2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)
June 28-30, 2023. Seattle, Washington, USA

Copy and Paste Augmentation for Deformable Wiring Harness Bags Segmentation

Bare Luka Žagar*, Alessio Caporali*, Amadeusz Szymko, Piotr Kicki,
Krzysztof Walas, Gianluca Palli and Alois C Knoll

***Abstract*—Wiring harnesses, i.e. a collection of electrical cables organized into branches, are vastly present in the automotive industry. Moreover, the number of wires and overall weight of automotive wiring harnesses are steadily increasing over time. Deformable wiring harness bags were introduced by manufacturers to simplify assembly operations. However, this task is still entirely performed manually by human labor. Despite the efforts, the degree of automation in wiring harness assembly is still close to zero. Due to the lack of task-specific datasets, modern state-of-the-art computer vision approaches are not commonly employed in the wiring harness industrial processes. In this work, we propose an approach to generate a dataset of a specific object of interest, i.e. deformable wiring harness bags, with minimal effort employing the copy and paste technique. The obtained dataset is validated on the semantic segmentation task in a real-world test setup, consisting of laboratory and automotive factory environments. An overall IoU of 53.8% and Dice score of 65.6% is obtained, demonstrating the capability of the proposed method.**

***Index Terms*—Deformable Objects, Segmentation, Data Augmentation, Industrial Manufacturing**

I. Introduction

According to a recent report by the European Environment Agency (EEA), passenger cars account for more than 60% of road transportation emissions [1]. Hence, to meet the climate targets set by the European Green Deal [2], an increase in the share of electric vehicles used in road transportation is needed [1]. However, the production of electric vehicles encounters a hard-to-tackle bottleneck, namely, the still manual manufacturing of wiring harnesses. Therefore, the production process of automotive wiring harnesses requires a drastic increase in automation.

A wiring harness is a collection of electrical cables organized into branches, where single wires and cables are grouped together with adhesive tape, cable ties, straps, or cable ducts [3]. Wiring harnesses are vastly present especially in the automotive sector, with a single car containing several wiring harnesses each devoted to specific functionalities, e.g. interior electronics, dashboard, and engine [3]. Nowadays, the number of wires and total weights of wiring harnesses in the car industry has steadily increased over time. Aiming at simplifying the assembly operations, manufacturers are starting to organize sub-sections of the wiring harness into bags, like the ones depicted in Fig. 1. Despite the manufacturing process of wiring harnesses has achieved some level of automation, the assembly operation during the production cycle of a car is still performed manually [3]. Indeed, difficulties in both manipulation and perception of these objects are affecting the introduction of a higher degree of automation [4], [5].

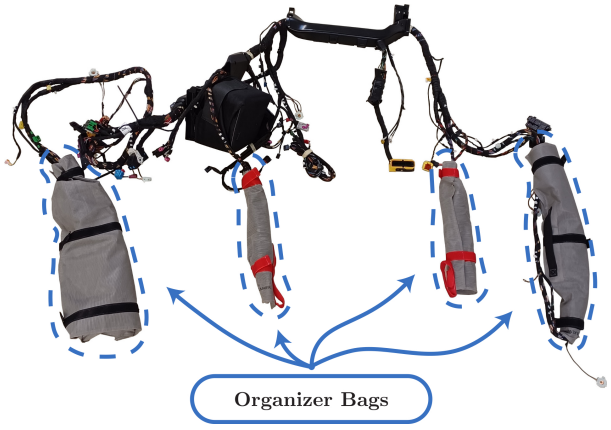

Fig. 1: Example of an automotive cockpit wiring harness. Subsections of the wiring harness are grouped and put into bags to ease the handling during the assembly process.

From the perception point of view, deploying a state-of-the-art detection system in a novel environment still consists of many impeding aspects, the major one being the lack of an annotated dataset consisting of *specific* objects of interest. This problem is usually addressed by either manually annotating the images and expanding the dataset with standard augmentation techniques [6], or by relying on synthetic environments and CAD data [7], [8]. However, when dealing with deformable objects, such solutions are less practical due to the problem of simulating effectively the deformability. Moreover, synthetic setups suffer from the *sim2real* gap, particularly affecting dense prediction settings, like semantic and instance segmentation tasks [9].

