A Supplementary Material

We first go over neural network architecture details in Section A.1. We then present additional quantitative results on the Bot-and-Objects (B&O) dataset in Section A.2. Next we show quantitative results on both datasets on an additional baseline. Next we present an additional result where we inspect the image reconstruction qualities from discovered Keypoint Pyramids. Next we describe the network architectures for the action category recognition on H3.6M experiments in Section A.5 where we show that improved keypoints discovery by our approach improves downstream task performance. Next, we demonstrate the robustness of our method to multi-object scenes. Lastly, we show more visualizations and examples on both datasets in Section A.7.

A.1 Architecture Implementation Details

In Section 3.5 of the main paper, we overviewed the architectural choices. Here we provied a more detailed description of the implementation details.

The FPN networks have four bottom-up layers and three top-down layers. Each bottom-up layer is consists of six Conv-BatchNorm layers. The filter size is 1×1 followed by 3×3, alternatively. For each top-down layer, it consists of one convolutional layer with a skip connection from the output of its corresponding bottom-up layer. The same FPN design is used as the feature extraction network $\Phi(x_{ref})$ and $\Phi_l(x)$ and to provide input to our keypoint network. The keypoint encoder at level l takes in FPN features $f_l(x)$ and $\Psi_{l-1}(x)$, followed by two Conv-BatchNorm-ReLu layers. The keypoint-only reconstruction decoders and feature refinement decoders are the reverse of their corresponding encoders with bilinear-upsampling layers to undo striding. We have also provided our implementation code for further implementation details.

A.2 Detailed Quantitative Results on B&O Dataset

In the main paper's Section 4.3, we reported keypoint regression results on H3.6M which comes with comprehensive annotations of 17 keypoints corresponding to main joints of the human body. For B&O, such comprehensive annotation is not possible because each scene contains three deformable objects and they have infinite degrees of freedom. Here, we flesh out the quantitative evaluation of discovered keypoint representations on B&O in two ways: first, by annotating 5 coarse key points as a direct analogue to the H3.6M results, and next, by reporting the errors for reconstructing the full image from the keypoints alone, for which we showed qualitative visualization results in Figer 7 in the main paper.

Coarse keypoint regression: On our B&O dataset, we randomly select 1000 images from the held-out evaluation set and annotate five keypoints: one on each of the three deformable objects, one on the QR code attached to the end effector

	Methods ↓ / Level (num. keypts.) →	level 1(10)	level 2(20)	flattened(30)
	Transporter	**113.74**	116.57	119.71
	KeyNet	114.41	118.10	119.53
	Keypoint Pyramids (Ours)	117.35	**115.85**	**113.76**
Ablations	(KP) No-Reconstruction	119.19	116.98	114.98
	(KP) No-Transport	119.26	117.15	115.69
	(KP) No-Spring	117.44	**115.36**	114.26
	(KP) All-Transport	120.54	117.68	115.28
	(KP) All-Reconstruction	119.39	117.01	114.50
	(KP) Unconditioned	117.73	**115.44**	**113.67**

Table 3. Keypoint regression RMSE error on B&O for regressing to coarse annotations of just 5 keypoints (2 on robot, 1 on each object), compared to prior flat unsupervised keypoint discovery baselines and ablations. Lower is better.

of the robot arm and one on the last revolute joint of the robot arm (these correspond to the two consistently visible joints of the arm). We further split these labeled test images: we train a linear regressor from discovered keypoints to our annotations on 500 samples and test on the remaining 500. We report the RMSE error in Table 3. We observe similar trends as in H3.6M dataset (Table 4.3). Flattened Keypoint Pyramids has much lower RMSE than baselines. The ablations on the objective functions (the No-X rows) show that dropping any one of them leads to worse performance, thus all terms in the designed objective function are important. In B&O, dropping the transport loss hurts the performance the most. All-Reconstruction again works slightly better than All-Transport, but both are worse than our method. Finally, Unconditioned, which removes the forward connections from coarser to finer keypoint levels produces marginally worse flattened representation than our full-approach, confirming that our training objective alone is sufficient to enforce the discovered keypoint hierarchy.

