Energy Proceedings
Vol 36, 2023
ISSN 2004-2965

Leveraging Generative AI for Renewable Energy: Photovoltaic Panel Semantic Segmentation Case Study

Zhengyuan Lin[1], Zhiling Guo[23*], Dou Huang[3], Chenchen Song[5], Hongjun Tan[2], Xiaoya Song[6], Haoran Zhang[4], Jinyue Yan[2]

1 South China Normal University

2 The Hong Kong Polytechnic University

3 The University of Tokyo

4 Peking University

5 Beijing Information Science and Technology University

6 Harbin Institute of Technology

(*Corresponding Author: zhiling.guo@polyu.edu.hk)

ABSTRACT

As solar energy gains prominence, the demand of photovoltaic (PV) panels has increased. To assess photovoltaic power capacity, it is vital to derive accurate distribution information of PV panels. Common cost-effective approach involves deep learning technique such as semantic segmentation. However, available datasets remain scarce and expensive. Fortunately, Generative Artificial Intelligence (Generative AI), specifically text-conditioned diffusion models, exhibits the potential to automatically generate high-resolution synthetic images paired with annotations created from cross-attention maps, serving as training datasets for photovoltaic panel semantic segmentation. In this study, we employ the off-the-shelf Stable Diffusion model to explore the power of Generative AI to address dataset limitations and curtail data collection and annotation expenses. From the outcomings, we believe that Generative AI will play a revolutionary role in renewable energy systems.

Keywords: solar energy, Generative AI, semantic segmentation, photovoltaic panel

1. INTRODUCTION

The application of solar energy as a kind of renewable energy source has gained significant attention in recent years, leading to an increasing demand for photovoltaic (PV) panels that can efficiently convert solar energy into electricity. For accurately assessing the capacity of PV panels, deep learning technique has been adopted as an effective mean. Nevertheless, the ability of segmentation models to recognize PV panels precisely is hindered due to the paucity of adequate training data. The traditional method of obtaining training datasets and annotations through remote sensing techniques and manual masking is expensive and laborious. For practical tasks, we must face challenges about the scarcity and absence of available datasets and corresponding annotations, which is frequent barrier when harnessing deep learning technique in the realm of renewable energy applications. As an innovative alternative, Generative AI[1], e.g., text-conditioned diffusion models[2, 3], provides a possibility to create free training datasets paired with annotations[4], significantly reducing the cost of making datasets. In this study, we leverage the off-the-shelf Stable Diffusion model[3] to investigate the potential of synthetic datasets with annotations created from cross-attention maps for PV segmentation.

Our approach shown in Fig.1, addresses the challenge of insufficient data via text-to-image generation, text-guided image-to-image generation, and text-guided image inpainting. By these techniques, we evaluate the quality of synthetic datasets generated by varying prompt combinations encompassing text, image, and mask, for proving the feasibility of the application in practical tasks as real data and annotations are exceedingly scarce or even absent. For text-to-image generation and text-guided image-to-image generation, we make annotations based on cross-attention maps[4,

This is a paper for the 9th Applied Energy Symposium: Low Carbon Cities and Urban Energy Systems (CUE2023), Sep. 2-7, 2023, Matsue & Tokyo, Japan.

