2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

WasteGAN: Data Augmentation for Robotic Waste Sorting through Generative Adversarial Networks

Alberto Bacchin[1†], Leonardo Barcellona[1], Matteo Terreran[1], Stefano Ghidoni[1], Emanuele Menegatti[1] and Takuya Kiyokawa[2]

***Abstract*— Robotic waste sorting poses significant challenges in both perception and manipulation, given the extreme variability of objects that should be recognized on a cluttered conveyor belt. While deep learning has proven effective in solving complex tasks, the necessity for extensive data collection and labeling limits its applicability in real-world scenarios like waste sorting. To tackle this issue, we introduce a data augmentation method based on a novel GAN architecture called wasteGAN. The proposed method allows to increase the performance of semantic segmentation models, starting from a very limited bunch of labeled examples, such as few as 100. The key innovations of wasteGAN include a novel loss function, a novel activation function, and a larger generator block. Overall, such innovations helps the network to learn from limited number of examples and synthesize data that better mirrors real-world distributions. We then leverage the higher-quality segmentation masks predicted from models trained on the wasteGAN synthetic data to compute semantic-aware grasp poses, enabling a robotic arm to effectively recognizing contaminants and separating waste in a real-world scenario. Through comprehensive evaluation encompassing dataset-based assessments and real-world experiments, our methodology demonstrated promising potential for robotic waste sorting, yielding performance gains of up to 5.8% in picking contaminants. The project page is available at `https://github.com/bach05/wasteGAN.git`.**

I. INTRODUCTION

Robotic waste sorting systems aim to recognize and separate materials (such as paper, plastic, metal, and glass) in a chaotic waste stream with industrial robots [1]. Beyond mere automation, this technology can make an impact at multiple levels. It not only relieves humans from highly repetitive and burdensome tasks but also supports the advancement of a more circular economy. Consequently, there is a growing interest among companies [2] and the research community [3], [4] in developing robotic waste sorting technologies.

However, industrial waste sorting presents numerous challenges, including the irregular shapes of objects, high variability, and cluttered environments, all of which pose significant hurdles for perception and manipulation. Furthermore, the availability of comprehensive public datasets tailored for industrial waste management remains limited. To the best of our knowledge, ZeroWaste [5] is the only extensive dataset, collected from a paper sorting facility in the US. However, leveraging such data effectively poses significant challenges due to the inherent variations in object classes across sorting plants and the differing appearances of objects based on geographical locations [6]. This challenge is particularly evident when dealing with dense annotations such as semantic segmentation masks, despite their ability to accurately represent irregular object boundaries compared to other annotation techniques like bounding boxes.

To reduce the dependency on large training sets and boost the applicability in waste sorting facilities, our work investigates the opportunity of leveraging synthetic data. Previous attempts using procedural data generation [7] proved ineffective. Therefore, we propose to use Generative Adversarial Networks (GANs) to generate a large amount of synthetic examples and augment the real-world data to train semantic segmentation models. GANs have been preferred due to the possibility of being trained with limited computational resources and faster compared to other generative models,

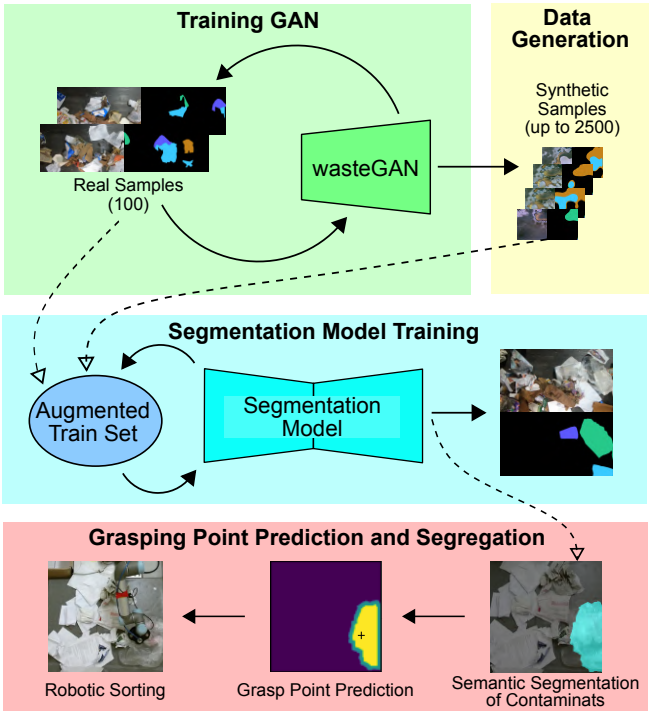

Fig. 1: A general overview of the proposed pipeline for robotic waste sorting.

† Corresponding Author `bacchinalb@dei.unipd.it`

This work was supported by the New Energy and Industrial Technology Development Organization (NEDO) grant number JPNP20012 and the Italian Minister for University and Research (MUR) under the initiative "PON Ricerca e Innovazione 2014 - 2020", CUP C95F21007870007

[1]Intelligent Autonomous System Lab, Department of Information Engineering, University of Padova, Padua, Italy

[2] Department of Systems Innovation, Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka, Japan.

such as Diffusion Models [8]. Additionally, GANs have proven successful in data augmentation, although primarily in the contexts of image classification [9], [10], [11], [12] and object detection [13], [14] tasks. While there have been some efforts in applying GANs to semantic segmentation tasks, such as *SemanticGAN* [15], these approaches have not been extensively explored in industrial settings characterized by unstructured object clusters, such as waste sorting facilities, and constrained by data scarcity.

