Leveraging Foundation Models To learn the shape of semi-fluid deformable objects

Omar El Assal, Carlos M. Mateo, Sebastien Ciron, David Fofi

▶ **To cite this version:**

Omar El Assal, Carlos M. Mateo, Sebastien Ciron, David Fofi. Leveraging Foundation Models To learn the shape of semi-fluid deformable objects. 4th workshop on RObotic MAnipulation of Deformable Objects: beyond traditional approaches (ROMADO), IROS, Oct 2024, Abu Dhabi, United Arab Emirates. hal-04790483

HAL Id: hal-04790483
https://hal.science/hal-04790483v1

Submitted on 21 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Leveraging Foundation Models To learn the shape of semi-fluid deformable objects.

Omar El Assal[1,2], Carlos M. Mateo[3], Sebastien Ciron[1] , David Fofi[2]

***Abstract*— One of the difficulties imposed on the manipulation of deformable objects is their characterization and the detection of representative keypoints for the purpose of manipulation. A keen interest was manifested by researchers in the last decade to characterize and manipulate deformable objects of non-fluid nature, such as clothes and ropes. Even though several propositions were made in the regard of object characterization, however researchers were always confronted with the need of pixel-level information of the object through images to extract relevant information. This usually is accomplished by means of segmentation networks trained on manually labeled data for this purpose. In this paper, we address the subject of characterizing weld pool to define stable features that serve as information for further motion control objectives. We achieve this by employing different pipelines. The first one consists of characterizing fluid deformable objects through the use of a generative model that is trained using a teacher-student framework. And in the second one we leverage foundation models by using them as teachers to characterize the object in the image, without the need of any pre-training and any dataset. The performance of knowledge distillation from foundation models into a smaller generative model shows prominent results in the characterization of deformable objects. The student network was capable of learning to retrieve the keypoitns of the object with an error of 13.4 pixels. And the teacher was evaluated based on its capacities to retrieve pixel level information represented by the object mask, with a mean Intersection Over Union (mIoU) of 75.26%.**

I. INTRODUCTION

In order to accomplish manipulation tasks successfully, a robot has to perceive and comprehend the manipulated object. This consists of analysing information from several sources like images, points clouds or the model of the object [1], [2], [3]. Most of the research found in the literature focused on modeling and characterising rigid bodies [4]. However, the ever-increasing need for robots to interact with day-day objects, alongside the recent advances in machine learning and computer vision have raised an interest in the subject of deformable objects. For instance, the topic of modeling and estimating the state of these objects have attracted the interest of several researchers [5], [6], [7], [8], [9]. Which indicates that the manipulation of deformable objects presents several challenges imposed by the deformations of the manipulated object [10]. This paper targets fluid-like objects of deformable nature, and that presents high-dynamical topological surface changes. A texture rarely discussed in state of the art deformable objects manipulation of fabric and ropes [10], [11], [12], [13]. The shape of these fluid bodies is often governed by complex physical modeling, and their shape is difficult to predict. For instance, weld pools are fluid deformable objects, often described as visco-elastic, and their shape is influenced by different process parameters: robot speed, welding position, and the type of welded geometry [14], [15].

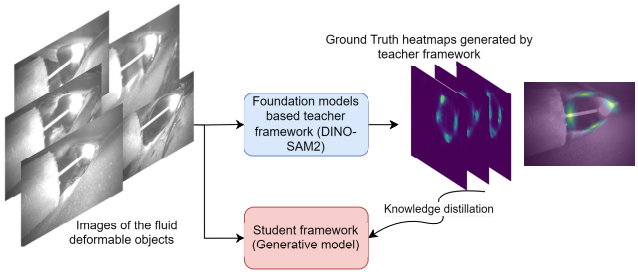

Fig. 1: Teacher-Student framework adopted for training a student network to characterize fluid deformable objects.

We tackle the issue of characterizing fluid deformable objects by describing them according to the shape of their contours. Our approach is based on a Teacher-student framework in which the knowledge of a teacher that consists of foundation models is distilled into a student lighter generative network as shown in Fig. 1. The teacher network consists of two foundation models, DINO [16] and SAM2 [17] to retrieve the shape of the object. We describe the object through heatmaps representing the keypoints of the object. These heatmaps are first generated by the teacher, and the knowledge is later transferred to the student network. The proposed approach targets an industrial application which is the state and shape estimation of weld pool in robotic welding task. Because modelling the deformations of molten metal pool is a key element to perform advanced control of the process such as feature based visual servoing. We also test and validate our proposal on this task and on a real-world dataset, generated in harsh industrial conditions.