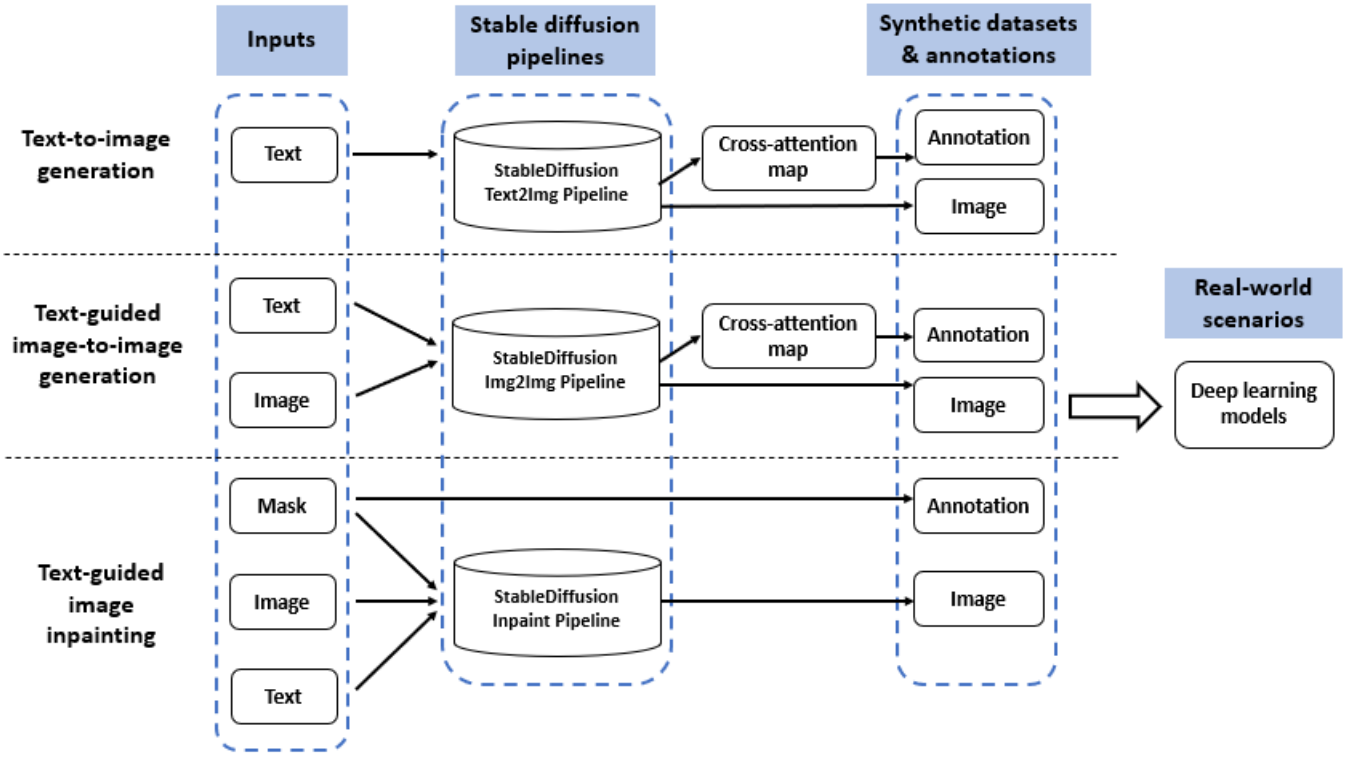

Fig. 1. The overview of the proposed approach

5, 6], a pivotal component of text-guided image generation, which reflects the intrinsic connection between text prompt and vision content, serving as potential annotations, shown in Fig.2.

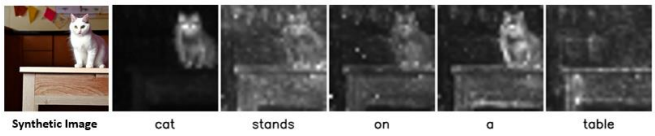

Fig. 2. Synthetic image generated with "cat stands on a table" as text prompt via text-to-image generation, and the cross-attention maps of different text tokens.

By refining cross-attention map of particular class such as "cat" shown in Fig.2, we can obtain high-resolution synthetic images with annotations, which is one of our basic ideas.

To our best knowledge, the application of synthetic images paired with annotations generated by Generative Ai in the renewable energy domain, specifically for recognizing PV panels, remains largely unexplored and challenging. Since there are notable disparities between synthetic and real data, encompassing factors like distribution, features, perspective, and size of PV panels[7], it is worthy of a comprehensive evaluation of segmentation method performance trained on synthetic data. Furthermore, this study delves into the prospect of augmenting datasets through text-guided image inpainting, effectively reconstructing original images by reimagining backgrounds while preserving vital objects in need of recognition.

In this study, we try to figure out these questions:

- Can synthetic datasets created by Generative AI effectively facilitate PV semantic segmentation tasks?
- As only using text as prompts to generate datasets, whether performance of segmentation methods trained by synthetic datasets generated via text-to-image generation is satisfactory enough?
- Can more variety of prompts improve the performance?

