SD-Pose: Semantic Decomposition for Cross-Domain 6D Object Pose Estimation

Zhigang Li,[1] Yinlin Hu,[2] Mathieu Salzmann,[2,3] and Xiangyang Ji [1*]
[1] Tsinghua University
[2] EPFL
[3] ClearSpace SA
lzg15@mails.tsinghua.edu.cn, yinlin.hu@epfl.ch, mathieu.salzmann@epfl.ch, xyji@tsinghua.edu.cn

Abstract

The current leading 6D object pose estimation methods rely heavily on annotated real data, which is highly costly to acquire. To overcome this, many works have proposed to introduce computer-generated synthetic data. However, bridging the gap between the synthetic and real data remains a severe problem. Images depicting different levels of realism/semantics usually have different transferability between the synthetic and real domains. Inspired by this observation, we introduce an approach, SD-Pose, that explicitly decomposes the input image into multi-level semantic representations and then combines the merits of each representation to bridge the domain gap. Our comprehensive analyses and experiments show that our semantic decomposition strategy can fully utilize the different domain similarities of different representations, thus allowing us to outperform the state of the art on modern 6D object pose datasets without accessing any real data during training.

Introduction

Accurately estimating the rotation and translation of a 3D object from a single RGB image is a fundamental problem in computer vision. It has broad applications in the real world, such as augmented reality, mobile robotics, and autonomous navigation. As such, this task has attracted continuous attention in the research community.

While much progress has been made (Kehl et al. 2017; Sundermeyer et al. 2018; Peng et al. 2019; Li, Wang, and Ji 2019), the current leading approaches are data-hungry and rely heavily on annotated real-world data, which are costly to obtain, because 6D object pose annotations have to be done in the 3D space. A straightforward solution to this problem consists of exploiting virtual environments. In principle, given the 3D object model, one can generate large amounts of synthetic data with perfect pose annotations. However, this typically results in a significant domain gap between the synthetic and real images, limiting the performance of methods trained on such synthetic data.

To this date, this problem is mainly alleviated by either employing a physically-based renderer (Hodan et al. 2019) or resorting to domain randomization techniques (Kehl et al. 2017; Sundermeyer et al. 2018) . Nevertheless, to achieve reasonable accuracy, these methods still require having access to some annotated real images. In short, accurate pose estimation without any annotated real-world data remains an open problem, for which a solution is urgently needed.

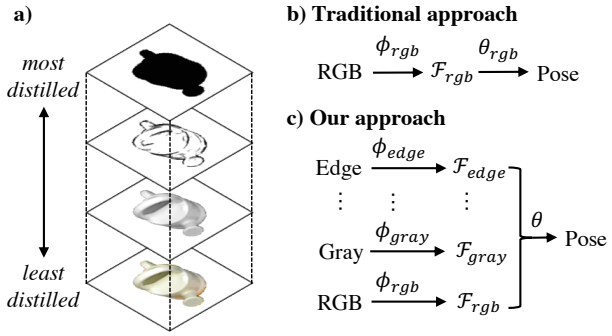

Figure 1: a) A single input image can be decomposed into diverse representations, thus constructing a hierarchical representation with different semantic levels. b) Traditional RGB-based pose estimation only uses the raw RGB representation. Here, ϕ and θ are functions represented by some neural networks. c) By contrast, we leverage the different properties of the diverse semantic representations so as to bridge the domain gap between the synthetic and real data and produce high-quality 6D object pose estimates when only synthetic data is available during training.

In this paper, we address this by observing that different semantic representations of an image have different levels of transferability between the synthetic and real domains. Inspired by this, we propose to explicitly decompose the input image to multi-level semantic representations, as shown in Fig. 1. We then propose to fuse the features extracted from these individual semantic representations so as to jointly leverage their complementary transferability power.

Our contributions can be summarized as follows:

- We propose a novel pose estimation approach, SD-Pose, which effectively leverages the domain transferability benefits of diverse representations for cross-domain 6D object pose estimation.
- Our SD-Pose relies on the following novel components:

*Corresponding author.

I. We introduce a siamese pose estimation module, Cross-semantic Coordinates Net (CCNet), which efficiently handles multiple representations with siamese training; II. We propose an adaptive feature fusion module, Context-aware Aggregation Net (CANet), able to integrate the contributions of each representation adaptively; III. We introduce a learnable ensemble module, Coordinates Ensemble Net (CENet), which improves the performance by ensembling the diverse semantic representations.

As evidenced by our experiments, SD-Pose outperforms the state of the arts on both the LineMOD and Occluded-LineMOD datasets without accessing any real data during training.

Related Work

Object Pose Estimation

We focus on estimating object pose from a single RGB image without depth (Wang et al. 2020a; He et al. 2020). Traditionally, this task was regarded as a geometric problem, which can be solved with a two-stage pipeline: I. features matching between the 2D image and the 3D object model; II. geometric verification of the matched features via a Perspective-n-Point (PnP) algorithm. While traditional approaches handle textured objects well, they do not generalize to poorly-textured ones.

