [2] and Data Augmentation methods [3] (image rotation, flipping, color adjustment, etc).

However, these methods modify images with simple and often content-agnostic transformations, limiting their ability to introduce completely new information into the dataset. This limitation led us to the exploration of Generative AI models like Stable Diffusion [4] that can generate entirely new images. Through the usage of ControlNet [5] with predefined features extracted from the original image, we can tailor the generation process to meet specific task requirements. This creates an unprecedented opportunity to augment datasets beyond traditional methods. Concurrently, when dealing with

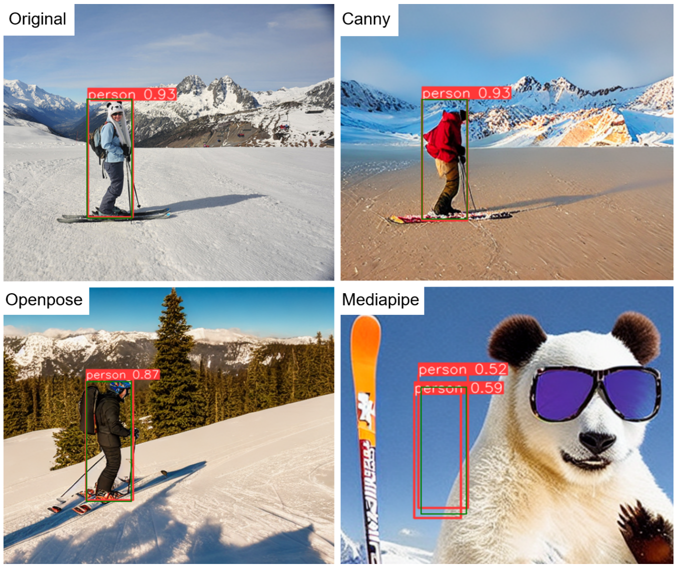

Fig. 1: CIA-generated images from an image taken from the COCO dataset for different ControlNets. Either efficient (*Openpose*, *Canny Edge*) or inefficient (*Mediapipe*) for an object-detection task. Prediction of YOLOv8n trained on the dataset corresponding to the image is shown in red, and ground truth in green.

synthetic data augmentation, we want to generate the most useful images for model training. This raises the question of how to assess the quality and relevance of the generated data. Finally, a pivotal question arises:

Can the quality of region of interest vision datasets be enhanced to ensure better model performance through the incorporation of images generated with Controlled Stable Diffusion ?

To answer that question, this work introduces CIA, a modular framework for data augmentation. It integrates Stable Diffusion with Control Net models and is able to :

(1) Generate synthetic images for dataset augmentation using both generative and classic data augmentation methods.
(2) Filter out low quality samples with defined metrics.
(3) Control the generative process to create specific patterns in the generated images for region dependent computer vision tasks (object detection, segmentation, etc.)
(4) Easily preform parallel training, testing, and comparison of multiple augmentation methods

We prove the efficacy of CIA, through a case study on human object detection, in a scenario where the low amount of data severely limits the performances of the trained model. Examples of images generated with the proposed framework can be seen in Fig.1.

II. Related Works

Data augmentation has become an indispensable strategy for enhancing the quality and diversity of visual datasets and improving models' performances in Computer Vision tasks. Shorten and Khoshgoftaar's comprehensive review [6] extensively explores the diverse range of techniques employed, spanning from fundamental geometric transformations [3] to sophisticated generative methods such as Stable Diffusion [4] and ControlNet [5].

These advanced techniques can generate novel content and scene conditions. For example, introducing new variations in weather, people position, object appearance, image style, etc. This essential for mitigating the limitations posed by inadequate datasets, ultimately enhancing the performance and reliability of models.

Chen et al. [7] preform scale-aware data augmentation strategies for region dependent tasks. They focus on adding objects of different scales through computationally efficient, zoom-in/out operations. Ghiasi et al. [8] expanded on this approach by copy-pasting the zoomed objects at various scales into different backgrounds.

A. Stable Diffusion for Data Augmentation

Eliassen and Ma [9] demonstrated how Stable Diffusion, combined with Active Learning, can effectively re-balance classification datasets, notably outperforming traditional oversampling methods on CIFAR-10. Trabucco et al. [10] demonstrated the efficacy of text-to-image diffusion models in creating synthetic images for data augmentation. Similarly, Azizi et al. [11] highlighted how synthetic data from diffusion models can enhance ImageNet [12] classification. By fine-tuning text-to-image models, they achieved class-conditional models with impressive fidelity.