In response to these challenges, we introduce a novel GAN architecture called *WasteGAN*. Building upon *SemanticGAN*, our architecture is specifically tailored for rapid convergence with minimal labeled data in the training set. Unlike traditional GANs, our focus is not on producing photo-realistic samples, but rather on generating synthetic samples that enhance the generalization capabilities of semantic segmentation models in real-world scenarios. Experimental results demonstrate the effectiveness of the proposed GAN-based augmentation method when used to train different off-the-shelf semantic segmentation models on the ZeroWaste dataset [5], the most popular benchmark for waste sorting. Compared to the baseline *SemanticGAN* [15], our approach demonstrates superior performance across most cases.

To further showcase the effectiveness of the augmentation achieved through *WasteGAN* in improving waste segregation, we integrated a semantic segmentation model trained on the augmented datasets into a real-world waste sorting system, composed of an industrial robot equipped with a vacuum gripper. Leveraging semantic segmentation of the scene, we determine optimal grasping points for target contaminant objects, enabling a robot equipped with a vacuum gripper to remove them from the waste stream. Notably, our approach addresses cluttered scenarios akin to those encountered in actual waste sorting facilities, distinguishing it from previous studies [16] that usually assume well-separated objects. The overall scheme of the proposed approach is illustrated in Fig. 1, starting from the top you can appreciate our data augmentation technique based on *WasteGAN* and the consequent grasping point generation for robotic sorting.

In summary, our contributions are (1) the introduction of *WasteGAN*, a GAN architecture designed for data augmentation to address the constraints of limited annotated data and cluttered unstructured scenes inherent of the waste sorting domain; (2) an extensive validation of the proposed approach by evaluating different state-of-the-art segmentation models trained on *wasteGAN* output, demonstrating an overall improvement in semantic segmentation performance when compared with other GANs synthetic outputs; (3) integration of the proposed augmentation model in a real robotic waste sorting system, exploiting the segmentation mask to extract a suitable grasping point for an industrial manipulator; (4) evaluation of the entire proposed methodology in a real paper sorting application developed in a mockup environment in our laboratory, highlighting the robustness and generalization capabilities of the model trained with *wasteGAN* when used in a different setting than the one of the training data.

II. RELATED WORKS

A. Robotic Waste Sorting Systems

A Robotic Waste Sorting System (RoWSS) is composed of two key elements: (1) a perception system tasked with identifying various types of waste within the working area and (2) a robotic arm to physically pick and segregate waste. RoWSS can be broadly categorized into (a) mobile platforms [3] able to collect waste from the environment and (b) industrial sorters intended for installation in waste management facilities [2].

Our focus in this study is primarily on the latter category. Koskinopoulou *et al*. [17] acknowledges the challenge posed by data scarcity and the difficulty of generalizing to new scenarios. The proposed solution involves a data augmentation technique that randomly pastes patches of waste images onto random backgrounds. Despite the promising results, their testing scenario lacks cluttered scenes and the variance of objects remains limited to those present in the original dataset. Conversely, Kiyokawa *et al*. [18] design a semi-automatic data collection procedure, but again applicable for single object detection. Using a generative model, we want to go beyond these limitations, dealing with cluttered scenarios and increasing the variance of the original dataset. In addition to perception, manipulating waste poses a significant challenge due to the large variability in the shape and dimension of objects. Ku *et al*. [19] propose to use depth images from an RGB-D camera to sample grasp poses, represented as grasp rectangles [20], and score them with a neural network to extract the most suitable one. However, this method is effective only when the target object is clearly distinguishable from the background in the depth image, thus impractical for highly cluttered scenarios. Um *et al*. [16] instead focuses on suction gripper technology, introducing a scoring algorithm to evaluate the quality of grasping poses on irregular surfaces. Despite the good results, this method passes over the semantics of the scene and solely focuses on computing the most suitable grasping pose. Recognizing the importance of waste classification in the sorting process alongside grasp success rate, we propose leveraging semantic segmentation masks to compute suitable grasping poses in our work.

B. Data Augmentation with GAN

Generative Adversarial Networks (GAN) have been introduced by Goodfellow *et al*. [21]. The proposed model was meant to generate new samples that were indistinguishable from a set of training examples. Many milestone improvements have been developed, such as *Conditional GAN* (C-GAN) [22], *Wasserstein GAN* [23], *Progressive-Growing GAN* [24] and *StyleGAN* [25], [26]. These models have demonstrated the capability to generate good-quality images in various scenarios, inspiring data augmentation techniques based on GANs. Most of the works are focused on classification because applying GANs in this domain is straightforward: it is possible to train multiple networks on different classes to generate per-class synthetic samples and

train a classifier [27]. However, this approach is very time-consuming when the number of classes increases. *Conditional GAN* solves the issue by using the label as a prior to synthesize images of a specific class, allowing to generate a labelled classification dataset [28], [29]. When it comes to more complex tasks, such as semantic segmentation, GANs have been applied in different ways. In the agriculture scenario, Fawakherji *et al*. [30] used segmentation labels to condition a C-GAN in order to generate new samples of crops of the rare classes in the dataset to reduce the class imbalance. However, this strategy is not suitable in the waste sorting scenario where many objects of different shapes and classes are cluttered. Li *et al*. [15] proposed a *StyleGAN*-based architecture, named *semanticGAN*, to achieve weak supervised semantic segmentation. Since images and labels are strictly correlated, the authors wanted to exploit the knowledge accumulated by the GAN trained on the unlabelled dataset to ease the generation of labels for an unknown query image. To do that, authors performed GAN inversion [31] to map the query image into the latent space of *StyleGAN* and then they drove the network to reconstruct the original image alongside a coherent segmentation mask. While this strategy has been demonstrated effective, it is slow in inferring segmentation masks, reducing its applicability in waste sorting scenarios where real-time processing is required. To overcome these limitations, in this work we investigate several architectural changes to *SemanticGAN*, focusing on generating images and segmentation masks to be used as training data for semantic segmentation models. A similar approach was explored in *DatasetGAN* [32], where the latent code of a *StyleGAN*-based architecture was used to generate semantic segmentation labels using a lightweight ensemble of classifiers. However, this idea relies on the assumption that the latent code reflects certain structures in the image (e.g., the position of eyes in a face) and fails in unstructured scenarios, such as randomized clusters of waste.