II. RELATED WORK AND MOTIVATION

Describing deformable objects have always been a research question that accompanied the manipulation of these objects [18]; In many cases they were represented as objects as continuous geometrical entities like surfaces and

[1]Omar El Assal And Sebastien Ciron are with Alstom transports, 105 allée Albert Einstein – 71200 LE CREUSOT – France, omar.el-assal@alstomgroup.com, and sebastien.ciron@alstomgroup.com

[2]David Fofi and Omar El Assal with the IMVIA EA 7535 laboratory, university of Burgundy, 720 Av. de l'Europe, 71200 Le Creusot, France david.fofi@u-bourgogne.fr

[3]Carlos M. Mateo with the ICB UMR CNRS 6303, Universite de Bourgogne, 9 Av. Alain Savary, 21000 Dijon, carlos-manuel.mateo-agullo@u-bourgogne.fr

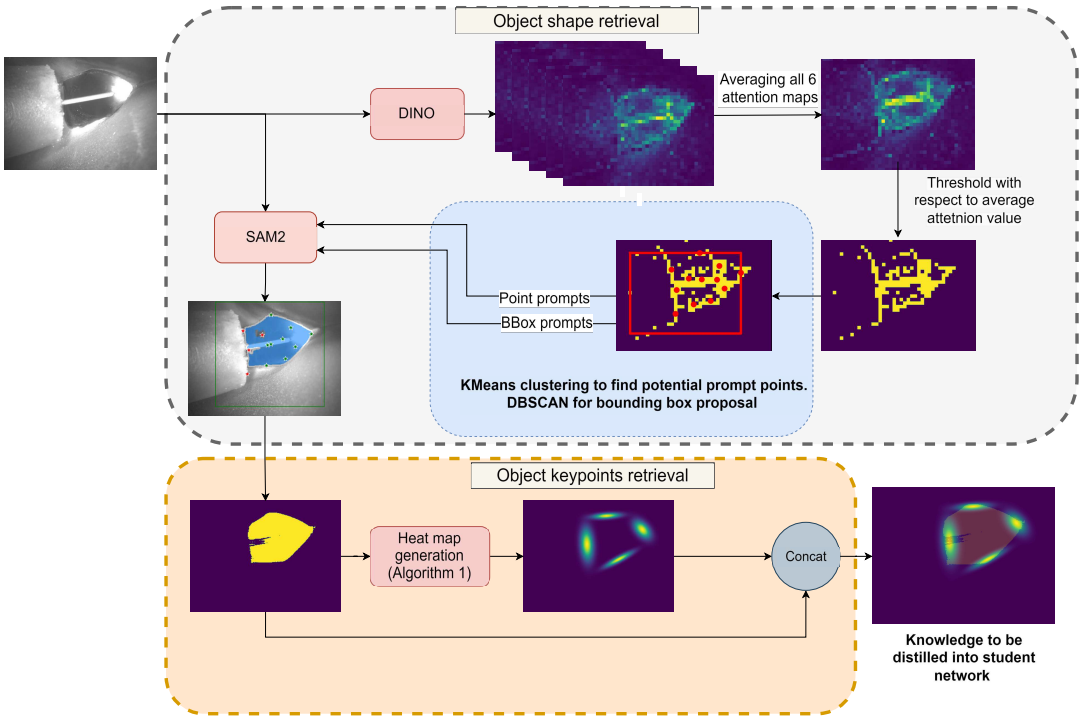

Fig. 2: Framework of the proposed ResNet-VAE method

curves [19], [20]. Some researchers however described them in a discrete representation through meshes, skeletons or landmark points [21]. And these objects were of different shapes and natures. In some cases they were of a linear deformable objects nature, and their characterization serves as a feedback for for later use in a geometric optimal control problem [22]. In other cases they were of more complex shapes to represent such as clothes [23]. The problematic of deformable clothing grasping in this case was resolved by assimalting clothes to polygonal shapes described by their corners. From a different perspective, [24] relied on the detection of extremed edges and points to represent objects such as towels and ropes. They presented a method that is based on the use of Convolutional Neural Networks. While this approach worked well for fabric, it still does not take into account the deformable nature of fluid objects. In [25], Coherent Point Drift (CPD) is applied after the points are registered with a Gaussiant Mixture model (GMM), to maintain feature from one frame to another. Softer objects such as dough characterization and manipulation have also been studied by researchers, in [26] visual images with depth maps are analyzed by Graph Neural Netwroks to learn a particle-based model of the object. Besides that, tissues are another type of soft deformable objecs that were explored by researchers. Feature extraction and state tracking of deformable of soft tissues by deep neural networks approach exist in the literature [27]. Nevertheless, manipulation and characterization of softer deformable objects, such as fluid and fluid like objects have been introduced [28], [5]. Where in [28] the manipulation of semi-fluid objects in a wok is addressed. These objects are considered as a single particle represented by the center of the object. In [5] a model-based method was used to manipulate water.