For region dependent tasks such as object detection and instance segmentation, Ge et al. [13] introduced a text-to-image synthesis paradigm leveraging DALL-E [14]. Their method generates diverse labeled data by utilizing segmentation masks to separately produce foregrounds (objects) and backgrounds (scenes). However, the approach exhibits limitations in the quality of the generated samples due to the artificial merging of generated objects onto backgrounds, resulting in a noticeable discontinuity between the foreground and background elements.

Wu et al. [15] focused on image augmentation with Controlled Diffusion. Their method significantly boosts performance with minimal training data.

Although these studies collectively showcase the transformative impact of Stable Diffusion models, none of them offer a complete and reliable pipeline for generative data augmentation. They do not provide an easy to set up tool for quickly and efficiently testing the different augmentation strategies, or for employing quality metrics to evaluate synthetic data.

B. Synthetic Images Quality Assessment

Assessing the quality of the synthesized samples presents a notable challenge. The complexity stems from the subjective and multidimensional nature of ”quality”, as its definition can defer depending on the intended application of the generated data. The literature often highlights the use of Image Quality Assessment (IQA) metrics for evaluating visual quality. Active Learning metrics can be used for assessing the potential impact of the data on model training.

1) IQA Metrics: IQA metrics focus on different features and patterns in the image to quantify its quality. Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [16], measures contrast, luminosity, distortion, etc, to quantify anomalies in generated images. The Neural Image Assessment (NIMA) model [17] employs a CNN trained to measure the aesthetics and realism of synthetic images, and outputs a distribution of scores that represent different criteria. As introduced in [18], ClipIQA leverages the power of large-scale pre-trained vision-language models to predict image quality without reference images. It was trained on specific features related not only to quality but to the general look, feel, content, and context of the image.

2) Active Learning metrics: Unlike Model-Agnostic metrics, Model-Aware quality metrics rely on the discriminative task's model for quality assessment. Active learning sampling strategies are often employed to assess the quality of the synthetic data by predicting its impact on model performance. Uncertainty based and diversity based methods [2] are the most common. In the example of object detection using YOLO, we can use the detection confidence score as a measure of the distance between a new synthetic image and the average distribution of real images used to train the baseline model.