III. METHOD

Our hypothesis posits that by augmenting a small set of labeled data (*e.g*., 100) with generated samples, we can alleviate the burden of data collection while simultaneously improving model performance. In pursuit of this objective, the proposed *wasteGAN* introduces architectural innovations, including an optimized activation function, an enhanced generator, and a novel loss function. Leveraging semantic segmentation masks, we compute grasping poses to effectively separate waste with an industrial robot.

A. GAN Architecture

We designed a GAN architecture, evolved from previous work [15], [26], to effectively leverage limited amount of real-world data. The architecture of the proposed *wasteGAN* is shown on Fig. 2. Starting from *SemanticGAN* [15], we refine the architecture of the generator $G(w) : \mathcal{W} \rightarrow \mathcal{X}_\mathcal{G} \times \mathcal{Y}_\mathcal{G}$, where w is a latent variable and X_G , Y_G are the synthetic image-label couples. Since we do not need to reconstruct the original image, we removed the unnecessary GAN-inversion encoder while retaining the fully connected mapping network $FC : \mathcal{Z} \rightarrow \mathcal{W}$ introduced in [26]. Instead, we enhanced the generator's capabilities by augmenting the size of its blocks, achieved by stacking an additional style layer atop the original two. This refined generator is denoted as *GenXL*.

We made use of two discriminators as in *SemanticGAN* [15]: (1) $D_{rgb} : \mathcal{X}_\mathcal{R} \cup \mathcal{X}_\mathcal{G} \rightarrow \mathbb{R}$ takes real ($\mathcal{X}_\mathcal{R}$) or generated ($\mathcal{X}_\mathcal{G}$) images and gives a score to discriminate between them; (2) $D_{seg} : (\mathcal{X}_\mathcal{R}, \mathcal{Y}_\mathcal{R}) \cup (\mathcal{X}_\mathcal{G}, \mathcal{Y}_\mathcal{G}) \rightarrow \mathbb{R}$ takes the concatenation of images and labels. While D_{rgb} favors the generation of high quality images, D_{seg} encourages the generation of consistent image-label pairs. We used a residual architecture for both discriminators, as in [26].

Originally, the last layer was a simple linear combination of previous layer's activations. Since discriminators are binary classifiers, the higher the confidence in the prediction, the larger the output score. Linear activation implies a quicker divergence and the overfitting of the discriminator, especially when the training set is small. To reduce this effect it is possible to use bounded activation like $\tanh()$, which however suffers of the zero-gradient problem. For these reasons, we designed the *Symmetric Logarithm* function, defined as:

$$\mathrm{slog}(x) = \begin{cases} \log(ax+1) & x \geq 0 \\ -\log(-ax+1) & x < 0 \end{cases}, \tag{1}$$

where $a > 0$ is a hyperparameter. The $\mathrm{slog}()$ function is differentiable in $\mathbb{R}$, and it better controls divergence in case of large input activation since it follows a logarithm trend, while mitigating the zero-gradient problem because the gradient never goes to zero.

B. Loss Function

The *WasteGAN* adversarial loss comes from the hinge loss [33], [34], we refer to this as *cADV*. For the discriminators, we introduce the following hinge losses:

$$\mathcal{L}_{D_{rgb}} = \mathop{\mathbb{E}}_{x \in \mathcal{X}_\mathcal{R}} \Big[\mathrm{ReLU}\big(k - D_{rgb}(x)\big) \Big] + \mathop{\mathbb{E}}_{x \in \mathcal{X}_\mathcal{G}} \Big[\mathrm{ReLU}\big(k + D_{rgb}(x)\big) \Big], \tag{2}$$

and

$$\mathcal{L}_{D_{seg}} = \mathop{\mathbb{E}}_{x \in \mathcal{X}_\mathcal{R} \cup \mathcal{Y}_\mathcal{R}} \Big[\mathrm{ReLU}\big(k - D_{seg}(x)\big) \Big] + \mathop{\mathbb{E}}_{x \in \mathcal{X}_\mathcal{G} \cup \mathcal{Y}_\mathcal{G}} \Big[\mathrm{ReLU}\big(k + D_{seg}(x)\big) \Big], \tag{3}$$

with $k \in [0, 1]$. The optimal value of $\mathcal{L}$ is $D^*(x) = \mathrm{sign}(P_R(x) - P_G(x)) = \pm k$. Usually k is set to 1, such that the discriminator tends to converge to the $D^*(x) = \pm 1$. We instead set k to 0.5 to shift the equilibrium point towards $D^*(x) = \pm 0.5$ to stay further from the low-gradient area.

For the generator, we used a loss composed of 3 terms:

$$\mathcal{L}_G = \mathcal{L}_h + \mathcal{L}_q + \mathcal{L}_{imc}, \tag{4}$$

where $\mathcal{L}_h$ is an hinge adversarial loss, $\mathcal{L}_q$ is the *quality loss* and $\mathcal{L}_{imc}$ is the *image-label correlation loss*. The first