However, unlike the H3.6M keypoint regression results reported in the main paper, these new keypoint regression errors on B&O reported in Table 3 do not quite tell the full story: (1) regressors are trained and evaluated on much smaller datasets (500 images), (2) we are only evaluating the recovery of 5 coarse keypoints in a scene with many more degrees of freedom: three deformable objects and one articulated 5-DOF robot arm.

Indeed, these drawbacks cause some anomalous results. Both the flat baselines, Transporter and KeyNet benefit on this metric when trained with only 10 keypoints, compared to when they are trained with 20 or 30. This happens because (1) linear regression from a small number of discovered keypoints involves learning fewer parameters, and therefore can more effectively avoid overfitting on the small training set, and (2) further, the additional detail captured by larger representations provides no advantage when trying to recover merely the 5 hand-labeled coarse keypoints. This latter reason is also why Transporter with 10 keypoints is able to match the best performance of Flattened Keypoint Pyramids.

	Methods ↓ / Level (num. keypts.) →	level 1(10)	level 2(20)	flattened(30)
	Transporter	0.0061	0.0050	0.0060
	KeyNet	0.0041	0.0023	0.0027
	Keypoint Pyramids (Ours)	**0.0039**	**0.0021**	**0.0019**
Ablations	(KP) No-Reconstruction	0.0081	0.0067	0.0069
	(KP) No-Transport	0.0061	0.0031	0.0028
	(KP) No-Spring	0.0043	0.012	0.0024
	(KP) All-Transport	0.013	0.0071	0.0070
	(KP) All-Reconstruction	0.0039	0.0027	0.0025
	(KP) Unconditioned	0.0038	0.0025	0.0021

Table 4. Image reconstruction MSE error on B&O, compared to prior flat unsupervised keypoint discovery baselines and ablations. Lower is better.

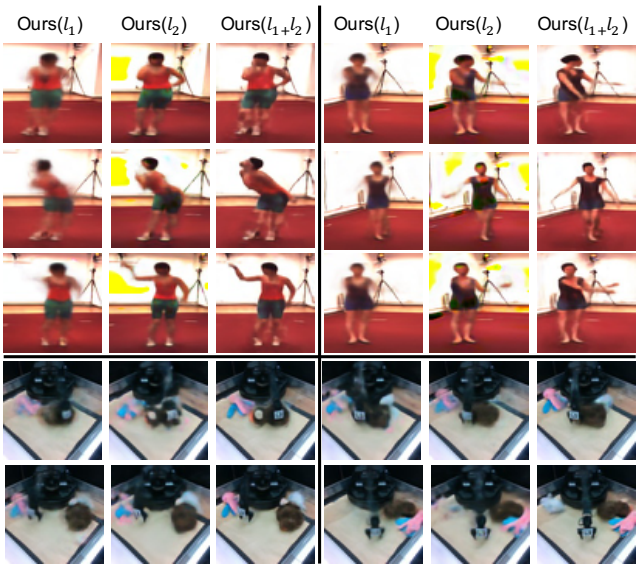

Fig. 7. Image reconstructions from keypoints discovered using Keypoint Pyramids.

A.3 Additional Baseline

In the main paper, we evaluated against Transporter (Neurips 2019) and KeyNet (Neurips 2018) because they remain the most widely used object keypoint discovery methods. Another reason is that we use components of these flat keypoint approaches in our hierarchical method, making them natural baselines for evaluating our key contributions. Here we include results on a more recent baseline PermaKey [12] in Table 5: while they report improved results compared to Transporter on Atari images, we find that it performs poorly on our more complex datasets.

	Method	level 1(10)	level 2(20)	flattened(30)	GT(17)
H3.6M	**KP(Ours)**	**52.81**	**43.97**	**43.30**	-
	PermaKey [12]	72.26	70.02	69.12	-
	Supervised	-	-	-	**37.97**
B&O	**KP(Ours)**	**117.35**	**115.85**	**113.76**	-
	PermaKey [12]	181.73	177.48	173.30	-

Table 5. Keypoint Regression RMSE error on H3.6M and B&O.

