Shape-biased Texture Agnostic Representations for Improved Textureless and Metallic Object Detection and 6D Pose Estimation

Peter Hönig¹ Stefan Thalhammer² Jean-Baptiste Weibel¹
Matthias Hirschmanner¹ Markus Vincze¹

¹TU Wien, ²UAS Technikum Vienna

Abstract

Recent advances in machine learning have greatly benefited object detection and 6D pose estimation. However, textureless and metallic objects still pose a significant challenge due to few visual cues and the texture bias of CNNs. To address this issue, we propose a strategy for inducing a shape bias to CNN training. In particular, by randomizing textures applied to object surfaces during data rendering, we create training data without consistent textural cues. This methodology allows for seamless integration into existing data rendering engines, and results in negligible computational overhead for data rendering and network training. Our findings demonstrate that the shape bias we induce via randomized texturing, improves over existing approaches using style transfer. We evaluate with three detectors and two pose estimators. For the most recent object detector and for pose estimation in general, estimation accuracy improves for textureless and metallic objects. Additionally we show that our approach increases the pose estimation accuracy in the presence of image noise and strong illumination changes. Code and datasets are publicly available at github.com/hoenigpeter/randomized_ texturing.

1. Introduction

Object detection and 6D pose estimation are the foundation of robotic manipulation [1] and scene understanding [26, 31]. Computational perception and reasoning are generally more challenging for textureless and metallic objects, as they have minimal or no texture cues [30]. These objects suffer from inferior detection and pose estimation accuracy due to varying illumination conditions resulting in significantly altered surface appearances.

The standard method for solving this problem is to train with a large amount of synthetic training data, assuming that the data distribution to be expected at runtime is adequately represented [14, 17, 21, 27, 33, 35]. However, this

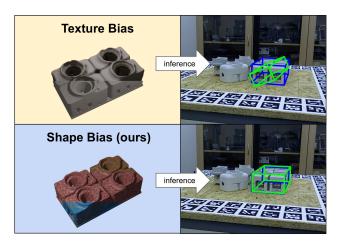


Figure 1. **Induction of Shape Bias.** 6D pose of a textureless object is visualized with 3D bounding boxes; the ground truth (blue) and estimate (green) of GDR-Net trained with conventional data (top) and when inducing a shape bias (bottom).

approach does not consider ambiguities caused by the interaction of the light with the object surface. Consequently, pose estimation approaches designed for such challenging materials focus on retrieving geometrical object representations [11, 12, 36, 38]. This strategy is supported by [7], where the authors demonstrate that CNNs exhibit a texture bias and that ImageNet [29] classification improves when a shape bias is induced. This fact actually exacerbates the aforementioned phenomenon of increased perception difficulty when processing textureless and metallic objects.

We show that inducing a shape bias during training will be particularly effective for textureless and metallic objects. For this purpose we conceptualize texture randomization as a UV-mapping approach. The surface textures of objects are randomized to render scene-level training data. This approach offers three key advantages over [7]: a) UV-mapping of the texture provides geometrically accurate visible cues for observing the object geometry, b) since encoded representations are agnostic to texture, the trained models are more robust with respect to changes in appearance due to il-

lumination, and c) training data is effectively repurposed for multiple vision problems, such as detection, segmentation, and pose estimation.

Experimental results demonstrate that our strategy improves over [7] for the classification of textureless objects. Experiments were conducted with three object detectors, YOLOx [6], Faster R-CNN [28], and RetinaNet [23], and with two pose estimators, GDR-Net [35] and Pix2Pose [27]. The experiments provide precise information regarding the improvements and limitations for textureless and metallic objects. Furthermore, experiments are presented showing enhanced robustness against image perturbations and regarding the object mesh origin.

We summarize our contributions as follows.

- We show that the induction of a shape bias to learned representations through the application of our UVmapping approach improves textureless object classification outperforming style transfer.
- The shape bias can be applied to any detection and pose estimation method and we report improvements for three object detection and two pose estimation approaches for textureless and metallic objects.
- Our strategy provides a partial alleviation for the need for online data augmentation. In particular, when recognizing textureless objects with YOLOx, the shape distortion leads to an object recognition accuracy comparable to that achieved by a grid search using the hyperparameters of the online data augmentation.
- The induction of the shape bias also robustifies the estimation of the object pose against typical real-world image perturbations, such as noise and significant illumination alterations.