Recently, the advent of deep learning has brought a leap forward in this field. In this context, some methods still follow the traditional strategy. Concretely, a deep neural network is trained to build 2D-3D correspondences by either a) detecting pre-defined semantic keypoints in the input image (Pavlakos et al. 2017; Rad and Lepetit 2017; Tekin, Sinha, and Fua 2018; Hu et al. 2018; Peng et al. 2019; Song, Song, and Huang 2020; Hu et al. 2020), or b) predicting the corresponding 3D coordinates of each pixel belonging to the object (Hodan, Barath, and Matas 2020). For a), one solution consists of detecting the semantic keypoints on the object surface from the image (Peng et al. 2019; Pavlakos et al. 2017). However, the need for distinct semantic keypoints across object categories limits the generality of this approach. To overcome this, other methods estimate the location of the 8 corners of the 3D object bounding box (Rad and Lepetit 2017; Tekin, Sinha, and Fua 2018; Hu et al. 2018). In this case, however, the detection is usually less accurate since the keypoints are far from the object. For b), because the resulting correspondences are dense, the pose must be obtained via a RANSAC-based procedure, which significantly reduces speed. Another strategy (Kendall, Grimes, and Cipolla 2015; Kendall and Cipolla 2017; Brahmbhatt et al. 2018; Kehl et al. 2017; Su et al. 2015; Tulsiani and Malik 2015; Massa, Marlet, and Aubry 2016) consists of training the deep model to estimate the pose directly from the image, so as to fully exploit the power of end-to-end learning. While the resulting models trained apply easily to various objects, they yield limited performance because the rotation $R \in SO(3)$ is challenging to regress. The limitations of such pose regression approaches were analyzed in (Sattler et al. 2019), showing that there is still a long way before pose regression becomes practical.

Domain Adaptation

In this paragraph, we focus on the domain adaptation techniques that were developed in the context of 6D object pose estimation, so as to better leverage synthetic data.

SSD6D (Kehl et al. 2017) is an early explorer of predicting the object pose with synthetic-only data. They estimated the rotation by training a classifier based on the SSD (Liu et al. 2016) detector, while the translation was calculated from the 2D bounding box. However, their performance is not remarkable. To reduce the domain gap, some works (Hodan et al. 2019) proposed to generate realistic images using a well-designed pipeline exploiting physically-based rendering. The generation of such plausible synthetic data, however, remains cumbersome and labor intensive. To overcome this, other methods (Sundermeyer et al. 2018; Wang et al. 2019) resort to domain randomization. Concretely, a series of data augmentation and randomization techniques are introduced to help a model trained in a synthetic-only scenario to perform well in the real world. AAE (Sundermeyer et al. 2018) trained an augmented autoencoder for pose retrieval. While they utilized a series of domain randomization techniques to help bridge the domain gap, the performance is still inferior. DPOD follows the coordinate-based approach to solve the pose from dense 2D-3D correspondences via the RANSAC-style PnP algorithm, and it constitutes the state of the art for pose estimation in the synthetic-only case currently.

While domain randomization improves pose accuracy compared to naive training on synthetic data, the resulting performance remains far from satisfying. In (Rambach et al. 2018), a single-channel sketch representation was leveraged instead of RGB for better object pose estimation, showing that a suitable representation offers the promise to improve performance. However, the sacrificed information (e.g., color) in the sketch makes the approach ill-suited to, e.g., color-dominant cases. More generally, in principle, any approach based on partial information may fail in certain circumstances. To circumvent the need for annotated real data, some works (Georgakis et al. 2019; Rad et al. 2018; Wang et al. 2020b) have proposed to leverage unlabeled real RGB-D data to bridge the synthetic-real domain gap. Self6D (Wang et al. 2020b) proposed a self-supervised approach to improve the pose estimation performance via refining the initially estimated pose by comparing the rendered RGB and depth with the observed ones. However, it fails to obtain remarkable results. Others (Georgakis et al. 2019; Rad et al. 2018) rely on the consistency between RGB and depth to improve the performance. Concretely, the depth and RGB models were trained to extract features from the corresponding modality, respectively. Then, an RGB-to-depth feature mapping (Rad et al. 2018) or an RGB-depth consistency loss (Georgakis et al. 2019) were applied by taking advantage of the consistency between RGB and depth. Benefiting from the small domain gap of depth, these methods were able to improve the domain transferability of the RGB model. Nevertheless, while these methods do not require real pose annotations, they cannot handle the situations where real-world depth training data are not available. In this paper, we introduce an RGB-based pose estimation approach that does not

