Distilling Semantic Features for 3D Cloth Representations from Vision Foundation Models

Alberta Longhini[1], Marcel Büsching[1], Bardienus P. Duisterhof[2],
Jeffrey Ichnowski[2], Mårten Björkman[1], and Danica Kragic[1]

***Abstract*— This study explores the use of vision foundation models to enhance 3D representations of cloth-like deformable objects. By focusing on the distillation of semantic information from RGB images, we examine the potential of pre-trained Visual-Language Models in capturing complex folded configurations of cloth. Our investigation reveals the challenges and preliminary successes in leveraging semantic information to improve the understanding and tracking of deformable object states.**

I. Introduction and Related Work

Manipulation of deformable objects, such as clothes, has been a long-standing challenge in robotic manipulation [1]. The deformable nature of these objects poses significant challenges for state estimation and tracking under self-occlusion. These often lead to incomplete information about the cloth state and hinders the robot's ability to plan optimal manipulation strategies [2], [3].

A common way to describe cloth states is using 3D representations such as graphs or point clouds, due to their potential to improve generalization and sim-to-real transfer over traditional image representations [4], [5]. However, point clouds are often ambiguous due to self-occlusions, making it difficult to distinguish between different layers of the cloth when being folded [6]. Graph-based representations, while theoretically capable of capturing the underlying structure of the cloth under occlusions, prove difficult to be effectively used in real-world scenarios due to the challenges of tracking the deformation [7], [8]. Recent applications of vision foundation models for manipulation tasks have demonstrated their potential in augmenting 3D representations of rigid objects with semantic information, facilitating improved understanding and interaction with these objects [9], [10]. However, the application of these models to deformable objects, particularly for extracting and distilling semantic information, remains underexplored.

In this work, we evaluate the capability of vision foundation models to distill semantic information of cloth-like deformable objects. We explore two commonly used models in robotic manipulation, Grounded SAM [11] and DINOv2 [12], [13]. We evaluate these models through image-level and pixel-level visual tasks, focusing on semantic segmentation and dense feature extraction. Since object-level semantics may not fully capture the cloth's deformed state (e.g. Fig 1), our study delves into two abstraction levels: layer descriptors (upper and lower layers) and dense pixel-level descriptors. Specifically, we assess the performance of Grounded SAM in visually segmenting the cloth's upper and bottom layers. We also investigate the temporal consistency of the dense semantic descriptors of DINOv2, by using it to track keypoints while the cloth undergoes deformations.

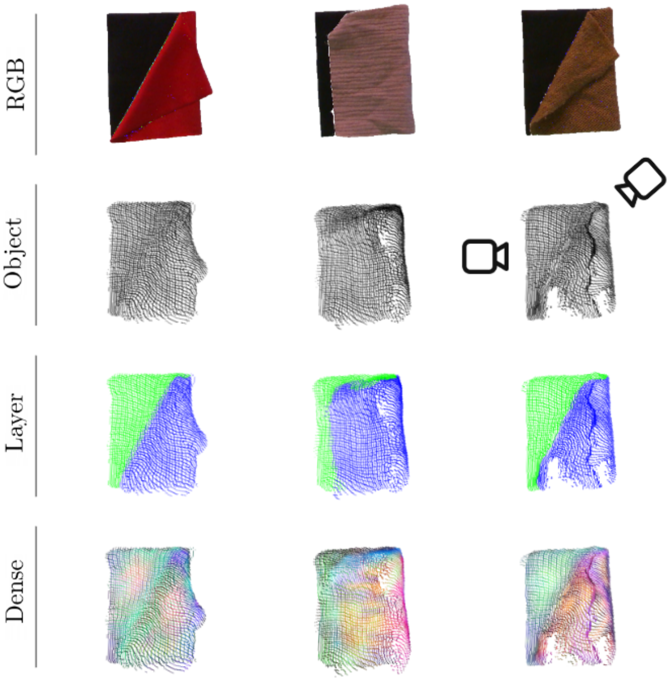

Fig. 1: The top row displays RGB images of cloths in various folded states, while the following rows depict their 3D representations at different abstraction levels (Object-Layer-Dense). Notably, the object-level representations fall short in accurately depicting the folded states due to their inability to precisely show the alignment between the top and bottom layers..