The lack of available datasets of deformable objects, like

Bare Luka Žagar and Alois C. Knoll are with the Department of Informatics, Technical University of Munich (TUM), Bolzmanstr. 3, 85748 Garching, Germany.

Alessio Caporali and Gianluca Palli are with DEI - Department of Electrical, Electronic and Information Engineering, University of Bologna, Viale Risorgimento 2, 40136 Bologna, Italy.

Amadeusz Szymko, Piotr Kicki and Krzysztof Walas are with the Institute of Robotics and Machine Intelligence, Poznan University of Technology, Poznan, Poland.

This work was supported by the European Commission's Horizon 2020 Framework Programme with the project REMODEL - Robotic technologies for the manipulation of complex deformable linear objects - under grant agreement No 870133. This research was also conducted in cooperation with Volkswagen Poznań, which provided the cockpit wiring harnesses.

Corresponding author: *bare.luka.zagar@tum.de*

*These authors contributed equally

wiring harness bags, is heavily affecting the introduction of automation in the assembly and manufacturing processes [10]. Even though there are few research works addressing deformable objects in industrial settings [11], [12], however, the perception problem is still present and far from being solved [13].

In this work, we propose a method to obtain with minimal effort a dataset composed of instances of deformable bags commonly found in automotive wiring harnesses, as shown in Fig. 1. We employ a *copy-and-paste* technique to generate training samples where the deformable bag instances are previously collected and annotated, and the backgrounds are taken from publicly available datasets [14], [15]. We validate the dataset on the semantic segmentation task [16] employing real-world test sets consisting of laboratory and automotive factory scenes.

Our key contributions are as follows:

- Deformable wiring harness bags dataset, which is generated using the *copy-and-paste* technique and image augmentations.
- Semantic segmentation validation of different state-of-the-art backbone architectures on the deformable wiring harness bags dataset.
- Extensive quantitative and qualitative evaluation in a real-world factory environment using different background styles.

II. Related Works

A. Data Augmentation

Data augmentation is a technique widely used in the deep learning domain to increase the size of the training dataset by slightly modifying the data samples. This technique acts as a regularization term and helps avoid overfitting during the training stages, improving generalization [17]. Standard data augmentation strategies can be grouped into geometric approaches, e.g. re-scaling and flipping, and photometric approaches, e.g. changing pixel values in terms of contrast, sharpness, blurring, brightness, and color.

Among the spread horizon of augmentation techniques, *copy-and-paste* [18]–[20] emerged as a key tool to supplement the original training dataset, achieving improved performances in common detection and segmentation tasks. *Copy-and-paste* is a method to compose *real* images by pasting object masks in backgrounds [18], thus avoiding the common drawbacks of synthetically rendered images and a dataset distribution shift. Usually, *copy-and-paste* is performed by randomly pasting the objects. The key insight exploited by the *copy-and-paste* technique, is that common deep learning methods pay a lot of attention to local region-based features as opposed to global scene layout [18]. Hence, by randomizing the object scale, viewpoint, and style of utilized backgrounds, a more uniform distribution of the object instances in the training dataset is achieved leading to a boost in performance. In [19] this approach is used in combination with random scale jittering for the training of instance segmentation models, which resulted in a better dataset distribution followed by improved accuracy. Instead, [20] applied *copy-and-paste* in a self-supervised contrastive pretraining scheme for improved segmentation performance of downstream tasks. This augmentation technique can be also employed on point cloud data, as presented in [21], where it is used to improve the performance of LiDAR-based semantic segmentation by realistically modeling the inserted objects' scan lines and shadows. These works demonstrate the capability of the *copy-and-paste* approach for a diverse set of tasks and frameworks.

Apart from *copy-and-paste*, other advanced augmentation approaches emerged in the literature. For instance, *Scale-aware AutoAug* [22] tackles the problem of discrepancy in scales in the dataset distribution. This discrepancy weakens the generalization capabilities of object detection and segmentation models across the diverse scales of objects. Thus, the authors introduce a method to learn a data augmentation policy to achieve scale invariance. Alternatively, *SemAug* [23] proposed a method to inject objects in contextual meaningful scenes, thereby augmenting the original training dataset. The trained network avoids becoming invariant to contextual information, e.g. as for random pasting of objects, but instead learns to exploit that contextual information as done by humans.