The paper is structured as follows: Section 2 provides an overview of related literature, Section 3 introduces the rationale behind the induction of a shape bias, followed by the comprehensive description of the implementation and the experimental setup. Section 5 presents the empirical evidence supporting our findings, while Section 6 summarizes the observations made and outlines future work.

2. Related Work

This section summarizes the state of the art for object detection, pose estimation, and data representation.

2.1. 2D Object Detection and 6D Pose Estimation

Throughout this paper, the term object detection defines the 2D detection of an object, defined by the class label c and bounding box b. Pose estimation refers to the instance-level 6D pose estimation of an object, defined by the translation T and rotation R.

The Benchmark for 6D Object Pose Estimation (BOP) [30] challenge ranks the best-performing object detection and pose estimation methods, evaluated with standardized metrics and datasets. Successful object detectors from the last years' challenges include two-stage models such as Mask R-CNN [9] and Faster R-CNN, and single-stage ones, e.g. YOLOx.

Pose estimators are either composed of a single stage [32,33] or multiple stages [27,35], where subsequent stages build upon location priors, provided by arbitrary object detectors. For each location prior, the subsequent pose estimator performs feature extraction, learns 2D-3D correspondences, and regresses translation and rotation of the object pose. Successful CNN-based methods [27,35] from the last years' challenges use multi-stage approaches employing one of the abovementioned object detectors.

2.2. Data Representation

State-of-the-art benchmarking datasets for object detection and pose estimation of textureless and metallic objects include TLESS [16] and ITODD [4] which consist of physically-based-rendered (PBR) training sets and real-world test sets. Training with synthetic and evaluating on real-world data leaves a domain gap [34]. Data rendering and model training offer different techniques for bridging this domain gap. During rendering, domain randomization [34] can be applied, which includes the variation of backgrounds, camera views, object poses, and lighting. After rendering, during the training procedure data augmentation can be applied online, either with pre-trained augmentations [2, 39], style-transfer [8] or handcrafted combinations of varying image perturbation types.

The domain gap is also influenced by the quality of the available object geometry and texture. Depending on the target domain, meshes for the datasets are either provided via CAD models or reconstructions [25,37]. TLESS and ITODD, for example, feature plastic and metallic objects for industrial purposes, where CAD models are usually available. CAD models are geometrically accurate but often lack texture information, while reconstructions have added texture information but less geometric accuracy due to reconstruction noise.

While existing methods for bridging the domain gap focus on textured objects with distinct visual cues, less attention has been paid to textureless and metallic objects. Since lighting, reflections and shadows alter the appearance of these objects, they do not have continuous textures in the real-world domain. We hypothesize that texture should be treated as an unknown and object detection as well as pose estimation models should be biased toward shape. We therefore want to explore the randomization of object textures, to force CNNs to learn exclusively from object geometries from 2D RGB images only.

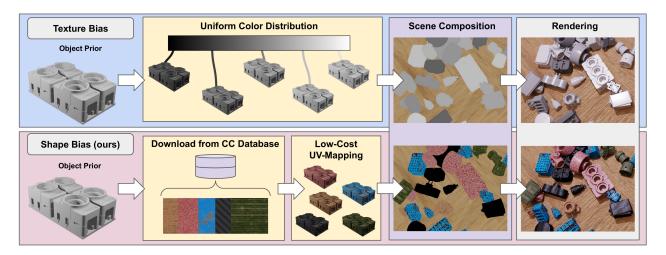


Figure 2. **Synthetic Data Generation with Randomized Texturing Pipeline.** Instead of sampling color values from a uniform color distribution we download textures from a creative-commons database and perform low-cost UV-mapping, a method that we call texture randomization. While the figure shows the scene composition and rendering steps with five exemplary textures, the texture randomization method can be used with an arbitrary number of texture files.

3. Induction of Shape Bias

This section presents the rationale of biasing representation learning toward shape. The authors of [7] present evidence that CNNs exhibit a bias toward texture. By employing style transfer with AdaIN [5] the learned representations of the network are shape biased, thereby enhancing classification on ImageNet [29]. Particularly for textureless and metallic objects, where the texture is sparse the phenomenon of texture bias is disadvantageous. This has also been observed in the context of object pose estimation for textureless and metallic objects. As a result, these objects are typically handled by estimating their shape, for example by estimating edges and silhouettes [11, 12, 36, 38].