2. BACKGROUND

2.1 PV semantic segmentation

Semantic segmentation stands as a foundational task within the realm of computer vision, involving the assignment of every pixel in an image to a specific class or object, culminating in the creation of a detailed pixel-wise segmentation map. Leveraging the advancements in Convolutional Neural Networks (CNNs) [6, 8, 9], numerous groundbreaking methods have surfaced, such as FCN[9, 10], U-Net[11] and DeepLabV3+[12], achieving remarkable outcomes. However, considering the application in specific field, the availability of both sufficient and high-quality training datasets remains a persistent challenge. For PV semantic segmentation, the acquisition of superior datasets via remote sensing techniques proves to be an expensive endeavor,

consequently contributing to the scarcity of adequate training data. This scarcity is further exacerbated by the unique attributes of PV panels within aerial imagery. Typically, PV panels appear relatively minor in size and amount, leading an insufficient representation of positive class instances within datasets. Taking aerial imagery obtained from Heilbronn as an example, after data preparation, sections containing PV panels constitute merely approximately 9.7% of the total imagery. Impressively, sections where photovoltaic panels comprise more than 10% of an image are exceedingly sparse, accounting for a mere 1.1% of the entire imagery. Evidently, employing aerial datasets from specific regions for PV segmentation seems inherently inefficient. Conversely, Generative AI introduces a transformative avenue by enabling the creation of synthetic training datasets, enriched with an ample positive class representation, addressing the limitation of adequate real training data.

2.2 Generative AI

Generative Artificial Intelligence (AI)[1] embodies the capability of producing diverse forms of content, including text, images, and other media. Recently, Generative AI has gained much attention in society. At the forefront of this advancement, advanced Generative AI techniques have revolutionized content creation by automating the generation of extensive content volumes in a short time. For instance, large language model chatbots, such as InstructGPT[13] and ChatGPT can adeptly decipher and respond to human language inputs, forging meaningful interactions. Text-to-image artificial intelligence models, such as DALL-E-2[2] and Stable Diffusion[3], is capable of creating high-quality images from textual prompts in a few minutes.

In this paper, we leverage the Stable Diffusion model[3] specially to generate synthetic dataset. Prior works like DatasetGAN[14] and BigDatasetGAN[15] can generate synthetic image and precise mask based on a few labeled real images, while Stable Diffusion model can generate synthetic images with annotations only relying on text supervision, and have the ability to utilize few images and masks to refine results to make their style closer to real imagery, which make it possible that generate free and high-quality datasets. Innovatively, Diffumask[4] introduces a methodology combined with DenseCRF[16] and AffinityNet[17] to create the pixel-level semantic mask of generative images via text-to-image generation. By contrast, we leverage three kinds of techniques, including text-to-image generation, text-guided image-to-image generation, and text-guided image inpainting, to evaluate the potential of Generative AI for PV segmentation. Moreover, our annotations are refined from cross-attention maps by simple binarization processing without complicated techniques, but gain a commendable result in segmentation models.

3. METHODOLOGY

3.1 Mask Generation

In this paper, we explore simultaneously generating images and the annotations via text-guided diffusion model. Central to our approach is the strategic harnessing of the cross-attention map, acting as the link between textual token and visual content, containing rich spatial localization information and semantic information[4, 5, 6]. Specifically, we utilize Stable Diffusion model as our experimental foundation model, comprising a text encoder[18], a variational autoencoder(VAE)[19], and a U-shaped network. The fusion of textual and visual information process within the U-Net architecture, where cross-attention layers blend the embeddings of visual and textual characteristics and produce cross-attention maps for each textual token. To make annotations, we use the average cross-attention map calculated by aggregating the multi-layer and multi-time cross-attention maps, then binarize the average cross-attention map. The process of binarization is orchestrated by identifying pixels surpassing a certain threshold, typically defined as the maximum pixel value multiplied by a predetermined threshold. Pixels that breach this threshold are assigned a value of 1, while the rest are set to 0. By applying the strategy, we can get synthetic dataset paired with annotations efficiently, shown in Fig. 3.

