CVAM-Pose: Conditional Variational Autoencoder for Multi-Object Monocular Pose Estimation

Jianyu Zhao
jzhao12@uclan.ac.uk
Wei Quan
wquan@uclan.ac.uk
Bogdan J. Matuszewski
bmatuszewski1@uclan.ac.uk

Computer Vision and Machine Learning (CVML) Group,
The University of Central Lancashire,
Preston, UK

Abstract

Estimating rigid objects' poses is one of the fundamental problems in computer vision, with a range of applications across automation and augmented reality. Most existing approaches adopt one network per object class strategy, depend heavily on objects' 3D models, depth data, and employ a time-consuming iterative refinement, which could be impractical for some applications. This paper presents a novel approach, CVAM-Pose, for multi-object monocular pose estimation that addresses these limitations. The CVAM-Pose method employs a label-embedded conditional variational autoencoder network, to implicitly abstract regularised representations of multiple objects in a single low-dimensional latent space. This autoencoding process uses only images captured by a projective camera and is robust to objects' occlusion and scene clutter. The classes of objects are one-hot encoded and embedded throughout the network. The proposed label-embedded pose regression strategy interprets the learnt latent space representations utilising continuous pose representations. Ablation tests and systematic evaluations demonstrate the scalability and efficiency of the CVAM-Pose method for multi-object scenarios. The proposed CVAM-Pose outperforms competing latent space approaches. For example, it is respectively 25% and 20% better than AAE and Multi-Path methods, when evaluated using the AR_{VSD} metric on the Linemod-Occluded dataset. It also achieves results somewhat comparable to methods reliant on 3D models reported in BOP challenges. Code available: https://github.com/JZhao12/CVAM-Pose

1 Introduction

The rapid and precise estimation of rigid objects' poses with six degrees of freedom (6-DoF) is crucial for a wide range of real-world applications, including explorative navigation, augmented reality, and automated medical intervention. The introduction of deep learning techniques [10, 43] marked a significant evolution in computer vision, yielding remarkable outcomes in 6-DoF pose estimation. Notably, most deep learning-based methods [6, 7, 22, 23, 24, 25, 30, 31, 33, 39, 40, 41, 44] tend to train individual networks for each object to

obtain higher pose accuracy. However, these approaches are resource-consuming compared to training a unified multi-object network, as memory usage increases with the number of objects (networks). Additionally, most methods typically require 3D models [4, 20, 22, 40, 41] or establish 2D-3D correspondences based on the models [15, 23, 24, 25, 26, 30, 31, 36, 42], where the need for 3D models can be seen as one of the limiting factors for broader applications.

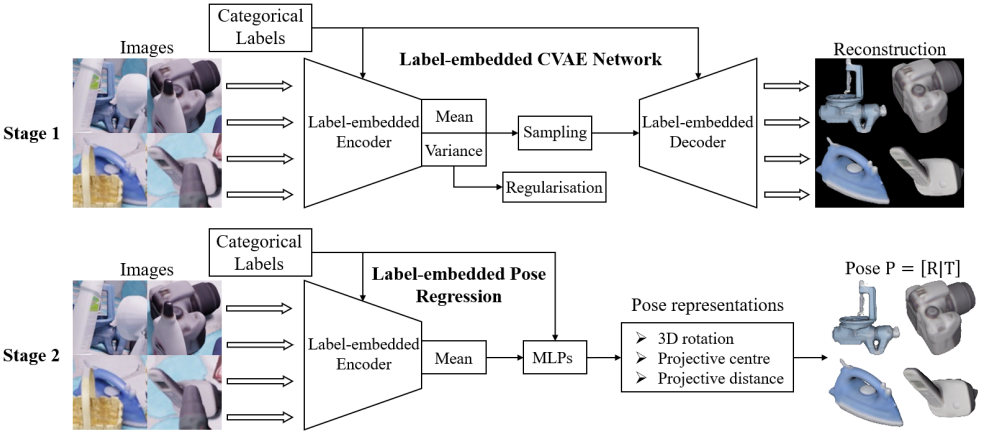

Figure 1: During training, the label-embedded CVAE network abstracts information from both images of objects and the corresponding categorical labels in the latent space, which are then interpolated to multi-object 6-DoF poses using MLPs. The images of objects are taken from the Linemod PBR dataset [12, 14, 16, 17].