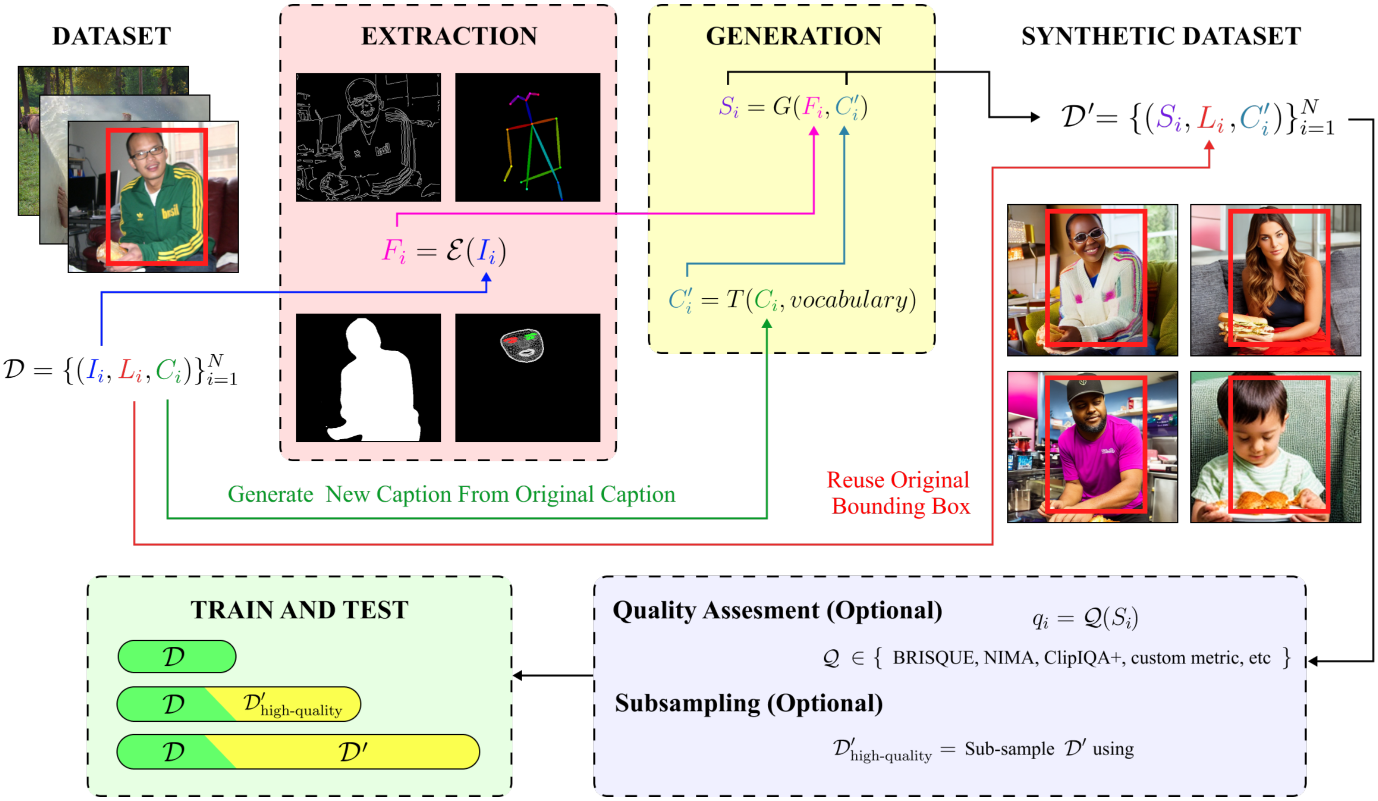
DATASET
EXTRACTION
GENERATION
SYNTHETIC DATASET
$S_i = G(F_i, C_i')$
$\mathcal{D}' = \{(S_i, L_i, C_i')\}_{i=1}^N$
$F_i = \mathcal{E}(I_i)$
$C_i' = T(C_i, vocabulary)$
$\mathcal{D} = \{(I_i, L_i, C_i)\}_{i=1}^N$
Generate New Caption From Original Caption
Reuse Original Bounding Box
TRAIN AND TEST
$\mathcal{D}$
$\mathcal{D}$ $\mathcal{D}'_{\text{high-quality}}$
$\mathcal{D}$ $\mathcal{D}'$
Quality Assesment (Optional)
$q_i = \mathcal{Q}(S_i)$
$\mathcal{Q} \in \{$ BRISQUE, NIMA, ClipIQA+, custom metric, etc $\}$
Subsampling (Optional)
$\mathcal{D}'_{\text{high-quality}} =$ Sub-sample $\mathcal{D}'$ using

Fig. 2: The CIA Framework for improving object detection accuracy through data augmentation using Stable Diffusion and ControlNet. Real images are taken from the COCO dataset. Notations used in the figure are further explained in the text.

III. Proposed CIA Framework

CIA is composed of four modules, as seen in Fig.2.

Initially, an *Extraction* module performs feature extraction from original images, to acquire the control features that maintain the integrity of the dataset's intrinsic characteristics. These features are used in the next phase by the ControlNet to condition the output of Stable Diffusion, thus adding an extra control sequence beyond the conventional text prompt.

The *Generation* module takes in the extracted features combined with text prompts to synthesize new images. The prompts are either manually specified, or automatically generated. Optionally, In order to put constraints on the resulting dataset quality, the *Quality Assessment* module can filter the generated images using chosen quality metrics, which allows for retaining only the highest-quality images.

The final stage of the pipeline is the *Train and Test* block. Through training different models, we can explore the effects of using various combinations of original and synthetic data on task performance.

A. Extraction

We begin by extracting features specific to the chosen ControlNet. Although custom extractors can be added, a few are implemented by default in CIA and cover some popular domains, from extracting human features through poses [19] or faces [20], to broader generic features like edges [21] and segmentation masks [22].

Let $\mathcal{D}$ denote the original dataset of N real images, where each image I_i, has a label L_i, and a caption C_i. Then, with the selected extractor $\mathcal{E}$, the feature image F_i extracted from I_i is given by: $F_i = \mathcal{E}(I_i)$.