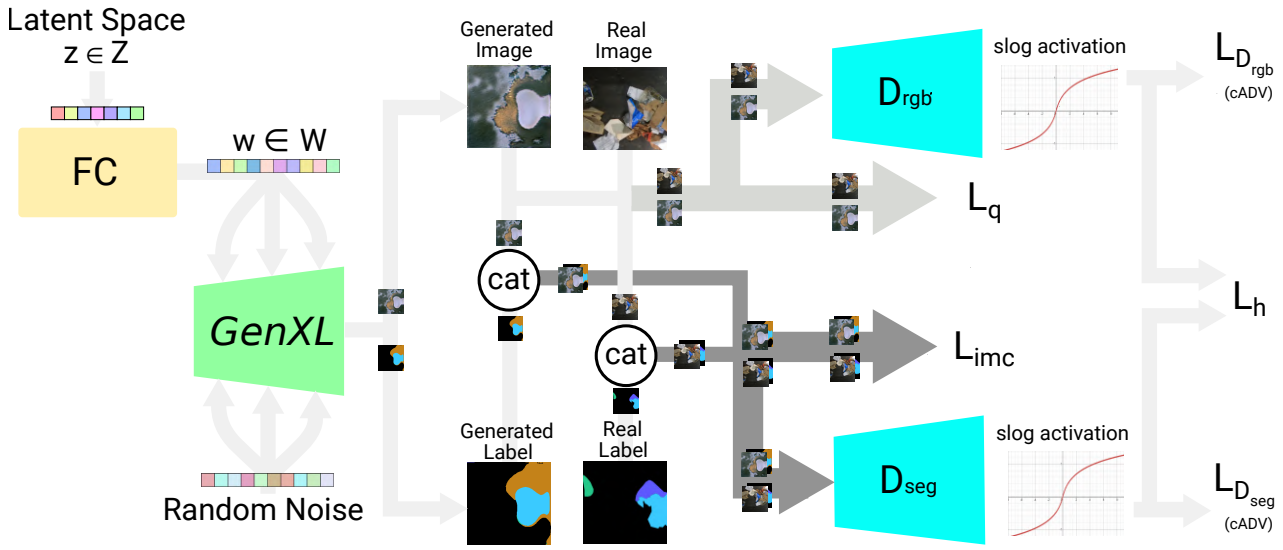

Fig. 2: The overall structure of the proposed *wasteGAN*. In particular, we highlight the usage of the newly proposed methods, *i.e.*,the improved generator (*GenXL*), the *image-label correlation loss* ($\mathcal{L}_{imc}$), the custom adversarial loss (*cADV*) and the *quality loss* ($\mathcal{L}_q$).

term in Eq 4 is an hinge loss, taking in account of both discriminators.

$$\mathcal{L}_h = -\alpha \mathop{\mathbb{E}}_{x \in \mathcal{X}_\mathcal{G}} \Big[D_{rgb}(x) \Big] + (1-\alpha) \mathop{\mathbb{E}}_{x \in \mathcal{X}_\mathcal{R}} \Big[D_{rgb}(x) \Big] + \\ - \alpha \mathop{\mathbb{E}}_{x \in \mathcal{X}_\mathcal{G} \cup \mathcal{Y}_\mathcal{G}} \Big[D_{seg}(x) \Big] + (1-\alpha) \mathop{\mathbb{E}}_{x \in \mathcal{X}_\mathcal{R} \cup \mathcal{Y}_\mathcal{R}} \Big[D_{seg}(x) \Big]. \tag{5}$$

The hinge loss $\mathcal{L}_h$, expanded in Eq 5, penalizes the generator when the discriminator correctly classify an image, whether it is real or synthetic. we employed an unbalanced hinge loss to better manage overfitting on the real data. We empirically set $\alpha = 0.8$, placing more emphasis on the first term to achieve a better balance in training.

The second loss term in Eq. 4, $\mathcal{L}_q$ or *quality loss*, is used to encourage high quality on synthetic images and is composed of two terms: $\mathcal{L}_q = \mathcal{L}_p + \mathcal{L}_s$. The first term $\mathcal{L}_p$ is a perceptual loss [35] based on VGG features. Since GAN tends to generate image with smooth edges, we also added $\mathcal{L}_s = \text{MAE}\big(\mathfrak{S}(x_R) - \mathfrak{S}(x_G) \big)$, where $\text{MAE}(\cdot)$ is the mean absolute error and $\mathfrak{S}(x) = \sqrt{(x \circledast S_x)^2 + (x \circledast S_y)^2}$ is a measure of the sharpness through Sobel operator.

Finally, $\mathcal{L}_{imc}$ or *image-label correlation* loss has been designed to promote the correlation between synthetic images and labels. It is defined as a mean absolute error as follows:

$$\mathcal{L}_{imc} = \text{MAE}\big(P_R(p_i \,|\, l_j) - P_G(p_i \,|\, l_j) \big), \tag{6}$$

where the conditional distributions describes the probability that a pixel with value p_i is assigned to the label l_j, respectively in the real and generated image-mask couples.

By minimizing $\mathcal{L}_{imc}$ loss, we aim to maintain the relationship between labels and masks we have in real samples also in the generated samples. Since $P_G(p_i \,|\, l_j)$ cannot be computed a priori, we compute on the fly the probability distributions using the image-mask couples in each mini-batch and then we updated the actual distribution estimations with exponential moving average (EMA). This term in the loss function is also meant to retain annotations for the infrequent classes, balancing the convergence to produce more varied labels.

C. Training procedure

Our model has been implemented in *PyTorch*, starting from the implementation of *StyleGAN2* [26] from [36]. We trained our GAN model with ADAM optimizer with learning rate 10^{-4} for the discriminators and the generator and 10^{-6} for the mapping network W. At each training iteration, we update discriminators D_{rgb} and D_{seg} first and generator G afterwards. Since discriminators are updated using the weights of G, in Eq. 5 the gradient of second and forth terms is non-zero.

We trained our model for 80K steps on a single NVIDIA RTX 4090 GPU. We did not apply gradient penalty [37] since the architectural changes proposed in Sec. III can control gradients, avoiding the slow down in training time. It is worth noticing that in the previous work of Li *et al.* [15], the gradient propagation from D_{seg} to G was stopped since the authors aimed to align the generated label with the synthesized image, as opposed to altering the image creation process to fit the labels, in order to generate a correct label for the input image. In our case, however, we do not need this constraint since we want to generate both images and labels.