B. Deformable Objects Segmentation

In the last decade, the progress and impact of deep learning in the field of visual recognition were immense. In particular, Convolutional Neural Networks (CNN) established themselves as the primary framework for several vision tasks, thanks to some important properties like translation equivariance and efficiency in terms of the number of parameters [24]. In this regard, the *ResNet* family [25] was the most widely used backbone in vision tasks for several years. Recently, the field was altered with the advent of *Transformers* for vision tasks, showing superior performances compared to ResNet and CNNs in general, especially with larger models and datasets. The *Swin-Transformer* [26] spreads widely as a state-of-the-art backbone thanks to its efficiency and scalability to dense predictions compared to alternative *Transformers* architectures [24]. With *ConvNeXt* [24], there was a successful attempt to match the performances of transformers employing only convolution layers. Thus, the *ConvNeXt* family of backbone was demonstrated as a valid alternative to the *Swin-Transformer*.

Specific to the segmentation task, several popular networks were developed in the past [16], [27], [28]. Additionally, successful application to the context of deformable linear objects like cables and wires were obtained employing *DeeplabV3+* [16], both in a pure semantic segmentation task [9], [29] or as pre-processing step for downstream tasks [13], [30], [31].

III. Method

As mentioned in Sec. I, we are aiming toward the application of state-of-the-art deep learning methods for specific real-world applications, which require the collection of new data. Indeed, plenty of unique real-world problems are characterized

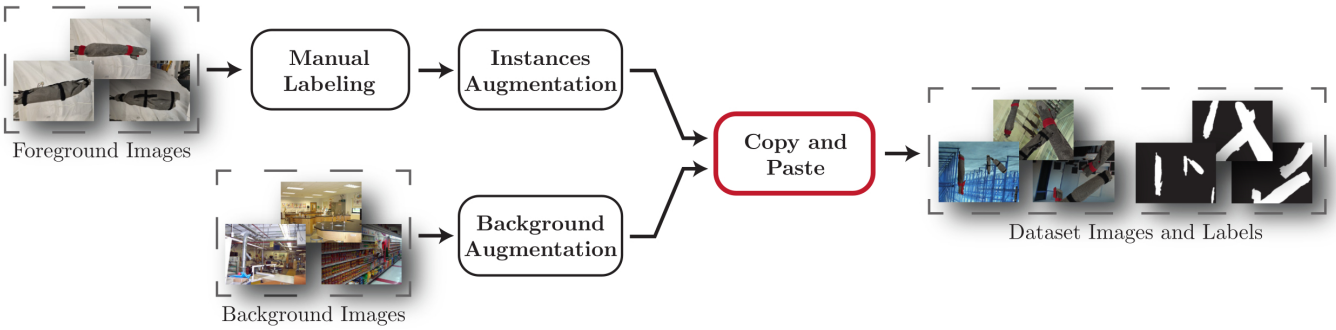

Fig. 2: Deformable wiring harness bags dataset generation pipeline based on *copy-and-paste* paradigm and data augmentation techniques.

by the lack of public data. Thus, we propose a dataset generation pipeline that requires minimal human intervention. We showcase our approach using the semantic segmentation task of wiring harness deformable bags, see Fig. 1. We organize the data into foreground images, i.e. the object of interest that we want to segment, and background images. Only a small set of foreground images is required to be annotated. Moreover, we leverage the simple and effective *copy-and-paste* [18] and data augmentation techniques to combine foreground and background images obtaining a broad and diverse dataset distribution.

The key steps of our dataset generation pipeline, shown in Figure 2, can be summarized as follows:

- **A)** Manual labeling of foreground images containing deformable wiring harness bags instances;
- **B)** Randomized selection and augmentation of background images;
- **C)** *Copy-and-paste* of a random amount of bag instances on each background image.

In the following, the details about the foreground and background image sampling are provided in Sec. III-A and Sec. III-B. The copy-and-paste method is discussed in Sec. III-C. Finally, the semantic segmentation framework used to validate the approach is presented in Sec. III-D.