The capacity of networks to generalize from synthetic to real object primitives with random color has been demonstrated by [34]. The authors consider various scene composition parameters, including background texture and camera, object, and light source poses as random variables. Similarly [3] observed that replacing the scene background with textures is an effective method for foreground-background separation in various vision problems. This strategy is considered the standard for creating training data for object pose estimation. We hypothesize that the application of random textures to objects themselves results in encoding the object shape in feature spaces, a process analog to that observed by [7]. This hypothesis is tested through the formulation of a learning strategy that biases estimators toward shape through rendering, thus:

 Encoding geometrically accurate cues. Localization tasks, such as object detection and pose estimation, are susceptible to geometry and illumination effects such as shadows [20]. Rendering allows us to accurately model and control the distribution of such effects.

- 2. Being robust against strong surface appearance changes. Metallic objects present a particular challenging case as their appearance is strongly influenced by the nature of the incident light. By learning shape representations, it is possible to achieve a certain degree of agnosticism with regard to surface effects.
- 3. Obtaining training data for different problems. Although the authors of [7] demonstrate that style transfer is a viable strategy for object classification, we present a strategy for inducing a shape bias for classification, detection and pose estimation. While not yet tested, our strategy can theoretically be applied to the same set of vision problems as that of [3].
- 4. **Being more universally applicable.** A significant limitation of instance-level approaches is that the trained models lack the capacity to generalize effectively to changes of the object. While this remains unsolved, our approach offers a degree of mitigation. As our trained models are color and texture agnostic, we are able to accommodate instances where the object color is unknown.

Next, we present the experimental setup and implementation detail to validate our hypothesis.

4. Data Rendering for Inducing a Shape Bias

We employ physically based rendering (PBR) to accurately model and project the object geometry to the image

Table 1. **Domain Randomization Overview.** Parameters and functions denoting the rendering specifications for the TLESS and ITODD objects

Method	Parameter	Function		
Texture Bias	Specularity Roughness Color	$S \sim \mathcal{U}(0,1)$ $R \sim \mathcal{U}(0,1)$ $(R,G,B) \sim \mathcal{U}(0.1,0.9)$		
Shape Bias (ours)	Specularity Roughness Color	$S \sim \mathcal{U}(0,1)$ $R \sim \mathcal{U}(0,1)$ $(R, G, B) = T_{n \sim \mathcal{U}(1,1226)}$		

space and cast realistic shadows to induce a shape bias in estimators. This concept has been demonstrated to be beneficial for detection and pose estimation under the synthetic-to-real setting [14,30].

The standard tool for rendering data for object pose estimation is BlenderProc [3]. The procedure is delineated in Figure 2. Conventional approaches either assume the availability of a texture; obtained through object reconstruction or by a texture file, or a color prior for textureless and metallic objects. Instead of relying on these strategies, we hypothesize that they are detrimental to the representations learned during training. Therefore, we uniformly sample from a set of n textures and apply those via UV-mapping for PBR rendering of the training data.

Scene composition and rendering are executed using [3] with the standard configuration of BOP [30]. Table 1 illustrates the list of object material parameters and their ranges utilized for scene composition. The specular and roughness values are assigned identical values to those used within the standard configuration¹. However, to induce a shape bias, color randomization is replaced with our own texture randomization strategy. Let $T = \{T_1, T_2, \dots, T_n\}$ be a set of textures and $M = \{M_1, M_2, \dots, M_k\}$ a set of objects. The random texture is drawn from T, with n being a hyperparameter denoting the number of textures. The training dataset for M is generated according to the number of scenes $N = \{N_1, N_2, \dots, N_j\}$ and that for the rendered views per scene. For each sampled scene O_m and object of M the texture is drawn from a uniform distribution $\mathcal{U}_{1,n}$ over T, for each sampled object of M_k . For every scene O_m , l images are rendered with randomized object and camera poses, specular, roughness, and color values or textures.

According to the standard configuration of [30] the obtained set of I is 50,000, obtained from l=25 images for each of the m=2000 scenes in O. Textures are taken from a Creative Commons online texture database (cc-textures)².

The number of random textures is n=1226. This number represents the complete set of available opaque textures in the database and has negligible computational overhead in comparison to a smaller number. As test sets, we use the ones provided by BOP.

