Visual Robot Pose Tracking through Counter-Hypothetical Nonparametric Belief Propagation

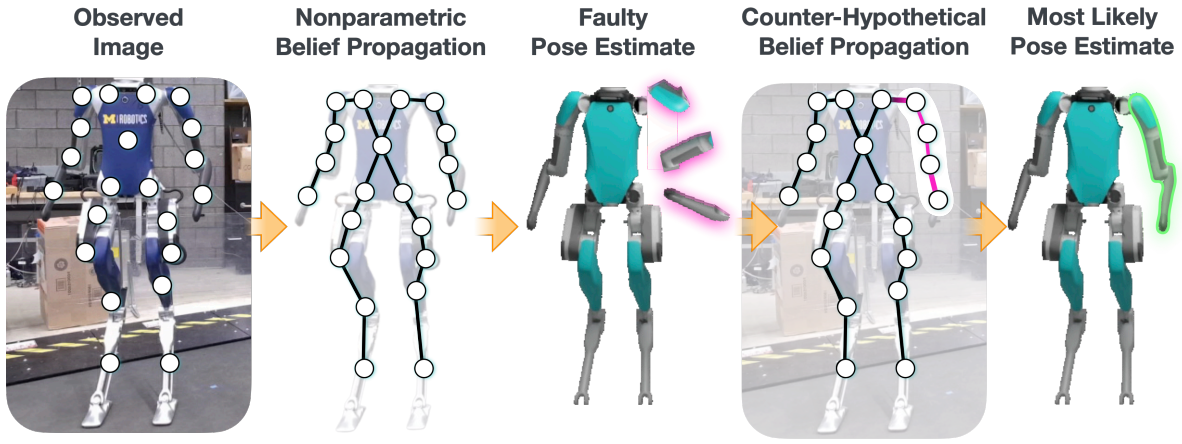

Fig. 1: Nonparametric belief propagation is a distributed graph algorithm that uses observed sensor data (e.g. camera images) to infer continuous poster distributions (e.g. articulated robot pose tracks) in noisy, unstructured environments. Due to its computational constraints, belief propagation often converges to faulty and incorrect estimates due to its discrete sample-based approximation of continuous state spaces. This paper investigates the potential for *Counter-Hypothetical reasoning* to be introduced within the belief propagation algorithm to overcome faulty estimates and produce the most likely posterior samples in real-world, continuous pose estimation and tracking tasks.

Elizabeth A. Olson J. Arden Knoll Anthony Opipari Grant Gibson Odest Chadwicke Jenkins

***Abstract*— For safe and efficient collaborative environments, co-located robots must be able to visually estimate and track the movements of surrounding robots. Nonparametric belief propagation provides a probabilistic framework for autonomous systems to reason about uncertainty and multiple hypotheses at once, as well as leverage portions of the robot that are better observed. Due to its computational constraints on the sample size, belief propagation often causes the filter to converge to incorrect estimates, because the particles are ineffectively representing the true belief. Our work seeks to maintain particle diversity by re-initializing the particles from a variety of proposal distributions as needed. We extend promising work of adaptive particle reinvigoration from a single distribution in the particle filter domain, which introduced *Counter-Hypothetical reasoning* to independently estimate when the filter was in failure mode. Our proposed framework explores the usefulness of this reasoning within belief propagation to manage sampling from multiple distributions and propagating its information through the standard graphical model. We present preliminary qualitative results for this method on tracking 21 links on a humanoid robot, Digit.**

I. Introduction

As we deploy robots into collaborative and unstructured environments, their fidelity in visual state estimation and tracking will become increasingly vital. Future autonomous systems must track the pose of other robots or humans in their workspace, the state of which will be of a high dimension and usually exclusively estimated through perception. Visual robot pose tracking is an appropriate task for validating and testing the efficacy of varying frameworks in high dimensional state tracking, because proprioceptive sensors can provide ground truth labels in spite of heavy occlusions. This domain also circumvents some open computer vision problems, such as intra-class variance or lack of known models, which allows research to focus on the underlying reasoning of tracking systems.