This work aims to evaluate how existing vision foundation models perform in distilling features for 3D representations of cloth-like deformable objects. By doing so, we seek to extend the utility of these models beyond rigid objects and explore their potential in more complex, real-world applications involving deformable objects such as cloths. This work also highlights rich research opportunities in modifying and extending foundation models for deformables.

II. Instance Segmentation

To evaluate Grounded SAM on an instance segmentation task, we consider multiple rectangular cloths, and query the

[1] The authors are with the Division of Robotics, Perception and Learning, EECS, at KTH Royal Institute of Technology, Stockholm, Sweden {albertal, busching, celle, dani}@kth.se
[2] The authors are with are with Carnegie Mellon University, Pittsburgh, USA {bduister, jeffi}@andrew.cmu.edu

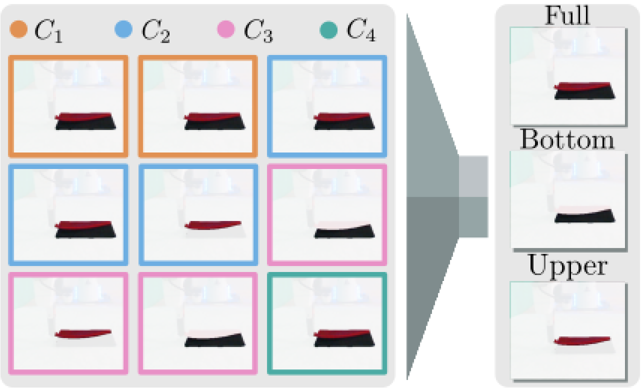

Fig. 2: Example of the mask selection process. For each prompt $C = \{C_1, C_2, C_3, C_4\}$, Grounded SAM provides one or multiple masks with different confidence values. The colors of each box match the prompt that generated the mask, where we used as prompts $C = \{$ "rectangular cloth", "upper colored half cloth", "black bottom half cloth", "grasped central cloth" $\}$. Our ensemble approach aggregates all these masks and filters them to obtain the corresponding masks of the full, bottom, and upper layers of the cloth.

model to segment the *upper* and *bottom* halves of the cloth at every time step of a folding interaction. We color one of the two halves of each cloth in black to better distinguish them during the manipulation. The goal consists of processing RGB-D observations into a semantically labeled point cloud $P = P^U \cup P^B$ that represents the *upper* and *bottom* halves of the cloth, respectively. A planner could leverage these labels to achieve the desired fold, otherwise impossible with object-level representations such as unlabeled point clouds.

A. Point Cloud Labeling

Given an RGB-D image I_t, we first find the segmentation mask of the cloth from the RGB observation using Grounded SAM [14], [15]. Then, we transform the masked depth into a point cloud corresponding to the observable points of the cloth P_t using the camera's intrinsic matrix.

We introduce a novel mask selection strategy that considers an ensemble of prompts $C = \{C_1, \ldots, C_n\}$, with n the number of prompts. We set $C = \{$ "rectangular cloth", "upper colored half cloth", "black bottom half cloth", "grasped central cloth" $\}$, and aggregates the obtained masks based on the following heuristics:

Full: We obtain the mask of the full cloth M_f by aggregating all the masks from the set of prompts C, where the aggregation corresponds to a logical OR operation.

Bottom: We assume that the bottom layer has a predefined color. Leveraging this assumption, we select the masks corresponding to the bottom layer by initially imposing a color threshold r on each resulting masked RGB image. Subsequently, we aggregate only the masks with many pixels satisfying this threshold, thereby obtaining M_b. Differently from standard color threshold techniques, we utilize this threshold as a *soft* voting mechanism without accurately choosing r and tuning it for every different cloth.

