2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

DODUO: Learning Dense Visual Correspondence from Unsupervised Semantic-Aware Flow

Zhenyu Jiang[1], Hanwen Jiang[1], Yuke Zhu[1]

***Abstract*— Dense visual correspondence plays a vital role in robotic perception. This work focuses on establishing the dense correspondence between a pair of images that captures dynamic scenes undergoing substantial transformations. We introduce DODUO to learn general dense visual correspondence from in-the-wild images and videos without ground truth supervision. Given a pair of images, it estimates the dense flow field encoding the displacement of each pixel in one image to its corresponding pixel in the other image. DODUO uses flow-based warping to acquire supervisory signals for the training. Incorporating semantic priors with self-supervised flow training, DODUO produces accurate dense correspondence robust to the dynamic changes of the scenes. Trained on an in-the-wild video dataset, DODUO illustrates superior performance on point-level correspondence estimation over existing self-supervised correspondence learning baselines. We also apply DODUO to articulation estimation and zero-shot goal-conditioned manipulation, underlining its practical applications in robotics. Code and additional visualizations are available at https://ut-austin-rpl.github.io/Doduo/**

I. INTRODUCTION

Dense visual correspondence involves identifying the corresponding pixel in a target image for any given pixel in a source image. Correspondence is a cornerstone of robot perception, driving a multitude of robotic applications, such as dense visual tracking [1]–[5], articulation estimation [6]–[8], and deformable object modeling [9]–[11]. These applications typically require fine-grained correspondence reasoning of *dynamic scenes undergoing substantial transformations*.

In recent years, supervised deep learning methods [12]–[19] have made great strides in learning correspondences from annotated datasets. These methods perform well in constrained scenarios, such as with rigid objects/scenes and minor viewpoint changes. However, their efficacy under large, non-rigid motions is curtailed by the prohibitive costs (thus limited availability) of ground-truth annotations. Alternatively, research exemplified by DenseObjectNets [20]–[24] employs robot interaction to capture multi-view observations of objects and generate correspondence ground truth for training. However, such procedures only yield category-level visual descriptors that fail to generalize beyond the category of the captured objects. Self-supervised learning methods like DINO [25] enable large model training on in-the-wild images and videos, offering a more scalable and generalizable approach to semantic correspondence learning without manual annotations. Nevertheless, how to learn a generalizable model to robustly establish fine-grained correspondence required by robotic tasks remains an open question.

[1] Department of Computer Science, the University of Texas at Austin. Correspondance to zhenyu@utexas.edu

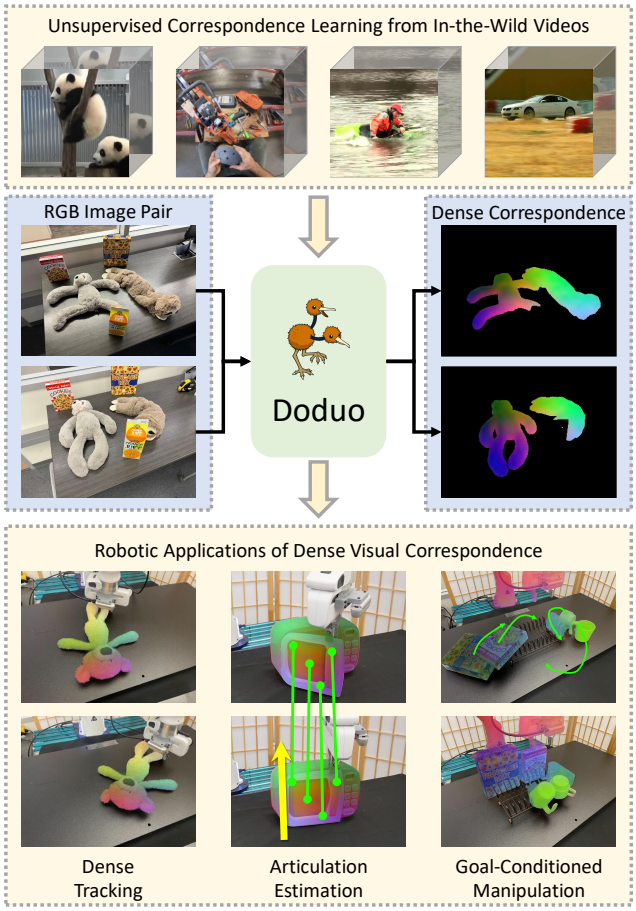

Fig. 1: DODUO is trained on in-the-wild videos without manual annotations. It can predict accurate dense correspondence from pairs of images to enable a diverse set of robotic applications with robustness to lighting and occlusion. The visualized dense correspondences are actual predictions from DODUO.

Two notable categories of self-supervised techniques have been studied in the visual correspondence learning literature. The first category is **self-supervised optical flow** [26]–[31]. The key idea revolves around predicting dense optical flow between adjacent video frames and utilizing a photometric loss to minimize the color discrepancy between the matched pixels. The pixel-level supervision from the photometric loss allows for learning fine-grained correspondences. However, the color-matching property of the photometric loss limits the model's resilience against large appearance variations and changes in illumination. The second category is **self-supervised semantic feature learning** [25], [32], which generates semantically consistent feature representations and exhibits robustness to large lighting and viewpoint variations.

Despite this, in cases where pixels have similar semantics, such as points on a cabinet door, the estimated correspondence is coarse due to indistinguishable semantics. These two categories of methods possess complementary strengths and weaknesses. The former offers satisfying fine-grained correspondence but struggles to generalize under appearance changes, while the latter provides robust coarse matching. We argue that a more general correspondence learning method should integrate the advantages of both categories.