Probabilistic filters are a popular mechanism for pose estimation and tracking of such highly articulated objects. Bayesian filters leverage prior information of the object's pose when it becomes less observable. Their uncertainty estimation also aids in overcoming noisy or occluded observations. In such occluded viewpoints, a factor graph representation of the object is particularly useful. Belief propagation allows for components of the object that are better observed to inform the pose of less viewable parts. More specifically, nonparametric belief propagation enables the consideration of several hypotheses simultaneously when multiple poses could be plausible in the given observation, due to its ability to represent multi-modal belief distributions.

Unfortunately, exhaustively sampling across the state space is intractable. Render and compare methods typical in the likelihood function of pose estimation and tracking, as well as the quadratic computational burden of belief propagation with respect to sample size, limit particle-based methods to a relatively small number of samples at each node. This constraint leads to the filter inefficiently representing the true distribution in its belief, because particle deprivation in

E. A. Olson, J. A. Knoll, A. Opipari, G. Gibson, and O. C. Jenkins are with the Robotics Department, University of Michigan, Ann Arbor, MI, USA, {lizolson, knolljoh, topipari, grantgib, ocj}@umich.edu.

regions of likely pose causes the filter to explore implausible regions of the state space. Previous methods [1], [2] address the issue of particle diversity by augmenting the particle set by sampling from additional random distributions. Though Monte Carlo localization has had extensive work of adaptive particle reinvigoration [3], [4], [5], there has yet to be an exploration of this task in belief propagation, or insight on useful techniques to adaptively reinvigorate from several candidate distributions.

This paper presents a preliminary framework for combating particle deprivation in belief propagation through adaptive particle reinvigorating at each node from multiple random candidate distributions. Our work expands on Counter-Hypothetical particle filters, which presented the idea of independently quantifying the doubt associated with each sample to inform the rate of adaptive particle reinvigoration from a single random distribution in the domain of rigid object tracking. We similarly implement a Counter-Hypothetical likelihood function to produce a confidence score that a given hypothesis of an individual link of the robot *significantly wrong* based on the observation. Our work demonstrates how this estimation can be leveraged throughout the graphical structure to inform the amount of particles at each node that should be reinitialized from a uniform random distribution, or sampled off of a hypothesis pose of a neighboring node.

II. Related Work

A. Robot Pose Estimation and Tracking

Works for visual robot pose estimation have explored localizing joints [6] or keypoints [7] within the image, and leveraging knowledge of the kinematic structure and camera parameters to determine the pose of the robot. Other methods have examined rendering each individual hypothesis for a given link, and comparing against the observation for iterative improvement. This can be done in a probabilistic fashion with belief propagation[8], or through passing the observation and rendered hypothetical robot pose through a neural network [9]. Most datasets in this domain have no obstacles, and very little self-occlusion. Our work also uses render and compare for robot pose tracking, but we specifically focus on how to reliably track pose despite occlusions.

B. Belief Propagation for Articulated Objects

For tracking articulated objects in continuous state spaces, nonparametric belief propagation (NBP) methods [10], [11] have been proposed. In contrast to traditional sum-product belief propagation [12], which requires exact integral computations, NBP algorithms approximate continuous posterior distributions using graph-based message passing with discrete sample sets. NBP algorithms have been applied successfully to visual parts-based face localization tasks [11] as well as human hand tracking tasks [13]. Moreover, NBP methods have been shown to be effective in challenging human pose estimation and tracking problems [14]. Notably, the parts-based representation used by NBP algorithms in pose estimation tasks can be more efficient than alternative filtering algorithms, such as particle filters, for articulated objects with high-dimensional state spaces [2]. For an articulated object of interest, these approaches encode the known articulation constraints in a factor graph representation then use local message passing operations to infer the posterior distribution over each part's pose given access to some observed sensor data (e.g. images from a camera). While traditional NBP algorithms require substantial domain knowledge to encode an object's articulation constraints, recent work has demonstrated the potential for deep learning to be integrated with the factor-graph representation used by the NBP algorithms for learning these inter-part relationships [15], [16].