Upper: Given the two previous steps, the upper mask is easily obtained as the difference between the full and the bottom mask: $M_u = M_f - M_b$. In Fig. 2, we present an example of the masks selected by our procedure.

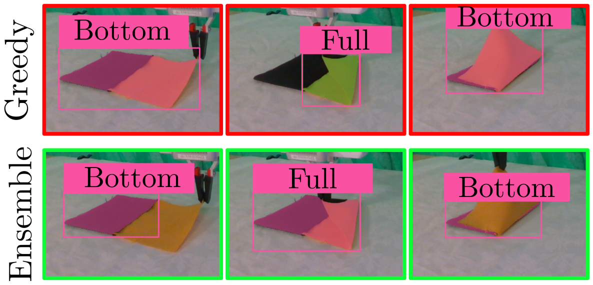

Fig. 3: Visualization of cases where the Greedy approach wrongly segments (red frames) the desired mask. The box label specifies the desired mask. The Ensemble approach, on the other hand, is able to obtain the correct segmentations (green frames).

TABLE I: Evaluation of the MAE of the IoU estimated with different point cloud representations. Arrows indicate the direction of improvement of the metric.

Cam.	`No Semantic` ↓	`Greedy` ↓	`Ensemble` ↓
Front	0.40 ± 0.17	0.21 ± 0.20	0.10 ± 0.11
Back	0.29 ± 0.15	0.20 ± 0.21	0.07 ± 0.11
Both	0.32 ± 0.14	0.22 ± 0.21	$\mathbf{0.04 \pm 0.08}$

We find this approach to be more robust compared to relying on SAM's confidence values to select masks. Common failure instances involved improper segmentation of either the upper or lower half of the cloth, resulting in a complete cloth mask rather than the intended half, or conversely, segmenting only one half when the entire cloth was desired.

We repeat this process for images $\{I_t^1, I_t^2\}$ recorded from two calibrated cameras to reduce the effect of self-occlusions. The resulting point clouds are then transformed into a common reference frame, for example, the robot base frame. This step aggregates the point clouds to derive the final cloth state $P_t = P_t^U \cup P_t^B$, which can be further voxelized to make the density of the points uniform across the cloth surface.

B. Evaluation

We use the intersection over union (IoU) of the two halves as a quantitative description of the cloth state, where the higher the IoU, the better folded the cloth is. We used a set of the real-world trajectories recorded while folding the cloths in half, and we manually annotated the mask belonging to the upper or bottom half of the cloth to extract P^U and P^B and compute the ground truth IoU.

We compare the representations with semantic descriptors obtained through our mask selection approach (`Ensemble`) against the one derived from a greedy selection of the highest confidence mask from the Grounded SAM (`Greedy`). We further consider a point cloud lacking semantic descriptors as object-level baseline representation (`No Semantic`). Additionally, we integrate the comparison between one and two camera points of view. As an evaluation metric, we use the mean absolute error (MAE) between the ground truth IoU and the one estimated from the different cloth representations.

TABLE II: Final MSE (10^{-3}) between the tracked keypoints and the ground truth keypoints. We repeated the tracking 5 times for each combination of method and configuration of cameras (Cam.). Arrows indicate the direction of improvement of the metric.

Cam.	**DINO** ↓	**D3F** ↓	**C** ↓	**C+DINO** ↓
Front	10.0 ± 0.3	11.7 ± 0.5	16.5 ± 0.3	15.6 ± 0.1
Back	6.6 ± 0.3	6.2 ± 1.0	2.9 ± 0.2	2.9 ± 0.3
Both	7.1 ± 0.6	5.3 ± 0.4	$\mathbf{1.4 \pm 0.3}$	1.5 ± 0.1

As shown in Table I, the `No Semantic` representation obtained the worst results, confirming that object-centric representations do not allow the state of the cloth to be accurately quantified. While the `Greedy` approach performed better than the `No Semantic`, considering only the highest confident mask was still prone to error. On the other hand, the `Ensemble` approach consistently achieved the lowest MAE, suggesting that the proposed filtering technique improved the quality of the representation due to a more accurate segmentation, as visualized in Fig. 3. Optimal results were obtained using observations from both cameras, underscoring the impact of occlusions on the cloth state estimation.