To this end, we introduce DODUO (**D**ense Visual C**o**rrespon**d**ence from **U**nsupervised Semantic-Aware Fl**o**w), designed to efficiently learn fine-grained dense correspondences between a pair of images with substantial variations. Using a transformer-based neural network [33], [34], DODUO extracts correspondence-aware feature maps from the image pairs and estimates the flow from source to target based on feature similarities. Learning to find correspondence from scratch with all the pixels in the target image as matching candidates is difficult, especially when direct supervision of ground truth correspondence is not available. We mitigate this challenge by incorporating robust semantic features from DINO [25]. When finding the matching of a query point in the source image, we use DINO feature correspondence to select a subset of points on the target image with higher semantic similarities to the query point. Incorporating the coarse correspondence from DINO allows our model to identify fine-grained correspondences inside this subset of matching candidates.

One noteworthy challenge in self-supervised flow-based methods is the occlusion across frames. We tackle this issue by using predicted flow to select region masks from an off-the-shelf image segmentation model, thereby identifying common regions visible in both images. As our correspondence model improves, the visible region localization becomes more precise and, in turn, facilitates model training.

We train DODUO on frames from an in-the-wild Youtube-VOS video dataset [35] and evaluate it on the TAP-Vid [36], a benchmark with fine-grained point correspondence annotations. Our model demonstrates superior performance across all metrics compared with baselines. Furthermore, we apply DODUO to the articulation estimation task and show its ability to estimate articulation without the need for training with any ground truth articulated data. We further conduct real-robot experiments of zero-shot goal-conditioned manipulation, highlighting how accurate dense visual correspondence enables precise actions in fine-grained object manipulation. These results affirm the broad applicability of DODUO to downstream robotic tasks without the need for training with domain-specific data.

II. Related Work

Unsupervised Visual Correspondence Learning. Recent years have seen considerable advancements in unsupervised visual correspondence learning, which capitalizes on a plethora of unannotated data for training [37]–[44]. Despite the progress, the practical applicability of these models remains limited due to various inherent model design constraints. Most unsupervised visual correspondence learning frameworks belong to three main types.

The first type of method learns from synthesized self-supervision [45], [46]. These methods generate synthetic warping between image pairs and harness the reverse warping as the supervisory signal. However, its inability to simulate the dynamic aspects of real-world scenes, such as moving objects, results in suboptimal performance.

The second is unsupervised optical flow [29], [30], [38], [41], [47]–[50]. These methods can deal with dynamic scenes, as they learn from natural motions in the videos. However, they train models with the photometric loss, assuming the matched pixels possess the same color. This assumption requires the training image pairs to have small illuminance changes, typically neighboring frames in a video. Consequently, they struggle to handle image pairs with significant appearance or content changes.

The third is the semantic representation learning [25], [32], [40], [41], [51]–[56]. To learn semantic features, these methods leverage self-supervised learning objectives, including contrastive learning, self-distillation, and cycle consistency. These features are robust to appearance variations, including illumination, viewpoint, and spatial transformations. Nevertheless, features of points that belong to the same semantic regions tend to be indistinguishable, resulting in coarse correspondences in these regions. Such coarse matching is insufficient for downstream robotics applications.

Visual Descriptor Learning from Interaction. There is a branch of research using robot interaction to learn dense visual descriptors [20]–[24], [57]. DenseObjectNets [21] is one of the pioneer works in this direction. DenseObjectNets uses a camera mounted on a robot arm to collect multi-view observations and extract ground truth correspondence with 3D reconstruction. The learned visual descriptors are robust to variations in object pose, lighting, and deformation. However, the learned descriptor cannot generalize to objects from novel categories, limiting its applicability to general robot manipulation tasks with diverse object categories.

In contrast to existing works, DODUO learns a general correspondence model that predicts dense correspondence between image pairs capturing scenes undergoing substantial variations. Since it is trained on large-scale in-the-wild frame pairs containing diverse visual contents, DODUO generalizes to different objects and scenes without finetuning.

III. Method

We now present DODUO, a self-supervised learning approach for dense visual correspondence based on semantic-aware flow. Fig. 2 illustrates the DODUO model. We introduce a semantic-aware flow estimation model (Sec. III-A), leveraging robust coarse matching from pre-trained semantic features for correspondence prediction. Moreover, to avoid the ill-posed problem of estimating the flow of occluded pixels, we introduce a bootstrapping strategy to locate visible regions in both images and apply supervision only on the visible regions (Sec. III-B). In addition, we introduce a novel pixel-level feature-metric loss as the self-supervision