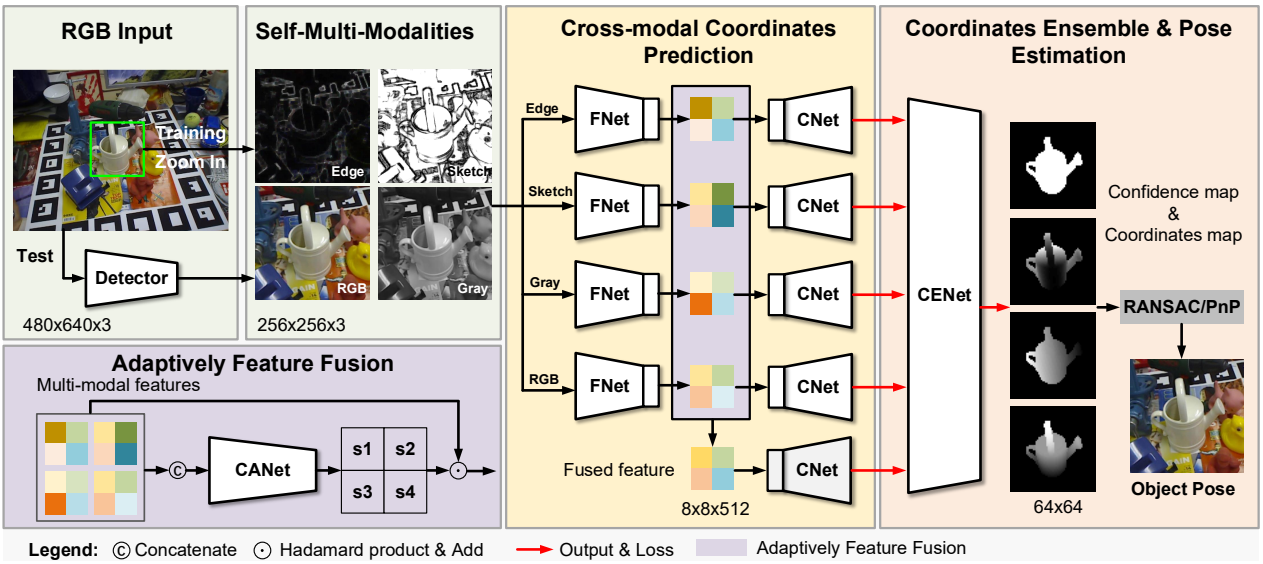

Figure 2: The framework of our SD-Pose. We first use a 2D detector to zoom in on the object, from which the multi-level semantic representations are distilled. Then, we employ FNet to extract semantic features, which can be further adaptively fused by CANet. All features are fed to CNet to predict the coordinates & confidence maps, from which the pose can be solved via PnP & RANSAC. The CENet is proposed for ensembling to achieve a better estimation.

require access to any real-world data during training. To this end, we leverage diverse semantic representations distilled from RGB to bridge the domain gap.

Method

Let us now first present our analysis of multi-level semantic representations. Based on this analysis, we then introduce our SD-Pose approach.

Multi-Level Semantic Representations (MSR)

Given an RGB image, various semantic representations $\mathbb{I}_{MSR}$ (gray, hue, sketch, edge, mask, etc.) can be extracted to construct a multi-level semantic representation structure, as shown in Fig. 2. In particular, we sort the resulting representations according to the amount of information they carry, from high to low. We then express the resulting hierarchical MSR as

$$\mathbb{I}_{MSR} = \{\mathcal{I}_{gray}, \mathcal{I}_{sketch}, \mathcal{I}_{edge}, \cdots, \mathcal{I}_{mask}\} \,. \quad (1)$$

Domain Transferability. In MSR, from RGB to mask, as the amount of information decreases, so does the synthetic-real domain gap. For example, (Rambach et al. 2018) has shown that a single-channel sketch has better domain transferability than the RGB image. Here, we conduct further analysis by comparing $\mathcal{I}_{rgb}$ with $\mathcal{I}_{mask}$ in MSR. Given an object in a certain pose, the RGB image in the synthetic and real domains differ in terms of color, texture, brightness, contrast,... By contrast, the mask should be the same in both domains, even in the presence of occlusion. In theory, the highly-distilled semantic representations in MSR can help to narrow the domain gap. To evidence this, in Table 1, we compare RGB and mask on the LineMOD dataset (see details in the supplementary material). Given the ground-truth mask at test time, the mask-based approach achieves dramatically better results than RGB-based one (accuracy 94.3% *vs.* 46.3%). We also generate a synthetic test set to perform this comparison without a domain gap. While the RGB-based model improves significantly in this case (from 46.3% to 83.5%), it remains inferior to the mask-based one (94.3%), showing that highly-distilled semantic representations may also reduce the learning difficulty.

Robustness. While high-level semantic representations in MSR help to narrow the domain gap, they can also be less robust due to the sparse information they contain. For example, for mask-based pose estimation, when we use a mask obtained with MaskRCNN, which contains segmentation errors, the performance drops significantly. Additionally, the lost information in such highly abstract representations may also contain critical clues of the pose. For instance, since the mask only reflects the 2D projected shape, it is insufficient for objects with symmetric shape but asymmetric (unambiguous) texture. Therefore, a single highly-distilled modality is not suitable for pose estimation. We, therefore, propose to exploit multi-level semantic representations, ranging from low to high in MSR, to improve and robustify pose estimation.

Pose Estimation based on Semantic Decomposition

Here, we introduce our pose estimation approach combining multiple semantic representations to improve accuracy.

Overview. We adopt a detection-based framework, which first locates the object in 2D in the image and then estimates the 6D pose from the detected region. For pose estimation, we leverage coordinates-based approaches (Li, Wang, and Ji 2019; Park, Patten, and Vincze 2019), which have proven to

Test Data	Ape	Bv.	Cam.	Can	Cat	Dril.	Duck	Eggb.	Glue	Hol.	Iron	Lamp	Ph.	**Avg**
RGB-real	7.0	70.1	31.6	70.2	28.5	39.0	37.3	68.1	91.2	12.7	68.3	38.8	38.8	46.3
RGB-syn	21.4	96.8	95.0	**98.5**	79.2	98.1	73.4	**100**	**99.7**	36.7	99.0	95.2	92.2	83.5
Mask-gt	**60.8**	**98.6**	**99.1**	97.6	**98.6**	**98.7**	**82.1**	**100**	95.9	**97.0**	**99.4**	**98.6**	**99.3**	**94.3**
Mask-det	6.9	39.9	23.5	18.8	0.0	66.8	9.9	53.8	48.4	22.9	45.2	72.5	40.9	34.6

Table 1: RGB *vs.* Mask on the LineMOD dataset (ADD metric, higher is better). We only use synthetic RGB/mask data during training. For testing, RGB-real and RGB-syn represent the official real test data and our generated synthetic test data, respectively. Mask-gt and Mask-det are the ground-truth mask from synthetic data and the detected mask (using MaskRCNN (He et al. 2017)) from real data, respectively.

