

VideoPose: Estimating 6D object pose from videos

Apoorva Beedu*
Georgia Institute of Technology
abeedu3@gatech.edu

Varun Agrawal
Georgia Institute of Technology
varunagrawal@gatech.edu

Zhile Ren
Georgia Institute of Technology
jrenzhile@gmail.com

Irfan Essa
Georgia Institute of Technology
irfan@gatech.edu

Abstract

We introduce a simple yet effective algorithm that uses convolutional neural networks to directly estimate object poses from videos. Our approach leverages the temporal information from a video sequence, and is computationally efficient and robust to support robotic and AR domains. Our proposed network takes a pre-trained 2D object detector as input, and aggregates visual features through a recurrent neural network to make predictions at each frame. Experimental evaluation on the YCB-Video dataset show that our approach is on par with the state-of-the-art algorithms. Further, with a speed of 30 fps, it is also more efficient than the state-of-the-art, and therefore applicable to a variety of applications that require real-time object pose estimation.

1. Introduction

Estimating the 3D translation and 3D rotation for every object in an image is a core building block for many applications in robotics [27, 31, 6] and augmented reality [21]. The classical solution for such 6-DOF pose estimation problems utilises a feature point matching mechanism, followed by Perspective-n-Point (PnP) to correct the estimated pose [26, 29, 25, 12]. However, such approaches fail when objects are texture-less or heavily occluded. Typical ways of refining the 6DOF estimation involves using additional depth data [33, 3, 9, 15] or post-processing methods like ICP or other deep learning based methods [36, 13, 19, 28], which increase computational costs. Other approaches treat it as a classification problem [32, 13], resulting in reduced performance as the output space is not continuous.

In robotics, augmented reality, and mobile applications, the input signals are usually videos rather than a single image. Li *et al.* [18] utilize multiple frames from different viewing angles to estimate single object poses, which does

not work robustly in complex scenes. Wen *et al.* [35] and Deng *et al.* [5] use tracking methods to estimate the poses, however these methods do not explicitly exploit the temporal information in the videos. The idea of using more than one frame to estimate object poses has seen limited exploration. As the object poses in a video sequence are implicitly related to camera transformations and do not change abruptly between frames, and as different viewpoints of the objects aid in the pose estimation [17, 4], we believe that modelling a temporal relationship can only help the task.

Motivated by this, in our proposed approach, we utilize a simple CNN-based architecture to extract useful features, and subsequently aggregate the information across consecutive frames using a recurrent neural network (RNN). The training is performed on the YCB-Video dataset [36] and the approach achieves comparable performance to state-of-the-art approaches, while requiring lower computational costs. We also conduct extensive ablation studies and demonstrate the effectiveness of our network design.

The primary contributions of our paper are:

- We introduce a simple yet effective neural network architecture for estimating 6-DOF object poses from videos. Our system first extracts image features and estimates depth and labels, then use a temporal module to aggregate information across frames and estimate 6-DOF pose of every object in the current frame.
- We perform extensive ablation studies on different design choices of the system, and show that using videos, as opposed to using single images, can improve the predictions significantly at an improved computational speed of 30 fps.

2. Related Work

Estimating the 6-DOF pose of objects in the scene is a widely studied task. The classical methods either use