B. Generation

Several generators could be obtained by combining the chosen ControlNet model with any compatible Stable Diffusion model. Once the Diffusion model is chosen, the generator G is able to generate the synthetic image S_i, for each extracted feature F_i from an image I_i, and the text prompt C_i (the caption of the original image). Such that S_i is given by $S_i = G(F_i, C_i^{'})$.

To introduce more diversity in generated images, we use modified captions $C_i^{'}$. Many methods could be used to generate prompts, such as LLMs (e.g., LLama2 [23]). However, the default prompt generator T of CIA follows a simple implementation. It takes a prompt C_i and a *vocabulary* to produce a new prompt $C_i^{'}$. For example, T modifies $C_i =$ *a man in a red shirt*, by substituting words from a *vocabulary*: $\{v0$: [*man, woman, child*] , $v1$: [*red, black, yellow*]$\}$. a possible modified caption could be $C_i^{'} =$ *a woman in a yellow shirt*. We can generate many modified captions $C_{ij}^{'}$, meaning j possibilities of synthetic images generated from a single real image where $j \in \{0, 1, 2, \ldots, (\prod_{i=1}^{n} v_i) - 1\}$.

The new text prompts are the input prompts of Stable Diffusion. Its output is conditioned by the control features from the *Extraction* module. Finally, we get the new $\mathcal{D}'$ dataset of generated images. By default, the labels of the original images are conserved in the generated ones.

C. Quality Assessor and Sampler

To assess the quality of synthetic images in $\mathcal{D}'$, we introduce a *Quality Assessment* module that filters out low quality images. Here, the quality of S_i can be defined according to any metric suitable for the task. The quality score q_i is then computed using the selected metric $\mathcal{Q}$ from the set of Quality Metrics Q such that $q_i = \mathcal{Q}(S_i)$. The quality metrics implemented in CIA includes IQA and Active Learning metrics.

D. Train and Test

We can train and test multiple models for the task at hand. Through modifying generation parameters, we can choose the amount of synthetic data in this training set. Optionally, if the *Quality Assessment* module is used, we can control the quality thresholds of the added synthetic data. Performances of the model are evaluated on a validation set during training, and on a test set after training. Both sets are constituted of real images only.

IV. Experimental Setup

We preformed a case study on human object detection to prove the framework's effectiveness. In this toy example, we only have access to a limited dataset that leads to suboptimal performances. The goal is to study how to optimally improve performances, by adding CIA-generated synthetic images. The Generation parameters of Stable Diffusion were not optimized and kept constant. YOLOv8n [24] was used as the object detection model. For each experiment, it was trained for 300 epochs using the training parameters from [25] with the SGD optimizer. Experiments were conducted on subsets of Common Objects in Context (COCO) [26] and Flickr30k Entities [27].

A. Datasets

COCO was processed to focus on a subset of images containing only one instance of the "PERSON" class, where objects take an area between 5% and 80% of the image. In Flickr, objects are labelled with textual segments without consistent class annotations. We processed the textual descriptions to automatically annotate the images with the "PERSON" class. The inconsistency in Flickr's annotations provided a robust stress testing ground for CIA, simulating the variability and imperfection common in real-world datasets. Three types of training sets were created for both datasets :

1) Baseline: Contains real images. One lower $\mathcal{D}_{250}$ (250 images) and one upper $\mathcal{D}_{500}$ (500 images) baselines were used as basis for comparison.

2) Synthetic: To evaluate the impact of adding synthetic images on object detection performance, a larger synthetic dataset D'_{1250} (1250 images) was generated by using five distinct auto-generated captions $(C'_1, ..., C'_5)$ for each sample in $\mathcal{D}_{250}$. Multiple datasets were then created with different proportions (250,500,...,1250) of synthetic images sampled from $\mathcal{D}'_{1250}$ and added to $\mathcal{D}_{250}$.

3) Ablation: To compare the addition of new synthetic data to simply training on real data for more epochs. We duplicated the images from $\mathcal{D}_{125}$ to obtain $\mathcal{D}^{\text{ablation}}_{375}$, $\mathcal{D}^{\text{ablation}}_{500}$, ..., $\mathcal{D}^{\text{ablation}}_{1500}$.

B. Experiments

With these datasets, three experiments presented hereafter were conducted. The first one was done using both COCO and Flickr images, while only COCO was used for the two others.