In this paper, a novel multi-object pose estimation method called CVAM-Pose is proposed (Fig. 1), which contains two main stages. The first stage involves training a label-embedded conditional variational autoencoder (CVAE) network that incorporates the one-hot encoding technique to facilitate the learning of regularised and constrained representations of multi-object poses in the latent space. Different from the original CVAE network proposed in [29], the adapted layer-wise one-hot encoding technique encodes categorical labels as complete feature maps across every layer within the network, enhancing the learning of high-level representations. The second stage applies label-embedded pose regression that avoids the discretisation of poses. This involves concatenating the learnt multi-object representations with one-hot encoded label vectors, and training multilayer perceptrons (MLPs) to regress these concatenated features into continuous pose representations. The contributions of the CVAM-Pose method are summarised as follows:

1. The method enhances the scalability and efficiency for multi-object pose estimation using a single CVAE network. To the best of our knowledge, it is the first time a conditional generative model is employed to efficiently characterise multi-object poses. The adapted label-embedding technique also improves the capability of learning high-level representations.

2. The method does not require object 3D models, depth data, and post-refinement during inference, which can facilitate real-time processing. It achieves results comparable to the state-of-the-art approaches on the Linemod-Occluded benchmark dataset and outperforms those based on latent space representation.

3. The method effectively addresses various challenging scenarios, including texture-less objects, occlusion, truncation [25], and clutter.

2 Related Work

Deep Learning-based Approaches With the rapid development of deep learning techniques, numerous state-of-the-art pose estimation methods employing convolutional neural networks (CNNs) have been proposed. These methods can be categorised into three distinct groups based on their approach to utilising CNNs: direct, indirect, and latent representation methods.

The direct methods train CNNs to regress 3D rotation and translation from images directly. which either construct loss functions using 3D model points [4, 40, 41], or iteratively match the image rendered from a 3D model at its estimated pose with the observed input image [20, 22]. Typically, these methods reparameterise rotation into representations more suitable for network training, such as unit quaternion [5], axis-angle [37], or continuous representation [45]. The indirect methods focus on learning 2D-3D model correspondences via CNNs, with the 6-DoF poses subsequently estimated using PnP [9, 21] and RANSAC [8]. The model correspondences can be in the form of pixel-wise dense mapping [11, 15, 23, 24, 30, 31, 42], or a selection of sparse keypoints [25, 26, 36]. The latent representation methods learn implicit latent space representations using specific network architectures, typically autoencoders. The pose of a test instance is often retrieved using a lookup table (LUT) technique, which includes finding nearest neighbours [32, 33] and computing observation likelihoods [6, 7].

Both direct and indirect methods explicitly require accurate 3D models for training CNNs or establishing 2D-3D correspondences. The latent representation methods, despite using only images from single perspective camera, often suffer from the pose discretisation problem due to the nature of LUT.

Conditional Variational Autoencoder The variational autoencoder (VAE) [18, 19] was introduced in the context of generative models, which is different from typical autoencoder models [13, 27, 28, 38]. The primary objective of the VAE is to generate new, typically highly dimensional data points, with the generation process controlled by a low-dimensional latent code randomly drawn from a prior distribution, such as Gaussian. However, a notable limitation is its inability to specify the characteristics of the generated data. To address this issue, Sohn et al. [29] introduced the conditional variational autoencoder (CVAE), which extends the VAE framework to incorporate conditional parameters, thereby enabling the generation of data with desired attributes.

In the context of object 6-DoF pose estimation, Zhao et al. [44] proposed a VAE-based method called CVML-Pose, which was restricted to single-object predictions. Similar methods, such as [35], have also been developed, but using RGB-D images as input. We extend the CVML-Pose method by training a CVAE model with a layer-wise one-hot encoding technique. This adaptation facilitates the learning of multi-object representations in a single latent space, significantly improving scalability and efficiency in the prediction of multi-object poses.

3 Methodology

3.1 Implicit Learning of Multi-Object Representations

To effectively learn multi-object representations, a label-embedded CVAE network is trained to encode images of objects x_i and their corresponding one-hot encoded categories y_i as label conditions in a regularised latent space, subsequently outputting clean reconstructions x_i'.