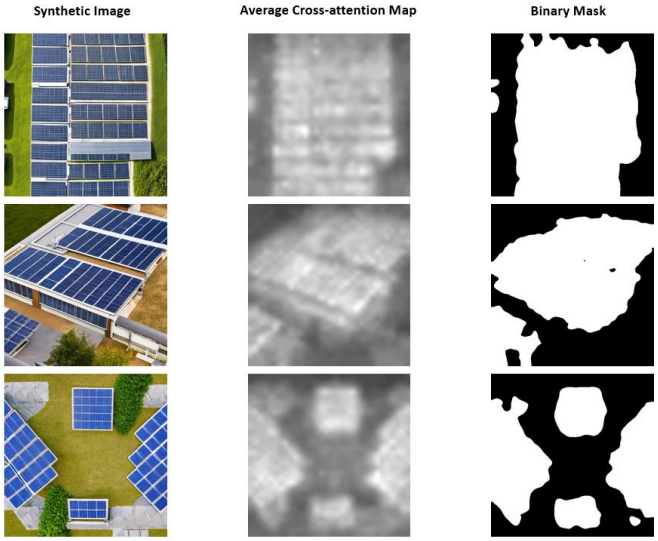

Fig. 3. Synthetic Images generated with “an overhead aerial image of photovoltaics panels mounted on rooftops” as text prompt via text-to-image generation, the cross-attention maps of the “panels” token, and the corresponding binary results, serving as weak semantic mask.

4. EXPERIMENTS

4.1 Datasets

For each technique, we created 3k synthetic images paired with annotations as training datasets for PV semantic segmentation training. To evaluate their performances, the high-resolution aerial imagery of Heilbronn, paired with accurate PV panel annotations, is used for semantic segmentation validation.

4.2 Evaluation metrics

In this study, we use the metric of Intersection-Over-Union (IoU) based on confusion matrix for PV semantic segmentation to evaluate the performance of models trained on each synthetic datasets generated via above techniques to segment PV panels.

4.3 Implementation details

4.3.1 Text-to-image generation

We ran Stable-Diffusion-2-1-base model using 30 inference steps per image with the text prompt as "an overhead aerial image of photovoltaics panels mounted on rooftops", to generate 3k synthetic images with annotations for PV segmentation. From Fig. 3, we can notice that the synthetic images are high-resolution with rich positive class.

4.3.2 Text-guided image-to-image generation

For text-guided image-to-image generation, based on the setting of text-to-image generation, we add one real image from Heilbronn imagery as image prompt, to make the style of the synthetic images closer to real data, shown in Fig. 4.

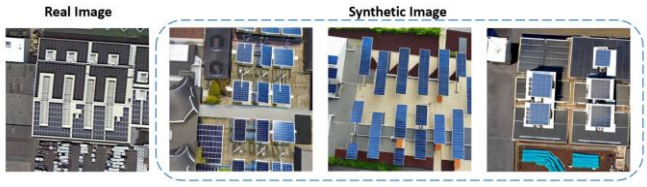

Fig. 4. The leftmost one is the real image provided as image prompt and the others are synthetic images. Synthetic Images are generated with "an overhead aerial image of photovoltaics panels mounted on rooftops" as text prompt via text-guided image-to-image generation.

4.3.3 Text-guided image inpainting

We utilize stable-diffusion-2-inpainting model resumed from stable-diffusion-2-base model to reconstruct the negative class of image to augment the diversity of datasets. In this experiment, we take 50 images with refined inversed masks as prompts, using "an overhead aerial photograph, extremely detailed" as text prompt, and set inference steps for per image as 30. The sample of our image inpainting results are shown in Fig. 5, which adeptly preserve the distinctive features of PV panels while reconstructing background.

Fig. 5. The sample of text-guided image inpainting results.

4.4 Semantic segmentation model

We selected a typical state-of-the-art semantic segmentation model, U-Net, as the evaluation model, which adopts a combination of U-Net and FCN as the backbone in the experiments.

4.5 Training strategies

We train each model for 160000 iterations. The models are optimized by using the Stochastic Gradient Descent (SGD) algorithm with a learning rate of 0.01. The momentum and weight decay coefficients are set to 0.9 and 0.0005 respectively.

5. RESULTS AND DISCUSSION

Here, Fig. 6 illustrates representative sets of segmentation results in different proportions of PV panels based on U-Net model exclusively trained using synthetic datasets generated from each technique. Notably, it exhibits unexpected results in recognizing PV panels, substantiating the feasibility of harnessing generative AI to generate synthetic datasets for PV segmentation tasks with few prompts.