1) ControlNet effect: To analyze the effects of choosing a good ControlNet that fits the task, we compared several models. Four models were chosen. Some tailored for people detection (*OpenPose* or *MediaPipe*), and others are more generic (*Canny Edge* and *Segmentation*), and suitable for various types of datasets and contexts. All four are compatible with Stable Diffusion v1.5 (runwayml/stable-diffusion-v1-5) from the Hugging Face platform. Example of CIA-generated images can be observed on Fig.1 for the first three.

We added a deficient *extraction* module to the case study, to understand how the usage of bad conditions affects the results. We used the *Segmentation* extractor, with a transposed segmentation mask as a condition instead of the true segmentation mask. As a result (see Fig.3), this new *extraction* module, *False-Segmentation*, generated bad quality images. Not only the shape and position of the label bounding box are affected[1], but the content of the image is not necessarily coherent with the label anymore.

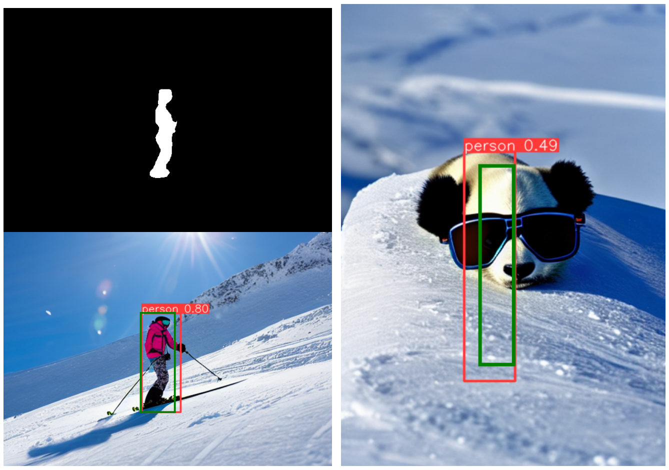

Fig. 3: Examples of synthetic images generated with ControlNets Segmentation and False-Segmentation from the same real image as in Fig.1. Left: YOLOv8m-seg's segmentation mask of the real image (top) and synthetic image generated (bottom). Right: synthetic image generated using the transposed segmentation mask.

2) Data Augmentation additivity: This experiment aims at illustrating the first claim stated in Section I, i.e. CIA augmentation can independently be used along with other data augmentation methods. Let's call this property additivity.

[1] Bounding boxes coordinates are defined in relative coordinates and depends on the Height and Width of the image.