As depicted in Fig. 2, an asymmetric architecture is proposed, consisting of an encoder network E_ϕ and a decoder network D_θ, with learnable parameters ϕ and θ respectively. The encoder $E_\phi(x_i, y_i)$ processes both x_i and y_i, where y_i are embedded as complete feature maps in every convolution block (block-wise), until the latent variables are obtained in the latent space, including $\mu_\phi(x_i, y_i) \in \mathbb{R}^n$ and $\log(\sigma^2_\phi(x_i, y_i)) \in \mathbb{R}^n$, where $(\mu_\phi, \sigma^2_\phi)$ represent mean and variance vectors of the multivariate normal distribution. Due to the non-differentiability of sampling from $\mathcal{N}(z_i; \mu_\phi(x_i, y_i), diag(\sigma_\phi(x_i, y_i)))$, a reparameterisation trick [18] is employed. The latent sampling $z_i \in \mathbb{R}^n$ is reparameterised as $\mu_\phi(x_i, y_i) + diag(\sigma_\phi(x_i, y_i)) \cdot \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, I)$. After sampling, the decoder network $D_\theta(z_i, y_i)$ reconstructs the complete and clean view x_i' from both z_i and y_i, where y_i is also embedded in every convolution layer (layer-wise) in the decoder.

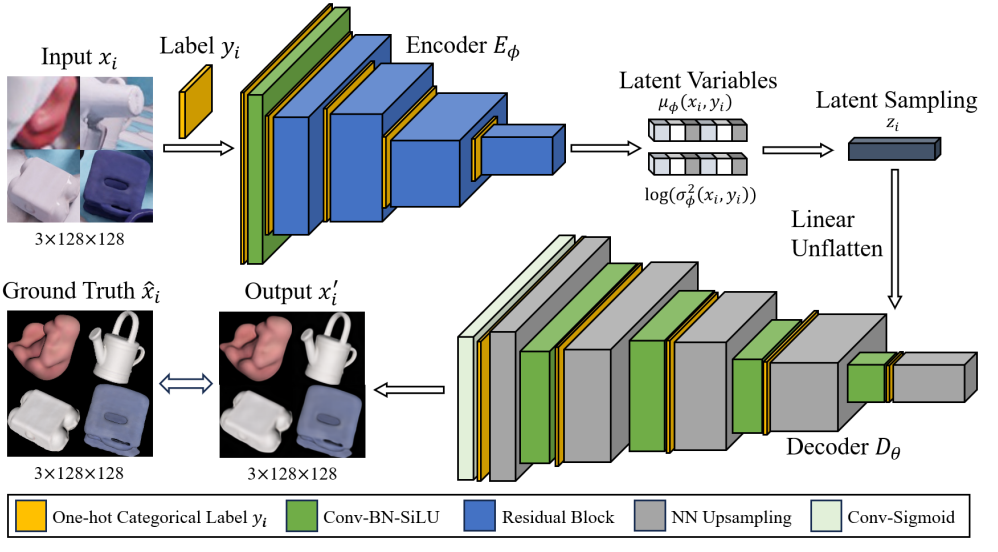

Figure 2: The proposed label-embedded conditional variational autoencoder network. The images of objects are taken from the Linemod PBR dataset [12, 14, 16, 17].

For network training, the evidence lower bound ($ELBO$) loss [44] is used, assuming a Gaussian prior distribution $p(z) = \mathcal{N}(z; 0, I)$. This loss comprises two components: i) a pixel-wise squared L2 norm between the output image x_i' and the ground truth reconstruction image $\hat{x}_i$; ii) a Kullback-Leibler (KL) divergence loss with a scalar α, which controls the regularisation of the latent space.

$$ELBO \simeq -\sum_{i=1}^{m} \left(||\hat{x}_i - x_i'||^2 - \alpha \cdot \sum_{j=1}^{n} \left(1 + \log(\sigma_{ij}^2) - \mu_{ij}^2 - \sigma_{ij}^2\right) \right) \tag{1}$$

where μ_{ij} refers to the j-th element of the vector μ_i, σ_{ij}^2 refers to the j-th element of the vector σ_i^2, $\mu_i = \mu_\phi(x_i, y_i)$, $\sigma_i^2 = \sigma_\phi^2(x_i, y_i)$, m represents the number of training data, and n is the dimensionality of the latent space.

After training, the label-embedded CVAE network has learnt robust representations of objects, possibly including poses. This can be evidenced by the clean and complete reconstructions produced from the trained network based on the test data. As illustrated in Fig. 3, these reconstructions not only preserve complete views of objects but also diminish irreverent information such as occlusion and cluttered background.