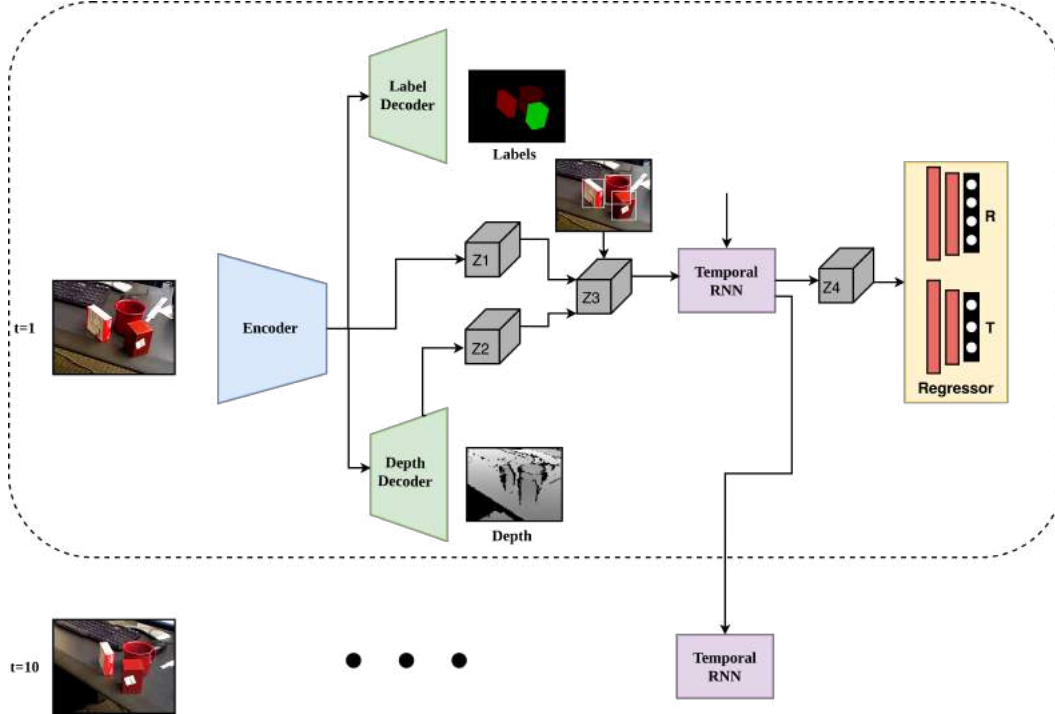


Figure 1. Overview of our VideoPose framework for 6D object pose estimation. We use the same encoder as in [36]. Z_4 is the fused features from Fig. 2

template-based or feature-based approaches. In template-based methods, a template is matched to different locations in the image, and a similarity score is computed [9, 8]. However, these template matching methods could fail to make predictions for textureless objects and cluttered environments. In feature based methods, local features are extracted, and correspondence between known 3D objects and local 2D features is established using PnP to recover 6D poses. However, these methods also require sufficient textures on the object to compute local features and face difficulty in generalising well to new environments as they are often trained on small datasets.

Convolutional Neural Networks (CNNs) have proven to be an effective tool in many computer vision tasks. However, they rely heavily on the availability of large-scale annotated datasets. Thus, the YCB-Video dataset [36], T-LESS [10], and OccludedLINEMOD dataset [16, 25] were introduced. These datasets have enabled the emergence of novel network designs such as PoseCNN, DPOD [37] and PVNet [36, 25]. To further increase the amount of accurate annotated data, Tremblay *et al.* [30] introduced synthetically generated photo-realistic data, which when trained on, gave improved performances on the estimation task [31]. In this paper, we use the challenging YCB-Video dataset, as it is a popular dataset that serves as a testbed for many recent algorithms.

Building on those datasets, various CNN architectures

have been introduced to learn effective representations of objects, and thus estimate accurate 6D poses. Kehl *et al.* [13] extends SSD [20] by adding an additional viewpoint classification branch to the network. Rad *et al.* [26] and Telkin *et al.* [29] predict 2D projections from 3D bounding box estimations. However, these methods fail to deal with pose ambiguities and objects under heavy occlusion. Most notably, PoseCNN [36] uses a Hough voting scheme to vote for the center of the object and then use the bounding boxes to estimate the 3D rotation. To address the problems of heavy occlusions and ambiguities, [25, 12, 23, 22] learn to detect keypoints and then perform PnP. However, these methods also encounter similar problems of pose ambiguities for symmetric and partially occluded objects.

Other methods involve a hybrid approach where the model learns to perform multiple tasks. Song *et al.* [28] enforce consistencies among keypoints, edges, and object symmetries. Billings *et al.* [2] predict silhouettes of objects along with object poses. There is also a growing trend of designing model agnostic features [34] that can handle novel objects. These directions are beyond the scope of our paper, as our goal is to find the best practice for pose estimation in videos when the objects of interest are given.

To refine the predicted poses, several works use additional depth information and perform a standard ICP algorithm [36, 13], or directly learn from RGB-D inputs [33, 19, 37]. We argue that since the input signals to robots

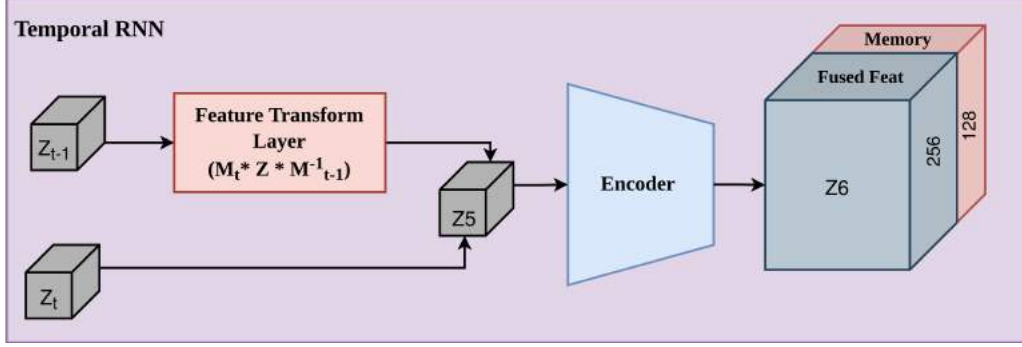


Figure 2. Overview of a simple temporal RNN network: Z_{t-1} is the feature from the previous time-step and Z_t is the $Z3$ from Fig 1.