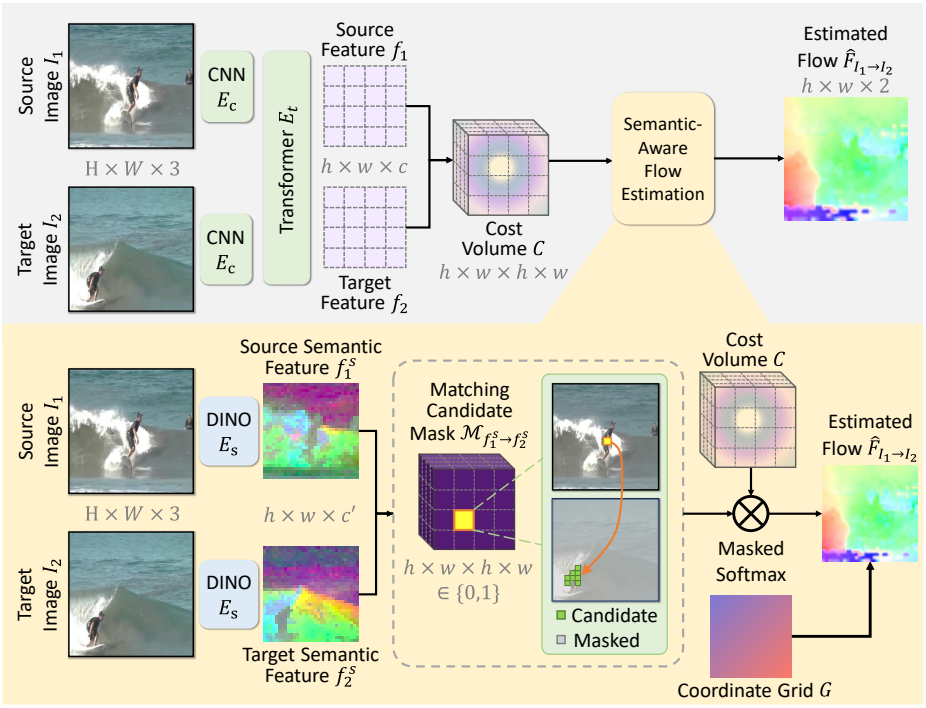

Fig. 2: **DODUO Model Architecture.** Our model takes two images as inputs and uses a Transformer-based network to extract features of both images. Then, we construct a 4D cost volume of all pairs of feature pixels and predict a dense semantic-aware flow field given the cost volume. For semantic-aware flow estimation, we use a DINO encoder [25] to extract semantic feature maps of both input frames. According to the similarity of the semantic feature map, we compute a matching candidate mask for each of the feature points of the source frame and integrate this information during flow estimation using masked Softmax.

for training DODUO (Sec. III-C). The pixel-level losses encourage the model to predict fine-grained correspondences.

A. Flow Estimation

Problem Formulation. We study the problem of pixel-level dense correspondence between a source image $I_1 \in \mathbb{R}^{H \times W \times 3}$ and a target image $I_2 \in \mathbb{R}^{H \times W \times 3}$. Given the pair of input images I_1, I_2, we aim to predict a flow field $F_{I_1 \to I_2}$. The flow field $F_{I_1 \to I_2}$ contains a 2D vector for each pixel in the source image I_1, representing the offset from the pixel coordinate in I_1 to the corresponding pixel coordinate in I_2, as $F_{I_1 \to I_2}(p_x^1, p_y^1) = (p_x^2 - p_x^1, p_y^2 - p_y^1)$. (p_x^1, p_y^1) stand for the pixel coordinates of a point p^1 in image I_1, while (p_x^2, p_y^2) denote the pixel coordinates of the corresponding point p^2 in image I_2.

Cost Volume Prediction. As shown in Fig. 2 (top), DODUO first extracts image features from I_1, I_2 with a convolutional encoder independently. We then use a Transformer-based neural network to correlate their image features. The output image features of the two images I_1 and I_2 are denoted as f_1 and f_2, respectively. f_1, $f_2 \in \mathbb{R}^{h \times w \times c}$, $h = H/8$ and $w = W/8$. Subsequently, we compute the cost volume between two frames based on feature similarity $C = \frac{f_1 f_2^T}{\sqrt{c}} \in \mathbb{R}^{h \times w \times h \times w}$. The cost volume C represents the similarity of each pixel pair between the source and target images.

Semantic-Aware Flow Estimation. We exploit the coarse semantic correspondence from the pretrained DINO image encoder [25] to provide priors for learning fine-grained correspondence. As shown in Fig. 2 (bottom), we use DINO to construct a matching candidate mask. The matching candidate mask is applied to the cost volume, narrowing down the matching space for dense flow estimation. In detail, we use the DINO encoder E_s to acquire semantic feature maps $f_i^s = E_s(I_i), i = 1, 2$ and $f_i^s \in \mathbb{R}^{h \times w \times c'}$. For each feature pixel $p \in f_1^s$, we identify the top 1% similar pixels in f_2^s. The resulting feature pixels in f_2^s constitute a mask for the feature pixel p, denoted as $\mathcal{M}_p \in \{0,1\}^{h \times w}$, representing the matching candidates of p in f_2^s. We compute the mask for each $p \in f_1^s$ to obtain a matching candidate mask $\mathcal{M}_{f_1^s \to f_2^s} \in \{0,1\}^{h \times w \times h \times w}$.

Next, we apply a masked softmax to the last two dimensions (belonging to the target image I_2) of the cost volume C with the matching mask $\mathcal{M}_{f_1^s \to f_2^s}$, yielding a normalized matching distribution $\tilde{C} = \text{MaskedSoftmax}(C, \mathcal{M}_{f_1^s \to f_2^s})$, where $\tilde{C} \in \mathbb{R}^{h \times w}$. The dense correspondence G' can then be attained by calculating the weighted average of the matching distribution with the 2D coordinates of pixel grid $G \in \mathbb{R}^{h \times w \times 2}$. And the flow field can be derived as the difference between the corresponding pixel coordinates as $\hat{F}_{I_1 \to I_2} = \tilde{C}G - G \in \mathbb{R}^{h \times w \times 2}$. Finally, we use bilinear upsampling to obtain a flow field of original resolution $\hat{F}^{up}_{I_1 \to I_2} \in \mathbb{R}^{H \times W \times 2}$.