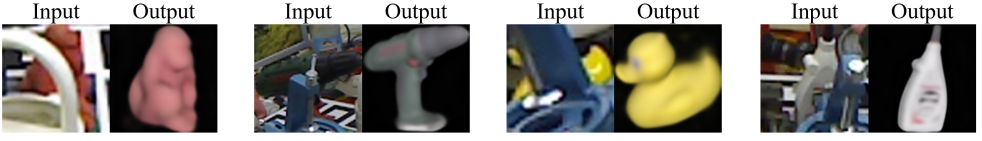

Figure 3: The output images from decoder show objects' representations with occlusion and clutter removed. The test input images also shown are taken from the Linemod-Occluded dataset [2, 3].

3.2 Continuous Regression for Multi-Object Pose Estimation

The subsequent stage employs a continuous pose regression strategy [44], with an adaptation to handle the multi-object scenario. As detailed in Fig. 4, this strategy utilises one-hot encoded object labels y_i to train MLPs, enabling smooth interpolation of poses.

For estimating 3D rotation, the rotation MLP is trained to regress from (μ_i, y_i) to the continuous 6D rotation representation $r \in \mathbb{R}^6$ [45], and the output rotation $\mathbf{R} \in \mathrm{SO}(3)$ is derived from r through a process similar to Gram-Schmidt orthogonalisation. This continuous representation has proven more effective than others such as unit quaternion and axis-angle, and has been successfully implemented in [20, 39].

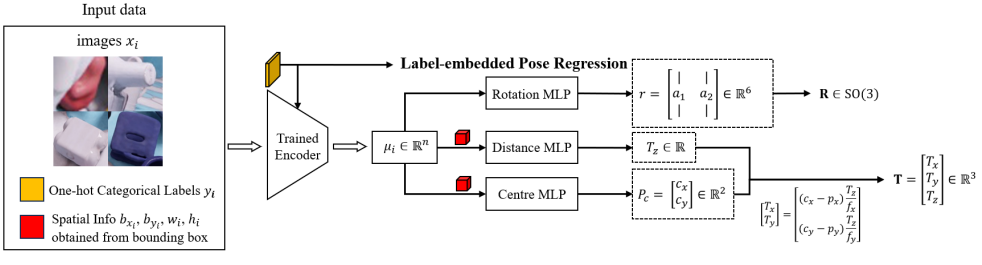

Figure 4: The proposed label-embedded pose regression approach interpolates multi-object representations to continuous pose representations using multiple MLP heads.

The estimation of 3D translation $\mathbf{T} = \begin{pmatrix} T_x & T_y & T_z \end{pmatrix}^T \in \mathbb{R}^3$ is disentangled into estimating the 2D projective centre coordinates $P_c = (c_x, c_y)^T \in \mathbb{R}^2$ and the projective distance $T_z \in \mathbb{R}$. Specifically, the latent vector μ_i is concatenated with spatial information obtained from the object's bounding box, including its width w_i, height h_i, and top-left corner coordinates $P_{bbox} = (b_{x_i}, b_{y_i})^T$, as well the label y_i. The dedicated MLP is then trained to regress these concatenated features to $(c_x, c_y)^T$. The regression procedure is also applied to T_z by training

the distance MLP, but only utilises μ_i, w_i, h_i, and y_i. Once T_z and $(c_x, c_y)^T$ are determined, T_x and T_y are calculated using the pinhole camera model (Eq. 2).

$$\begin{bmatrix} T_x \\ T_y \end{bmatrix} = \begin{bmatrix} (c_x - p_x)\frac{T_z}{f_x} \\ (c_y - p_y)\frac{T_z}{f_y} \end{bmatrix} \tag{2}$$

where f_x and f_y denote the focal lengths, $(p_x, p_y)^T$ is the principal point, and all these parameters can be obtained from camera calibration.

4 Experiments

The CVAM-Pose method is benchmarked in two aspects. The first involves conducting a series of ablation tests to determine favourable configurations of the method. The second way is to follow the evaluation methodologies proposed in the BOP challenges [16, 34].

4.1 Experimental Setup

Data All the experiments are conducted using the Linemod-Occluded dataset [2, 8], as it presents a wide range of challenging scenarios, such as texture-less objects with significant occlusion and background clutter. To facilitate a fair comparison with methods participating in the BOP challenges, the same training and test data are employed. The physically based rendering (PBR) images [12, 14, 16, 17] are used for training, and the BOP version test set is chosen for evaluation.

Evaluation pipeline All the results, including those from the ablation tests and the main evaluation, are reported using the metrics specified in the BOP challenges: VSD, MSSD, and MSPD [16]. The overall performance score, AR, is calculated based on the average recall of these three metrics, defined as $\text{AR} = (\text{AR}_{\text{VSD}} + \text{AR}_{\text{MSSD}} + \text{AR}_{\text{MSPD}})/3$.