be robust to occlusion. Given an input $\mathcal{I}$, the model is trained to recognize the pixels $\mathcal{P}$ belonging to the object and then estimate the corresponding 3D coordinates $\mathcal{C}$ for these pixels. This yields dense 2D-3D correspondences between $\mathcal{P}$ and $\mathcal{C}$, from which the pose can be obtained by RANSAC-based PnP algorithm. Our work then differs from existing approaches in that we aim for our model to learn to combine the strengths of multiple semantic representations. To this end, we propose a semantic decomposition pose estimation approach (SD-Pose), which consists of three modules: I. Cross-semantic Coordinates Net (CCNet); II. Context-aware Aggregation Net (CANet); and III. Coordinates Ensemble Net (CENet). First, diverse semantic representations are fed to CCNet to extract features for each of them. These individual features are then adaptively fused by CCNet into an integrated feature map that combines the power of each representation. This integrated feature map can be directly used to predict the coordinates & confidence maps, from which the pose can be obtained by PnP+RANSAC. Nevertheless, we further introduce CENet to ensemble the different representations and produce better pose estimates.

Detection. We employ MaskRCNN (He et al. 2017) for detection and introduce the zoom-in operation of (Li, Wang, and Ji 2019) to scale the object patch to a unified resolution. Compared with (Rambach et al. 2018; Zakharov et al. 2019) that directly estimate the object pose from the whole image, our detection and zooming in operations significantly reduce the learning difficulty not only for small objects in the image, but also particularly when using multiple semantic representations, since dealing with a zoomed-in object facilitate extracting unified-style representations from RGB, as illustrated by Fig. 3.

We emphasize the superiority of extracting multi-level semantic representations from the zoomed-in local object patch. By contrast, (Rambach et al. 2018) extracted a sketch from the whole image with a pre-defined filter. However, the arbitrary object size in the image makes the filtered sketch have diverse styles. Thanks to our zooming in operation, we obtain semantic representations that have a unified style, which greatly reduces the learning difficulty.

Semantic Representations Distillation. After zooming in on the object, we need to extract semantic representations from the local image patch for pose estimation, which is achieved by filtering out the redundant information from RGB for each representation. In this process, we leverage traditional algorithms (e.g., hand-crafted filters) instead of learning-based ones to circumvent the subjective process and ensure the data's authenticity. Thus, the mask in MSR is unused since extracting semantic object mask needs to learn the prior-knowledge of the object. Ultimately, in addition to $\mathcal{I}_{rgb}$, our semantic representations include $\mathcal{I}_{gray}$, $\mathcal{I}_{sketch}$, $\mathcal{I}_{edge}$, all of which can directly be obtained from the RGB image.

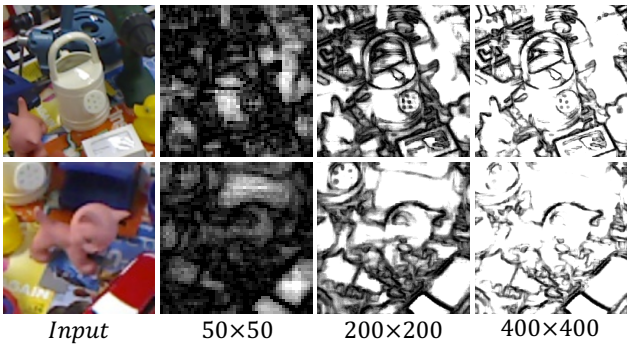

Figure 3: The sketch generated from the same filter varies dramatically for different resolutions, which greatly increases the learning difficulty. We circumvent this by resizing all detected objects to a uniform resolution (256×256) before pose estimation.

Cross-semantic Coordinates Net (CCNet). To deal with these diverse semantic representations, an intuitive solution consists of introducing an individual model for each one. This, however, makes the system less efficient and precludes any interactions across the different representations. Instead, to efficiently handle the semantic representations, we introduce Cross-semantic Coordinates Net (CCNet), which consists of two modules, FNet and CNet, which we describe in detail below. CCNet is trained in a siamese manner across the different representations, that is, a single CCNet can handle multiple representations. The multiple supervision signals, and the shared parameters and optimization process in CCNet also enable transferring information across the representations, making training more effective.

The FNet module consists of a 34-layers residual network for feature extraction. To enable FNet to be trained across semantic representations, we use a unified format for all representations, i.e., $h_{in} \times w_{in} \times 3$, obtained by repeating the channel when necessary. We set $h_{in} = w_{in} = 256$. The

CNet module then serves to predict the coordinate & confidence maps from the features extracted by FNet. We adopt a classification-based training strategy. Concretely, for the coordinate maps, CNet outputs 3 heatmap volumes, each of which representing an axis and having size $h_{out} \times w_{out} \times n_{bin}$, where n_{bin} is the number of bins used to discretize the axis range. We set $h_{out} = w_{out} = n_{bin} = 64$. For the confidence maps, CNet predicts a heatmap of size $h_{out} \times w_{out} \times 2$, where binary classification is performed at each position.

To train our model, we use the masked cross-entropy loss for the coordinate maps, which only calculates the cross-entropy on the object foreground region. For the confidence maps, we compute the binary cross-entropy on the whole region. This can be expressed as

$$\ell(\mathcal{M}, \mathcal{C}) = \tau \cdot \sum_{j=1}^{n_c} \ell_{ce}(\tilde{\mathcal{M}} \circ \mathcal{C}^j, \tilde{\mathcal{M}} \circ \tilde{\mathcal{C}}^j) + \eta \cdot \ell_{ce}(\mathcal{M}, \tilde{\mathcal{M}}) \ , \tag{2}$$

where $\tilde{*}$ and $*$ represent the ground truth and the prediction, respectively; $\mathcal{M}$ and $\mathcal{C}$ represent the confidence maps and coordinates maps, respectively; $n_c = 3$ is the number of coordinate axes; $\circ$ is the Hadamard product; and ℓ_{ce} is the cross-entropy loss.