B. Bootstrapping Visible Region Discovery

Self-supervised flow prediction contends with the challenge of occlusions. Applying loss on occluded pixels encourages the network to correspond pixels devoid of genuine matchings in the other frame, which in turn diminishes its performance [48], [59], [60]. Instead of detecting occlusions,

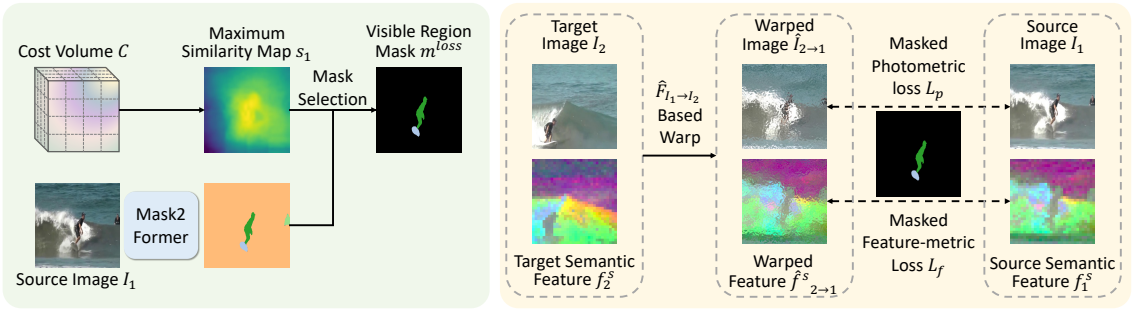

Fig. 3: **DODUO Training.** (Left) Bootstrapping Visible Region Discovery. We use Mask2Former [58] to get object segments of the source frame and select the segments that are most likely to be matched in the target frame using the predicted cost volume. (Right) We use the estimated flow to warp the pixels and the DINO features of the target image. The training objective is to minimize the photometric and feature-metric loss between the warped image and the source image in located visible regions.

we introduce a method of identifying instance-level visible regions in both frames. The visible regions are used as a mask for computing the loss, bringing instance-level priors for correspondence learning. This technique is particularly beneficial for image pairs exhibiting significant differences where occluded pixels are prevalent.

As shown in Fig. 3 (left), we employ an off-the-shelf image segmentation model Mask2Former [58] to get image region masks. Specifically, we attain segmentation masks of I_1 as $m_1 = \text{Mask2Former}(I_1) \in \{0,1\}^{H\times W\times N}$, where N is the maximum number of image regions.

To identify regions that are most likely to be matched in I_2, we leverage the cost volume C produced by DODUO. For each feature point of f_1, we compute the maximum similarity score, the similarity between this feature point and the closest feature point in f_2. A higher maximum similarity score indicates that the feature point is more likely to be matched in f_2. We obtain the maximum similarity map by maximizing out the last two dimensions of the cost volume $s_1 = \max(C, \text{dim} = (3,4)) \in \mathbb{R}^{H\times W}$. Next, we average the maximum similarity score for each pixel in a segment. We identify the top k segments with the largest scores, where the resulting segments are visible regions between the two images with high possibility, denoted as $m^{loss} \in \{0,1\}^{H\times W\times k}$.

C. Self-Supervised Losses

Photo-Metric and Feature-Metric Losses. As shown in Fig. 3 (right), following self-supervised optical training, we warp I_2 with the predicted flow field to obtain $\hat{I}_{2\to1} = \text{warp}(I_2, \hat{F}_{I_1\to I_2})$. We then employ the photo-metric loss to minimize the difference between $\hat{I}_{2\to1}$ and I_1. Specifically, we only apply the loss on the visible regions in both frames, identified as mask m^{loss}. And we use the Charbonnier loss ψ [61] to calculate the pixel-level discrepancy. Photometric loss offers fine-grained pixel-level supervision. However, it presumes color consistency between frames. This assumption holds for neighboring video frames but falters for frame pairs with appearance changes. Consequently, we introduce a feature-metric loss to motivate the network to concentrate on matching semantic features, which is more robust under appearance changes. We warp the semantic feature map of I_1 with the predicted flow field to get $\hat{f}^s{}_{2\to1} = \text{warp}(f_2^s, \hat{F}_{I_1\to I_2})$. Next, we acquire photo-metric loss L_p and feature-metric loss L_f by computing $\psi(I_1 - \hat{I}_{2\to1})$ and $\psi(f_1^s - \hat{f}^s{}_{2\to1})$ inside the visible regions m^{loss}.

Flow Regularizer. Prior works on self-supervised optical flow [49], [62] regularize the flow prediction using a smoothness term, assuming the motion between image pairs is translational. To relax the unrealistic assumption, we propose the distance consistency loss. We enforce the distance between neighboring pixels to remain consistent after warping by the flow. For a pair of neighboring pixels (p_i^1, p_j^1) in I_1, we find the corresponding pixels, denoted as $(\hat{p}_i^2, \hat{p}_j^2)$, in I_2 using our estimated flow field. We minimize $D(i,j) = \psi(||p_j^1 - p_i^1|| - ||\hat{p}_j^2 - \hat{p}_i^2||)$ inside each identified image regions separately and acquire regularization loss L_d. Our final training loss is defined as $L = L_p + L_f + L_d$.

Training DODUO. We train DODUO on video frames from Youtube-VOS dataset [35]. We randomly choose frame pairs with temporal intervals between 1-3 seconds for controllable variations between inputs. We apply random crop augmentation on the chosen frames.

IV. EXPERIMENTS

We first introduce the baselines for comparison. We compare DODUO with WarpC [45], SMURF [50], and DINO [25]. WarpC is an unsupervised warp-based method that predicts the dense flow field. SMURF is a method for unsupervised learning of optical flow. DINO (and recently updated version DINOv2 [32]) is a self-supervised semantic representation learning method. For DINO, we predict the dense flow in the same way as DODUO (Sec. III-A).