4.2 Ablation Study

To obtain effective configurations of the CVAM-Pose method, extensive ablation tests are conducted using the BOP version of the Linemod-Occluded dataset [2, 8, 16, 17]. These tests include evaluations of the adapted label embedding technique, the regularisation of the label-embedded CVAE network, and the dimensionality of the latent space.

Label Embedding Technique The effectiveness of the adapted layer-wise one-hot encoding technique is assessed in both the CVAE network and the MLPs. The original CVAE network [29], where the label conditions only exist in the initial layer in both the encoder and decoder, is trained under the same conditions as our proposed label-embedded network. The tests on MLPs involve determining whether the representations learnt from the label-embedded CVAE network can be effectively regressed without the labels.

The results presented in Table 1 demonstrate that the adapted label-embedding technique enhances the ability to learn and regress pose representations. The proposed label-embedded CVAE network yields the most promising results compared to the original CVAE and MLPs, showing improvements of 10% and 13% respectively in AR. This improvement is attributed to its capability to abstract high-level features related to object pose within the latent space.

Network	original CVAE	MLPs without labels	**Ours**
AR_{VSD}	0.256	0.251	0.346
AR_{MSSD}	0.255	0.243	0.362
AR_{MSPD}	0.630	0.554	0.714
AR	0.380	0.349	0.475

Table 1: Ablation results on the adapted label embedding technique.

In contrast, the original CVAE network primarily incorporates low-level features, which is less effective for learning distinct multi-object representations because, with the increasing depth of the network, the conditioned features introduced at the initial layers may not be evident in the latent space. Similarly, the MLPs benefit from the label-embedding technique that helps regress distinct poses for each object. However, it is observed that even in the absence of label conditions, the learnt representations still retain certain categorical information, leading to reasonable results.

Latent Space Regularisation When training the CVAM-Pose, the weighting factor α of the KL regularisation term plays a crucial role in determining the smoothness of the latent space. To find a good balance that allows for both informative latent space and robust generalisation, the proposed CVAE network is trained with different values of $\alpha \in [0, 0.1, 0.5, 1]$. The pose estimation results corresponding to these values are reported in Table 2.

Regularisation	$\alpha = 0$	$\boldsymbol{\alpha = 0.1}$	$\alpha = 0.5$	$\alpha = 1$
AR_{VSD}	0.316	0.346	0.319	0.336
AR_{MSSD}	0.340	0.362	0.352	0.353
AR_{MSPD}	0.712	0.714	0.686	0.685
AR	0.456	0.475	0.452	0.458

Table 2: Ablation results on the regularisation of the latent space.

Based on the reported results, an α value of 0.1 is identified as most effective, providing sufficient regularisation of the latent space without overly constraining it. Specifically, when $\alpha = 0$, the CVAE network lacks control over the assumed prior Gaussian distribution in the latent space, leading to unrestricted μ, and σ^2 approaching 0, which results in suboptimal performance in comparison to $\alpha = 0.1$. Conversely, when α is set to 0 or 1, the latent space is greatly affected by the KL divergence, which seems to overly smooth the distribution. This excessive smoothing may cause a loss of critical pose-related information, as the network focuses on minimising KL divergence over retaining distinctive features of the input data.

Dimensionality of the Latent Space The dimensionality of the latent space, denoted as n, determines the capacity of the proposed CVAE network to encapsulate information about objects. Previous research, such as that by Sundermeyer et al. [33], has explored the effect of latent space dimensionality on pose estimation; however, their ablation tests were limited to $n \leq 128$ and focused solely on single-object scenarios. This limitation prompts further investigation into the performance impact of $n > 128$ across a broader range of objects, as a single object may not be sufficient to reflect the complexities of an entire dataset.

To determine an effective size of the latent space for multi-object pose estimation, comprehensive experiments are conducted using all the Linemod-Occluded objects with dimensionalities set at $n \in [32, 64, 128, 256, 512, 1024]$. Results, as detailed in Table 3, show that the highest accuracy is observed at $n = 256$. However, notably, the accuracy decreases at

Dimensionality	$n = 32$	$n = 64$	$n = 128$	$\boldsymbol{n = 256}$	$n = 512$	$n = 1024$
AR_{VSD}	0.226	0.301	0.283	0.346	0.306	0.263
AR_{MSSD}	0.224	0.318	0.303	0.362	0.317	0.284
AR_{MSPD}	0.528	0.654	0.681	0.714	0.706	0.695
AR	0.326	0.424	0.422	0.475	0.443	0.414

Table 3: Ablation results on the dimensionality of the latent space.

higher dimensionalities such as 512 and 1024, suggesting that an overly large latent space may not effectively contribute to pose encoding, and could potentially lead to diminished performance due to the CVAE network capturing too much variability in the latent space, overfitting to the specific training set.