and/or mobile devices are typically video sequences, instead of heavily relying on additional depth information, estimating poses in videos by exploiting the temporal data could already refine the single pose estimations. A notable work from Deng *et al.* [5] introduces the PoseRBPF algorithm that uses particle filters to track objects in video sequences. This state-of-the-art algorithm provides accurate estimations at a high computational cost. Wen *et al.* [35] also perform tracking, but use synthetic rendering of the object at the previous time-step. In contrast to the above papers, we show that a simple temporal module that aggregates information across different frames at a high computational speed performs comparable or better than using single frames.

3. Approach

Given an RGB video stream, our goal is to estimate the 3D rotation and 3D translation of all the objects in every frame of the video. We assume the system has access to the 3D model of the object. In the following sections, R denotes the rotation matrix with respect to the annotated canonical object pose, and T is the translation from the object to the camera.

3.1. Overview of the network

Our pipeline, shown in Figure 1 consists of two stages. The first stage corresponds to the feature extractor, depth estimator and semantic label predictor. The second has a temporal RNN and a Regressor. For extracting image features, we use a simple VGG-16 model similar to [36], and fine-tune the last 2 layers to encode features from the depth and semantic prediction tasks.

Pose Estimation relies on accurate object detection. The object detection module is responsible for giving the class-id and Region-Of-Interest (ROI). During training, we use the ground truth bounding box and during testing, we use the predictions and bounding box from the PoseCNN model. This can ideally be replaced with any lightweight feature extraction model such as MobileNet [11] to make

the inference faster, but we choose bounding boxes from PoseCNN for fair comparison to prior works. We learn the depth and semantic segmentation using a Decoder consisting of 3 CNN layers. The final layer from the feature extractor and the penultimate layer of the depth estimator are concatenated and pooled together by using ROI Align [7] which is passed through the temporal layer. We believe that adding depth features can aid in the estimation, when depth information is not available.

3.2. Temporal block

To use the features from the previous step, we project the image features from the previous step to current step using the camera transformation matrix M as $M_t * feat * M_{t-1}^{-1}$, which is subsequently concatenated with the features from the current time-step. This is passed to an Encoder network and its output is divided into two parts – the first 128 layers representing the memory, and the remaining 256 layers representing the fused features.

The fused features are further passed through a regressor module to estimate the poses, and the memory features are used for the next time-step. We perform the transformation based on the ground truth camera transformations. This is illustrated in Figure 2.

3.3. 6D Pose Regression

The translation vector T is the object location in the camera coordinate system. A naive way of estimating T is to directly regress to it. However, doing so cannot handle multiple object instances or generalise well to new objects. To tackle this problem, Xiang *et al.* [36] estimate T by localising the 2D object center in the image and estimating object distance from the camera. Suppose $c = (c_x, c_y)^T$ are the centers of the object in the frame and T_z is either learnt or estimated from the depth image, then T_x and T_y can be es-

	PoseCNN		PoseRBPF (50 particles)		VideoPose (GT BBox)		VideoPose (PoseCNN BBox)	
	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S
002_master_chef_can	50.9	84	56.1	75.6	55.5	85.0	<u>52.1</u>	84.3
003_cracker_box	51.7	76.9	73.4	85.2	10.9	63.3	6.9	58.6
004_sugar_box	68.6	84.3	73.9	86.5	47.1	71.9	41.2	68.6
005_tomato_soup_can	66	80.9	71.1	82.0	62.6	83.3	61.0	83.2
006_mustard_bottle	79.9	90.2	80.0	90.0	67.9	85.9	73.7	88.7
007_tuna_fish_can	70.4	87.9	56.1	73.8	56.1	83.3	53.1	<u>82.10</u>
008_pudding_box	62.9	79	54.8	69.2	56.7	76.6	48.9	<u>71.32</u>
009_gelatin_box	75.2	87.1	83.1	89.7	76	87.2	70.8	84.6
010_potted_meat_can	59.6	78.5	47.0	61.3	45.9	77.7	41.4	<u>75.6</u>
011_banana	72.3	85.9	22.8	64.2	40.6	70.7	<u>43.4</u>	<u>72.1</u>
019_pitcher_base	52.5	76.8	74.0	87.5	60	82.5	48.8	<u>77.30</u>
021_bleach_cleanser	50.5	71.9	51.6	66.7	41	59.9	29.5	50.4
024_bowl	6.5	69.7	26.4	88.2	1.5	73.2	1.6	67.1
025_mug	57.7	78	67.3	83.7	56.3	85.4	43.2	77.9
035_power_drill	55.1	75.8	64.4	80.6	26.4	63.9	16.5	61.9
036_wood_block	31.8	65.8	0.0	0.0	0.00	16.30	0.00	<u>13.6</u>
037_scissors	35.8	56.2	20.6	30.9	29.5	62.5	<u>27.9</u>	<u>72.1</u>
040_large_marker	58	71.4	45.7	54.1	21.6	55.6	20.5	<u>54.2</u>
051_large_clamp	25	49.9	27.0	73.3	14	61.5	16.5	<u>55.3</u>
052_extra_large_clamp	15.8	47	50.4	68.7	55.3	49.9	4.1	46.0
061_foam_brick	40.4	87.8	75.8	88.4	43.1	80.7	<u>40.9</u>	77.5
ALL	53.7	75.9	57.1	74.8	39.7	71.2	35.4	68.3