As discussed in earlier sections, the methods present in the state of the art do not apply to fluid and semi-fluid deformable objects. Although these methods are efficient for their specific task, they fail to be generalized for other types of applications and objects, since many of the proposed solutions are application oriented [29]. Providing model-free methods to estimate the shape and to characterize deformable objects is the main motivation of this work. This motivation is backed by the unpredictable nature of semi-fluid objects, such as weld pool and glue, and the difficulties of predicting their deformations. To address the problem of generalization, we propose the use of foundation models. Foundation models are large-scale models trained on a large scale of data and tasks, which allows for better analysis and understanding of this data [31]. We use these models as teachers for a smaller student network as depicted in Fig. 1.

As for the student network, we use a generative model. Generative models are mathematical models that analyze the underlying patterns in the streams of data. They have the capacity of generating new data with similar characteristics by learning underlying representations [30]. Several architectures are present in the literature such as Generative Adversarial Networks [32], Auto-encoders architectures [33], Gaussian Mixture Models (GMM), and Hidden Markov Models (HMM)) [30]. As for visual image generation the most commonly used are diffusion models [34], GANs [32], auto-encoders [33] and Variational Auto-encoders [35]. These models often need heavy computing powers and have complex convergence criteria [36]. Amongst these models,

the lighter, the fastet to covnerge and the most convenient at learning latent representation are Variational Autoencoders. The only cost is the clarity of the output data [37]. Thus, we define the architecture of the student network to be Variational Auto-Encoder (VAE).

III. Feature points extraction using Generative Models

We employ the teacher-student architecture of Fig. 1 in our framework. We consider the output of the teacher to be the corresponding ground truth. In this framework, the teacher takes as input the corresponding image of the deformable object and outputs a heatmap describing its shape. This heatmap is stacked with the mask of the object in the image to represent the output of the teacher. On the other hand, the student network has a variational auton-encoder architecture with a ResNet backbone. As illustrated in Fig. 2, the teacher consists of two different pipelines. The first one consists of finding the mask of the object by combining the two foundation models discussed earlier, DINO and SAM2, and the second one is dedicated to generate relevance heatmaps describing the keypoints of the object. In first place, the image is passed through DINO model. DINO is a self distillation network that is capable of generating attention maps describing the observed scene. After retrieving the six attention maps, their average is thresholded with respect to the mean value of attention to propose a 2D array with highest attention values. After that, we employ a pipeline to extract SAM2 prompts. SAM2 is another prompt based foundation model, that takes as input an image and possible prompts, and proposes a possible mask for these prompts on the output. We pass the original image through the SAM2 model, alongside with the extracted prompts from the DINO attention map in order to retrieve the pixel-level information of the object represented by the mask. After that, a heatmap that describes the keypoints of the object is proposed by Algorithm 1, this heatmap represented the most probable location of the keypoints, according to the normal lines on the contour of the object. Subsequently, the extracted heatmap alongside the previously defined mask are concatenated to represent the ground truth label which is distilled into the student network.

The generation of the heatmap goes through the different steps that are highlighted in Algoirhtm 1 and depcited in Fig 3. The contour of the object is extracted from the output mask of DINO-SAM. And Sobel operator is applied to find the values of normal lines. We refer to these lines according to their angle θ. Let C be the set of these points p_i,

$$[(x_i, y_i), \theta_i] = p_i \tag{1}$$

the pair (x_i, y_i) represents the coordinates of each point p_i in the contour C of the object, and θ_i the corresponding angular coordinate, codifying its normal angle.