A. Evaluation Datasets and Metrics

We evaluate DODUO on TAP-Vid [36] dataset and D3D-HOI [63] dataset for evaluating fine-grained visual correspondence and articulation estimation, respectively.

TAP-Vid DAVIS is a long-term point-tracking dataset containing 30 videos from the DAVIS dataset [64] and providing point-level annotations. For each point, we use the frame where it first appears as the source image and use each of the following frames as the target image. We use the predicted flow to warp the source image points to the target image. The warped points are considered as their corresponding points in the target image. We follow the

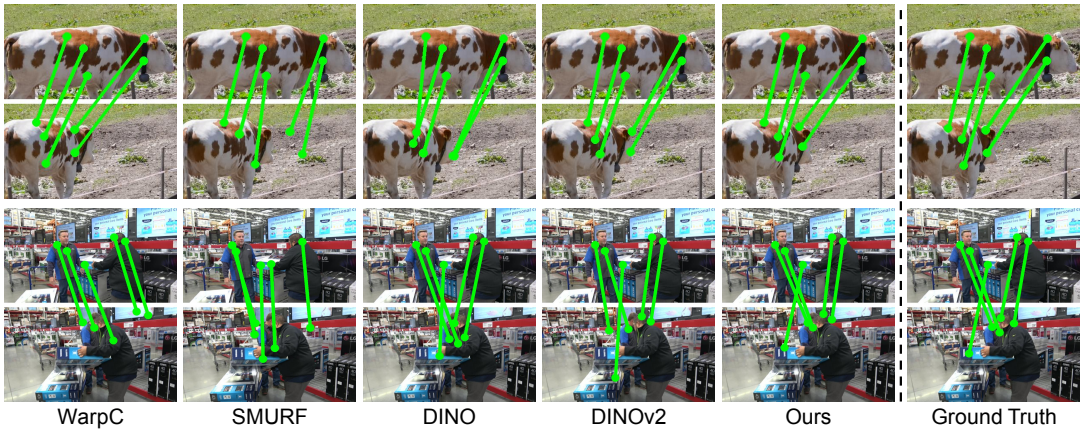

Fig. 4: Visualization of pixel-level correspondence results on TAP-Vid DAVIS dataset.

	Fine-Grained Corr.			Articulation Estimation Error				Computation	
	AJ$_\uparrow$	AD$_\downarrow$	$< \delta^x_{avg\uparrow}$	Angle$_\downarrow$	Pos$_\downarrow$	State$_\downarrow$	Dist$_\downarrow$	FPS$_\uparrow$	Size$_\downarrow$
WarpC [45]	25.8	28.1	35.8	11.41	0.182	13.43	0.148	**155**	**3M**
SMURF [50]	29.8	27.4	42.7	10.66	0.130	8.94	0.132	14	5M
DINO [25]	25.7	15.2	35.5	10.29	0.161	11.56	0.146	79	21M
DINOv2 [32]	27.2	13.4	36.0	12.14	0.144	8.06	0.126	47	87M
DODUO (ours)	**33.2**	**12.3**	**43.5**	**9.60**	**0.111**	**6.10**	**0.110**	55	4M

TABLE I: Quantitative comparison with baselines on TAP-Vid (left) and D3D-HOI (right).

Ablations	AJ (↑)	AD (↓)	$< \delta^x_{avg}$ (↑)
w/o visible region mask	29.9	13.0	39.8
w cycle-consistency mask	32.4	13.2	42.6
w/o feature-metric loss	32.5	**12.3**	42.9
w/o photometric loss	32.0	12.4	42.1
w/o flow regularizer	24.7	18.8	34.3
w smoothness regularizer	32.8	13.4	43.2
w/o semantic prior	29.1	22.7	40.2
Full model	**33.2**	**12.3**	**43.5**

TABLE II: Ablation study on model design choices.

official metrics for evaluation, including 1) *Average Distance* (AD) between the prediction and the ground truth in pixels; 2) $< \delta^x_{avg}$, which measures the average percentage of points within the distance threshold of 1, 2, 4, 8, and 16 pixels; 3) *Average Jaccard* (AJ), measuring the precision under the mentioned distance thresholds.

D3D-HOI is a video dataset with annotations of 3D object pose and part motion during human-object interaction. We filter out the videos with severe occlusions and noisy annotations, resulting in a subset of 159 videos and 4 object categories, all with revolute joints. We choose two frames from each video that capture half of the articulated motion as inputs. We estimate pixel correspondence on RGB frames and get their 3D points using depth images. We use the least square algorithm [65] to estimate the articulation parameters using the predicted 3D point correspondence. We evaluate *angle*, *position* and *state* errors of articulation parameters. To evaluate the predicted correspondence without ground truth annotation, we transform the points of the source image with the ground truth articulation parameters and compute the *distance* between the transformed points and the corresponding points in the target image.

Comparison with Baselines. We evaluate the accuracy of predicted correspondence on TAP-Vid. As shown in Tab. I (left) and Fig. 4, our model DODUO outperforms baselines on all metrics with a significant margin. In detail, DINO and DINOv2 achieve a slightly worse AD than our model while getting much worse AJ and $< \delta^x_{avg}$. The reason is that they can only provide coarse semantic correspondence for relating points with similar semantics, leading to coarse correspondence. Meanwhile, WarpC, trained for accurate dense matching, achieves slightly better AJ than DINO with a much worse AD. As shown in Fig. 4, SMURF excels in matching pixels with little movement between the two frames and fails to establish long-range correspondence. Therefore, it has the best AJ and the worst AD among the baselines. We also report the inference speed and number of trainable parameters to evaluate the computation cost.