The complete CCNet loss then sums this loss over the different semantic representations and further includes a fusion loss computed on the predictions from the fused features, which we discuss in more detail below. Altogether, this yields the CCNet loss function

$$\mathcal{L}_{CC} = \ell(\mathcal{M}_{fuse}, \mathcal{C}_{fuse}) + \sum_{i \in \mathbb{I}_{MSR}} \ell(\mathcal{M}_i, \mathcal{C}_i) \ , \tag{3}$$

where $*_{fuse}$ and $*_i$ represent the predictions from fused branch and other semantic branches respectively.

Context-aware Aggregation Net (CANet). Inspired by attention mechanisms (Bahdanau, Cho, and Bengio 2015; Vaswani et al. 2017), we propose Context-aware Aggregation Net (CANet) to fuse the multiple semantic features and combine the benefits of each modality. Given an input sample, the choice of the most appropriate semantic representation can easily be affected by many factors (e.g., object, pose, blur, lighting, etc.). Thus, the model should be able to learn the importance of each representation according to the input. Our CANet achieves this by regressing a score s_i for each representation, which we use to fuse them as

$$\mathcal{F}_{fuse} = \sum_{i \in \mathbb{I}_{MSR}} s_i \cdot \mathcal{F}_i \ . \tag{4}$$

During inference, we concatenate the semantic features and feed them to the CANet, which consists of a sequence of convolutional layers and several linear layers to predict the scores. For training, since ground-truth supervision for CANet is unavailable, we follow the attention mechanism to embed the CANet into the CCNet and thus implicitly train it in an end-to-end fashion.

Coordinates Ensemble Net (CENet). Since each feature from FNet or CANet yields a candidate coordinate & confidence map, we rely on an ensembling process to fuse them all into a refined prediction. In contrast to traditional ensemble approaches, we propose Coordinates Ensemble Net (CENet) to integrate all the candidates to achieve better results. Specifically, all candidates are stacked and fed into CENet, and a sequence of convolutions are applied to combine and refine them. We write the CENet loss as

$$\mathcal{L}_{ensemble} = \ell(\mathcal{M}_{ensemble}, \mathcal{C}_{ensemble}) \ , \tag{5}$$

where ℓ is defined in Eq. 2.

Training

Our SD-Pose is trained end-to-end. That is, all modules, including CCNet, CANet and CENet, are optimized jointly. We apply the coordinate & confidence loss to each output from CANet and CENet. This yields the overall loss

$$\mathcal{L}_{total} = \mathcal{L}_{CC} + \mathcal{L}_{ensemble} \ . \tag{6}$$

Data Preparation

Dataset. We conduct our experiments on both the LineMOD and Occluded-LineMOD dataset. LineMOD (Hinterstoisser et al. 2012) is the *de facto* standard benchmark for 6D pose estimation of textureless objects in cluttered scenes. It comprises 13 texture-less objects for evaluation. Each object appears in about 1000 test images covering the upper view hemisphere at different scales on a cluttered desk. The Occluded-LineMOD dataset was proposed by (Krull et al. 2015), and shares the same images as LineMOD. 8 heavily occluded objects in one video sequence are annotated for testing purposes. We follow (Brachmann et al. 2016; Li et al. 2018) to split the dataset. Concretely, the test set consists of all occluded images. For both LineMOD and Occluded-LineMOD, we only use synthetic data for training.

Synthetic Training Data. We employ the widely used OpenGL-based renderer (Kehl et al. 2017; Sundermeyer et al. 2018; Li et al. 2018) to generate synthetic data. It is lightweight and can achieve real-time rendering, while the generated synthetic images present an evident domain gap with the real data. Additionally, we utilize Blender for physically-based rendering to generate high-quality realistic synthetic images to evaluate our approach in the situation that the synthetic-real domain gap is narrow. In this context, we designed a series of settings (e.g., lighting, shadow, ground plane, etc.) to generate realistically rendered data. For either case, during rendering, we randomly generate 10000 synthetic images for each object according to the pose distribution of the training set. Concretely, during rendering, the rotation is uniformly sampled from the angle range of the training set, and the translation is randomly generated according to the mean and variance calculated from the training set. For Occluded-LineMOD, we randomly chose 3-8 objects to render one image to introduce occlusions among the objects. During training, for all synthetic images, the background is randomly replaced with indoor images from the PASCAL VOC2012 dataset.

Experiments

Metrics

To evaluate performance, we report three common metrics: 5cm 5°, Proj. 2D, and ADD. For 5cm 5°, a pose is correct

Row	Methods				5cm 5°	ADD	2D Proj.
	Syn	Inp	Fus	Ens			
1	Op	RGB	-	-	37.6	46.3	57.3
2	Op	MSR	Con	✗	46.7	53.2	64.1
3	Op	MSR	Ada	✗	48.0	54.5	66.7
4	Op	MSR	Ada	✓	52.5	56.3	70.3
5	Bl	RGB	-	-	71.5	63.2	83.8
6	Bl	MSR	Con	✗	72.1	64.1	84.8
7	Bl	MSR	Ada	✗	72.8	65.3	85.9
8	Bl	MSR	Ada	✓	73.3	67.3	86.1

Table 2: Ablation study on the input type (Inp), the synthetic data type (Syn), CANet and CENet on the LineMOD dataset. Op: OpenGL-based; Bl: Blender-based; MSR: multi-level semantic representations; Con: feature fusion with constant value, i.e., 1; Ada: adaptive feature fusion via CANet; Ens: Ensemble via CENet. Note that our SD-Pose bridges the domain gap between synthetic and real data, especially in the case of OpenGL-based rendering, where the gap is particularly large.

if the translation error and rotation error are smaller than $5cm$ and $5°$, respectively. For Proj. 2D, the estimated pose is correct if the average 2D projection error is smaller than 5 pixels. For ADD, a pose is deemed correct if the average vertex-to-vertex distance in 3D space is below $0.1d$, where d is the object diameter. For symmetric objects, the nearest points are used to compute the distance.