Table 1. Quantitative Evaluation of 6D poses (ADD and ADD-S) on the YCB-Video Dataset. **Bold** values compare between PoseCNN, PoseRBPF and VideoPose using the PoseCNN bbox. Values in **green** compares the columns with PoseRBPF, underlined values compare with PoseCNN

timated as:

$$\begin{bmatrix} c_x \\ c_y \end{bmatrix} = \begin{bmatrix} f_x \frac{T_x}{T_z} + p_x \\ f_y \frac{T_y}{T_z} + p_y \end{bmatrix}, \quad (1)$$

where f_x and f_y are focal lengths and $(p_x, p_y)^T$ are principal points. Since we have rough estimates of object locations from the noisy object detection inputs, we train our model to estimate Δc_x , Δc_y , and T_z . We then estimate T_x and T_y using the following equation:

$$\begin{bmatrix} c_x + \Delta c_x \\ c_y + \Delta c_y \end{bmatrix} = \begin{bmatrix} f_x \frac{T_x}{T_z} + p_x \\ f_y \frac{T_y}{T_z} + p_y \end{bmatrix}. \quad (2)$$

The fused features from the temporal module are then fed to 2 disconnected regressor blocks - a 2 layer FCN Regressor module, with 512 dimensions and $3 \times n$ where n is the number of objects for the translation and a 2 layer FCN regressor the temporal features are fed to a regressor with 512 dimensions and $4 \times$. We disconnect T_x , T_y and T_z by training two separate linear layers to account for the different dimensions learnt. Similar to [36], we represent the rotation R using quaternions.

3.4. Training Strategy

We use the L_1 loss to learn depth (L_{depth}), and cross entropy loss for semantic segmentation (L_{label}). The pose estimation loss is obtained by projecting the 3D points using the estimated and ground truth pose, and then computing their distance:

$$L_{\text{pose}}(\tilde{\mathbf{q}}, \mathbf{q}) = \frac{1}{m} \sum_{x \in M} \|(R(\tilde{\mathbf{q}})x + \tilde{\mathbf{t}}) - (R(\mathbf{q})x + \mathbf{t})\|^2, \quad (3)$$

where M denotes the set of 3D points, m is total number of points. $R(\tilde{\mathbf{q}})$ and $R(\mathbf{q})$ indicate the rotation matrix computed from the quaternion representation as in [36]. In addition, we also add a cosine loss on the quaternions, and regularisation loss to force the norm of the quaternion to be 1. Quaternions that represent rotations are unit norm, and forcing the norm to be bounded by 1 helps in the learning process by reducing the scope.

$$L_{\text{reg}} = \|1 - \text{norm}(\tilde{\mathbf{q}})\|, \quad L_{\text{inner.prod}} = 1 - \langle \tilde{\mathbf{q}}, \mathbf{q} \rangle. \quad (4)$$

The total loss can be defined as

$$L(\tilde{\mathbf{q}}, \mathbf{q}, \tilde{\mathbf{t}}, \mathbf{t}) = L_{\text{depth}} + L_{\text{label}} + L_{\text{pose}} + L_{\text{reg}} + L_{\text{inner.prod}}. \quad (5)$$

4. Experiments

Now we compare our method with PoseCNN[36] and PoseRBPF[5]. We also conduct ablation studies on the choice of model architecture, and the number of video frames the model requires to perform well.

4.1. Dataset

We evaluate the proposed method on the YCB-Video dataset [36]. Details of which are explained in sec 4.3. **The YCB-Video Dataset** contains 92 RGB-D video sequences of 21 objects. The dataset contains textured and textureless objects of varying shape, and different levels of occlusion where about 15% of objects are heavily occluded. Objects are annotated with 6D poses, segmentation masks and depth images. For our purposes, We create smaller video sequences of 10 RGB images by taking every alternate frame from the video.