Fig. 7 demonstrates the performances of each technique visually. Considering the characteristic of PV panels in synthetic image, our evaluation centers on images where the PV panel proportion exceeds 0.1. As

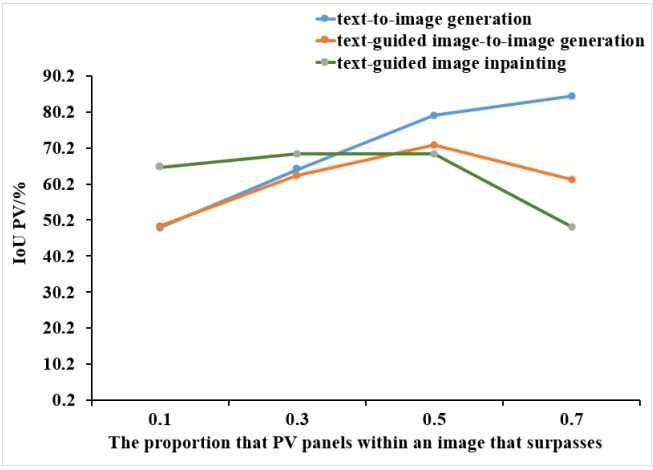

Fig. 7. Performance of PV segmentation with synthetic datasets for each technique using U-Net model, expressed by IoU comparison.

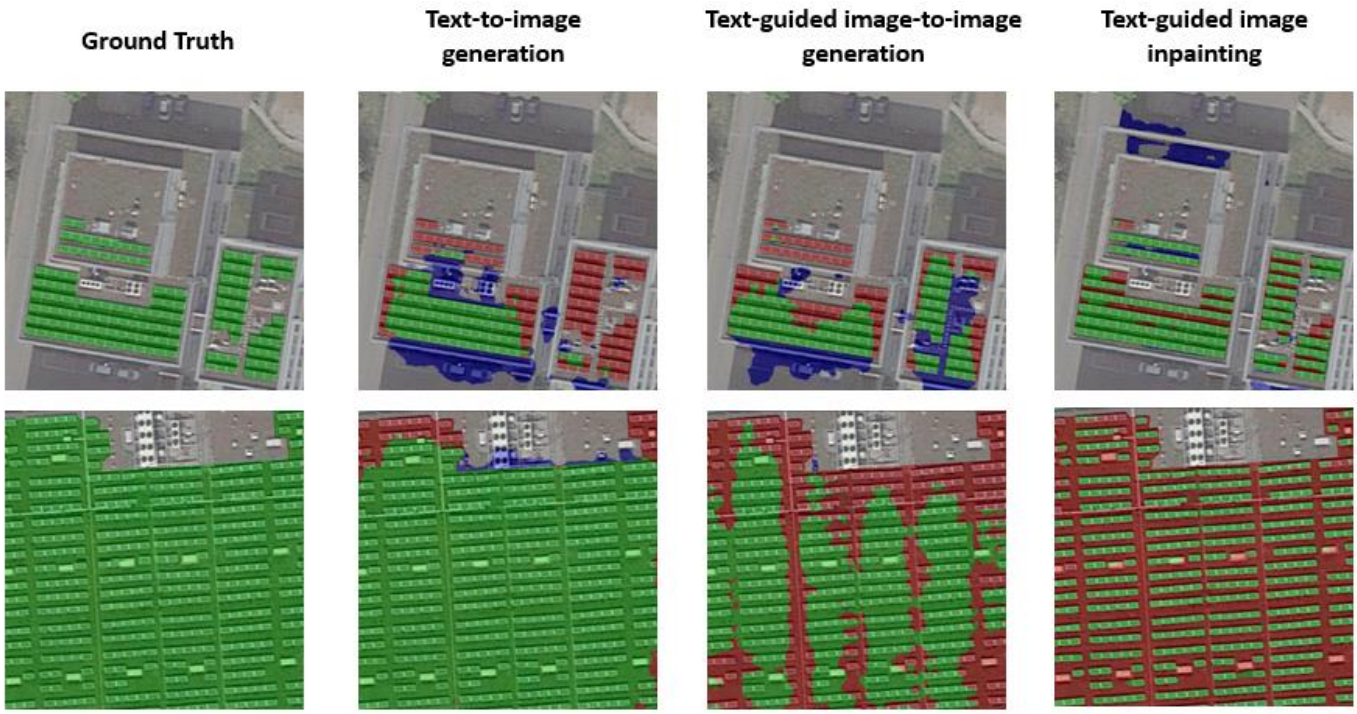

Fig. 6. Representative results to indicate the performance of the semantic segmentation models trained by synthetic datasets in different proportions of PV panels. In the ground truth, green color refers to positive class. In the results, green, red, blue colors refer to tp , fn , fp, respectively.