A.4 Image Reconstructions from Discovered Keypoint Pyramids

To fix the aforementioned shortcomings of our keypoint regression on B&O and perform a more thorough evaluation, we also report an image reconstruction MSE on all the held-out data in Table 4. For these results, we train stop-gradient decoders that takes in the discovered keypoint heatmaps $\Psi_l^n(x)$ or $\Psi(x)$, along with the "appearance" features $\Phi(x_{ref})$, to reconstruct the input image x. These decoders has the same architecture as the keypoint-only reconstruction decoder described in Section 3.2, and are trained along with the main architecture. We use these stop-gradient decoders for generating the visualizations in Figure 7. Unlike coarse keypoint regression, this metric rewards methods that more comprehensively capture the object configuration. The reconstructions help to easily visualize which information about the pose is correctly captured at various levels. Both datasets show a clear progression in detail from level 1 to 2 to the combination. For example, on H3.6M person images, level 1 omits arms nearly entirely and produces coarse estimates of the rest of the pose. Reconstructions get progressively sharper with more detailed keypoints, and the flattened keypoint representation in both cases produces the sharpest, least blurry reconstructions with quite detailed object poses. On B&O, reconstructions from the flattened representation (l1 + l2) are near-perfect, capturing the complex shapes and articulations of the octopus toy, and even permitting representing the QR code on the robot wrist. These results agree with our main observations both from the H3.6M results in the main paper, and from the results above: namely, Keypoint Pyramids outperforms flat baselines, and all components contribute to its performance. Additionally, the anomalous results from Table 3 do not hold up under this more comprehensive metric: coarser representations such as Transporter and KeyNet trained with 10 keypoints now perform poorly as expected, since they do not capture sufficient information for image reconstruction.

A.5 Architecture Details for H3.6M Action Category Recognition

To evaluate how our keypoint pyramids help with downstream usecases, we design an action classification task for action category recognition on H3.6M dataset. We train a 2-layer GRU network for this task. The input to this recurrent network is a sequence of predicted keypoints from a video snippet sampled

from a random time within the original video. The sampled video snippet is initially 25 frames long, and we subsample it 5x to contain only 5 frames before keypoint encoding. The keypoint representations of these five frames are the inputs to the GRU.

A.6 Robustness of Keypoint Pyramids in Multi-object Scenes

In the main paper, we show results on H3.6M which is a single-object scene and B&O, a real-world multi-object scene. Our choices of number of keypoints are only loosely influenced by the number of objects we expect: we set 10 and 20 keypoints for the two levels for both B&O and H3.6M datasets. Further, our method is robust to varying numbers of objects: in images with a missing/occluded object, we find that its corresponding keypoints are predicted with very low confidence, and could therefore be plausibly discarded (Fig 8).

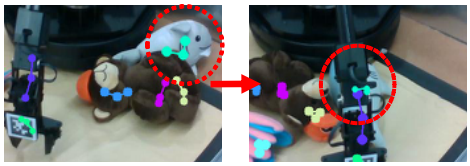

Fig. 8. Keypoint confidence with and without occlusion. Smaller dots means lower confidence.

A.7 Additional Qualitative Results

In Figure 9, 10, 11 and 12, we show more visualizations of our flattened keypoint representation versus the other two baseline methods that also predict 30 keypoints. Detailed analysis are in the captions. In summary, our Keypoint Pyramid consistently binds to parts of subjects under occlusions (Figure 9), large movements (Figure 10), and perspective changes (Figure 11), and can efficiently represents articulated objects (Figure 12).