C. Particle Diversity

Sample diversity in particle-based inference spans several domains and works. Example works from Monte Carlo Localization sample a portion of the next particle set from a distribution auxillary to the particle set distribution. The adaptive rate of this reinvigoration was first calculated by a user-determined upper-bound threshold of acceptable weights [3], which was then dynamically set by comparing against historic average weightings over time [17]. More recently, a fixed portion of new particles have been augmented to the particle set by sampling off of output from a CNN [5]. We expand Counter-Hypothetical particle filters [18], which argue the invalidity of directly measuring doubt as the absence of likelihood due to the potential presence of ignorance and ambiguity in the sensor model. To this end, we similarly inform the rate of particle reinvigoration at a given time step based any glaring inconsistencies between the observation and potential hypotheses.

Additionally from the tracking domain, simulated annealing function cyclically softens the shape of the likelihood function to allow for more exploration of the state space at intermittent iterations of the filter [19], [20]. However, annealing adds another component of hand-tuning for its scheduling, and will not recover belief when a large region of the state space near the true state is deprived of samples. Specifically for belief propagation, reinitializing particles from sampling off of a set of proposal distributions, as opposed to one, has been proposed [1] and implemented for improved performance in pose estimation for highly articulated objects [2]. We build off of this idea by combining it with adaptive particle reinvigoration, in order to determine the rate each distribution should be drawn from online.

III. Methodology

Given a sequence of t RGB-D images $z_{1:t}$, we seek to localize the 6D pose, x_{st}, of an link s at time t. The marginal belief distribution of X_s at time t, $bel^t(X_s)$, can be approximated by

$$bel^t(X_s) \propto \phi_s(X_s, Z_s) \prod_{r \in \rho(s)} \hat{m}^t_{rs}(X_s) \tag{1}$$

where $\phi_s(X_s, Z_s)$ is the unary potential of the latent state X_s and its corresponding observable state, Z_s. $\hat{m}^t_{rs}(X_s)$ represents the message passed from r to s, where r is a neighboring node of s as indicated with $r \in \rho(s)$. The messaged passed from r to s at is defined as:

$$\hat{m}^t_{rs}(X_s) = \sum_{X_r \in \mathbb{X}_r} \phi_t(X_r, Z_r)\psi_{r,s}(X_r, X_s) \prod_{u \in \rho(r) \setminus s} \hat{m}_{ur}(X) \tag{2}$$

These equations demonstrates the chain of message passing coming into a given node through belief propagation. Specifically with nonparametric belief propagation, the belief distribution of $bel^t(X_s)$ is represented by a set of particles $\mathbb{X}_s$:

$$\mathbb{X}_s = \{(x^1_s, \pi^1_s), (x^2_s, \pi^2_s), \ldots, (x^N_s, \pi^N_s)\} \tag{3}$$

where x^i_s is the ith sample of the particle set, and π^i_s is its corresponding normalized importance weighting, given from Equation 1, and N is the number of particles at the given node. In traditional nonparametric belief propagation, the next set of particles would be sampled off of the current set, and the probability of a given particle being selected would be based on its importance weighting. However, this causes mode collapse in the underlying belief distribution, and can push the filter into failure mode. In practice, $\mathbb{X}_s$ for the next iteration is often a combination of samples from the current set, $\mathbb{X}^{prop}_s$, as well as randomized particles sampled of a set of sampled off of other candidate proposal distributions, $\mathbb{X}^{aug}_s$. With $\mathbb{X}_s = \mathbb{X}^{prop}_s \bigcup \mathbb{X}^{aug}_s$, the ratio from which to sample off of each distribution needs to be addressed.