5. Experiments

This section presents the empirical validation of our contributions. The experiments on object classification demonstrate the advantage of biasing CNNs toward shape with our approach over that of [8]. Following this, an analysis regarding the advantages and limitations of our approach for object detection and pose estimation of textureless and metallic objects is presented. Subsequently, the robustness to typical real-world image perturbations is demonstrated. Ultimately, a series of ablations are presented to ascertain the significance of the number of textures utilized for rendering, to effectively integrate the presented shape bias with online data augmentation, and to investigate the influence of the mesh origin, texture, and color.

5.1. Experimental Setup

The following paragraphs present the set of methods and the datasets utilized for the validation our hypothesis, and the metrics employed for comparison.

5.1.1 Methods

The classification experiments are conducted using ResNet50 [10]. The object detection results are provided with Faster R-CNN [28], RetinaNet [23], and YOLOx [6]. These three approaches were considered the state of the art at different points in time over the last few years. The same applies to our object pose estimation experiments where results are presented with GDR-Net [35] and Pix2Pose [27].

5.1.2 Datasets

In order to validate our hypothesis that training models for a shape bias is particularly beneficial for textureless and metallic objects, evaluations on TLESS [16] and ITODD [4] are presented. The TLESS dataset comprises 30 highly symmetrical textureless industrial objects. The ITODD dataset contains 28 metallic industrial objects. In both cases, the BOP test sets are employed for the evaluation, and the results for the texture bias baselines are obtained using the PBR training sets of the BOP [30]. These were generated using the rendering hyperparameters outlined in Section 4. The ground truth of the ITODD test set is not available to prohibit users from optimizing their data augmentation hyperparameters for the test data. Thus, evaluations can only be performed with the online evaluation tool

Ihttps://github.com/DLR-RM/BlenderProc/tree/
main/examples/datasets/bop_challenge

²https://cc0-textures.com/

Table 2. Object Detection. mAP ^{0.50:0.95} scores for Faster R-CNN [28], RetinaNet [23], and YOLOx [6] trained on TLESS and ITODD.
Results are provided for learning a texture and a shape bias, with and without tuned online data augmentations.

Dataset	Data		Faster R-CNN		RetinaNet		YOLOx	
	Augmentation	Bias:	texture	shape (ours)	texture	shape (ours)	texture	shape (ours)
TLESS	default		20.20	70.10	23.10	69.50	52.60	79.70
1LESS	tuned		80.50	77.90	78.40	74.10	80.20	80.60
ITODD	default		21.10	23.50	15.00	13.90	25.70	41.20
	tuned		46.20	46.60	41.50	44.20	53.10	56.30

of the BOP, which only provides metrics for object detection and pose estimation. Consequently, experiments for classification and the ablations are done using TLESS.

5.1.3 Metrics

For the classification experiments the top-1 and the top-5 accuracy in percentage are reported. The COCO evaluation metric [22], mean average precision (mAP), is reported for object detection. The intersection over union (IoU) range emploed is 0.5:0.95. For pose estimation, we present the Average Recall (AR) [13], the combination of the visible surface discrepancy (VSD) [13], maximum symmetry-aware surface distance (MSSD) [15], and the maximum symmetry-aware projection distance (MSPD) [15].

5.2. UV-Mapping versus Style Transfer

Table 3. **UV-Mapping versus Style Transfer.** Comparing top-1 and top-5 accuracy (Acc.) of style transfer for shape bias induction [7] and our approach on the TLESS dataset.

Bias	Top-1 Acc. (%)	Top-5 Acc. (%)
texture	63.22	86.27
shape [7]	75.88	91.16
shape (ours)	87.27	97.37

We conduct an experiment to verify our hypothesis that our UV-mapping approach is improving over style transfer for learning a shape bias. Table 3 presents results for ResNet50 [10] trained for object classification on TLESS. For classification, all object instances of the test set are cropped using the ground truth bounding box with a scaling factor of 1.0. This value correlates with the cropping ratio of ImageNet, which [7] has been designed for. Our experiments show that learning classification with a texture bias, using uniform color sampling, results in the worst top-1 and top-5 accuracy. Using AdaIN [19] for learning a shape bias with style transfer, as done by [7] improves over the texture bias variant. However, inducing a shape bias through our UV-mapping approach results in a 15.01% relative top-1 and a 6.81% relative top-5 accuracy improvement over [8].