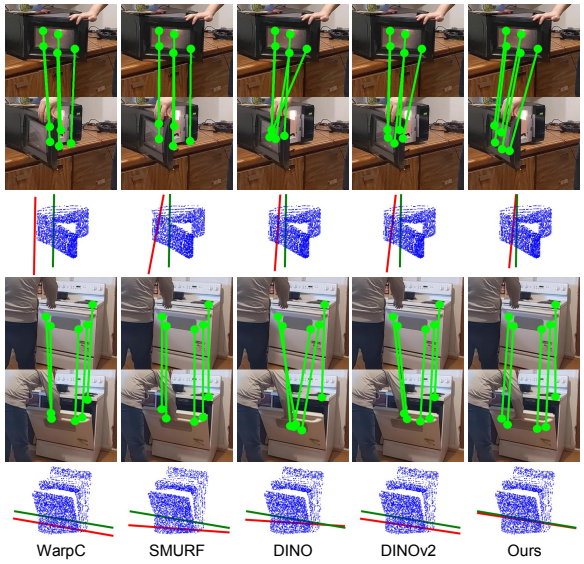

Fig. 5: **Visualization on D3D-HOI dataset.** The top two rows show pixel-level correspondence. The last row shows articulation estimation results, where the green and red lines are ground truth and estimation.

B. Evaluation of Fine-grained Correspondence

Ablation Study. To validate the design choices of DODUO, we conduct comprehensive ablation studies on the TAP-Vid dataset. The quantitative results are shown in Tab. II.

Method	Monkey	Sloth	Peg Insertion
DON [21]	4/10	2/10	1/10
DINOv2 [32]	5/10	8/10	4/10
DODUO (ours)	**10**/10	**9**/10	**8**/10

TABLE III: Quantitative results of zero-shot goal-conditioned object manipulation.

F1irst, we evaluate the importance of the visible region discovery (Sec. III-B) by applying the losses on the entire image. We observe that applying losses on regions invisible in the target frame forces the network to find correspondence for pixels with no genuine matching. It provides a false supervisory signal, leading to worse performance. We also implement a cycle-consistency-based loss mask, which also gives inferior performance. Then, we evaluate the effect of feature-metric and photometric loss (Sec. III-C). Both ablated versions give a worse AJ, indicating both losses contribute to the accuracy of the estimated correspondence. In addition, we investigate the flow regularizer (Sec. III-C). Removing the flow regularizer leads to a significant performance drop. Replacing the proposed distance consistency loss with a smoothness loss also leads to inferior performance, validating the efficacy of the distance consistency loss. Finally, we validate the importance of semantic prior by removing the matching candidate mask (Sec. III-A), which leads to a considerable drop in performance.

C. Evaluation of Articulation Estimation

As shown in Tab. I (right), our model gives better performance over the baselines on all metrics with significant margins. Visualizations in Fig. 5 show that WarpC cannot establish correct correspondences. SMURF tends to match points in the source frame to the points with the same coordinates in the target frame, while the DINO series can only match points with similar semantics. In contrast, DODUO demonstrates much more accurate correspondence prediction. The reason is that DODUO preserves the local structure of neighboring pixels during dense correspondence prediction, thanks to the self-supervised flow training with distance consistency regularizer. The accurate correspondence further improves the performance of articulation estimation.

D. Evaluation of Goal-Conditioned Object Manipulation

We demonstrate that DODUO can be applied to zero-shot goal-conditioned object manipulation. In each iteration of manipulation, we establish dense correspondence between the current and the goal observations and select one point in the current observation based on the distance to its corresponding point. Then, we back-project the selected point and its corresponding target point into 3D space, producing a manipulation action to move the object closer to the goal.

We conduct quantitative experiments on manipulating two deformable objects and one peg insertion task as in Fig. 6. We compare with DINOv2 [32], the strongest baseline in the other two evaluations, and DenseObjectNets (DON) [21], a well-established dense visual descriptor for robotics manipulation. Since DON produces category-level descriptors and we want to test the zero-shot generalizability of the visual correspondence model, we use the official model weights of DON pretrained on caterpillar observations, which are closest to our tested deformable objects.

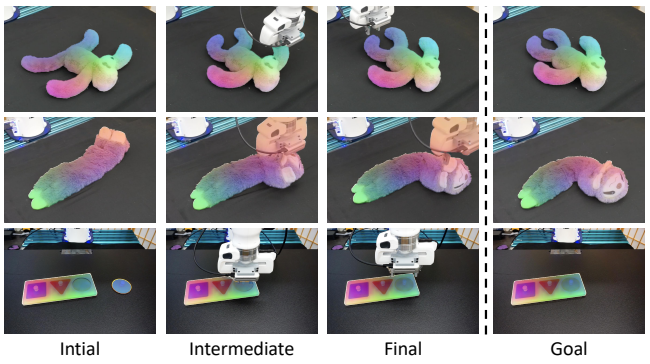

Fig. 6: Illustrations of DODUO in goal-conditioned manipulation tasks: "Monkey", "Sloth", and "Peg Insertion", ordered from top to bottom. We visualize the dense correspondence of manipulated objects using PCA of image features.

We evaluate the success rate of goal-conditioned manipulation driven by these correspondence models. For the two deformable objects, we follow ACID [66] and use Chamfer Distance as the success metric. We use a 7-DoF Franka Emika Panda arm with an Intel RealSense D435i RGBD camera to execute the manipulation action.