From the Counter-Hypothetical Particle Filter, we estimate α_s, the ratio of samples from $\mathbb{X}^{prop}$ for the particle set at the s node:

$$\alpha_s = \frac{\sum_{i=1}^{N} \mathcal{L}(x^i_t)}{\sum_{i=1}^{N} \mathcal{C}(x^i_t) + \sum_{i=1}^{N} \mathcal{L}(x^i_t)} \tag{4}$$

where $\mathcal{L}(x^i_t)$ is the unnormalized likelihood weighting for the given particle, and $\mathcal{C}(x^i_t)$ is the unnormalized weighting from the Counter-Hypothetical likelihood. Note that the ratio of particles sampled from distributions other than the previous particle, comprising $\mathbb{X}^{aug}$, set would be $1-\alpha$. This formulation is incomplete for adaptive particle reinvigoration within belief propagation, as the there are multiple candidate distributions from which the samples can be reset. They may be reinitialized form a random uniform distribution, similar to initialization, which we'll denote $\mathbb{X}^{rand}$. Otherwise, they may be sampled off of a particle from a neighboring node, as described in [1], [2], denoted here as $\mathbb{X}^{pair}$. To extend the notation of particle reinvigoration to this case, we find $\mathbb{X}^{aug} = \mathbb{X}^{rand} \bigcup \mathbb{X}^{pair}$, leaving us to determine the ratio between $\mathbb{X}^{rand}$ and $\mathbb{X}^{pair}$.

We then introduce β, which is the ratio of augmented particles sampled off of neighboring samples, $\mathbb{X}^{pair}_s$. Intuitively, this ratio should be in accordance with our confidence that the neighboring nodes are containing plausible samples. Therefore, it is based on the α scores of each of the neighboring nodes:

$$\beta_s = (1 - \alpha_s) * \frac{1}{M} \sum_{r \in \rho(s)} \alpha_r \tag{5}$$

For the particles reinvigorated in $\mathbb{X}^{pair}$, the frequency each neighboring node's pairwise distribution is sampled off of is proportional to its α score relative to the other neighbors.

Lastly, the ratio of particles to be sampled from a uniform random distribution is defined as γ. This is then calculated from the other ratios and the fact that they must sum to 1:

$$\gamma_s = 1 - \alpha_s - \beta_s \tag{6}$$

The size of $\mathbb{X}^{prop}_s$, $\mathbb{X}^{pair}_s$, and $\mathbb{X}^{rand}_s$ are then $\alpha_s N$, $\beta_s N$, $\gamma_s N$ respectively. Note $\alpha_s N$, $\beta_s N$, and $\gamma_s N$ are all integers that would be rounded to sum to N.

With this formalization, the filter is able to maintain the Bayesian prior in nodes where there is little evidence that the current particles are wrong. Additionally, samples that appear to be wrong at a given node can be sampled off of their neighboring nodes when those exhibit promise, so it more quickly produces valid configurations. When the current node and neighboring nodes do not seem correct, the filter remains in the global localization stage for more iterations.

IV. Experiments

To analyze the ability of Counter-Hypothetical belief propagation to overcome poor initialization and maintain particle diversity, we validate its performance on tracking the pose of Digit, a humanoid from Agility Robotics [21]. We collected 60 sequences of 15 seconds capturing Digit moving within a workspace. Though several sequences were fully observable similar to other benchmark datasets, most of our sequences were heavily occluded, as seen in Fig. 3. Though we currently only present qualitative results on an unoccluded sequence, this collection is a stepping stone to a benchmarked dataset for testing performance on tracking complex and occluded movements. A series of fidicual markers are attached to Digit's torso to continually track the 6D pose of the torso. This data, along with the recorded encoder readings at each joint allow for full pose annotation at each time step and quantitative results for later work.

Our implementation of belief propagation seeks to track the 6D pose of Digit's 21 links. The pairwise potential for the algorithm measures the compatibility between a particle in a given node with the particles in its neighboring nodes. The compatibility measurement rewards closeness between the appropriate end points of neighboring links, as well as relative rotation transformations within the joint's known limits and axes. We currently use simple render and compare method to estimate the uncertainty associated with each sample based on the observation. For each hypothesis of a given link, it is rendered based on the camera's known parameters, providing a synthetic depth image for the hypothesis, as well as a segmentation of the hypothesis. The captured depth