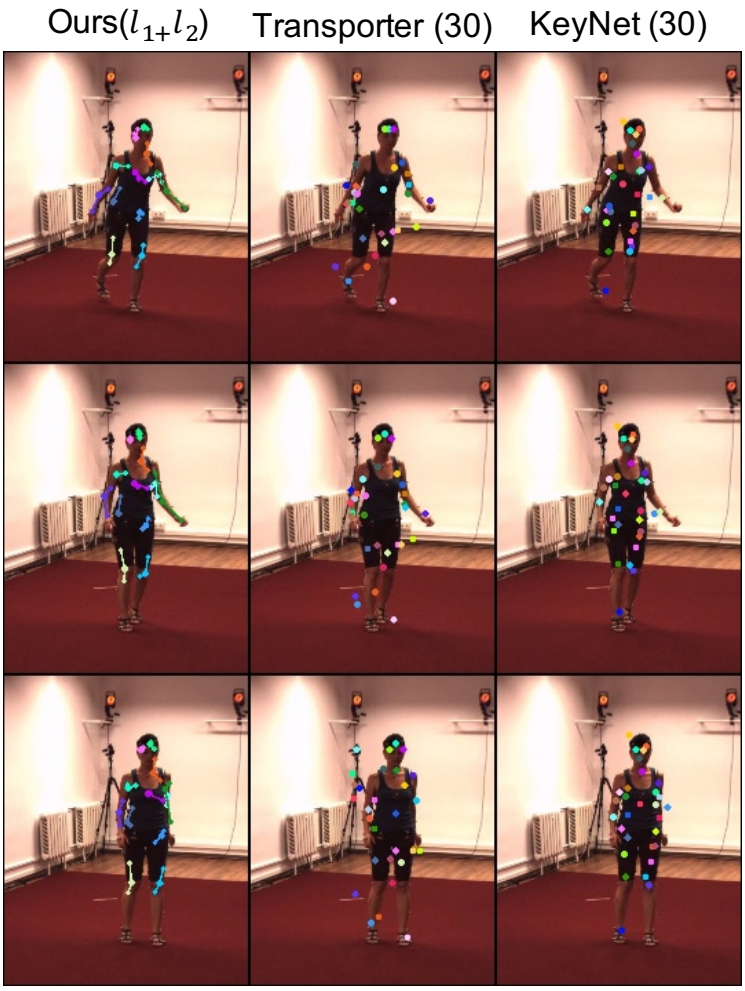

Fig. 9. (Best seen in pdf) More visualization of discovered keypoints from our method and baselines Transporter [24] and KeyNet [19]. For our method, parent and children keypoints are illustrated with the same color, as well as their connections. Even with significant occlusions, the discovered keypoint hierarchy (green) consistently maps to the left arm of the subject. However, for the two flat baselines, keypoints don't bind to the left arm consistently with occlusions (purple, yellow and blue points on the left arm for Transporter(30), light yellow point on the left arm for KeyNet(30))