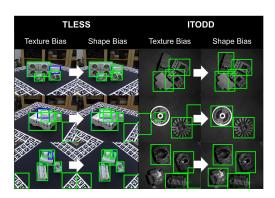


Figure 3. **Detection Example.** Examplary test images from TLESS and ITODD, with ground truth (blue) and predicted (green) bounding boxes using YOLOx. YOLOx yields more accurate bounding boxes, better recall and precision rates, when inducing a shape bias; detection and IoU thresholds are both set to 0.5.

5.3. Object Detection

Experiments using Faster R-CNN, RetinaNet, and YOLOx for object detection on TLESS and ITODD are conducted to identify the advantages and limitations of our method for shape bias induction.

For Faster R-CNN and RetinaNet the Image-Net pretrained ResNet50 backbone is used. For YOLOx the 1variant is used. All three object detectors are trained for 30 epochs, since validation accuracy converges in all cases until then. The remaining training parameters, are all identical to the default settings of MMDetection³, which are true to the original papers.

All experiments are conducted using the standard configuration for online data augmentation, the data augmentation of the winner of the BOP 2022 [24] when learning a texture bias. Tuned data augmentation is used to maximize the detection accuracy at the superposition with shape bias. More detail on the superposition of our shape bias and data augmentation is provided in Section 5.6.2.

Object detection results are presented in Table 2. For the default two-staged Faster R-CNN and the single-staged RetinaNet our shape bias improves by 247.03 and 200.87

³https://github.com/open-mmlab/mmdetection

Table 4. **Pose estimation performance.** Pose estimation of GDR-Net and Pix2Pose on TLESS and ITODD. Compared is the AR training with object color priors (default) and for texture agnosticity (none), for object detection (O.D.) and pose estimation (P.E.); YOLOx is used for object detection.

Dataset .	Bias		GDR-Net			Pix2Pose				
	O.D.	P.E.	AR _{MSPD}	AR _{MSSD}	AR _{VSD}	AR	AR _{MSPD}	AR _{MSSD}	AR_{VSD}	AR
	texture	texture	72.66	50.52	45.70	56.29	68.35	37.85	33.84	46.68
TLESS	texture	shape (ours)	73.54	52.94	48.09	58.19	67.71	38.25	34.10	46.69
-	shape (ours)	shape (ours)	73.73	53.05	48.17	58.32	71.22	41.86	37.30	50.13
ITODD	texture	texture	14.30	10.10	7.80	10.70	22.50	8.00	8.00	12.90
	texture	shape (ours)	14.30	10.10	7.80	10.70	22.70	10.10	10.30	14.30
	shape (ours)	shape (ours)	15.30	10.20	7.90	11.10	23.10	11.00	11.70	15.30

relative percent on TLESS, respectively. On ITODD the improvement in mAP for Faster R-CNN is 11.37%. However, the mAP of RetinaNet decreases by 7.91% when learning a shape bias for object detection on ITODD. Similarly, when using tuned data augmentations for TLESS our shape prior reduces the mAP by 3.34% for Faster R-CNN and 5.80% for RetinaNet. On ITODD the relative mAP improvement is 0.87% for Faster R-CNN and 6.51% for RetinaNet.

On the more recent YOLOx model, which is arguably the current standard for object detection, our shape bias improves the mAP for TLESS and ITODD using the default and the tuned data augmentation version. Comparing the default versions, our shape bias improves by a relative percentage of 51.52 for TLESS. The improvements on TLESS using a superposition of our shape bias and tuned online data augmentations are negligible to that of using the online data augmentations of [24] for training for a texture bias. However, on the metallic objects of ITODD our relative improvement is 60.31%, using the default, and 6.03% using the tuned training configuration. Interestingly, on TLESS using our method for learning a shape bias with all tested detectors is almost on par with using tuned data augmentations. As such, a valid alternative when detecting textureless objects. Metallic object detection on ITODD is more challenging, yet reaching the maximum mAP requires all tested detectors to learn a shape bias. The experiments demonstrates that YOLOx exhibits enhanced accuracy in comparison to the other two detectors. We conclude that approaches with higher learning capacities benefit more from the induced shape bias.

Figure 3 shows example images of TLESS and ITODD with bounding boxes predicted by YOLOx, with texture and shape bias. On TLESS the bounding box corner accuracy and the true positive rate are similar, yet the false negative rate is lower for the shape bias. On ITODD, the shape-biased YOLOx shows improved accuracy in bounding box corners, higher true positive rate, and lower false negative rate.