Ablation Study

Semantic Representations vs. raw RGB. Here, we conduct an ablation study on the input type to show the superiority of our multi-level semantic representations over the raw RGB representation. We first compare them in a simplified setting. Concretely, we only use CCNet to predict coordinates & confidence maps from different inputs, abandoning the adaptive feature fusion (CANet) and coordinates ensemble (CENet). For the case of multi-level semantic representations, without CENet, we only use the fused feature branch for training and prediction. Furthermore, without CANet, we use identical weights (i.e., 1) for feature fusion. As shown in Table 2, using OpenGL-based synthetic training data, the semantic representations (row 2) achieve significantly better results than the RGB one (row 1) on all metrics (accuracies of 46.7%, 53.2%, 64.1% *vs.* 37.6%, 46.3%, 57.3% in terms of 5cm 5°, ADD, and 2D Proj., respectively). This is true even without adaptive feature fusion and coordinates ensemble, but the use of CANet (row3) and CENet (row4) further improve the performance of the semantic representations.

Ablation Study on CANet. We analyze CANet to show the effectiveness of adaptive feature fusion. Without CANet, the multiple semantic features are fused using identical weights (i.e., 1). As shown in Table 2, our adaptive feature fusion (row 3) significantly outperforms the baseline (row 2) according to all metrics. Even when the domain gap becomes narrow (i.e., on Blender-based data), our CANet still improves the performance (rows 6 and 7).

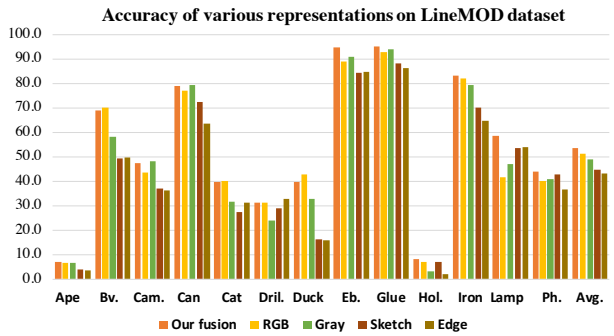

Figure 4: We compare the pose estimation accuracy of various modalities (metric: ADD, higher is better). The proposed fusion network performs the best on average, demonstrating that it can fully benefit from the domain transferability of different semantic representations.

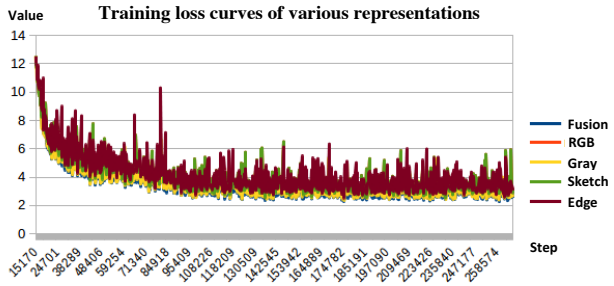

Figure 5: Training loss curves of various representations for 'Ape'. Across the different representations, the loss of the fused feature branch typically reaches the minimum value.

Ablation Study on CENet. Here, we analyze the effect of the coordinate ensemble of CENet. Without CENet, the output from the fused feature branch is used to compute the pose. As shown in Table 2, fusing predictions from multiple semantic representations significantly improves the pose estimation performance on both OpenGL-based (rows 3 and 4) and Blender-based (rows 7 and 8) data.

Ablation Study on Synthetic Type. We also conduct an ablation study on the different types of synthetic data to analyze the effectiveness of our approach with various synthetic-real domain gaps. As shown in Table 2, our SD-Pose achieves significant performance improvements over the RGB baseline using both OpenGL-based and Blender-based data.

Effectiveness of Representation Fusion

We further analyze our multi-level semantic representation learning to reveal the mechanism behind its performance improvement. Concretely, we remove CENet from SD-Pose and train CCNet and CANet using the $\mathcal{L}_{CC}$ loss to estimate the pose from each representation and from the fused features simultaneously. This allows us to compare the pose estimation performance of each semantic representation and of the fused feature on the same model. As shown in Figure 4, the fused features yield the best average performance

Data	Method	Ape	Bv.	Cam.	Can	Cat	Dril.	Duck	Eggb.	Glue	Hol.	Iron	Lamp	Ph.	Avg
Syn	SSD6D (Kehl et al. 2017)	2.6	15.1	6.1	27.3	9.3	12.0	1.3	2.8	3.4	3.1	14.6	11.4	9.7	9.1
	AAE (Sundermeyer et al. 2018)	4.0	20.9	30.5	35.9	17.9	24.0	4.9	81.0	45.5	17.6	32.0	60.5	33.8	31.4
	MHP (Manhardt et al. 2019)	11.9	66.2	22.4	59.8	26.9	44.6	8.3	55.7	54.6	15.5	60.8	-	34.4	38.8
	Self6D (Wang et al. 2020b)	37.2	68.9	17.9	50.4	33.7	47.4	18.3	64.8	59.9	5.2	68.0	35.3	36.5	40.1
	DPOD (Zakharov et al. 2019)	37.2	66.8	24.2	52.6	32.4	**66.6**	26.1	73.4	75.0	24.5	**85.0**	57.3	29.1	50.0
	Ours (OpenGL)	11.2	**83.5**	41.4	77.3	46.5	45.0	37.4	**94.3**	**95.6**	15.3	82.9	50.5	51.0	56.3
	Ours (Blender)	**54.0**	76.4	**50.2**	**81.2**	**71.0**	64.2	**54.0**	93.9	92.6	**24.0**	77.0	**82.6**	**53.7**	**67.3**
Real*	DomainTF (Rad et al. 2018)	19.8	69.0	37.6	42.3	35.4	54.7	29.4	85.2	77.8	36.0	63.1	75.1	44.8	51.6
	Self6D (Wang et al. 2020b)	38.9	75.2	36.9	65.6	57.9	67.0	19.6	99.0	94.1	15.5	77.9	68.2	50.1	58.9