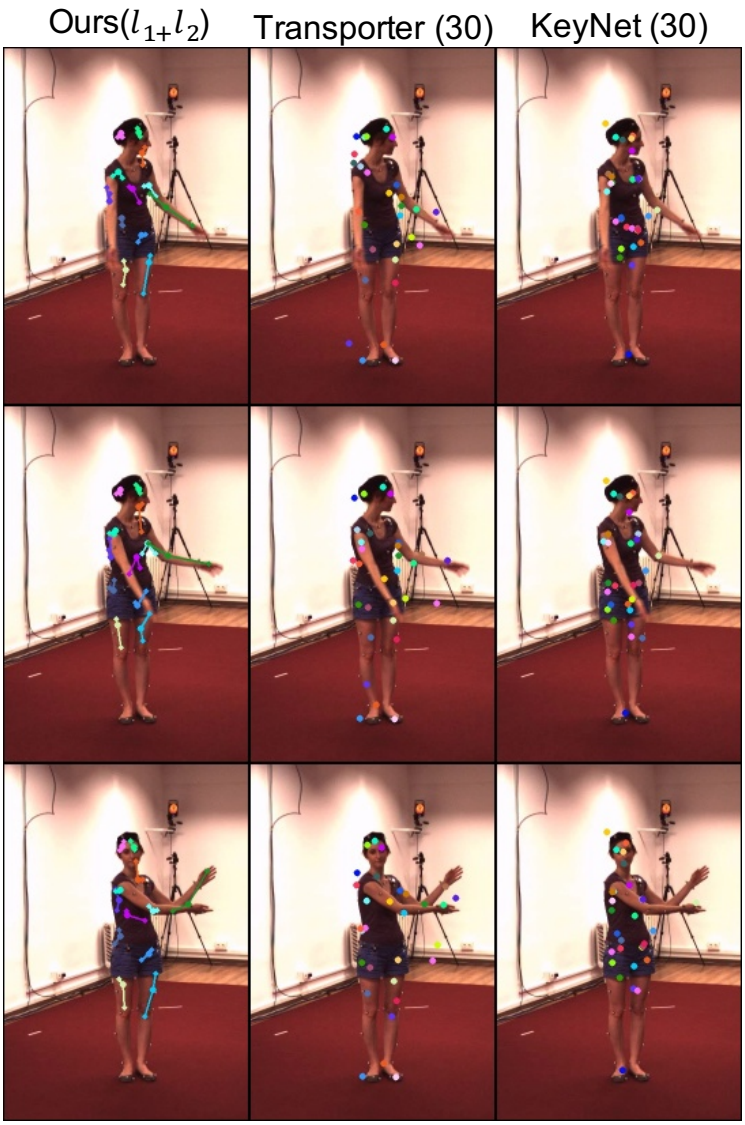

Fig. 10. (Best seen in pdf) More visualization of discovered keypoints from our method and baselines Transporter [24] and KeyNet [19]. When the subject exhibits large movements like raising both hands, it becomes easy to see that our Keypoint Pyramids manage to bind consistently to landmarks like shoulders (cyan and orange) and elbow(green). In contrast, for the flat baselines, keypoints that match to parts with large movements varies and it's hard to find consistent binding between keypoints and the body parts they represent.

Fig. 11. (Best seen in pdf) More visualization of discovered keypoints from our method and baselines Transporter [24] and KeyNet [19]. Even with large perspective changes, the keypoint pyramids consistently bind to certain body parts of the subject in the scene. For example, our blue pyramids binds to the left knee throughout the three frames presented here. In comparison, the dark pink point in Transporter(30) matches to the left feet in the first frame, lies in between two legs in the second frame and match to the left calf in the last frame. The light pink point in KeyNet(30) is on the left knee for the first and the third frame, but lies in between two legs in the second frame.

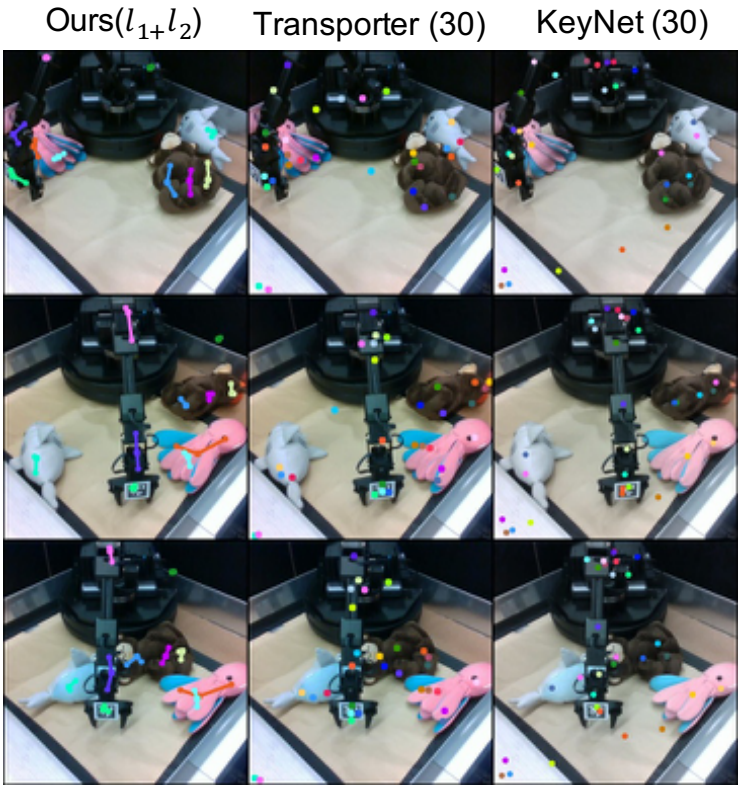

Fig. 12. (Best seen in pdf) More visualization of discovered keypoints from our method and baselines Transporter [24] and KeyNet [19]. Our discovered Keypoint Pyramids efficiently and consistently represent the articulated objects in the scene(pink, purple and light green). When the robot arm extends forward, the parent keypoints(pink, purple and light green) consistently map to each joint of the robot arm, while their children keypoints extend and capture the full pose of the robot arm. In contrast, the position and number of keypoints on the robot arm in the other baselines varies significantly.