A. Foreground Images

To generate a dataset for the segmentation of automotive wiring harness organizer bags, we first manually labeled 65 images constituting the set of bag instances to be used as the key ingredients of the pipeline. Due to the utilization of the *copy-and-paste* approach combined with standard data augmentation techniques, a small number of instances is sufficient and the time and effort for the human labeling are significantly reduced. In addition, better quality control of the annotation is obtained. For manual labeling, several tools are available, the most famous one being LabelMe[1]. From the pool of available tools, we selected the easy-to-use *segments.ai*[2] web interface which employs AI and a superpixels-based method to simplify labeling operations. In total, a labeling time of about 30 minutes was necessary to obtain accurate and reliable instance masks from the foreground input set. In Fig. 3 some labeled wiring harness bags with the background removed are shown.

Prior to performing the merging of the extracted instance bags with the background images, we exploit the possibility of performing data augmentation individually to each instance, thus enabling the creation of a greater variance in the final overall dataset. In particular, the following augmentation are performed: flip, noise, blur, rotation, shear, distortion, brightness and contrast, color jittering, and gamma.

B. Background Images

For good generalization on real-world data, it is necessary to have as similar as possible data distributions between the training data and the real-world environment. Therefore, the choice of background images is of fundamental importance. Since the application environment is expected to be a factory/industrial plant, we decided to rely on indoor scenes as opposed to outdoor ones. The need of collecting the set of background images is avoided by employing already publicly available datasets. In particular, the *MIT indoor* dataset [14] is selected to provide the background images. It contains 67 indoor categories and a total set of 15620 images. For comparison purposes, two additional datasets are tested: The first is the *HRSOD* dataset [15] containing 2010 high-resolution images; The second is a *complex* dataset containing 100 images displaying abstract and chaotic contents, similar to [29].

Additionally, to create diverse and broad training data, a set of standard augmentation techniques is applied to the

Fig. 3: Example of instance bags extracted from the foreground images.

[1]https://github.com/wkentaro/labelme
[2]https://segments.ai/

background images: rotation, cropping, shear, brightness, contrast, color jittering, blurring, and noise. By utilizing these augmentations, we ensure that every data sample has a different background which helps to create a broad training data distribution.

C. Copy & Paste combined with Augmentation

The *copy-and-paste* approach to enrich or generate new data is widely used and proved itself successful [19]. The idea is very simple. A random background I_{bg} is selected from the pool of available images and it is augmented as described in Sec. III-B. Similarly, an image of a wiring harness bag I_{fg} is randomly picked from the foreground set along with its annotation mask M_{fg}. Both I_{fg} and M_{fg} are augmented as detailed in Sec. III-A. Finally, I_{fg} is randomly pasted on I_{bg} as the following:

$$I_{\text{out}} = M_{\text{fg}} I_{\text{fg}} + (1_{h \times w} - M_{\text{fg}}) I_{\text{bg}} \tag{1}$$

where I_{out} is the obtained composed image and with $(1_{h \times w} - M_{\text{fg}})$ the inverse mask is computed, $1_{h \times w}$ being a unit matrix with the same size of the mask.

For a single background image, the operation described in eq. 1 is repeated for n number of bag instances, and the insertion location is randomized.

D. Semantic Segmentation Framework

As a general framework to test the proposed dataset generation approach and to validate the obtained dataset, the semantic segmentation task is used as a challenging benchmark. For the baseline method, the popular *DeepLabV3+* [16], an encoder-decoder architecture, is selected for its proven performances especially in decoding precise object boundaries.

As encoder, the original implementation of *DeepLabV3+* employs a modified *ResNet* [25] backbone with atrous convolutions, instead of the common convolutions, allowing the explicit control of the computed features resolution via the output stride parameter. In this work, we provide also a comparison of state-of-the-art backbone architectures, such as *Swin-Transformer* [26] and *ConvNeXt* [24], on the segmentation task.

The decoder consists of a simple yet effective module that refines the segmentation results along object boundaries. Here, the low-level features are concatenated to the bilinearly upsampled (4x) high-level features coming from the encoder. Several convolutions are performed to refine the features and a final upsampling (4x) is performed. This design choice of the decoder, compared to a direct bilinear 16x upsampling, provides improved performances [16].