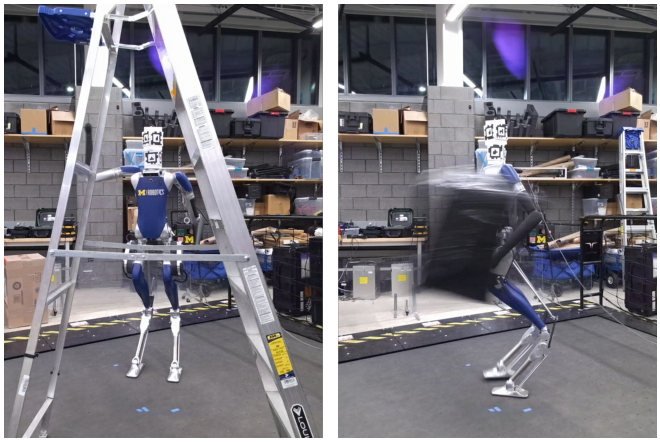

Fig. 3: Examples of collected sequences. (Left) We staged stationary occlusions in front of the robot, the ladder in particular has similar coloring and structure to some of the links of Digit. (Right) We have several sequences with dynamic obstacles quickly crossing the scene, as depicted with a black box, to test tracking.

image is masked to the segmentation, and compared against the rendered synthetic depth image. The unary likelihood of each particle is the ratio of pixels in the segmentation that have little difference between the synthetic rendered depth image and captured depth image for the corresponding pixel.

The Counter-Hypothetical likelihood function is also the masked rendered and captured depth images, similar to the likelihood function. Whereas the traditional unary likelihood is an inlier function, the Counter-Hypothetical likelihood function measures the ratio of pixels with depths lying outside of the margin of error. However, the Counter-Hypothetical likelihood function is not merely a function with a zero-sum relationship with the likelihood function. When the captured depth image has a distance shorter than the rendered depth image, that can be explained away by an occlusion of the link. This mismatch is not necessarily a signal misalignment, and can be caused by ignorance or ambiguity in the observation. Instead, the Counter-Hypothetical likelihood function only measures mismatch depths when the captured depth image has a *longer* distance than the rendered depth image. In this case, there is unambiguous misalignment

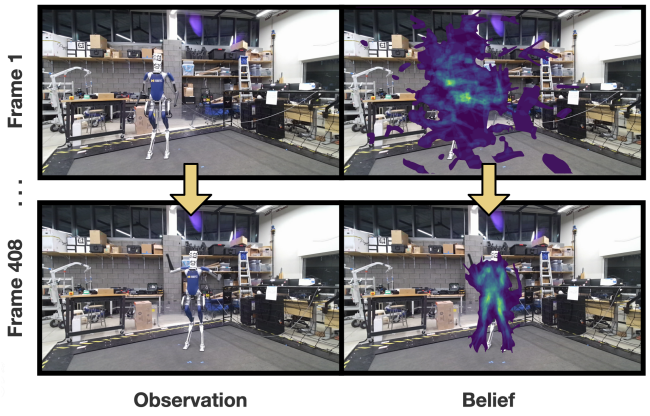

Fig. 4: Qualitative results: (Left) Examples of the RGB portion of the input data for the filter. (Right) Visualizing the belief at each time frame. (Top) At the beginning of the sequence, the initial belief is very random and incorrect. (Bottom) Towards the end of the sequence, it has begun to converge to a valid configuration, with less particle reinvigoration needed in the underlying distribution.

that cannot be explained by occlusion, and might be a red flag that the hypothetical link pose is significantly wrong.

For implementation details, these likelihood functions rely on segmentation masks provided by the YOLOv8 network [22]. We finetuned pretrained weights for the network on synthetic renderings of Digit which gave automatically generated ground truth masks. Each node had 50 particles.

V. Results and Future Work

An example qualitative result is provided in Figure 4. The belief distribution at the beginning of a sequence is very noisy before it has converged, and does not contain a valid configuration. Over time, our method is able to converge to a reasonable pose. We find that the high rates of reinvigoration at the beginning of the sequence are dampened as the maximum likelihood estimates of each node begin to become more plausible.