Table 3: Comparison with the state-of-the-art approaches on LineMOD without using real annotations. Metric: ADD; 'Syn': only synthetic data involved during training; 'Real*': *unlabeled* real data involved during training. Our SD-Pose trained on the simple OpenGL-based synthetic data yields on-par performance with the competitors. Furthermore, it outperforms most of the competitors that exploit real data by a large margin, even though it only relies on synthetic data generated using Blender.

Data	Method	Ape	Can	Cat	Dril.	Duck	Eggb.	Glue	Hol.	Avg
Syn	Self6D (Wang et al. 2020b)	*10.1*	16.5	6.2	*16.8*	12.8	25.2	21.6	7.5	14.6
	Ours (OpenGL)	4.3	49.5	8.9	16.4	25.4	30.3	37.1	7.6	22.4
	Ours (Blender)	**21.5**	**56.7**	**17.0**	**44.4**	**27.6**	**42.8**	**45.2**	**21.6**	**34.6**
Real*	Self6D (Wang et al. 2020b)	17.0	41.4	19.0	31.1	8.9	57.4	40.8	17.8	29.2

Table 4: Comparison with the state-of-the-art approaches on Occluded-LineMOD in the synthetic-only case. (Metric: ADD; 'Syn': only synthetic data involved during training; 'Real*': *unlabeled* real data involved during training.)

across 13 objects, surpassing the single-semantic competitors by a significant margin. More importantly, for each object, the performance of the fused features is on par with or outperforms the best single-semantic result. This constitutes evidence that our approach successfully guides the model to integrate the benefits of every single modality. In Fig. 5, we show the training loss curve of each prediction branch. The output of the fused feature branch yields the minimum loss in most cases, showing the benefit of the fused features over the individual single-semantic features when it comes to learning the pose.

Comparison with State-of-the-art Approaches

LineMOD dataset. We first compare our approach with competitors using only synthetic data during training. We focus on the ADD because most synthetic-only approaches only report results on this metric. As shown in Table 3, our SD-Pose trained on OpenGL-based synthetic data outperforms the competitors by a large margin (our 56.3% *vs.* 50.0%). Exploiting realistic rendering data further boosts our performance from 56.3% to 67.3% on ADD. On the LineMOD dataset, we achieve state-of-the-art performance in the synthetic-only case. We then compare our approach with those that exploit real data but without using real annotations. Specifically, in the bottom portion of the Table 3, we report the results of DomainTF (Rad et al. 2018) and Self6D (Wang et al. 2020b), both of which leverage unlabeled real-world RGBD data to help bridge the domain gap. Our approach trained on OpenGL-based synthetic-only data without any real-world data performs on par with these methods. When we utilize physically-based rendered data, our approach outperforms them by a large margin (67.3% *vs.* 58.9%). On the LineMOD dataset, we achieve state-of-the-art performance for pose estimation without using real annotations.

Occluded-LineMOD dataset. We then compare our approach with the state-of-the-art methods on the Occluded-LineMOD dataset. Self6D (Wang et al. 2020b) achieves the state-of-the-art performance on Occluded-LineMOD in synthetic-only case. However, as shown in Table 4, our approach with OpenGL data already outperforms it by a large margin. Training with physically-based synthetic data allows us to achieve the state-of-the-art synthetic-only performance on this dataset.

Conclusions

We have introduced a framework based on multi-level semantic decomposition for cross-domain 6D object pose estimation. Our approach has allowed us to bridge the gap between the synthetic and real domains. Our thorough experiments and ablation studies of each component of our approach have evidenced the effectiveness of our framework. We achieve state-of-the-art performance on both the LineMOD and Occluded-LineMOD datasets without requiring any real data during training, outperforming the competitors by a large margin. In principle, our semantic decomposition framework is not specific to pose estimation and should apply to other tasks that suffer from a synthetic-real domain gap. In the future, we will therefore verify its effectiveness in other contexts.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2017YFB1002202, the National Natural Science Foundation of China under Grant 61620106005, and the Swiss Innovation Agency (Innosuisse) under Grant 38398.1 IP-ICT.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. *International Conference on Learning Representations (ICLR)* .

Brachmann, E.; Michel, F.; Krull, A.; Ying Yang, M.; Gumhold, S.; et al. 2016. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3364–3372.

Brahmbhatt, S.; Gu, J.; Kim, K.; Hays, J.; and Kautz, J. 2018. Geometry-aware learning of maps for camera localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2616–2625.

Georgakis, G.; Karanam, S.; Wu, Z.; and Kosecka, J. 2019. Learning Local RGB-to-CAD Correspondences for Object Pose Estimation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* .