IV. Experiments

A. Training Runs and Dataset

All the models are implemented in PyTorch 1.10.1 and trained with an NVIDIA GeForce RTX 3090 with 24 GB VRAM and on an AMD Ryzen 2950X 16-Core CPU clocked at 3.50GHz.

The dataset is obtained as detailed in Sec. III and it is composed of a total of 5000 samples having a resolution of 640×480 pixels with the common split of 90% for training and 10% for validation. The number n of foreground instances in each training image is bounded between 1 and 4.

As backbones are selected *ResNet101*, *SwinS* and *ConvNeXtS*. All of them share a similar number of parameters and complexity, thus making the comparison fair [24], [26]. The training runs are performed using common hyperparameters allowing an unbiased comparison among the different backbones. In particular, a total of 50 epochs are conceived. The early stopping procedure is enabled after the first 20 epochs and it is configured to end the training process when the validation loss does not decrease for 5 epochs in a row. The final weights of each run are selected as those having the minimum validation loss. As optimizer AdamW is selected while a polynomial learning rate scheduler is employed with power 0.95 and with the learning rate initialized at $5 \cdot 10^{-6}$. A batch size of 16 is selected for *ConvNeXtS* and *ResNet101*, whereas for *SwinS* it is configured to 12 due to memory limitations.

During training, a standard data augmentation scheme is employed: channel shuffling; hue, saturation, and value randomization; flipping; perspective distortions; random cropping; random brightness, and contrast.

B. Testing Dataset and Metrics

The segmentation network produces a mask that corresponds to the predicted semantic segmentation of the bags. We evaluate and compare the outputs of the training runs by means of the *Dice* coefficient (Dice $= 2\frac{|M_p \cap M_{gt}|}{|M_p| + |M_{gt}|}$) and the *Intersection-over-Union* (IoU $= \frac{|M_p \cap M_{gt}|}{|M_p \cup M_{gt}|}$, where M_{gt} is the ground truth and M_p the predicted mask for both formulations.

The models of Sec.IV-A are evaluated with *Dice* coefficient and IoU on a real *test set*. It is composed of a total of 75 images accurately annotated and organized into 3 different sub-classes (25 images each), in particular:

- *C1*: scenes with the wiring harness placed on an almost uniform background, e.g. table or floor, in a laboratory environment. Lights are uniform and occlusions are only due to other parts of the wiring harness.
- *C2*: scenes with the wiring harness hanging from the main branch, replicating the starting configuration at the beginning of the assembly operations in a factory environment. Occlusions are mainly due to other parts of the wiring harness.
- *C3*: scenes with the wiring harness lying on the cockpit assembly station, replicating the configuration during the assembly operations in a factory environment. Occlusions are due to complex mounting structures and a cluttered environment with shiny surfaces, and difficult lighting conditions.

C. Results

The quantitative evaluation of the real-world test sets of different variations of the DeepLabV3+ semantic segmentation

TABLE I: The average dice coefficient and intersection over union computed for each network and dataset configuration, across the test sets and particularly for each subgroup, namely *C1*, *C2* and *C3*. In all the tests the predictions are thresholded at 0.5. **Bold** denotes the best-performing method network backbone.