These results confirm that vision foundation models can provide a relevant prior to segmenting different instances of cloths and improve 3D state representations. Still, they might provide overly confident results for this specific class of objects despite our simplification of introducing halves with different colors.

III. Keypoint Tracking

In this section, we explore pixel-level abstraction represented by dense semantic descriptors. We consider a keypoint tracking task to assess the temporal robustness of these descriptors while the cloth undergoes deformations during the manipulation.

A. *Tracking with Dense Descriptors*

Similarly to the previous section, given an RGB-D image I_t we obtain the point cloud P_t at time t from the masked depth observation. The semantic descriptors $f^{p^i} \in \mathbb{R}^{1024}$ of each point $p^i \in P$ are extracted from RGB observations using DINOv2 [13]. We merge descriptors from different camera views as proposed in [16], and select a set of keypoints $K \subset P$ from the observed point cloud using farthest point sampling. Let $f^{p^j}_t$ represent the semantic descriptor of keypoint $k^j_t \in K_t$ at time t. We formulate the problem of tracking the next position of the keypoint k^j_{t+1} as an optimization problem that finds the best matching descriptor $f^{p^j}_{t+1}$ across consecutive times:

$$k^j_{t+1} = \underset{p^i \in P_{t+1}}{\arg\min} \, \|f^{p^i}_{t+1} - f^{p^j}_t\|, \tag{1}$$

where $f^{p^i}_{t+1}$ is the semantic descriptors of the point $p^i_{t+1} \in P_{t+1}$. As suggested in [16], solely relying on semantic features for tracking leads to unstable training. To mitigate the issues, the authors introduce a loss term to minimize the distance of keypoints from the observed surface and a rigid constraint to maintain structural consistency. Given the inadequacy of rigid constraints for deformable objects, we modify the framework to optimize the displacement $\Delta k^i_t \in \mathbb{R}^3$ of the keypoints $k^i_t \in K_t$ at time t, such that $k^i_{t+1} = k^i_t + \Delta k^i_t$. Specifically, we minimize the Chamfer Distance [17] between the observed point cloud P_{t+1} and the set of displaced keypoints K_{t+1}.

B. *Evaluation*

We evaluate DINOv2 features on a synthetic dataset of cloths collected using SoftGym [18], as ground truth keypoints of deformable objects are challenging to extract from real-world data. Specifically, the dataset consists of RGB-D observations from two camera views of a rectangular cloth folded by randomly choosing pick and place positions.

To evaluate the effectiveness of DINOv2 descriptors, we compare four different loss combinations used to optimize the keypoint tracking. We denote `DINO` as the variant that uses only the loss over the semantic descriptors, `D3F` the loss proposed in [16], where we removed the rigid constraint and optimized the spatial displacement of each keypoint. We then consider the tracking using only the Chamfer loss as `C`, while the combination of the Chamfer loss with the loss over the semantic descriptors as `C+DINO`.

The quantitative results of the tracking shown in Table II confirm that solely relying on semantic descriptors leads to poor results. The `D3F` method showed lower performance than `C` as the distance loss used for the optimization does not directly optimize for the closest surface point, instead minimizing the distance to any visible surface point. The authors demonstrated their approach to be highly performant, as the rigid constraint helped retain geometric information, which is not the case for deformable such as cloth. Interestingly, integrating semantic features into the optimization loss, `C+DINO`, did not help improve the tracking performance. We attribute the loss in performance of `C+DINO` with respect to `C` to the inconsistency of DINOv2 features across different deformed states. In particular, we noted that the semantic features were not temporally consistent across various time intervals while the cloth was deformed.