We trained the proposed *wasteGAN* architecture on a dataset of 100 labeled examples taken from ZeroWaste [5]. The dataset contains images from a real-world paper sorting plant and manually labelled with semantic segmentation masks of 5 classes (background+paper, rigid plastic,

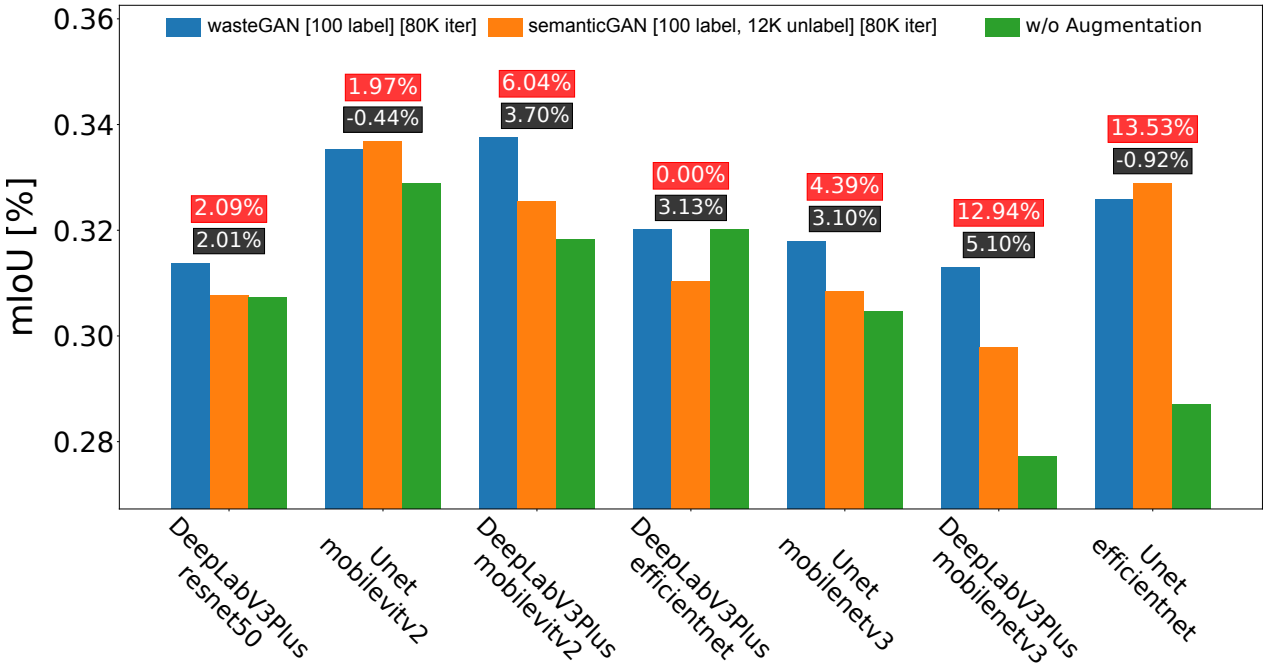

Fig. 3: Results from testing the models trained on synthetic datasets generated with *semanticGAN* [15] (orange), our approach (blue) and the dataset with only 100 real samples. Black values are the performance gains of *wasteGAN* with respect to *semanticGAN*. Red values are the performance gains with respect to training on 100 samples.

cardboard, metal, soft plastic). For comparison, we trained *semanticGAN* [15] too. Since it requires a weakly supervised setting, we prepared an additional unlabelled set of data of 11780 images obtained gathering the rest of ZeroWaste [5] and part of ZeroWaste-v2 [7] dataset. The latter is acquired in a paper sorting plant as well, but it does not provide any label.

D. Grasp Point Inference

The generator G produces synthetic samples $(x_G, y_G) \in X_G \cup Y_G$, which are combined with a limited amount of real-world data to train state-of-the-art real-time semantic segmentation networks to obtain better performance. The segmentation masks derived from this process are subsequently employed to determine suitable grasping points for a robot equipped with a suction cup, enabling the effective segregation of contaminants.

To compute grasping points, we considered two requirements: (1) the grasping point has to belong to the contaminant with high confidence and (2) the grasping point should be near the center of the object to facilitate a stable grip. To fulfill these requirements, we design a simple yet effective algorithm. To optimize the positioning of the suction cup within each segmented object, we apply erosion to the boundaries of the segmentation mask using a kernel size matching the radius of the suction cup. This ensures that the cup is accurately placed within the object boundaries. Next, to determine the centroid of irregularly shaped objects, we iteratively applied an erosion morphological operator until all clusters in the mask were reduced to single pixels. We associate a score to each point based on the logits predicted by the network. The final grasping point prediction will be the point with the highest score *i.e.*, the point with the higher confidence in category prediction. Some examples are shown in Fig. 6. Note that it is also possible to consider the top-k points with the highest scores to compute grasping points for k objects.

IV. Experiments

To evaluate the effectiveness of *WasteGAN* for data augmentation from limited annotated samples, we articulated experiments both on ZeroWaste [5] dataset and in real-world.

A. Evaluation on Dataset

The evaluation pipeline on dataset is composed of four phases: (A) the training of a GAN model on the small set of 100 random examples from ZeroWaste [5] training set; (B) the creation of a synthetic dataset by random sampling from the GAN generator G; (C) the training of several off-the-shelf semantic segmentation models [38] with different augmentation ratios (*i.e.*, the ratio between the synthetic and real examples used in training process); (D) the testing of the trained models on a set of unseen examples. For the latter, we employed the test set of *ZeroWaste* [5] and the mIoU as a metric. For testing, we considered 8 different segmentation models combining the popular Unet [39] and DeepLabV3+ [40] heads with Resnet50 [41], EfficientNet [42], MobileNetV3 [43] and MobileViTV2 [44] backbones, pretrained on ImageNet-1k [45].