The difference between θ_i and θ_{i-1} is calculated and filtered to find the points at which the variation is the highest. The set of found points P is denoted as,

$$P = \{(x_i, y_i) \quad | \quad |\theta_i - \theta_{i-1}| > \lambda\} \tag{2}$$

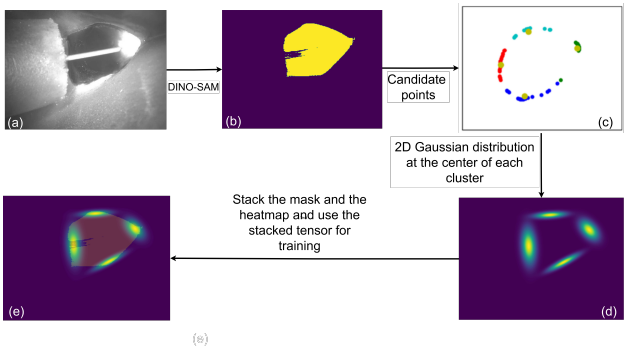

Fig. 3: The phases of finding the Gaussian heatmap for an image. ***(a)*** Initial image. ***(b)*** Output of DINO-SAM. ***(c)*** potential candidates clustered into 4 clusters. ***(d)*** The Gaussian heatmap at the center of each cluster. ***(e)*** Output of teacher : mask and heatmap.

λ is the threshold value for filtering out low frequency variations. K-means is later applied to cluster the points of P into k groups. And each of these groups P_j is a region of interest described by a 2D Gaussian distribution, of center μ and covariance matrix Σ

Following the clustering of the points and the calculation of the covariance matrices, the heatmap F_j of cluster P_j is defined at the center of this cluster, Thus the ground truth heatmap is finally computed as the mixture of all F_j,

$$F = \sum_{j=1}^{k} F_j \tag{3}$$

Algorithm 1 highlights the steps of the heatmap calculation discussed before.

The output of the teacher framework, Fig. 3e, represents the ground truth knowledge that is distilled into the ResNet-VAE stduent [39], [35], [40]. This type of networks encodes the input into a latent probabilistic space of continuous nature.

In our application, we encode the image of the weld pool using resnet [40], and we decode the probabilistic latent vector to reconstruct the heatmap and the mask using a fully connected convolutional layer, outputing a two channels matrix of shape (W,H,2). Here, W and H are the dimension of the input image, and each of the channels represent respectively the DINO-SAM mask and the calculated heatmap of Fig.3

The purpose of reconstructing the mask of the image instead of the image itself using the student network (the VAE), is that reconstructing a single element of the image instead of all three channels helps the network focus on a single zone and a single shape, which helps converging the network and accelerates it. t

IV. Data acquisition and dataset

An industrial application that consists of characterizing the shape of weld pool as taken as case in point. Characterizing weld pool is important to control the robotized welding

Algorithm 1: Heatmap calculation

Data: I and λ. Respectively, the input image and a threshold parameter
Result: F a matrix of the shape (W,H,2); first channel is the heatmap, second channel is the mask

$S = Block\_B(I)$;
$C = findContour(S)$;
$\nabla S_x = sobel_x(S)$;
$\nabla S_y = sobel_y(S)$;
$\Theta = \{\}$;
foreach $p_i \in C$ **do**
  $\theta_i = \nabla S_x \otimes \nabla S_y$;
  $\Theta = \Theta \cup \{\theta_i\}$;
end
$P = \{\}$;
foreach $\theta_i \in \Theta$ **do**
  if $|\theta_i - \theta_{i-1}| > \lambda$ **then**
    $P = P \cup \{(x_i, y_i)\}$
  end
end
$P_j = Kmeans(P)$;
$F = \{\}$;
foreach P_j **do**
  $\Sigma_j = covMatrix(P_j)$;
  $F+ = f_j(P_j; \mu_j, \Sigma_j)$
end
return $F \otimes S$

process and to analyze the quality of the weld. Since no existing benchmark or dataset exist for that purpose we create our own dataset by equipping a camera to an industrial robot, equipped with a welding source. The changes in geometry, robot speed, and process parameters accompanied with external perturbations impose deformations on the molten pool manipulated by the robot. Nine video sequences are taken of the weld pool in these conditions and by welding actual samples in a fillet joint configuration (Tee joint). Fig. 4 shows the different elements of the welding setup. Table I shows the different parameters used to acquire the dataset.

TABLE I: The conditions of dataset acquisitions. In addition to these conditions, positional offset is applied.