These results on deformable objects suggest that dense descriptors distilled from foundation models such as DINOv2 do not yet provide as strong results as shown for rigid objects for keypoint tracking. This points at a promising direction for future research.

IV. Conclusion and Future Work

In this work, we investigated the potential and limitations of current visual foundation models for distilling semantic features of 3D cloth representations during manipulation. Our investigation revealed that although models like GroundedSAM and DINOv2 are effective with rigid objects, they struggle with deformable objects like cloth. The introduction of structural priors, such as graph-based methods, emerges as a potential pathway to enhance the overall model performance. Furthermore, extending the capabilities of these models to out-of-distribution objects, like textureless cloth, will enable the next generation of autonomous agents.

References

[1] H. Yin, A. Varava, and D. Kragic, "Modeling, learning, perception, and control methods for deformable object manipulation," *Science Robotics*, vol. 6, no. 54, p. eabd8803, 2021.

[2] J. Zhu, A. Cherubini, C. Dune, D. Navarro-Alarcon, F. Alambeigi, D. Berenson, F. Ficuciello, K. Harada, J. Kober, X. Li, *et al.*, "Challenges and outlook in robotic manipulation of deformable objects," *Robotics & Automation Magazine*, vol. 29, no. 3, pp. 67–77, 2022.

[3] B. P. Duisterhof, Z. Mandi, Y. Yao, J.-W. Liu, M. Z. Shou, S. Song, and J. Ichnowski, "Md-splatting: Learning metric deformation from 4d gaussians in highly deformable scenes," 2023.

[4] Z. Huang, X. Lin, and D. Held, "Mesh-based dynamics with occlusion reasoning for cloth manipulation," *arXiv preprint arXiv:2206.02881*, 2022.

[5] Y. Wang, Z. Sun, Z. Erickson, and D. Held, "One policy to dress them all: Learning to dress people with diverse poses and garments," *arXiv preprint arXiv:2306.12372*, 2023.

[6] A. Longhini, M. C. Welle, Z. Erickson, and D. Kragic, "Adafold: Adapting folding trajectories of cloths via feedback-loop manipulation," *arXiv preprint arXiv:2403.06210*, 2024.

[7] A. Longhini, M. Moletta, A. Reichlin, M. C. Welle, D. Held, Z. Erickson, and D. Kragic, "Edo-net: Learning elastic properties of deformable objects from graph dynamics," *arXiv preprint arXiv:2209.08996*, 2022.

[8] Z. Huang, X. Lin, and D. Held, "Self-supervised cloth reconstruction via action-conditioned cloth tracking," *arXiv preprint arXiv:2302.09502*, 2023.

[9] Y. Zhu, Z. Jiang, P. Stone, and Y. Zhu, "Learning generalizable manipulation policies with object-centric 3d representations," in *7th Annual Conference on Robot Learning*, 2023.

[10] N. Di Palo and E. Johns, "Dinobot: Robot manipulation via retrieval and alignment with vision foundation models," *arXiv preprint arXiv:2402.13181*, 2024.

[11] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded SAM: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[12] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *ICCV*, 2021.

[13] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[14] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[16] Y. Wang, Z. Li, M. Zhang, K. Driggs-Campbell, J. Wu, L. Fei-Fei, and Y. Li, "D3 fields: Dynamic 3d descriptor fields for zero-shot generalizable robotic manipulation," *arXiv preprint arXiv:2309.16118*, 2023.

[17] H. Barrow, J. Tenenbaum, R. Bolles, and H. Wolf, "Parametric correspondence and chamfer matching: Two new techniques for image matching, tech," Note, Tech. Rep., 1977.

[18] X. Lin, Y. Wang, J. Olkin, and D. Held, "Softgym: Benchmarking deep reinforcement learning for deformable object manipulation," in *CoRL*. PMLR, 2021, pp. 432–448.