4.2. Metrics

We use two metrics to report on the YCB-Dataset. ADD is the average distance between the corresponding points on the 3D object at the ground truth and predicted poses. Given the estimated $[\hat{\mathbf{R}}|\hat{\mathbf{t}}]$ and the ground truth poses $[\mathbf{R}|\mathbf{t}]$, ADD-S, designed for symmetric objects, calculates the mean distance from each 3D point to a closest point on the target model.

4.3. Implementation

VideoPose is implemented using the PyTorch [24] framework. We use a learning rate of $5e^{-4}$ and the Adam optimiser [14] with a weight decay of $1e^{-5}$. Learning rate is multiplied by 0.8 after every 5 epochs, until it hits a lower bound of $1e^{-6}$. For the feature encoder, we freeze the VGG16 weights from PoseCNN, and train rest of the network from scratch. We create video samples of 10 frames and train our model for 100 epochs with the learning schedule described above.

During training, we augment the input images with colour-jitter and noise, and for the bounding box, we augment it by extending the height and width randomly between 0 and 10% of the height and width of the object. While training the temporal block, we create videos with random time jumps in between. For instance, given a large video sequence, we create video samples $1 : n : 10 * n$, where n is a random number between 1 and 10, thus forcing the model to account for small and large jumps between consecutive frames.

4.4. Evaluation

We compare our results with PoseCNN [36] for single frame prediction and PoseRBPF [5] for videos in Table 1.

We compare two different approaches for computing the ROI: 1) Using the ground truth; 2) Using the ROI predicted by PoseCNN. In order to maintain comparable FPS, the PoseRBPF is computed using 50 particles.

We observe that, VideoPose, when using the bounding boxes from PoseCNN, has a small drop in the accuracy, showing that our method is robust to noises in the ROIs. For objects where our method is not comparable to the sota, we further look into the AUC curves in Fig. 3 and note that our method outperforms PoseCNN in rotation for symmetric objects like foam_brick and large_marker, and in translation for scissors and mustard_bottle which are non-symmetric.

	PoseCNN		VideoPose (ConvGRU)		VideoPose (baseline)	
	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S
002_master_chef_can	50.9	84	36.34	80.86	55.5	85.0
003_cracker_box	51.7	76.9	22.51	62.14	10.9	63.3
004_sugar_box	68.6	84.3	40.73	68.76	47.1	71.9
005_tomato_soup_can	66	80.9	66.99	83.49	62.6	83.3
006_mustard_bottle	79.9	90.2	75.33	88.96	67.9	85.9
007_tuna_fish_can	70.4	87.9	60.58	86.35	56.1	83.3
008_pudding_box	62.9	79	49.56	75.77	56.7	76.6
009_gelatin_box	75.2	87.1	81.06	89.32	76	87.2
010_potted_meat_can	59.6	78.5	61.54	83.64	45.9	77.7
011_banana	72.3	85.9	22.31	69.59	40.6	70.7
019_pitcher_base	52.5	76.8	70.54	85.92	60	82.5
021_bleach_cleanser	50.5	71.9	46.93	62.11	41	59.9
024_bowl	6.5	69.7	12.78	80.08	1.5	73.2
025_mug	57.7	78	67.62	88.79	56.3	85.4
035_power_drill	55.1	75.8	35.99	71.21	26.4	63.9
036_wood_block	31.8	65.8	0.00	28.72	0.00	16.30
037_scissors	35.8	56.2	50.08	73.39	29.5	62.5
040_large_marker	58	71.4	36.81	53.89	21.6	55.6
051_large_clamp	25	49.9	20.72	69.33	14	61.5
052_extra_large_clamp	15.8	47	5.93	55.39	5.3	49.9
061_foam_brick	40.4	87.8	44.78	86.14	43.1	80.7
ALL	53.7	75.9	44.09	73.90	39.7	71.2

Table 2. Comparison of performance between different architectures. **Bold** values represent the best method for a given object.

Impact of different temporal blocks We also perform ablation studies on different architectures used to capture the temporality, as shown in Table 2. Instead of the baseline temporal RNN in fig. 2, we use a ConvGRU [1] as the temporal module and observe that it handles the temporal

Methods	[36]	[5] (50)	[5] (200)	Ours(RNN)	Ours(ConvGRU)
Time (fps)	5.88	20	5	30	25

Table 3. Comparison of frame rates for different methods: PoseCNN [36], PoseRBPF (50 particles) [5], PoseRBPF (200 particles) [5], VideoPose (baseline) and VideoPose(ConvGRU)

Methods	Position=2	Position=5	Position=10	Position=15	Position=19
ADD	41.6	41.61	44.08	41.3457	40.79
ADD-S	71.77	71.83	73.9	71.71	71.18

Table 4. Studying the number of previous frames required for a good estimate. Position refers to the location of the keyframe in a video sequence of 20 frames.