Following the aforementioned evaluation procedure, we trained the semantic segmentation models, with augmentation rates ranging from 0 (*i.e.*, only 100 real examples) to

Aug Mode	A_C	A_G	FPR
No Aug	0.44	0.29	0.3
semanticGAN [15]	0.32	0.18	0.7
wasteGAN (ours)	**0.45**	**0.35**	**0.2**

TABLE I: Comparison of the impact of different augmentation modalities on a real-world pick and place task. Bold indicates best scores. *No Aug* indicates a model trained on 100 real world samples. While *semanticGAN* and *wasteGAN* indicates that the model has been trained also with synthetic data.

25 (*i.e.*, 2500 synthetic examples in addiction to the 100 real samples already available). We trained *semanticGAN* [15] on *ZeroWaste* [5] for comparison. The *semanticGAN* architecture is one of the few works which rely on GANs to improve semantic segmentation performance under our same hypothesis of having only a few labelled data. Nevertheless, authors in [15] made use of an additional set of unlabelled samples which our *wasteGAN* does not require. After training *semanticGAN* with the additional unlabelled data, we followed the aforementioned pipeline from point (B).

Once completed, we collected the results in Fig. 3. Despite that the effect of the augmentation depends on the underlying architecture, we can observe that in most of the cases our *wasteGAN* performs better with an average gain of 2.2% with respect to *semanticGAN* and a 5.8% improvement against training on real samples only.

Evaluation on different semantic segmentation models showed the generalization capability of our approach while providing better results than state-of-the-art GAN-based solutions. We tried to investigate the reason behind this success by analyzing the label distribution in the real and synthetic datasets. In Fig. 4 we depict the label frequencies of the full training set of ZeroWaste [5] and of the datasets generated with *wasteGAN* and *semanticGAN*. It is easy to see that the proposed method can better reproduce the original distribution, even for less frequent classes of contaminants like "metal" or "rigid plastic".

The precise label distribution in *wasteGAN* is a key factor contributing to the improvements demonstrated by models trained on *wasteGAN* synthetic images. This advantage over *semanticGAN* primarily stems from the two modifications implemented in the *wasteGAN* generator. Firstly, the inclusion of an additional style block in the *GenXL* generator enhances the expressiveness of synthetic images. Secondly, the introduction of quality loss and image-label correlation loss terms in the generator's loss function enables a more accurate capture of the label distribution, especially for underrepresented classes.

Finally, it is worth to highlight that the absence of the encoder to perform GAN inversion in *wasteGAN* brings to a significant 20x speed-up in generation time, from $2.2s$ to $0.12s$, reducing the time needed to create new synthetic data and boosting the applicability in the real world scenarios.

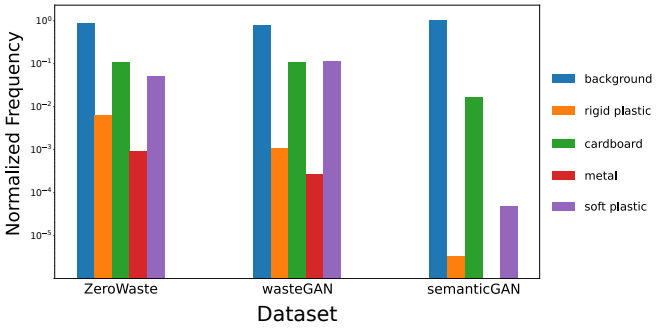

Fig. 4: Frequencies of the labels in the real-world dataset (ZeroWaste [5]) and the generated datasets with *semanticGAN* [15] and the proposed *wasteGAN*. Scale is logarithmic.

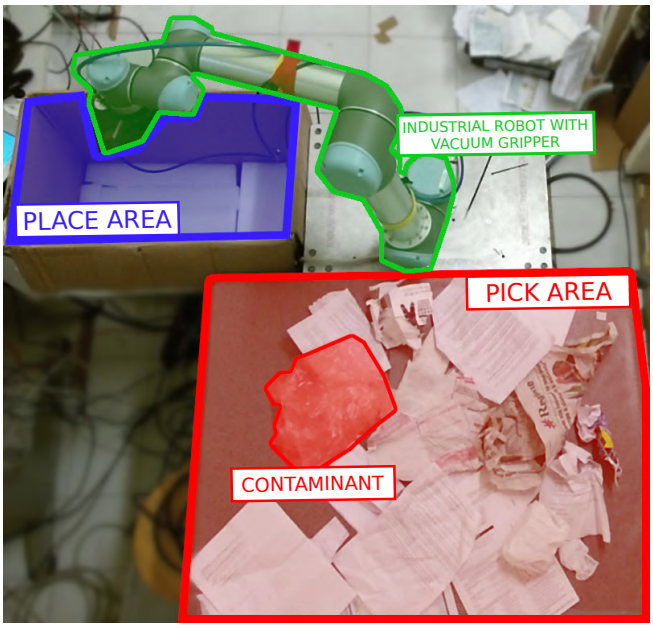

Fig. 5: Our experimental setup.

B. Evaluation with a Robotic Waste Sorting System

The experiments with the real robot had the objective to validate the grasping point prediction, in particular evaluating the impact of the GAN-based augmentation pipeline on the success rate of a pick and place task. The experiments have been held using a custom 3D-printed vacuum gripper applied to a UR5 robot and a KinectV2 RGB-D camera. We used the model showing best performances evaluated on ZeroWaste [5] *i.e.*, a DeepLabV3Plus + MobileViTV2 model.

After predicting the grasping point in the image space, we use the depth map to project the point into the robot space and execute the motion. Mirroring the composition of ZeroWaste [5], we collected real samples of waste belonging to the same five categories of the dataset. The sorting task consists in removing contaminants from paper waste in a cluttered scenario and drop into the place area, see Fig. 5.