shown in Fig. 7, when the proportion that PV panels within an image is over 0.1, the models can perform satisfactorily, attaining IoUs exceeding 47.67%. As the proportion of PV panels is modest, the model trained on the synthetic datasets provided via text-guided image inpainting outperforms others, while with increasing PV panel proportions, the model trained using the synthetic datasets created via text-to-image generation exhibits superior performance, which are exhibited visually in Fig. 6 as well. Evaluating these outcomes roundly, the text-guided image-to-image generation technique is relatively less impressive in this experiment. This may be attributed to the scarcity of image prompts, potentially limiting the stylistic diversity of dataset. These results implies that visual features (i.e., texture features and color features) of PV panels provided from synthetic datasets play an important role in segmentation model to recognize the object. Nonetheless, facing the restricted proportions, class imbalance, and dispersion of PV panels within real data[7], model performance often falls short. Via text-guided image inpainting technique, it is possible to provide the more detailed localization information and texture information for synthetic datasets, aligning them more closely with real data characteristics, but due to the invariance of positive class, it also highlights a challenge to address the scarcity of original images with masks.

Generally, synthetic datasets paired with annotations created by Generative AI have the competence in PV semantic segmentation task and synthetic datasets generated exclusively using text prompt can provide satisfactory segmentation results, especially when PV panels are major in an image. The variety of prompts serves to bring synthetic datasets into closer alignment with real-world data, enhancing the performance of segmentation models to recognize objects within actual datasets. However, these prompts also constrain the diversity of the synthetic datasets due to their restricted range of types, thereby resulting in relatively less satisfactory outcomes in certain cases when compared to datasets generated only using text prompt.

6. CONCLUSION

In this study, we explore the feasibility to leverage Generative AI for PV semantic segmentation task, proving that synthetic datasets paired with annotations created by cross-attention maps have potential to replace the real datasets for the application of deep learning technique in renewable energy field, reducing the cost of datasets collection and annotations significantly. In the experiments, it becomes evident that while diverse prompts facilitate a closer resemblance to real data, they concurrently impose limitations on the diversity of synthetic images. Therefore, there emerges a trade-off we should consider between dataset diversity and similarity to real data. Given the scarcity of image prompts and mask prompts in general, it might be more

effective to exclusively employ text prompts for dataset generation as the objects requiring identification prominently occupy the image. To enhance the quality and stability of synthetic images and annotations, our future work is to apply efficient techniques like ControlNet[20], or leverage pretraining via LoRA[21]. Additionally, there exists a promising prospect of fusing cross-attention map and feature map to enhance the performance of segmentation task, since the feature maps[22, 23] from U-Net decoder encompass rich semantic information.

REFERENCE

[1] Y. Cao *et al.*, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," *arXiv preprint arXiv:2303.04226,* 2023.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125,* 2022.

[3] 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*, 2022, pp. 10684-10695.

[4] W. Wu, Y. Zhao, M. Z. Shou, H. Zhou, and C. Shen, "Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models," *arXiv preprint arXiv:2303.11681,* 2023.

[5] R. Tang *et al.*, "What the DAAM: Interpreting Stable Diffusion Using Cross Attention," *arXiv preprint arXiv:2210.04885,* 2022.

[6] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626,* 2022.

[7] P. Li *et al.*, "Understanding rooftop PV panel semantic segmentation of satellite and aerial images for better using machine learning," *Advances in Applied Energy,* vol. 4, p. 100057, 2021/11/19/2021,doi: https://doi.org/10.1016/j.adapen.2021.100057.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM,* vol. 60, no. 6, pp. 84-90, 2017.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 2015: Springer, pp. 234-241.

[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801-818.

[13] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems,* vol. 35, pp. 27730-27744, 2022.

[14] Y. Zhang *et al.*, "Datasetgan: Efficient labeled data factory with minimal human effort," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10145-10155.

[15] D. Li, H. Ling, S. W. Kim, K. Kreis, S. Fidler, and A. Torralba, "BigDatasetGAN: Synthesizing ImageNet with pixel-wise annotations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21330-21340.

[16] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in neural information processing systems,* vol. 24, 2011.

[17] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4981-4990.

[18] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021: PMLR, pp. 8748-8763.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114,* 2013.

[20] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv preprint arXiv:2302.05543,* 2023.

[21] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685,* 2021.

[22] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," *arXiv preprint arXiv:2303.02153,* 2023.

[23] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," *arXiv preprint arXiv:2112.03126,* 2021.