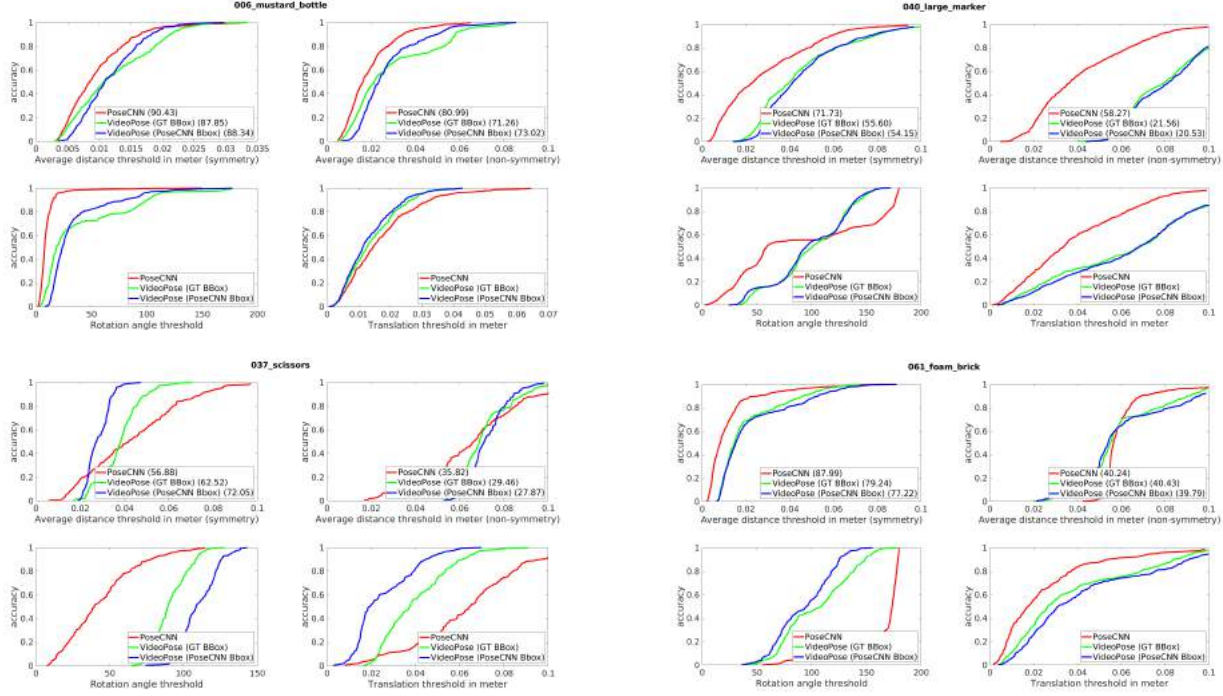


Figure 3. AUC with rotation and translation curves for few objects

information more effectively. The performance is comparable to or better than PoseCNN and PoseRBPF for more than 50% of the objects. We treat this ablation study as a proof that using previous estimates can aid in the pose estimation, regardless of the temporal module used.

Time efficiency We compare the time efficiency of our model with the baseline models. The run time for PoseCNN is taken from Wang *et al.* [33]. From Table 3, we see that VideoPose with ConvGRU, is slightly faster than PoseRBPF while providing an increase in the performance. As the baseline temporal RNN is more lightweight than ConvGRUs, we get about 50% increase in the speed compared to PoseRBPF, while maintaining the accuracy.

Qualitative Analysis of the 6D predictions We show three examples of the predictions by VideoPose, PoseCNN, and ground truth poses in Table 5 6. The columns represent the 2D projections of predictions using VideoPose, PoseCNN, and the ground truth poses, and rows specify the time-steps. More results are shown in the supplementary. We notice that the poses between frames are more consistent when estimated through videos as opposed to single images. We also observe that the initial frame estimation is as critical for our approach, as it is for other refinement methods.

Effect of number of previous frames used Table 4 shows the effect of number of previous frames used. We see that we get the best performance at position 10, and it reduces

a little for later positions. It is worth noting that the model was trained for a video sequence of length 10. So the little performance drop for 15th and 19th position shows that our model, even when trained for a video sequence of 10, can effectively model longer sequences. This is due to the ablation of the video samples during training as discussed in 4.3.

5. Conclusion

In this work, we introduce VideoPose, a simple convolutional neural network architecture to estimate object 6D poses from videos. We demonstrate that by using the 6D predictions from the previous frames, we can significantly improve 6D predictions in the subsequent frames. We also conduct an extensive ablation study on different design choices of the network, and show that our model is able to learn and utilise the features from previous predictions regardless of the network choices. Finally, the proposed network performs in real-time at 30fps, thereby improving the time efficiency over previous approaches. As a future work, we would like to further improve our architecture with a better temporal module and model the relationship with the camera transformation and the objects. Our method successfully maintains consistency in pose estimation between frames, however, still depends on the initial frame estimation. We would like to investigate further on improving this, while maintaining the computational efficiency.