4.3 Main Results and Discussion

Main Results The main results of the proposed CVAM-Pose method are reported in Table 4 and 5, where it is compared against a variety of state-of-the-art methods on the BOP version of the Linemod-Occluded dataset. These methods are categorised based on the criteria outlined in Sec. 2, with the CVAM-Pose method classified within the latent representation category. Symbol "*" next to a method indicates that it employs one network per object class strategy. Additionally, we produce the box plots in Fig. 5, which access how the visibility of objects (occlusion) in the scene images influences the pose estimation accuracy.

Method	**Ours**	CVML-Pose* [14]	AAE* [33]	AAE-ICP* [33]	Multi-Path [33]
AR_{VSD}	0.346	0.312	0.090	0.208	0.150
AR_{MSSD}	0.362	0.338	0.095	0.218	0.153
AR_{MSPD}	0.714	0.706	0.254	0.285	0.346
AR	0.475	0.452	0.146	0.237	0.217

Table 4: Comparison with latent representation methods.

Method	**Ours**	DPOD [42]	Pix2Pose* [20]	EPOS [11]	CDPN* [21]	PVNet* [25]	CosyPose [22]	SurfEmb [8]	ZebraPose* [31]	GDR-Net* [34]
AR_{VSD}	0.346	0.101	0.233	0.389	0.393	0.428	0.480	0.497	0.547	0.549
AR_{MSSD}	0.362	0.126	0.307	0.501	0.537	0.543	0.606	0.640	0.714	0.701
AR_{MSPD}	0.714	0.278	0.550	0.750	0.779	0.754	0.812	0.851	0.860	0.887
AR	0.475	0.169	0.363	0.547	0.569	0.575	0.633	0.663	0.707	0.713

Table 5: Comparison with direct and indirect methods that are reliant on 3D models.

Within the latent representation category, CVAM-Pose significantly outperforms methods such as AAE, AAE-ICP (AAE incorporates ICP refinement [1]), and Multi-Path (a multi-object AAE approach). For example, it is respectively 25%, 14%, and 20% better when evaluated using the AR_{VSD} metric. The overall performance margins across the three metrics are substantial, with improvements of 33%, 24%, and 26% in AR respectively. In comparison with methods from other categories, CVAM-Pose also surpasses some indirect methods like DPOD and Pix2Pose, by margins of 31% and 11%, respectively. Additionally, it achieves results comparable to EPOS, CDPN, and PVNet in specific metrics, e.g. AR_{VSD} and AR_{MSPD}.

Discussion The results on the challenging Linemod-Occluded dataset demonstrate that competitive performance for multi-object pose estimation can be achieved using a conditional generative model. The proposed label-embedded CVAE network with the continuous pose regression approach is more effective and accurate than other latent representation

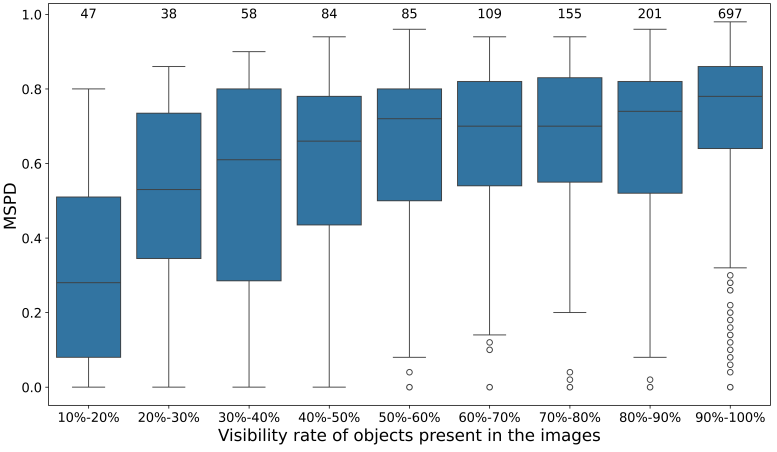

Figure 5: Box plots of the MSPD metric as a function of the objects' visibility rates. The number of data instances for each rate is shown above each box. Please note that for better visualisation, the MSPD metric is calculated using thresholds ranging from 1 to 50 with a step of 1, instead of using the thresholds (from 5 to 50 with a step of 5) defined in the BOP challenges.