For each run, we randomly place some paper waste as a background in the robot's workspace and we placed a

contaminant of specific category inside the clutter. To fairly compare the 3 different augmentation modalities, we replace the target object in the same position without changing the background. To evaluate each run we take in account of (1) the accuracy in recognizing the contaminant and its category through the semantic segmentation model (A_C) and (2) the accuracy in picking it and drop it in the place bin (A_G). We performed a total of 58×3 runs with different contaminants. Additionally, we performed 10×3 runs with solely paper waste in the working space to evaluate the false positive rate of the models (FPR).

Results are shown in Table I, alongside prediction examples in Fig. 6. It can be appreciated how the proposed augmentation method significantly enhances both recognition and grasping accuracy. While the quality of the images generated by *semanticGAN* seems to penalize the performance of the model, our *wasteGAN* is able to increase the generalization capabilities of the segmentator on a different scenario. Although the performance in recognizing the target object are comparable between *wasteGAN*-augmented and not augmented model, the higher mask quality provided by the former lead to a better grasping point prediction and so an higher accuracy for the whole pick and place task. Notably, the augmentation process with the proposed pipeline is also able to reduce the FPR which is important for real-world application in order to avoid useless robot working cycles.

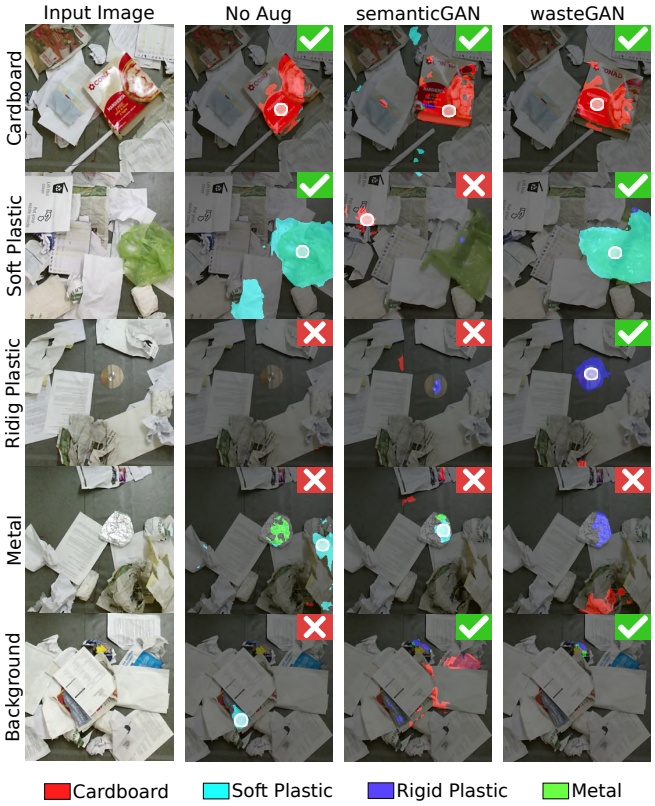

Fig. 6: Examples images taken from real world experiments. Each row represents a target contaminant or background. The white circle represents the predict grasp point (if any). Colors represent the segmented areas for each class. You can appreciate both success (✓) and failure (✗) cases.

V. Conclusions

This work presents an innovative approach for robotic waste sorting, addressing challenges of data scarcity and cluttered scenes. Our focus lies in optimizing a GAN architecture to facilitate training with limited data while preserving real distribution characteristics. By leveraging augmented datasets with synthetic samples, we train a semantic segmentation model and utilize the generated masks to compute semantic-aware grasping poses. This enables the deployment of a real robotic waste sorter, allowing us to evaluate both the model's generalization capabilities in real-world scenarios and the performance of the proposed semantic-aware grasping pose computation. The results confirm that the data augmentation process through *wasteGAN* significantly improve the performance in the real-world experiments, highlighting the capabilities of the proposed approach to tackle the domain shift between different settings.

Despite the promising results, we acknowledge the potential for further enhancements. To improve the effectiveness of data augmentation, our plan is to shift the augmentation process from the image space to the feature space. Working in a latent space has demonstrated a successful strategy with others generative models [8]. Additionally, we intend to explore advanced grasping point computation techniques, for example integrating our semantic-aware approach with existing class-agnostic methods [16] and better leveraging the degrees of freedom of a robotic arm.

References

[1] M. Koskinopoulou, F. Raptopoulos, G. Papadopoulos, N. Mavrakis, and M. Maniadakis, "Robotic waste sorting technology: Toward a vision-based categorization system for the industrial robotic separation of recyclable waste," *IEEE Robotics & Automation Magazine*, vol. 28, no. 2, pp. 50–60, 2021.

[2] H. Wilts, B. R. Garcia, R. G. Garlito, L. S. Gómez, and E. G. Prieto, "Artificial intelligence in the sorting of municipal waste as an enabler of the circular economy," *Resources*, vol. 10, no. 4, 2021.

[3] X. Chen, H. Huang, Y. Liu, J. Li, and M. Liu, "Robot for automatic waste sorting on construction sites," *Automation in Construction*, vol. 141, p. 104387, 2022.

[4] F. Raptopoulos, M. Koskinopoulou, and M. Maniadakis, "Robotic pick-and-toss facilitates urban waste sorting," in *Proc. IEEE International Conference on Automation Science and Engineering (CASE)*, 2020, pp. 1149–1154.

[5] D. Bashkirova, M. Abdelfattah, Z. Zhu, J. Akl, F. Alladkani, P. Hu, V. Ablavsky, B. Calli, S. A. Bargal, and K. Saenko, "ZeroWaste dataset: Towards deformable object segmentation in cluttered scenes," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 21 115–21 125.