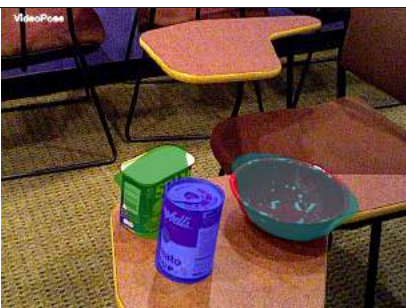





	VideoPose	PoseCNN	GT
t = 1			
t = 2			
t = 3			
t = 4			
t = 5			

Table 5. Visualisations of the estimated poses on YCB-Dataset for 3 different. Each row represents results at different time-steps. The columns are VideoPose, PoseCNN and Ground truth visualisations respectively.
















	VideoPose	PoseCNN	GT
t = 1			
t = 2			
t = 3			
t = 4			
t = 5			

Table 6. Visualisations of the estimated poses on YCB-Dataset. Each row represents results at different time-steps. The columns are VideoPose, PoseCNN and Ground truth visualisations respectively.

References

- [1] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*, 2015.
- [2] Gideon Billings and Matthew Johnson-Roberson. Silhonet: An rgb method for 6d object pose estimation. *IEEE Robotics and Automation Letters*, 4(4):3727–3734, 2019.
- [3] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6D object pose estimation using 3d object coordinates. In *European conference on computer vision*, pages 536–551. Springer, 2014.
- [4] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3D object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1907–1915, 2017.
- [5] Xinke Deng, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl, and Dieter Fox. Poserbpf: A rao-blackwellized particle filter for 6d object pose tracking. *Robotics: Science and Systems (RSS)*, 2019.
- [6] Clemens Eppner, Arsalan Mousavian, and Dieter Fox. A billion ways to grasp: An evaluation of grasp sampling schemes on a dense, physics-based grasp data set. *arXiv preprint arXiv:1912.05604*, 2019.
- [7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [8] Stefan Hinterstoisser, Cedric Cagniard, Slobodan Ilic, Peter Sturm, Nassir Navab, Pascal Fua, and Vincent Lepetit. Gradient response maps for real-time detection of textureless objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(5):876–888, 2011.
- [9] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 548–562. Springer, 2012.
- [10] Tomás Hodan, Pavel Haluza, Stepán Obdržálek, Jiri Matas, Manolis I. A. Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888, 2017.
- [11] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [12] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Segmentation-driven 6d object pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3385–3394, 2019.
- [13] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 1530–1538, 2017.
- [14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [15] Yoshinori Konishi, Kosuke Hattori, and Manabu Hashimoto. Real-time 6D object pose estimation on CPU. *arXiv preprint arXiv:1811.08588*, 2018.
- [16] Alexander Krull, Eric Brachmann, Frank Michel, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Learning analysis-by-synthesis for 6d pose estimation in rgb-d images. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 954–962, 2015.
- [17] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *European Conference on Computer Vision*, pages 574–591. Springer, 2020.
- [18] Chi Li, Jin Bai, and Gregory D Hager. A unified framework for multi-view multi-class object pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 254–269, 2018.
- [19] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. DeepIm: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 683–698, 2018.
- [20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [21] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12):2633–2651, 2015.
- [22] Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Making deep heatmaps robust to partial occlusions for 3d object pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–134, 2018.
- [23] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7668–7677, 2019.
- [24] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.
- [25] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4561–4570, 2019.
- [26] Mahdi Rad and Vincent Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 3848–3856, 2017.

- [27] Ashutosh Saxena, Justin Driemeyer, and Andrew Y Ng. Robotic grasping of novel objects using vision. *The International Journal of Robotics Research (IJRR)*, 27(2):157–173, 2008.
- [28] Chen Song, Jiaru Song, and Qixing Huang. Hybridpose: 6d object pose estimation under hybrid representations, 2020.
- [29] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 292–301, 2018.
- [30] Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2038–2041, 2018.
- [31] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stanley T. Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. In *Conference on Robot Learning (CoRL)*, 2018.
- [32] Shubham Tulsiani and Jitendra Malik. Viewpoints and key-points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1510–1519, 2015.
- [33] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [34] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019.
- [35] Bowen Wen, Chaitanya Mitash, Baozhang Ren, and Kostas E Bekris. se (3)-tracknet: Data-driven 6d pose tracking by calibrating image residuals in synthetic domains. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10367–10373. IEEE, 2020.
- [36] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [37] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1941–1950, 2019.
