Wire feed rate (10 m/min)	Voltage (V)	Current (A)	Arc Length Correction(%)	Robot Velocity (cm/min)
10	31.9	328	**1.5/3.5/4.2**	30
10	31.9	309	4	**30/35/40**
9.5/9.8/10	Synergic laws	Synergy lines laws	4	30

This dataset is manually labeled with the possible keypoints that represent the shape of the weld pool, to validate the performance of the student network. A subset of the acquired sequences are labele for segmentation task, for the purpose of validating the performance of the DINO-SAM2 framework for object shape retrieval. Some examples of the shapes encountered in the dataset are shown in Fig. 5

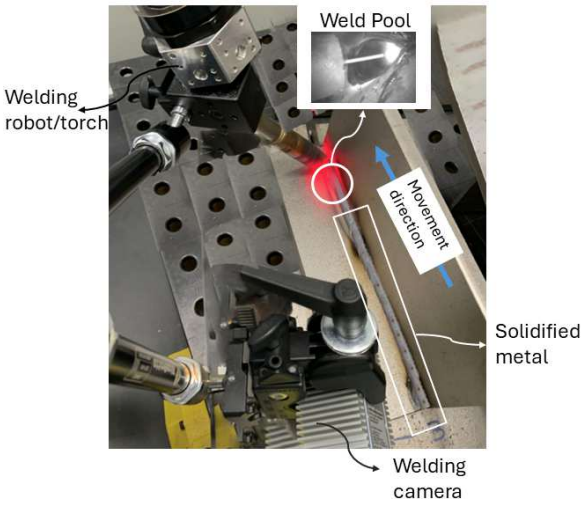

Fig. 4: The setup used for dataset generation (arc-off)

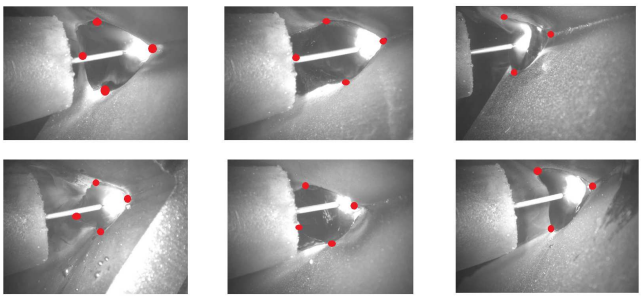

Fig. 5: Examples of the different shapes encountered in the dataset. In red are the features that should be used to control the robot

V. Results

The performance of the teacher network is validated according to its capacity of retrieving accurate masks of the designated object. We use the Intersection Over Union (IoU) for that objective. Several configuration are tested in an ablation study manner to decide the type and number prompts to optimize the performance of the SAM2 detections. We firstly experiment with the results of SAM2 without any prompts. We secondly introduce the prompts proposed by DINO networks: the center of the attention map and the different prompt values presented in Fig. 2

The performance of solely using SAM is first evaluated, however, the model in this case failed to achieve reliable values of mIoU with only 35.37%, with an upper quartile of 59%. Introducing DINO predicition with a single prompt ameliorated significantly the upper quartile without a significant impact on the value of mIoU that is 34.82%. On the other hand, after applying some heuristics such as clustering the thresholded mean attention map into different clusters, applying DBSCAN to propose a best fit bounding box and filtering the propmt points increased significantly the values of mIoU. After applying these heuristics, we obtained an mIoU of 75.26%, with an upper quartile of 85%. Fig. 6 depicts experimenting with different heuristics and different

configurations to retreive the mask of the object.

mIoU measures the similarity between the predicted segmentation mask and the ground truth mask by calculating the ratio of their intersection to their union.

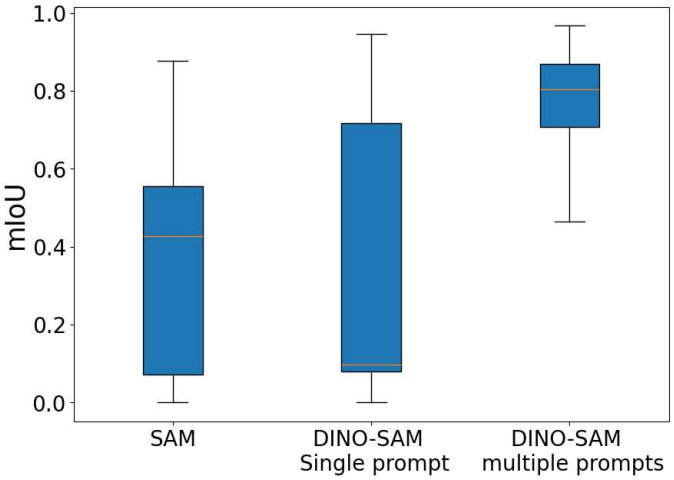

Fig. 6: The value of mIoU of DINO-SAM for diffrent configurations and with different heuristics.