backbone	dataset	augmentation	*C1*		*C2*		*C3*		*all*	
			dice ↑	IoU ↑	dice ↑	IoU ↑	dice ↑	IoU ↑	dice ↑	IoU ↑
ResNet101 [25]	MIT indoor	✗	0.464	0.374	0.477	0.368	0.109	0.067	0.350	0.269
	MIT indoor	✓	0.525	0.411	0.553	0.425	0.173	0.111	0.417	0.315
	HRSOD	✗	0.205	0.130	0.192	0.134	0.086	0.048	0.161	0.104
	HRSOD	✓	0.197	0.120	0.190	0.126	0.107	0.062	0.164	0.102
	complex	✗	**0.704**	**0.567**	**0.788**	**0.659**	**0.397**	**0.285**	**0.629**	**0.503**
	complex	✓	0.691	0.546	0.756	0.616	0.367	0.259	0.604	0.473
SwinS [26]	MIT indoor	✗	0.611	0.495	0.513	0.391	0.311	0.232	0.478	0.372
	MIT indoor	✓	0.622	0.495	0.509	0.390	0.308	0.227	0.479	0.370
	HRSOD	✗	0.451	0.339	0.255	0.181	0.182	0.118	0.296	0.212
	HRSOD	✓	0.524	0.404	0.325	0.241	0.210	0.144	0.353	0.263
	complex	✗	0.658	0.504	0.465	0.335	0.317	0.196	0.480	0.345
	complex	✓	**0.676**	**0.522**	**0.485**	**0.352**	**0.339**	**0.212**	**0.500**	**0.362**
ConvNeXtS [24]	MIT indoor	✗	0.760	0.635	0.734	0.596	0.396	0.299	0.630	0.510
	MIT indoor	✓	**0.782**	**0.656**	**0.780**	**0.648**	**0.406**	**0.310**	**0.656**	**0.538**
	HRSOD	✗	0.616	0.474	0.147	0.087	0.277	0.202	0.346	0.254
	HRSOD	✓	0.435	0.301	0.138	0.078	0.248	0.176	0.273	0.185
	complex	✗	0.647	0.494	0.539	0.385	0.349	0.243	0.511	0.374
	complex	✓	0.678	0.526	0.545	0.389	0.349	0.243	0.524	0.386

network is provided in Table I. A general observation is that the models trained with *MIT indoor* backgrounds outperform the ones trained with *HRSOD* and *complex* backgrounds, with *ConvNeXtS* being the best performing one. This is due to the more similar data distribution of the *MIT indoor* dataset and real-world *test set*. The *ResNet101* and *SwinS* backbone-based models trained on the *complex* dataset achieve the best results, according to Table I. The difference is significant for *ResNet101* and less substantial for *SwinS*. Among the different sub-groups of the test images, the *complex* backgrounds help in achieving good performances, especially on the challenging *C3 test set*. Overall, the difference between the results for *MIT indoor* and *complex* might be that, for the latter, the neural networks learn to distinguish chaotic and unstructured backgrounds from everything else, resulting in a more effective generalization to the *C3* test images.

The qualitative comparison of different variations of the DeepLabV3+ model on the real-world test sets is shown in Figure 4. Interestingly, *ResNet101* and *SwinS* generalize well on the real-world test sets but fail to precisely separate the organizer bags and tend to oversegment. On the other hand, the *ConvNeXtS* backbone-based model trained on *MIT indoor* backgrounds is capable of segmenting the organizer bags more accurately which can be attributed to multiple reasons: 1) the *MIT indoor* dataset is contextually more similar to the real-world test sets so that the network does not learn to eliminate the background but to understand the underlying context within the scene; 2) thanks to the design changes in the ConvNeXt architecture - depth-wise convolution and patchify layers - the backbone can learn better and more high-level features.

V. Conclusions

In this paper, we address the problem of segmenting *task-specific* objects of interest, such as wiring harness bags. The peculiarity of these objects, apart from the deformability, is given by the lack of public datasets. Thus, a dataset generation pipeline relying on minimal human effort and based on the *copy-and-paste* approach combined with data augmentation techniques is proposed. The obtained dataset is validated experimentally on real-world test sets and satisfying results are achieved.

In the literature, there are still confusing statements regarding the importance of blending approaches and background image selection [19]. Therefore, active research is still needed to verify the importance of the above-mentioned aspects. Additionally, the extension of the transfer learning approach presented in [32] could be evaluated for segmenting automotive wiring harness bags and other specific real-world applications.

References

[1] E. Commission and E. E. Agency, *Decarbonising road transport : the role of vehicles, fuels and transport demand.* Publications Office of the European Union, 2022.

[2] E. Commission and D.-G. for Communication, *European green deal : delivering on our targets*. Publications Office of the European Union, 2021.

[3] J. Trommnau, J. Kühnle, J. Siegert, R. Inderka, and T. Bauernhansl, "Overview of the state of the art in the production process of automotive wire harnesses, current research and future trends," *Procedia CIRP*, 2019.

[4] T. Hermansson, R. Bohlin, J. S. Carlson, and R. Söderberg, "Automatic assembly path planning for wiring harness installations," *Journal of manufacturing systems*, 2013.