[6] A. G. Satav, S. Kubade, C. Amrutkar, G. Arya, and A. Pawar, "A state-of-the-art review on robotics in waste sorting: scope and challenges," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 17, p. 2789–2806, 2023.

[7] D. Bashkirova, S. Mishra, D. Lteif, P. Teterwak, D. Kim, F. Alladkani, J. Akl, B. Calli, S. A. Bargal, K. Saenko, D. Kim, M. Seo, Y. Jeon, D.-G. Choi, S. Ettedgui, R. Giryes, S. Abu-Hussein, B. Xie, and S. Li, "VisDA 2022 Challenge: Domain adaptation for industrial waste sorting," in *Proc. Annual Conference on Neural Information Processing Systems (NIPS)*, 2023, pp. 104–118.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 684–10 695.

[9] W. Li, J. Chen, J. Cao, C. Ma, J. Wang, X. Cui, and P. Chen, "EID-GAN: Generative adversarial nets for extremely imbalanced data augmentation," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 3, pp. 3208–3218, 2023.

[10] Z. Du, L. Gao, and X. Li, "A new contrastive GAN with data augmentation for surface defect recognition under limited data," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2023.

[11] P. Ma, H. Lu, B. Yang, and W. Ran, "GAN-MVAE: A discriminative latent feature generation framework for generalized zero-shot learning," *Pattern Recognition Letters*, vol. 155, pp. 77–83, 2022.

[12] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "F-VAEGAN-D2: A feature generating framework for any-shot learning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 267–10 276.

[13] L. Liu, M. Muelly, J. Deng, T. Pfister, and L.-J. Li, "Generative modeling for small-data object detection," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6072–6080.

[14] M. Hammami, D. Friboulet, and R. Kechichian, "Cycle GAN-based data augmentation for multi-organ detection in CT images via YOLO," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 390–393.

[15] D. Li, J. Yang, K. Kreis, A. Torralba, and S. Fidler, "Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8296–8307.

[16] S. Um, K.-S. Kim, and S. Kim, "Suction point selection algorithm based on point cloud for plastic waste sorting," in *Proc. IEEE International Conference on Automation Science and Engineering (CASE)*, 2021, pp. 60–65.

[17] M. Koskinopoulou, F. Raptopoulos, G. Papadopoulos, N. Mavrakis, and M. Maniadakis, "Robotic waste sorting technology: Toward a vision-based categorization system for the industrial robotic separation of recyclable waste," *IEEE Robotics & Automation Magazine*, vol. 28, no. 2, pp. 50–60, 2021.

[18] T. Kiyokawa, H. Katayama, Y. Tatsuta, J. Takamatsu, and T. Ogasawara, "Robotic waste sorter with agile manipulation and quickly trainable detector," *IEEE Access*, vol. 9, pp. 124 616–124 631, 2021.

[19] Y. Ku, J. Yang, H. Fang, W. Xiao, and J. Zhuang, "Deep learning of grasping detection for a robot used in sorting construction and demolition waste," *Journal of Material Cycles and Waste Management*, vol. 23, pp. 84–95, 2021.

[20] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research (IJRR)*, vol. 34, no. 4-5, pp. 705–724, 2015.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, p. 139–144, 2020.

[22] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv:1411.1784*, 2014.

[23] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. International Conference on Machine Learning (ICML)*, 2017, pp. 214–223.

[24] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *Proc. International Conference on Learning Representations (ICLR)*, 2018.

[25] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4217–4228, 2021.

[26] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8107–8116.

[27] A. Bissoto, E. Valle, and S. Avila, "GAN-based data augmentation and anonymization for skin-lesion analysis: A critical review," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021, pp. 1847–1856.

[28] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, "CovidGAN: Data augmentation using auxiliary classifier GAN for improved Covid-19 detection," *IEEE Access*, vol. 8, pp. 91 916–91 923, 2020.

[29] J. J. Bird, C. M. Barnes, L. J. Manso, A. Ekárt, and D. R. Faria, "Fruit quality and defect image classification with conditional GAN data augmentation," *Scientia Horticulturae*, vol. 293, p. 110684, 2022.

[30] M. Fawakherji, C. Potena, A. Pretto, D. D. Bloisi, and D. Nardi, "Multi-spectral image synthesis for crop/weed segmentation in precision farming," *Robotics and Autonomous Systems*, vol. 146, p. 103861, 2021.

[31] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang, "GAN inversion: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3121–3138, 2023.

[32] Y. Zhang, H. Ling, J. Gao, K. Yin, J.-F. Lafleche, A. Barriuso, A. Torralba, and S. Fidler, "DatasetGAN: Efficient labeled data factory with minimal human effort," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 145–10 155.

[33] J. H. Lim and J. C. Ye, "Geometric GAN," *arXiv:1705.02894*, 2017.

[34] X. Chao, J. Cao, Y. Lu, Q. Dai, and S. Liang, "Constrained generative adversarial networks," *IEEE Access*, vol. 9, pp. 19 208–19 218, 2021.

[35] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[36] N. W. Varuna Jayasiri, "labml.ai annotated paper implementations," 2020. [Online]. Available: https://nn.labml.ai/

[37] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," in *Proc. Annual Conference on Neural Information Processing Systems (NIPS)*, 2017, p. 5769–5779.

[38] P. Iakubovskii, "Segmentation models pytorch," https://github.com/qubvel/segmentation_models.pytorch, 2019.

[39] C. Wang and C. Zhong, "Adaptive feature pyramid networks for object detection," *IEEE Access*, vol. 9, pp. 107 024–107 032, 2021.

[40] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. The European Conference on Computer Vision (ECCV)*, 2018, p. 833–851.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[42] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.

[43] A. Howard, M. Sandler, B. Chen, W. Wang, L. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324.

[44] S. Mehta and M. Rastegari, "Separable self-attention for mobile vision transformers," *arXiv:2206.02680*, 2022.

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.