As shown in Tab. III, DON gets relatively low success rates. It is because DON fails to generalize to novel objects, especially the blocks in the peg insertion task. DINOv2 achieves a decent success rate in deformable object manipulation but fails in peg insertion tasks where precise manipulation is required. Accurate visual correspondence from DODUO leads to fine-grained actions, giving the highest success rate.

V. CONCLUSION

In this work, we present DODUO, a self-supervised learning approach for dense visual correspondence. DODUO blends the advantages of self-supervised optical flow and semantic feature learning, establishing robust, dense correspondences between image pairs that capture scenes undergoing signification transformations. Results show that DODUO predicts more accurate point-level correspondence over baselines. Furthermore, we demonstrate the applicability of DODUO to robotic tasks such as articulation estimation and zero-shot goal-conditioned manipulation. These results manifest the potential of dense visual correspondence in robotics perception.

In the future, we would like to scale up the training of DODUO to more diverse in-the-wild images, including arbitrary pairs of images with any common content, leveraging the full potential of our self-supervised training paradigm.

ACKNOWLEDGMENT

We would like to thank Yifeng Zhu for his help with real robot experiments and Alan Sullivan and Yue Zhao for helpful discussions. This work has been partially supported by NSF CNS-1955523, the MLL Research Award from the Machine Learning Laboratory at UT-Austin, and the Amazon Research Award.

REFERENCES

[1] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *2011 international conference on computer vision*. IEEE, 2011, pp. 2320–2327.

[2] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, *et al.*, "Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011, pp. 559–568.

[3] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *2011 10th IEEE international symposium on mixed and augmented reality*. Ieee, 2011, pp. 127–136.

[4] T. Whelan, S. Leutenegger, R. Salas-Moreno, B. Glocker, and A. Davison, "Elasticfusion: Dense slam without a pose graph." Robotics: Science and Systems, 2015.

[5] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments," *The international journal of Robotics Research*, vol. 31, no. 5, pp. 647–663, 2012.

[6] M. J. Black and A. D. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *International Journal of Computer Vision*, vol. 26, pp. 63–84, 1996.

[7] K. Hausman, S. Niekum, S. Osentoski, and G. S. Sukhatme, "Active articulation model estimation through interactive perception," *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3305–3312, 2015.

[8] Z. Jiang, C.-C. Hsu, and Y. Zhu, "Ditto: Building digital twins of articulated objects from interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5616–5626.

[9] H. Yin, A. Varava, and D. Kragic, "Modeling, learning, perception, and control methods for deformable object manipulation," *Science Robotics*, vol. 6, 2021.

[10] S. Tulsiani, A. Kar, J. Carreira, and J. Malik, "Learning category-specific deformable 3d models for object reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 719–731, 2017.

[11] M. Liao, Q. Zhang, H. Wang, R. Yang, and M. Gong, "Modeling deformable objects from a single depth camera," *2009 IEEE 12th International Conference on Computer Vision*, pp. 167–174, 2009.

[12] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 402–419.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 2004.

[14] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4937–4946, 2019.

[15] K. M. Yi, E. Trulls, V. Lepetit, and P. V. Fua, "Lift: Learned invariant feature transform," *ArXiv*, vol. abs/1603.09114, 2016.

[16] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable cnn for joint description and detection of local features," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8084–8093, 2019.

[17] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker, "Universal correspondence network," *ArXiv*, vol. abs/1606.03558, 2016.

[18] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8918–8927, 2021.

[19] W. Jiang, E. Trulls, J. H. Hosang, A. Tagliasacchi, and K. M. Yi, "Cotr: Correspondence transformer for matching across images," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6187–6197, 2021.

[20] T. Schmidt, R. Newcombe, and D. Fox, "Self-supervised visual descriptor learning for dense correspondence," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 420–427, 2016.

[21] P. R. Florence, L. Manuelli, and R. Tedrake, "Dense object nets: Learning dense visual object descriptors by and for robotic manipulation," *arXiv preprint arXiv:1806.08756*, 2018.

[22] P. Florence, L. Manuelli, and R. Tedrake, "Self-supervised correspondence in visuomotor policy learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 492–499, 2019.

[23] L. Manuelli, Y. Li, P. Florence, and R. Tedrake, "Keypoints into the future: Self-supervised correspondence in model-based reinforcement learning," *arXiv preprint arXiv:2009.05085*, 2020.

[24] L. Yen-Chen, P. Florence, J. T. Barron, T.-Y. Lin, A. Rodriguez, and P. Isola, "Nerf-supervision: Learning dense object descriptors from neural radiance fields," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6496–6503.

[25] M. Caron, H. Touvron, I. Misra, H. J'egou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021.

[26] G. Long, L. Kneip, J. M. Álvarez, H. Li, X. Zhang, and Q. Yu, "Learning image matching by simply watching video," in *European Conference on Computer Vision*, 2016.

[27] S. Meister, J. Hur, and S. Roth, "Unflow: Unsupervised learning of optical flow with a bidirectional census loss," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[28] Z. Ren, J. Yan, B. Ni, B. Liu, X. Yang, and H. Zha, "Unsupervised deep learning for optical flow estimation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[29] J. J. Yu, A. W. Harley, and K. G. Derpanis, "Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 3–10.

[30] P. Liu, M. Lyu, I. King, and J. Xu, "Selflow: Self-supervised learning of optical flow," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4571–4580.

[31] H.-P. Huang, C. Herrmann, J. Hur, E. Lu, K. Sargent, A. Stone, M.-H. Yang, and D. Sun, "Self-supervised autoflow," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 412–11 421.