The student model represented by the VAE architecture is lastly trained on a 32GB Tesla V100s GPU. We train the model for 100 epochs before evaluating its performance with respect to manually labeled data.

Validation metrics. Our validation metric consist of the euclidean distance between the ground truth labels and the centers of predicted heatmap of the student network. We denote the average value of euclidean distance over the dataset, and for all the heatmaps as mED. And the average euclidean distance for each heatmap as mED_k. Ideally, this distance should not surpass 15 pixels.

Baselines. We compare our method with the most relevant works found in the state of the art for similar type of problems such as using active contours [42], Kernelized correlation filters (KCF) [43], fully connected Convolutional Neural Networks [44], and also without the use of foundation models [45]. Table II shows the results of experiments and comparison with different baselines.

TABLE II: Mean Euclidean distance using different baselines.

	mED1	*mED2*	*mED3*	*mED4*	*mED*	Standard Deviation
Active Contour	141.70	63.84	42.74	94.68	85.74	23.67
KCF	37.20	83.51	53.05	38.5	53.08	30.62
SIFT	32.05	26.87	35.32	36.97	32.8	3.85
CPD + Active Contour	30.59	30.74	24.36	32.73	29.61	9.84
CPD	33.77	23.30	24.34	24.96	26.59	8.62
CNN	26.29	16.42	12.75	9.59	18.92	12.99
VAE	9.55	15.62	13.07	19.96	14.55	3.4
Ours	6.01	15.35	12.89	19.34	13.4	3.1

VI. Discussion

Qualitative evaluation. As opposed to baseline methods, the student-teacher framework of our approach provided stable estimation of the shape of the deformable object without the need of any prior labeling. However, the automatically generated masks by the DINO-SAM2 framework were in some cases subject to ambiguity coming from erroneous prompt points, resulting low values of IoU. This ambiguity is due to the presence of impurities and intrusions in some cases, such as the presence of greases on the robot trajectory.

Quantitative evaluation In addition to qualitative evaluation, the calculated value of the Euclidean distance between the predicted points and the ground truth is relatively small and acceptable for such an application. However, this error is the result of two main factors: human factor and the lack of visual features in some images. We recall that validation images were labeled manually with the possible centers of the heatmap, which makes these annotations biased and prone to error. Since defining accurate labels for these tasks is not possible. To resolve this kind of uncertainty, we recommend adopting the automatically generated GT labels as ground truth for evaluation and not the manually defined GT. As a proposed solution to this, we propose the use of sequential generative models students, such as Long Shot Term Memory Variational Autoencodre (LSTM-VAE) [46] capable of preserving temporal information about the process.

On the other hand, the lack visual features in some frames, due to several factors; such as noise and process instability have also resulted low fidelity predicitions of the heatmap.

VII. Conclusion

A clear representation of the shape of the manipulated object is a required step for robotic manipulation task. In the case of deformable objects, this step is dependent on the shape of the deformations of the object. In this paper we address the case of highly-deformable fluid and viscoelastic objects. We propose a teacher-student framework that does not require any prior labeling to represent them. Our method is independent from the complex physical modeling requirements. During inference time, we use only the student of our framework. The student model, a generative model in our case, generates a heatmap that estimates the current state of the object. This heatmap represents the locations at which the most important deformation occurs, according to the contour of the object. The student network distils the knowledge of a teacher that is composed of two foundation models and the heatmap proposal algorithm 1. The two foundation models employed in the case of this paper are DINO and SAM2. We use DINO as a prompts generator for the SAM2 model.

We demonstrate first the capabilities of foundation models to retrieve pixel-level information about the deformable objects. Secondly, we show that the generative student network is capable of learning the shape of the fluid deformable object by absorbing the knowledge provided by the foundation models successfully.

Further research directions include ameliorating the performance of the teacher by including the human in the loop through techniques such as active learning on one hand. On the other hand, integrating the retrieved shape of the weld

pool in a control tasks such as visual servoing is also a perspective to this work.