[5] P. Kicki, M. Bednarek, P. Lembicz, G. Mierzwiak, A. Szymko, M. Kraft, and K. Walas, "Tell me, what do you see?—interpretable classification of wiring harness branches with deep neural networks," *Sensors*, 2021.

[6] M.-W. Liu, Y.-H. Lin, Y.-C. Lo, C.-H. Shih, and P.-C. Lin, "Defect detection of grinded and polished workpieces using faster r-cnn," in *Proc. of AIM*. IEEE, 2021.

[7] M. Denninger, M. Sundermeyer, D. Winkelbauer, D. Olefir, T. Hodan, Y. Zidan, M. Elbadrawy, M. Knauer, H. Katam, and A. Lodhi, "Blenderproc: Reducing the reality gap with photorealistic rendering," in *International Conference on Robotics: Sciene and Systems*, 2020.

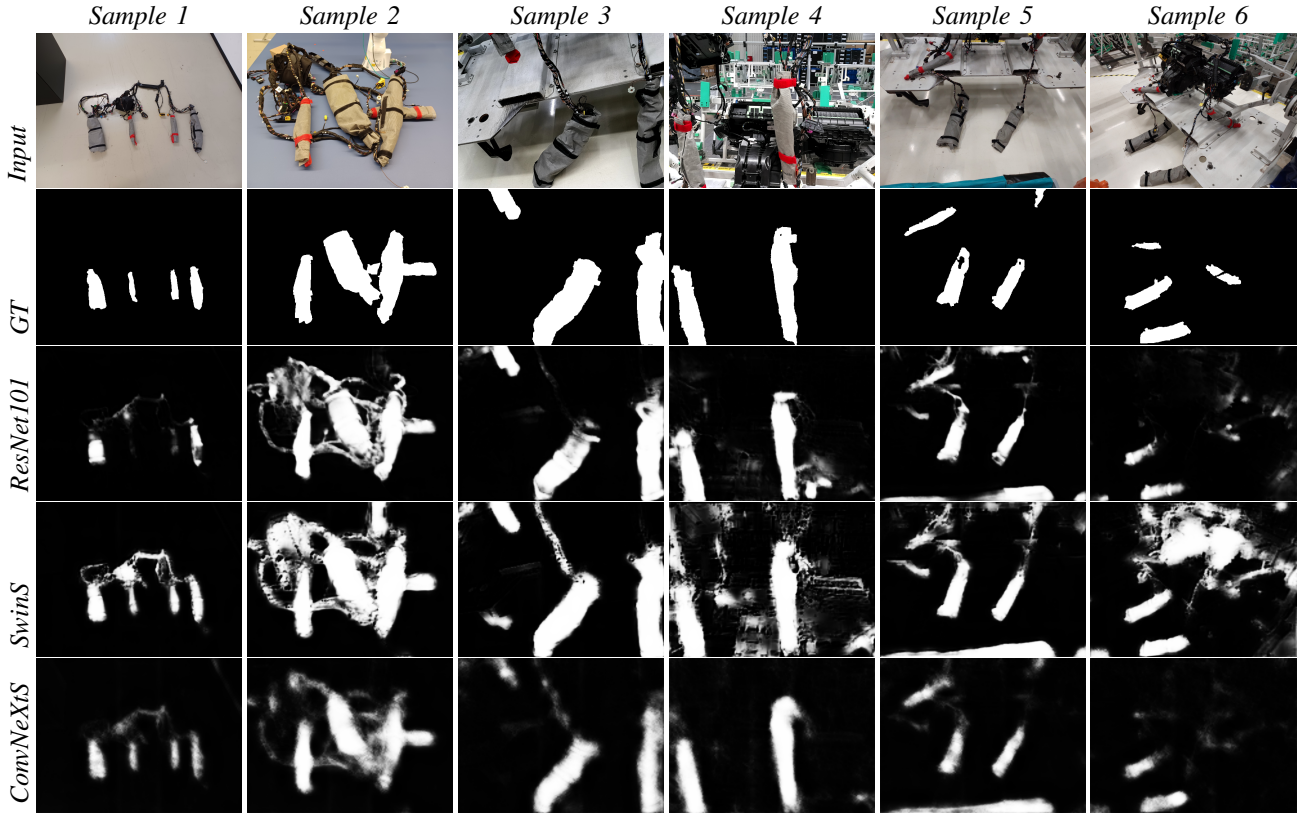

Fig. 4: Qualitative evaluation of DeepLabV3+ with the different backbones on six sample from the *test set*. From left to right, 2 samples for *C1*, *C2* and *C3* sub-classes. The results are chosen based on the best-performing model/dataset from Tab. I.