[32] M. Oquab, T. Darcet, T. Moutakanni, H. Q. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. G. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," *ArXiv*, vol. abs/2304.07193, 2023.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[34] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[35] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. S. Huang, "Youtube-vos: A large-scale video object segmentation benchmark," *ArXiv*, vol. abs/1809.03327, 2018.

[36] C. Doersch, A. Gupta, L. Markeeva, A. Recasens, L. Smaira, Y. Aytar, J. Carreira, A. Zisserman, and Y. Yang, "Tap-vid: A benchmark for tracking any point in a video," *ArXiv*, vol. abs/2211.03726, 2022.

[37] T. Zhou, M. A. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6619, 2017.

[38] Q. Wang, X. Zhou, B. Hariharan, and N. Snavely, "Learning feature descriptors using camera pose supervision," *ArXiv*, vol. abs/2004.13324, 2020.

[39] T. Schmidt, R. A. Newcombe, and D. Fox, "Self-supervised visual descriptor learning for dense correspondence," *IEEE Robotics and Automation Letters*, vol. 2, pp. 420–427, 2017.

[40] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. P. Murphy, "Tracking emerges by colorizing videos," in *European Conference on Computer Vision*, 2018.

[41] X. Wang, A. Jabri, and A. A. Efros, "Learning correspondence from the cycle-consistency of time," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2566–2576.

[42] N. Wang, Y. Song, C. Ma, W. gang Zhou, W. Liu, and H. Li, "Unsupervised deep tracking," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1308–1317, 2019.

[43] Y. Zhong, Y. Dai, and H. Li, "Self-supervised learning for stereo matching with self-improving ability," *ArXiv*, vol. abs/1709.00930, 2017.

[44] M. E. Banani, I. Rocco, D. Novotný, A. Vedaldi, N. Neverova, J. Johnson, and B. Graham, "Self-supervised correspondence estimation via multiview registration," *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1216–1225, 2022.

[45] P. Truong, M. Danelljan, F. Yu, and L. V. Gool, "Warp consistency for unsupervised learning of dense correspondences," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10 326–10 336, 2021.

[46] ——, "Probabilistic warp consistency for weakly-supervised semantic correspondences," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8698–8708, 2022.

[47] Y. Wang, Y. Yang, Z. Yang, L. Zhao, P. Wang, and W. Xu, "Occlusion aware unsupervised learning of optical flow," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4884–4893.

[48] J. Janai, F. Güney, A. Ranjan, M. J. Black, and A. Geiger, "Unsupervised learning of multi-frame optical flow with occlusions," in *European Conference on Computer Vision*, 2018.

[49] R. Jonschkowski, A. Stone, J. T. Barron, A. Gordon, K. Konolige, and A. Angelova, "What matters in unsupervised optical flow," in *European Conference on Computer Vision*, 2020.

[50] A. Stone, D. Maurer, A. Ayvaci, A. Angelova, and R. Jonschkowski, "Smurf: Self-teaching multi-frame unsupervised raft with full-image warping," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 3887–3896.

[51] J. Xu and X. Wang, "Rethinking self-supervised correspondence learning: A video frame-level similarity perspective," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10 055–10 065, 2021.

[52] Y. Xiong, M. Ren, W. Zeng, and R. Urtasun, "Self-supervised representation learning from flow equivariance," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10 171–10 180, 2021.

[53] A. Jabri, A. Owens, and A. A. Efros, "Space-time correspondence as a contrastive random walk," *ArXiv*, vol. abs/2006.14613, 2020.

[54] Z. Lai, E. Lu, and W. Xie, "Mast: A memory-augmented self-supervised tracker," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6478–6487, 2020.

[55] Z. Lai and W. Xie, "Self-supervised learning for video correspondence flow," *ArXiv*, vol. abs/1905.00875, 2019.

[56] Y. Hu, R. Wang, K. Zhang, and Y. Gao, "Semantic-aware fine-grained correspondence," in *European Conference on Computer Vision*, 2022.

[57] C. Graf, D. B. Adrian, J. Weil, M. Gabriel, P. Schillinger, M. Spies, H. Neumann, and A. G. Kupcsik, "Learning dense visual descriptors using image augmentations for robot manipulation tasks," in *Conference on Robot Learning*. PMLR, 2023, pp. 871–880.

[58] B. Cheng, A. Choudhuri, I. Misra, A. Kirillov, R. Girdhar, and A. G. Schwing, "Mask2former for video instance segmentation," *ArXiv*, vol. abs/2112.10764, 2021.

[59] Y. Wang, Y. Yang, Z. Yang, L. Zhao, and W. Xu, "Occlusion aware unsupervised learning of optical flow," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4884–4893, 2017.

[60] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *European Conference on Computer Vision*, 2004.

[61] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," *Proceedings of 1st International Conference on Image Processing*, vol. 2, pp. 168–172 vol.2, 1994.

[62] J. J. Yu, A. W. Harley, and K. G. Derpanis, "Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness," in *ECCV Workshops*, 2016.

[63] X. Xu, H. Joo, G. Mori, and M. Savva, "D3d-hoi: Dynamic 3d human-object interactions from videos," *ArXiv*, vol. abs/2108.08420, 2021.

[64] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. V. Gool, "The 2017 davis challenge on video object segmentation," *ArXiv*, vol. abs/1704.00675, 2017.

[65] Å. Björck, *Numerical methods for least squares problems*. SIAM, 1996.

[66] B. Shen, Z. Jiang, C. Choy, L. J. Guibas, S. Savarese, A. Anandkumar, and Y. Zhu, "Acid: Action-conditional implicit visual dynamics for deformable object manipulation," *Robotics: Science and Systems (RSS)*, 2022.