REFERENCES

[1] Xu, Z., He, Z., Wu, J., & Song, S. (2020). Learning 3D Dynamic Scene Representations for Robot Manipulation. Proceedings of Machine Learning Research, 155, 126–142.

[2] van der Merwe, M., Lu, Q., Sundaralingam, B., Matak, M., & Hermans, T. (2020). Learning Continuous 3D Reconstructions for Geometrically Aware Grasping. Proceedings - IEEE International Conference on Robotics and Automation, 11516–11522.

[3] Garg, R., Vijay Kumar, B. G., Carneiro, G., & Reid, I. (2016). Unsupervised CNN for single view depth estimation: Geometry to the rescue. 9912 LNCS, 740–756.

[4] Naseer, M., Khan, S., & Porikli, F. (2019). Indoor Scene Understanding in 2.5/3D for Autonomous Agents: A Survey. IEEE Access, 7, 1859–1887.

[5] Pan, Z., & Manocha, D. (2016). Motion planning for fluid manipulation using simplified dynamics. IEEE International Conference on Intelligent Robots and Systems, 2016-November, 4224–4231.

[6] Hou, Y. C., Sahari, K. S. M., & How, D. N. T. (2019). A review on modeling of flexible deformable object for dexterous robotic manipulation. International Journal of Advanced Robotic Systems.

[7] Herguedas, R., López-Nicolás, G., Aragüés, R., & Sagüés, C. (2019). Survey on multi-robot manipulation of deformable objects. IEEE International Conference on Emerging Technologies and Factory Automation, ETFA, 2019-September, 977–984.

[8] Nadon, F., Valencia, A. J., & Payeur, P. (2018). Multi-Modal Sensing and Robotic Manipulation of Non-Rigid Objects: A Survey. Robotics 2018, Vol. 7, Page 74, 7(4), 74.

[9] Mateo, C. M., Corrales, J. A., & Mezouar, Y. (2020). A manipulation control strategy for granular materials based on a Gaussian mixture model. In Robot 2019: Fourth Iberian Robotics Conference: Advances in Robotics, Volume 2 (pp. 171-183).

[10] Yan, M., Zhu, Y., Jin, N., & Bohg, J. (2020). Self-Supervised Learning of State Estimation for Manipulating Deformable Linear Objects. IEEE Robotics and Automation Letters, 5(2), 2372–2379.

[11] Borum, A., Matthews, D., & Bretl, T. (2014). State estimation and tracking of deforming planar elastic rods. 2014 IEEE International Conference on Robotics and Automation (ICRA), 4127–4132.

[12] Tzelepis, G., Aksoy, E. E., Borràs, J., & Alenyà, G. (2022). Semantic State Estimation in Cloth Manipulation Tasks.

[13] Malvido Fresnillo, P., Vasudevan, S., Mohammed, W. M., Martinez Lastra, J. L., & Perez Garcia, J. A. (2023). An approach based on machine vision for the identification and shape estimation of deformable linear objects. Mechatronics, 96, 103085.

[14] Hu, Z., Hua, L., Qin, X., Ni, M., Ji, F., & Wu, M. (2021). Molten pool behaviors and forming appearance of robotic GMAW on complex surface with various welding positions. Journal of Manufacturing Processes, 64, 1359–1376.

[15] Stadler, M., Freton, P., & Gonzalez, J.-J. (2017). Influence of Welding Parameters on the Weld Pool Dimensions and Shape in a TIG Configuration. Applied Sciences, 7(4), 373.

[16] Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., & Joulin, A. (n.d.). Emerging Properties in Self-Supervised Vision Transformers.

[17] Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Vasudev Alwala, K., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., Feichtenhofer, C., & Fair, M. (n.d.). SAM 2: Segment Anything in Images and Videos.

[18] Gu, F., Zhou, Y., Wang, Z., Jiang, S., & He, B. (2023). A Survey on Robotic Manipulation of Deformable Objects: Recent Advances, Open Challenges and New Frontiers.

[19] Gascuel, M.-P. (n.d.). An Implicit Formulation for Precise Contact Modeling between Flexible Solids.

[20] Palli, G. (2020). Model-based Manipulation of Deformable Linear Objects by Multivariate Dynamic Splines. 2020 IEEE Conference on Industrial Cyberphysical Systems (ICPS), 520–525.

[21] Arriola-Rios, V. E., Guler, P., Ficuciello, F., Kragic, D., Siciliano, B., & Wyatt, J. L. (2020). Modeling of Deformable Objects for Robotic Manipulation: A Tutorial and Review. Frontiers in Robotics and AI.