	VideoPose	PoseCNN	GT
t = 1			
t = 2			
t = 3			
t = 4			
t = 5			

Table 7. Visualisations of the estimated poses on YCB-Dataset. Each row represents results at different time-steps. The columns are VideoPose, PoseCNN and Ground truth visualisations respectively.

	VideoPose	PoseCNN	GT
t = 1			
t = 2			
t = 3			
t = 4			
t = 5			

Table 8. Visualisations of the estimated poses on YCB-Dataset. Each row represents results at different time-steps. The columns are VideoPose, PoseCNN and Ground truth visualisations respectively.

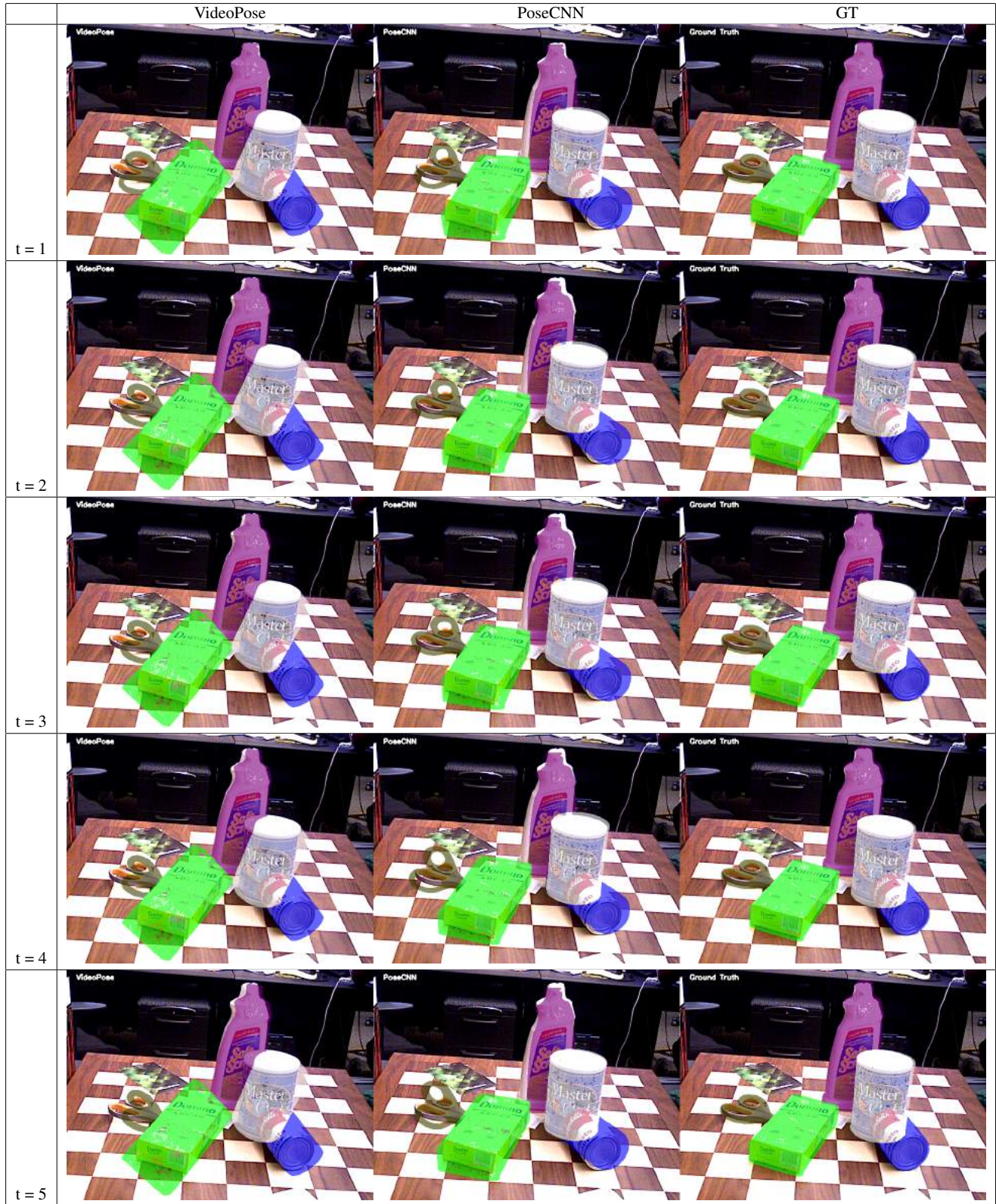


Table 9. Visualisations of the estimated poses on YCB-Dataset. Each row represents results at different time-steps. The columns are VideoPose, PoseCNN and Ground truth visualisations respectively.