methods. Compared to the single-object methods, such as CVML-Pose, AAE, and AAE-ICP, the proposed label-embedded CVAE network captures regularised and robust representations for multiple objects. Unlike CVML-Pose, the adapted label-embedding technique avoids training multiple VAE networks, thereby enhancing training efficiency without diminishing the performance. Different from the Multi-Path method, which employs multiple decoder networks demanding significant GPU resources, our single-encoder-single-decoder architecture achieves higher pose accuracy. Furthermore, the continuous pose regression approach, as opposed to the LUT technique used in AAE, AAE-ICP, and Multi-Path, effectively avoids errors associated with pose discretisation during inference. Our continuous regression on the 2D projective centre and distance also mitigates the effects caused by incorrect detection of occluded objects.

Performance against occlusion is illustrated in Fig. 5, which shows box plots quantifying the distribution of the MSPD metric (AR_{MSPD}) across different visibility rates from 10% to 100%. It is evident that as visibility increases, the median of pose estimation accuracy improves and eventually achieves a value of 0.78. Even under heavy (20%-30% visibility) or mild (50%-60% visibility) occlusions, our method still achieves reasonable results, indicating its robustness against challenging occlusion scenarios.

Compared to direct and indirect methods that rely on 3D models, the proposed CVAM-Pose method achieves higher results than some of them on the challenging occlusion data. This possibly suggests that approaches based on pixel-wise model correspondences, such as DPOD and Pix2Pose, suffer performance degradation due to an insufficient number of points available for assessing correspondences in heavily occluded scenes. In contrast, our proposed method benefits from reconstructing complete objects from partially obscured views, thereby robustly handling occlusions.

For the results reported in the paper, the proposed CVAM-Pose method is trained with 8 different objects. Experiments with larger numbers of objects were also conducted but not

reported. It was observed that increasing the number of objects in CVAM-Pose beyond 15 would lead to a decrease in pose estimation accuracy, without adjusting design parameters such as the size of the latent space.

Although the proposed CVAM-Pose method avoids training one network per object category, which enhances scalability and efficiency, it shows a certain gap in pose accuracy compared to leading methods like CosyPose, SurfEmb, and GDR-Net. These methods improve their accuracy through techniques such as iterative refinement using 3D model points (CosyPose), estimating continuous model correspondence distributions (SurfEmb), and combining pose regression with dense correspondences (GDR-Net). However, they all require precise 3D models for setting up 2D-3D correspondences or model point-based training. In contrast, the advantages of our method lie in addressing the 6-DoF pose estimation problem without relying on 3D models, depth measurements, and post-refinement processes, providing a novel solution in scenarios where such data are unavailable.

5 Conclusion

This paper addresses one of the key challenges in computer vision: finding multi-object 6-DoF poses from images captured by a perspective camera in real time (with fixed inference processing time of 0.02s with CVAM-Pose run on RTX3090). The proposed method demonstrates that competitive performance can be achieved using only a single perspective image, without reliance on 3D models, depth measurements, or iterative post-refinement. In particular, the scalability of a single latent space can be expanded to multi-object representations without compromising pose accuracy. The main contributions of the reported research include the proposed use of a conditional generative model, the adapted label-embedding technique, the construction of a regularised and constrained latent space for multiple objects, and the continuous pose regression algorithms, which facilitate fast and accurate multi-object pose estimation.

6 Acknowledgement

Data access statement: The study reported in this paper has been supported by two existing openly available datasets, namely Linemod and Linemod-Occluded. Both datasets are available from https://bop.felk.cvut.cz/datasets/.

References

[1] PJ Besl and Neil D McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 14(02):239–256, 1992.

[2] Eric Brachmann. 6d object pose estimation using 3d object coordinates [data], 2020. URL https://doi.org/10.11588/data/V4MUMX.

[3] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 536–551. Springer, 2014.

[4] Yannick Bukschat and Marcus Vetter. Efficientpose: An efficient, accurate and scalable end-to-end 6d multi object pose estimation approach. *arXiv preprint arXiv:2011.04307*, 2020.

[5] Erik B Dam, Martin Koch, and Martin Lillholm. *Quaternions, interpolation and animation*, volume 2. Citeseer, 1998.

[6] Xinke Deng, Yu Xiang, Arsalan Mousavian, Clemens Eppner, Timothy Bretl, and Dieter Fox. Self-supervised 6d object pose estimation for robot manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3665–3671. IEEE, 2020.