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

He, Y.; Sun, W.; Huang, H.; Liu, J.; Fan, H.; and Sun, J. 2020. PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 11629–11638.

Hinterstoisser, S.; Lepetit, V.; Ilic, S.; Holzer, S.; Bradski, G.; Konolige, K.; and Navab, N. 2012. Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. In *ACCV*.

Hodan, T.; Barath, D.; and Matas, J. 2020. EPOS: Estimating 6D Pose of Objects with Symmetries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Hodan, T.; Vineet, V.; Gal, R.; Shalev, E.; Hanzelka, J.; Connell, T.; Urbina, P.; Sinha, S.; and Guenter, B. 2019. Photorealistic Image Synthesis For Object Instance Detection. In *IEEE International Conference on Image Processing*.

Hu, Y.; Fua, P.; Wang, W.; and Salzmann, M. 2020. Single-Stage 6D Object Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Hu, Y.; Hugonot, J.; Fua, P.; and Salzmann, M. 2018. Segmentation-driven 6D Object Pose Estimation. *arXiv preprint arXiv:1812.02541* .

Kehl, W.; Manhardt, F.; Tombari, F.; Ilic, S.; and Navab, N. 2017. SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1521–1529.

Kendall, A.; and Cipolla, R. 2017. Geometric loss functions for camera pose regression with deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kendall, A.; Grimes, M.; and Cipolla, R. 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, 2938–2946.

Krull, A.; Brachmann, E.; Michel, F.; Yang, M.; Gumhold, S.; and Rother, C. 2015. Learning Analysis-by-Synthesis for 6D Pose Estimation in RGB-D Images. *2015 IEEE International Conference on Computer Vision (ICCV)* .

Li, Y.; Wang, G.; Ji, X.; Xiang, Y.; and Fox, D. 2018. DeepIM: Deep Iterative Matching for 6D Pose Estimation. In *European Conference on Computer Vision (ECCV)*.

Li, Z.; Wang, G.; and Ji, X. 2019. CDPN: Coordinates-based Disentangled Pose Network for Real-time RGB-based 6-DoF Object Pose Estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, 21–37.

Manhardt, F.; Arroyo, D. M.; Rupprecht, C.; Busam, B.; Navab, N.; and Tombari, F. 2019. Explaining the Ambiguity of Object Detection and 6D Pose From Visual Data. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* .

Massa, F.; Marlet, R.; and Aubry, M. 2016. Crafting a multi-task CNN for viewpoint estimation. *arXiv preprint arXiv:1609.03894* .

Park, K.; Patten, T.; and Vincze, M. 2019. Pix2Pose: Pixel-Wise Coordinate Regression of Objects for 6D Pose Estimation.

Pavlakos, G.; Zhou, X.; Chan, A.; Derpanis, K. G.; and Daniilidis, K. 2017. 6-dof object pose from semantic keypoints. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, 2011–2018. IEEE.

Peng, S.; Liu, Y.; Huang, Q.; Zhou, X.; and Bao, H. 2019. PVNet: Pixel-wise Voting Network for 6DoF Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4561–4570.

Rad, M.; and Lepetit, V. 2017. BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth. In *IEEE International Conference on Computer Vision (ICCV)*.

Rad, M.; Oberweger, M.; Lepetit, V.; Lepetit, V.; and Lepetit, V. 2018. Domain Transfer for 3D Pose Estimation from Color Images without Manual Annotations. *Asian Conference on Computer Vision (ACCV)* .

Rambach, J.; Deng, C.; Pagani, A.; and Stricker, D. 2018. Learning 6DoF Object Poses from Synthetic Single Channel Images. In *IEEE International Symposium on Mixed & Augmetned Reality*.

Sattler, T.; Zhou, Q.; Pollefeys, M.; and Leal-Taixe, L. 2019. Understanding the Limitations of CNN-based Absolute Camera Pose Regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3302–3312.

Song, C.; Song, J.; and Huang, Q. 2020. HybridPose: 6D Object Pose Estimation under Hybrid Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Su, H.; Qi, C. R.; Li, Y.; and Guibas, L. J. 2015. Render for CNN: Viewpoint Estimation in Images Using CNNs Trained with Rendered 3D Model Views. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2686–2694. doi:10.1109/ICCV.2015.308. URL https://doi.org/10.1109/ICCV.2015.308.

Sundermeyer, M.; Marton, Z.-C.; Durner, M.; Brucker, M.; and Triebel, R. 2018. Implicit 3D Orientation Learning for 6D Object Detection from RGB Images. In *The European Conference on Computer Vision (ECCV)*.

Tekin, B.; Sinha, S. N.; and Fua, P. 2018. Real-Time Seamless Single Shot 6D Object Pose Prediction. In *CVPR*.

Tulsiani, S.; and Malik, J. 2015. Viewpoints and keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1510–1519.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. *International Conference on Neural Information Processing Systems (NeurIPS)* .

Wang, C.; Xu, D.; Zhu, Y.; Martín-Martín, R.; Lu, C.; Li, F. F.; and Savarese, S. 2020a. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, G.; Manhardt, F.; Shao, J.; Ji, X.; Navab, N.; and Tombari, F. 2020b. Self6D: Self-Supervised Monocular 6D Object Pose Estimation. *The European Conference on Computer Vision (ECCV)* .

Wang, H.; Sridhar, S.; Huang, J.; Valentin, J.; Song, S.; and Guibas, L. J. 2019. Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation. *arXiv preprint arXiv:1901.02970* .

Zakharov, S.; Shugurov, I. S.; Ilic, S.; Ilic, S.; and Ilic, S. 2019. DPOD: 6D Pose Object Detector and Refiner. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* .