[22] Rastegarpanah, A., Howard, R., & Stolkin, R. (2022). Tracking linear deformable objects using slicing method. Robotica, 40(4), 1188–1206.

[23] Miller, S., van den Berg, J., Fritz, M., Darrell, T., Goldberg, K., & Abbeel, P. (2012). A geometric approach to robotic laundry folding. The International Journal of Robotics Research, 31(2), 249–267.

[24] Huang, S. H., Pan, J., Mulcaire, G., & Abbeel, P. (2015). Leveraging appearance priors in non-rigid registration, with application to manipulation of deformable objects. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 878–885.

[25] Tang, T., Fan, Y., Lin, H.-C., & Tomizuka, M. (2017). State estimation for deformable objects by point registration and dynamic simulation. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2427–2433.

[26] Shi, H., Xu, H., Huang, Z., Li, Y., & Wu, J. (2022). RoboCraft: Learning to See, Simulate, and Shape Elasto-Plastic Objects with Graph Networks. International Journal of Robotics Research.

[27] Lu, J., Jayakumari, A., Richter, F., Li, Y., & Yip, M. C. (2021). SuPer Deep: A Surgical Perception Framework for Robotic Tissue Manipulation using Deep Learning for Feature Extraction. Proceedings - IEEE International Conference on Robotics and Automation, 2021-May, 4783–4789.

[28] Liu, J., Chen, Y., Dong, Z., Wang, S., Calinon, S., Li, M., & Chen, F. (2022). Robot Cooking with Stir-Fry: Bimanual Non-Prehensile Manipulation of Semi-Fluid Objects. IEEE Robotics and Automation Letters, 7(2), 5159–5166.

[29] Yin, H., Varava, A., & Kragic, D. (2021). Modeling, learning, perception, and control methods for deformable object manipulation. Science Robotics, 6(54), 8803.

[30] GM, H., Gourisaria, M. K., Pandey, M., & Rautaray, S. S. (2020). A comprehensive survey and analysis of generative models in machine learning. Computer Science Review, 38, 100285.

[31] Awais, M., Naseer, M., Khan, S., Anwer, R. M., Cholakkal, H., Shah, M., Yang, M.-H., & Khan, S. (n.d.). Foundational Models Defining a New Era in Vision: A Survey and Outlook.

[32] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139–144.

[33] Makhzani, A., Shlens, J., Jaitly, N., Brain, G., Openai, I. G., & Frey, B. (2015). Adversarial Autoencoders. Elements of Dimensionality Reduction and Manifold Learning, 577–596.

[34] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[35] Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings.

[36] Oussidi, A., & Elhassouny, A. (2018). Deep generative models: Survey. 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), 1–8.

[37] El-Kaddoury, M., Mahmoudi, A., & Himmi, M. M. (2019). Deep Generative Models for Image Generation: A Practical Comparison Between Variational Autoencoders and Generative Adversarial Networks (pp. 1–8).

[38] Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking Atrous Convolution for Semantic Image Segmentation.

[39] Kargar, E., & Kyrki, V. (2021). Vision Transformer for Learning Driving Policies in Complex Multi-Agent Environments.

[40] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition (pp. 770–778).

[41] Wada, K. Labelme: Image Polygonal Annotation with Python.

[42] Liu, J., Fan, Z., Olsen, S. I., Christensen, K. H., & Kristensen, J. K. (2017). Boosting Active Contours for Weld Pool Visual Tracking in Automatic Arc Welding. IEEE Transactions on Automation Science and Engineering, 14(2), 1096–1108.

[43] Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2015). High-Speed Tracking with Kernelized Correlation Filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(3), 583–596.

[44] Huo, S., Duan, A., Li, C., Zhou, P., Ma, W., Wang, H., & Navarro-Alarcon, D. (2022). Keypoint-Based Planar Bimanual Shaping of Deformable Linear Objects under Environmental Constraints with Hierarchical Action Framework. IEEE Robotics and Automation Letters, 7(2), 5222–5229.

[45] Shape estimation of Viscoelastic fluid materials: an application for weld pool. El Assal, O., Mateo. C.M., CIRON, S., Fofi, D.

[46] Mahasseni, B., Lam, M., & Todorovic, S. (n.d.). Unsupervised Video Summarization with Adversarial LSTM Networks.