These estimates are still noisy, and we look to tweak the likelihoods and filter hyperparameters. There is also enough data across the 60 sequences to partition the data into training and testing, so that the traditional likelihood and Counter-Hypothetical likelihood can be trained end-to-end. We also will look to see if the Counter-Hypothetical likelihood function can be better learned by leveraging the global structure of the factor graph.

References

[1] J. Pacheco, S. Zuffi, M. Black, and E. Sudderth, "Preserving modes and messages via diverse particle selection," in *International Conference on Machine Learning*. PMLR, 2014, pp. 1152–1160.

[2] J. Pavlasek, S. Lewis, K. Desingh, and O. C. Jenkins, "Parts-based articulated object localization in clutter using belief propagation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 595–10 602.

[3] S. Lenser and M. Veloso, "Sensor resetting localization for poorly modelled mobile robots," in *International Conference on Robotics and Automation (ICRA)*, vol. 2, 2000, pp. 1225–1232.

[4] S. Thrun, D. Fox, W. Burgard, and F. Dellaert, "Robust monte carlo localization for mobile robots," *Artificial intelligence*, vol. 128, no. 1-2, pp. 99–141, 2001.

[5] N. Akai, T. Hirayama, and H. Murase, "Hybrid localization using model-and learning-based methods: Fusion of monte carlo and e2e localizations via importance sampling," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6469–6475.

[6] M. Stoiber, M. Sundermeyer, W. Boerdijk, and R. Triebel, "A multi-body tracking framework–from rigid objects to kinematic structures," *arXiv preprint arXiv:2208.01502*, 2022.

[7] J. Lu, F. Richter, and M. C. Yip, "Pose estimation for robot manipulators via keypoint optimization and sim-to-real transfer," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4622–4629, 2022.

[8] K. Desingh, S. Lu, A. Opipari, and O. C. Jenkins, "Factored pose estimation of articulated objects using efficient nonparametric belief propagation," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7221–7227.

[9] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, "Single-view robot pose and joint angle estimation via render & compare," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1654–1663.

[10] M. Isard, "PAMPAS: Real-valued graphical models for computer vision," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2003.

[11] E. B. Sudderth, A. T. Ihler, W. T. Freeman, and A. S. Willsky, "Nonparametric belief propagation," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2003, pp. 605–612.

[12] J. Pearl, "Chapter 4 - belief updating by network propagation," in *Probabilistic Reasoning in Intelligent Systems*, J. Pearl, Ed. San Francisco (CA): Morgan Kaufmann, 1988, pp. 143 – 237.

[13] E. B. Sudderth, M. I. Mandel, W. T. Freeman, and A. S. Willsky, "Visual hand tracking using nonparametric belief propagation," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*. IEEE, 2004, pp. 189–189.

[14] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard, "Tracking loose-limbed people," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2004, pp. 421–428.

[15] N. Heppert, T. Migimatsu, B. Yi, C. Chen, and J. Bohg, "Category-independent articulated object tracking with factor graphs," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2022, Kyoto, Japan, October 23-27, 2022*. IEEE, 2022, pp. 3800–3807.

[16] A. Opipari, J. Pavlasek, C. Chen, S. Wang, K. Desingh, and O. C. Jenkins, "Differentiable nonparametric belief propagation," *ICRA Workshop: Robotic Perception and Mapping: Emerging Techniques*, 2022.

[17] J.-S. Gutmann and D. Fox, "An experimental comparison of localization methods continued," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 1. IEEE, 2002, pp. 454–459.

[18] E. Olson, J. Pavlasek, J. Berry, and O. C. Jenkins, "Counter-hypothetical particle filters for single object pose tracking," in *2023 International Conference on Robotics and Automation (ICRA) (to appear)*.

[19] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. IEEE, 2000, pp. 126–133.

[20] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An introduction to mcmc for machine learning," *Machine learning*, vol. 50, pp. 5–43, 2003.

[21] A. Robotics, "Digit robot," https://agilityrobotics.com/robots.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.