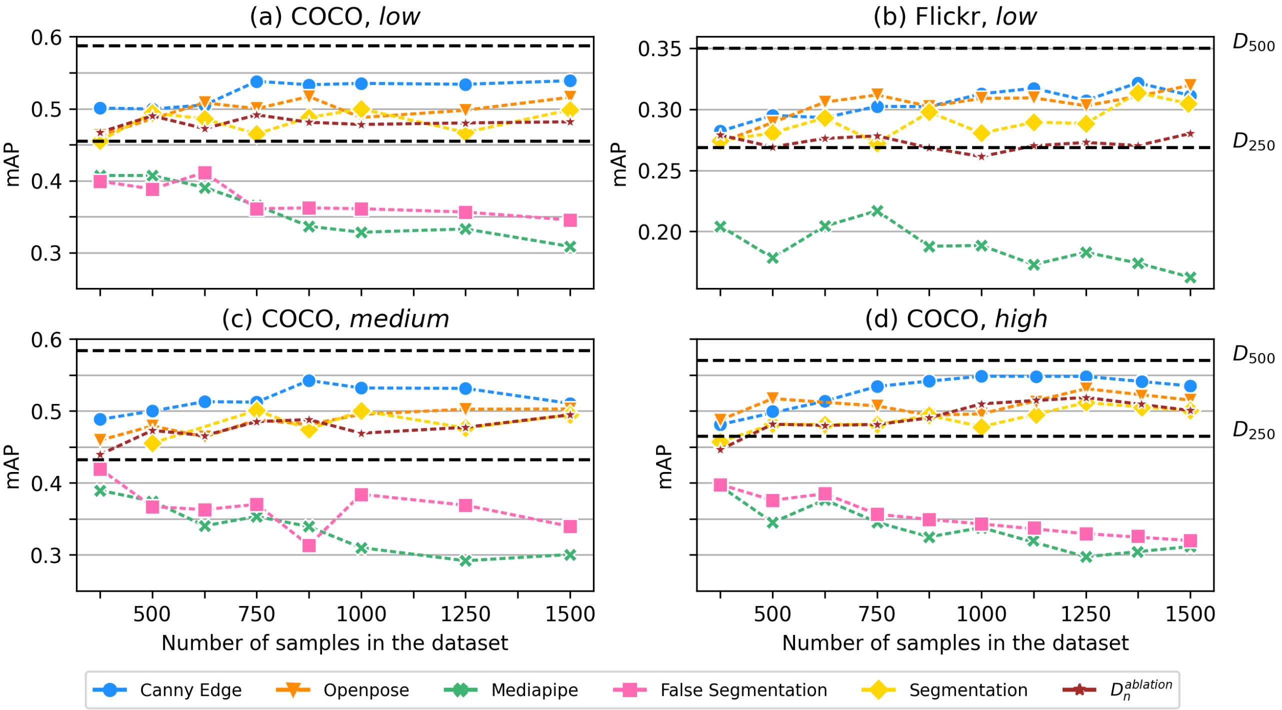

Fig. 4: Performance Evaluation of the trained YOLOv8 models on test set. **Influence of 5 ControlNets** (*Canny Edge*, *OpenPose*, *MediaPipe*, *Segmentation* and *False-Segmentation*) (a) on COCO dataset (b) on Flickr dataset. **Evaluation of gain using synthetic images in addition to data augmentation** on COCO dataset (c) *medium* (d) *high*.

We analyzed three levels of data augmentation already implemented in YOLOv8 [28]. (1) Low augmentation includes scale, translation, hue saturation and mosaic. (2) Medium augmentation, adds random shear and rotation with a maximum ±5 and ±10 degrees respectively, and a 10% probability of applying copy-paste. High augmentation has the same setting, with a 20% probability of applying copy-paste and mix up. In the latter, we named the models trained on those augmentation levels: *low*, *medium*, and *high* models.

3) Sampling with quality metrics: To showcase the ability of CIA to filter images according to predefined metrics, smaller synthetic datasets $\mathcal{D}'$ were refined. The top n images were selected from $\mathcal{D}'^{\mathcal{Q}}_{1250}$ according to the quality metric $\mathcal{Q}$. This process enabled the creation of high-quality synthetic subsets $\mathcal{D}'^{\mathcal{Q}}_{n-\text{high-quality}}$. This approach yielded datasets with varying sizes from $\mathcal{D}'^{\mathcal{Q}}_{125-\text{high-quality}}$ to $\mathcal{D}'^{\mathcal{Q}}_{875-\text{high-quality}}$, each comprising the highest quality images according to the quality metric. BRISQUE, ClipIQA, and NIMA were employed in addition to Model-Aware Active Learning metrics. Mainly, CORE-SET and confidence-score based selection. In this second method, images with the lowest confidence values predicted by the model are selected. The selection process unfolds over five rounds, with an incremental increase of 125 samples per round to produce 5 synthetic datasets $\mathcal{D}'^{\mathcal{Q}}_{n-\text{high-quality}}$.

V. Results

In this section, the results of the case study for the three aforementioned experiments are presented before being discussed to highlight the possibilities of CIA.

A. ControlNet effect

Evaluating the influence of different ControlNets on enhancing YOLOv8's object detection capabilities, focuses on variations in mAP. This evaluation, depicted in Fig.4 (a) and (b), reveals the significance of ControlNet choice on performance. While most ControlNets led to an increase in mAP compared to $\mathcal{D}_{250}$, and $\mathcal{D}_n^{\text{ablation}}$, none matched $\mathcal{D}_{500}$ performance. Notably, *Mediapipe* exhibits a decline in performance. This could be explained by images where the object deviates from the original bounding box (Fig.1). We then tested *False-Segmentation*, explained in section IV-B, and obtained similar results to *Mediapipe*. This confirms that the choice of the ControlNet needs to be consistent with the task domain.

On the contrary, *Canny Edge*, *OpenPose*, and *Segmentation* contributed positively to mAP. This improvement was notable up to 750 synthetic samples, beyond which mAP increase was considered not significant.