[8] W. Qiu and A. Yuille, "Unrealcv: Connecting computer vision to unreal engine," in *Proc. of ECCV*. Springer, 2016.

[9] A. Caporali, M. Pantano, L. Janisch, D. Regulin, G. Palli, and D. Lee, "A weakly supervised semi-automatic image labeling approach for deformable linear objects," *IEEE Robotics and Automation Letters*, 2023.

[10] R. Luque, E. Blanco, A. R. Galisteo, and E. Ferrera, "From augmented reality to deep learning-based cognitive assistance: An overview for industrial wire harnesses assemblies," in *Iberian Robotics conference*. Springer, 2023.

[11] X. Zhang, Y. Domae, W. Wan, and K. Harada, "Learning efficient policies for picking entangled wire harnesses: An approach to industrial bin picking," *IEEE Robotics and Automation Letters*, 2022.

[12] X. Jiang, K.-m. Koo, K. Kikuchi, A. Konno, and M. Uchiyama, "Robotized assembly of a wire harness in a car production line," *Advanced Robotics*, 2011.

[13] A. Caporali, R. Zanella, D. De Greogrio, and G. Palli, "Ariadne+: Deep learning–based augmented framework for the instance segmentation of wires," *IEEE Trans. on Industrial Informatics*, 2022.

[14] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. of the IEEE CVPR*. IEEE, 2009.

[15] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, and H. Lu, "Towards high-resolution salient object detection," in *Proc. of the IEEE/CVF CVPR*, 2019, pp. 7234–7243.

[16] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. of ECCV*, 2018.

[17] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, 2019.

[18] D. Dwibedi, I. Misra, and M. Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection," in *Proc. of IEEE ICCV*, 2017.

[19] 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 *Proc. of IEEE/CVF CVPR*, 2021.

[20] F. Wang, H. Wang, C. Wei, A. Yuille, and W. Shen, "Cp2: Copy-paste contrastive pretraining for semantic segmentation," *arXiv preprint arXiv:2203.11709*, 2022.

[21] Y. Ren, S. Zhao, and L. Bingbing, "Object insertion based data augmentation for semantic segmentation," in *Proc. of ICRA*. IEEE, 2022.

[22] Y. Chen, P. Zhang, T. Kong, Y. Li, X. Zhang, L. Qi, J. Sun, and J. Jia, "Scale-aware automatic augmentations for object detection with dynamic training," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2022.

[23] M. Heisler, A. Banitalebi-Dehkordi, and Y. Zhang, "Semaug: Semantically meaningful image augmentations for object detection through language grounding," in *Proc. of ECCV*. Springer, 2022.

[24] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proc. of the IEEE CVPR*, 2022.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE CVPR*, 2016.

[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. of the IEEE CVPR*, 2021.

[27] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. on pattern analysis and machine intelligence*, 2020.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015.

[29] R. Zanella, A. Caporali, K. Tadaka, D. De Gregorio, and G. Palli, "Auto-generated wires dataset for semantic segmentation with domain-independence," in *Proc. of ICCCR*. IEEE, 2021.

[30] A. Caporali, K. Galassi, G. Laudante, G. Palli, and S. Pirozzi, "Combining vision and tactile data for cable grasping," in *Proc. of AIM*. IEEE, 2021.

[31] A. Caporali, K. Galassi, and G. Palli, "3d dlo shape detection and grasp planning from multiple 2d views," in *Proc. of AIM*.

[32] B. L. Žagar, T. Preintner, A. C. Knoll, and E. Yurtsever, "Real-time instance segmentation of pedestrians using transfer learning," in *Proc. of ICAC*. IEEE, 2022.