[7] Xinke Deng, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl, and Dieter Fox. Poserbpf: A rao–blackwellized particle filter for 6-d object pose tracking. *IEEE Transactions on Robotics*, 37(5):1328–1342, 2021.

[8] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[9] Xiao-Shan Gao, Xiao-Rong Hou, Jianliang Tang, and Hang-Fei Cheng. Complete solution classification for the perspective-three-point problem. *IEEE transactions on pattern analysis and machine intelligence*, 25(8):930–943, 2003.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[11] Rasmus Laurvig Haugaard and Anders Glent Buch. Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6749–6758, 2022.

[12] 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 *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11*, pages 548–562. Springer, 2013.

[13] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[14] Tomáš Hodaň, Vibhav Vineet, Ran Gal, Emanuel Shalev, Jon Hanzelka, Treb Connell, Pedro Urbina, Sudipta N Sinha, and Brian Guenter. Photorealistic image synthesis for object instance detection. In *2019 IEEE international conference on image processing (ICIP)*, pages 66–70. IEEE, 2019.

[15] Tomas Hodan, Daniel Barath, and Jiri Matas. Epos: Estimating 6d pose of objects with symmetries. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11703–11712, 2020.

[16] Tomas Hodan, Martin Sundermeyer, Bertram Drost, Yann Labbe, Eric Brachmann, Frank Michel, Carsten Rother, and Jiri Matas. Bop challenge 2020 on 6d object localization. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 577–594. Springer, 2020.

[17] Tomas Hodan, Martin Sundermeyer, Bertram Drost, Yann Labbe, Eric Brachmann, Frank Michel, Carsten Rother, and Jiri Matas. Datasets., 2020. URL https://bop.felk.cvut.cz/datasets/.

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[19] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[20] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 574–591. Springer, 2020.

[21] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155–166, 2009.

[22] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. *International Journal of Computer Vision*, 128:657–678, 2020.

[23] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7678–7687, 2019.

[24] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7668–7677, 2019.

[25] Sida Peng, Xiaowei Zhou, Yuan Liu, Haotong Lin, Qixing Huang, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof object pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3212–3223, 2020.

[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. In *Proceedings of the IEEE international conference on computer vision*, pages 3828–3836, 2017.

[27] Marc'Aurelio Ranzato, Y-Lan Boureau, Yann Cun, et al. Sparse feature learning for deep belief networks. *Advances in neural information processing systems*, 20, 2007.

[28] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th international conference on international conference on machine learning*, pages 833–840, 2011.

[29] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[30] Chen Song, Jiaru Song, and Qixing Huang. Hybridpose: 6d object pose estimation under hybrid representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 431–440, 2020.

[31] Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6738–6748, 2022.

[32] Martin Sundermeyer, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O Arras, and Rudolph Triebel. Multi-path learning for object pose estimation across domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13916–13925, 2020.

[33] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, and Rudolph Triebel. Augmented autoencoders: Implicit 3d orientation learning for 6d object detection. *International Journal of Computer Vision*, 128:714–729, 2020.

[34] Martin Sundermeyer, Tomáš Hodaň, Yann Labbe, Gu Wang, Eric Brachmann, Bertram Drost, Carsten Rother, and Jiří Matas. Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2784–2793, 2023.

[35] Hiroki Tatemichi, Yasutomo Kawanishi, Daisuke Deguchi, Ichiro Ide, and Hiroshi Murase. Category-level object pose estimation in heavily cluttered scenes by generalized two-stage shape reconstructor. *IEEE Access*, 2024.

[36] 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*, pages 292–301, 2018.

[37] Carlo Tomasi. Vector representation of rotations. *Computer Science*, 527:2–4, 2013.

[38] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[39] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021.

[40] Jimmy Wu, Bolei Zhou, Rebecca Russell, Vincent Kee, Syler Wagner, Mitchell Hebert, Antonio Torralba, and David MS Johnson. Real-time object pose estimation with pose interpreter networks. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6798–6805. IEEE, 2018.

[41] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems (RSS)*, 2018.

[42] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1941–1950, 2019.

[43] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. *Dive into Deep Learning*. Cambridge University Press, 2023. https://D2L.ai.

[44] Jianyu Zhao, Edward Sanderson, and Bogdan J Matuszewski. Cvml-pose: Convolutional vae based multi-level network for object 3d pose estimation. *IEEE Access*, 11: 13830–13845, 2023.

[45] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019.