B. Data augmentation additivity

The second study aimed to analyze the impact of adding synthetic images to other data augmentation techniques to determine the value of synthetic data. The results are shown in Fig.4 (a), (c) and (d). We observe that using synthetic data never leads to lower performances for efficient ControlNets at all data augmentation levels, as demonstrated by the higher mAP compared to $D_n^{ablation}$. We can also see that the performances when using *Canny Edge* with low data augmentation level are the same as the *medium* baseline, coupled with the fact that *high* baseline is lower than *medium* baseline. Hence, classic augmentation methods are prone to cause overfitting, while CIA images guarantee good performance even at higher level of augmentation if the ControlNet is chosen correctly.

C. *Sampling with quality metrics*

The aim of this experiment was to refine the pipeline for optimal outcomes. The results for *Canny Edge* and *Mediapipe* are illustrated in Fig.5, but results are similar for all ControlNets: none of the sampling strategies significantly outperformed random sampling. This suggests that prioritizing images based solely on features like visual quality or diversity, may not be the most effective strategy for model improvement.

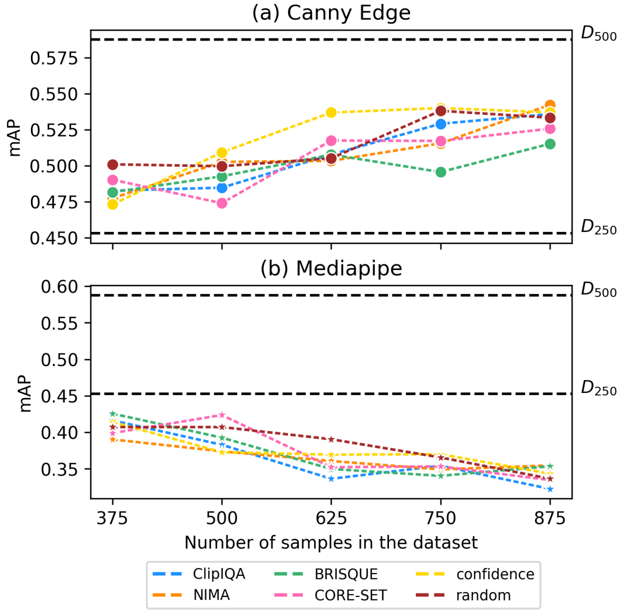

Fig. 5: Performance Evaluation of the trained YOLOv8 models on test set. **Influence of sampling methods** (ClipIQA, NIMA, BRISQUE, CORE-SET, confidence) on COCO dataset for ControlNet (a) *Canny Edge* (b) *MediaPipe*. ”random” sampling refers to plots (a) and (b) of Fig.4 for which the synthetic images are selected randomly.

VI. Discussion

This case study provides guidelines for using the CIA framework effectively. We demonstrated that adding synthetic images generated with the appropriate ControlNet can enhance detection performance. These images can also be used in conjunction with basic data augmentation. The analysis of the influence of sampling methods indicates that the diversity of the generated images may not be optimal. Exploring other hyperparameters during the generation process may lead to better results. Nonetheless, it is still far superior to classical methods even at very high levels of augmentation.

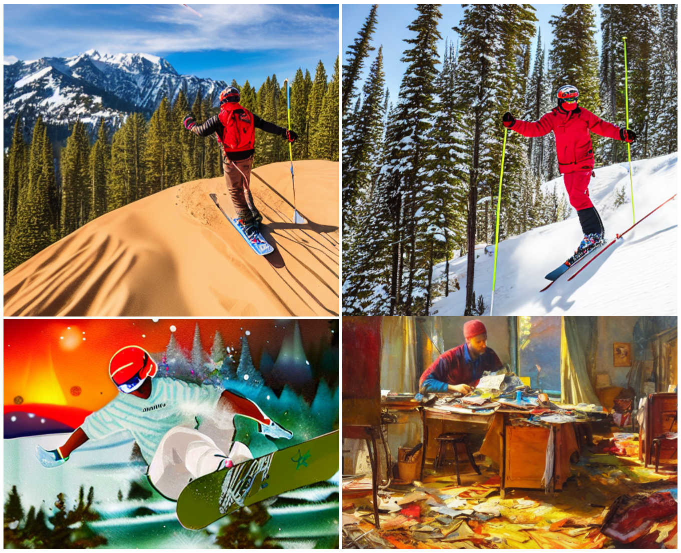

Fig. 6: Examples of synthetic images generated with CIA from different reference image. (Top) Changes in the point of view with the same reference image (turning back or turning away). (Bottom) Changes in style (from photography to poster or painting).

An overview of the different types of images that can be produced with the five studied ControlNets was already given in Fig.1, Fig.2 and Fig.3. However, Fig.6 displays the introduction of new patterns in the images. Changes in the background (snow, forest, sand, etc.), point of view (turning back or turning away), and style (realistic, drawing, painting, photography, etc.) can be observed. Such differences could be of high interest depending on the task. This is merely a glimpse of the generation possibilities, that can be tailored through the prompt and the Stable Diffusion model choice, both of which can be easily modified with the CIA framework.

VII. Conclusion

CIA offers a plug-and-play capability for developing, testing, and evaluating custom image generation pipelines. This framework has the potential to have a significant impact on the field of computer vision by providing researchers with a powerful tool for augmenting datasets and exploring new metrics and diffusion models. We demonstrated the capabilities of CIA in augmenting limited object detection datasets. But, the adaptability of the CIA framework allows for easy extension to other tasks like classification, segmentation or tracking. It allows for the incorporation of custom Diffusion models, ControlNet models and quality metrics to further adapt CIA to any application. Moreover, through its modularity, a module can easily be replaced or added. For example, adding other generative AI methods (not based on Stable Diffusion).

Acknowledgment

Thanks to ARIAC/TRAIL for funding this research.

References

[1] H. Su, J. Deng, and L. Fei-Fei, "Crowdsourcing annotations for visual object detection," in *Workshops at the twenty-sixth AAAI conference on artificial intelligence*, Citeseer, 2012.

[2] B. Settles, "Active learning literature survey," 2009.

[3] M. Xu, S. Yoon, A. Fuentes, and D. S. Park, "A comprehensive survey of image augmentation techniques for deep learning," *Pattern Recognition*, vol. 137, p. 109347, 2023.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[5] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

[6] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[7] Y. Chen, Y. Li, T. Kong, L. Qi, R. Chu, L. Li, and J. Jia, "Scale-aware automatic augmentation for object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9558–9567, 2021.

[8] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2917–2927, 2021.

[9] T. Ø. Eliassen and Y. Ma, "Data synthesis with stable diffusion for dataset imbalance-computer vision," 2022.

[10] B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov, "Effective data augmentation with diffusion models," *arXiv preprint arXiv:2302.07944*, 2023.

[11] S. Azizi, S. Kornblith, C. Saharia, M. Norouzi, and D. J. Fleet, "Synthetic data from diffusion models improves imagenet classification," *arXiv preprint arXiv:2304.08466*, 2023.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[13] Y. Ge, J. Xu, B. Nlong Zhao, L. Itti, and V. Vineet, "Dall-e for detection: Language-driven compositional image synthesis for object detection," *arXiv preprint arXiv:2206.09592v3*, 2022.

[14] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*, pp. 8821–8831, PMLR, 2021.

[15] W. Wu, T. Dai, X. Huang, F. Ma, and J. Xiao, "Image augmentation with controlled diffusion for weakly-supervised semantic segmentation," *arXiv preprint arXiv:2310.09760*, 2023.

[16] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/referenceless image spatial quality evaluator," in *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*, pp. 723–727, IEEE, 2011.

[17] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE transactions on image processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[18] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp. 2555–2563, 2023.

[19] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299, 2017.

[20] I. Grishchenko, A. Ablavatski, Y. Kartynnik, K. Raveendran, and M. Grundmann, "Attention mesh: High-fidelity face mesh prediction in real-time," *arXiv preprint arXiv:2006.10962*, 2020.

[21] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[22] W. S. Mseddi, R. Ghali, M. Jmal, and R. Attia, "Fire detection and segmentation using yolov5 and u-net," in *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 741–745, IEEE, 2021.

[23] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[24] G. Jocher, A. Chaurasia, and J. Qiu, "Yolo by ultralytics," jan 2023.

[25] G. Jocher, "Yolov8 hyperparameter config files."

[26] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Doll'a r, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014.

[27] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

[28] G. Jocher, "Yolov8 data augmentation docs of ultralytics," Nov 2023.