In conclusion, inducing object detectors with a shape bias not only improves object detection in the majority of the cases, but also presents a viable alternative to color augmentation tuning for such textureless objects as those of TLESS. For the industrial metallic objects of ITODD the combination of tuned online color augmentation and geometry bias is consistently demonstrated to be the optimal approach.

5.4. Object Pose Estimation

This section demonstrates the influence on object pose estimation of textureless and metallic objects when training for a shape bias. Results are presented for GDR-Net and Pix2Pose, integrated with the detections of YOLOx. The default training protocols as reported in [35] and [27] are employed, with GDR-Net using ResNet34 and Pix2Pose using ResNet50 as backbones. In both cases, one pose estimator is trained for each object. Table 4 shows that our shape-biased GDR-Net and Pix2Pose improve in AR using the texture-biased version of YOLOx for object detection. While the improvement with Pix2Pose is negligible, the improvement of GDR-Net is 3.38%. Further integrating the shape-biased YOLOx for detection enhances improvements to 3.61% for GDR-Net, and 7.39% for Pix2Pose. On ITODD, GDR-Net results in the same AR with texture and with shape bias. Integrating the detections of the shape-biased YOLOx improves AR by 3.74%. Using the texture-biased YOLOx, Pix2Pose improves by 10.85%. Integrating the detections of the shape-biased YOLOx the AR improvement is further enhanced to 18.60%. Pix2Pose shows a greater degree of accuracy improvement than GDR-Net. Analogous to our observations on object detection, Pix2Pose' results improve more due to the higher capacity of the feature extractor.

Figure 4 shows qualitative pose estimation results using YOLOx as detector and GDR-Net for pose estimation. On both, TLESS and ITODD, the alignment of the 3D bounding box, re-projected using the estimated pose, with the ob-

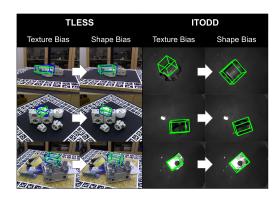


Figure 4. **Pose Estimation Examples.** Example images from the TLESS and ITODD datasets, showcasing the increased performance for occluded, dark and reflective object pose estimation; GDR-Net for pose estimation and YOLOx without color prior for detection; no detection ground truth for ITODD available.

ject improves. For TLESS the ground truth is additionally visualized in blue.

In conclusion, biasing representations learned by pose estimators for shape is either on par, or improves accuracy as compared to the standard, texture-biased versions. Integrating shape-biased object detectors with pose estimators consistently results in pose estimation accuracy improvement.

5.5. Robustness to Image Perturbations

To assess the robustness of pose estimators trained for shape bias, experiments with different common image perturbations are conducted. The perturbations are applied to the test images of TLESS, using ground-truth detections. Poses are estimated using GDR-Net. Figure 6 demonstrates that the shape bias yields higher robustness to Gaussian and Shot image noise. Pose estimates remain highly accurate even in the presence of considerable noise. In general, shape bias has little to no influence on the pose estimation accuracy when image blur, such as motion and Gaussian blur, is present. However, the presented shape bias increases robustness in the regime of high brightness changes, which is potentially useful for cases such as pose estimation in space [18].

5.6. Ablative Experiments

This section presents ablative studies. Reported are the influence of the number of random textures n, object detection and pose estimation at the superposition of our shape bias and online data augmentations, and the influence of the object geometry prior.

5.6.1 Number of Random Textures

Figure 5 shows the trend of YOLOx' mAP on TLESS when inducing a shape bias with a different number of random textures. With the availability of more textures mAP improves. Ultimately, using the n=1226 available cctextures yields the highest precision. However, using five random textures already results in 93.85% and 100 textures achieve 99.63% of the maximally achieved pose estimation accuracy.

5.6.2 Superposition of Shape Bias and Data Augmentation

This section provides a detailed analysis of the interplay of our shape bias with online data augmentation. We perform a grid search over the augmentation probabilities used by [24] and [35]. Table 5 lists the types and probabilities of augmentations applied to object detection and pose estimation. As done by GDR-Net [35], and adopted for YOLOx by [24], a coefficient $\lambda \in [0,1]$ for scaling the probability of the applied data augmentations is defined. The left part of Figure 7 demonstrates results for object pose esti-

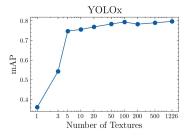


Figure 5. **Number of Textures.** Ablation on the influence of the number of random textures for YOLOx on TLESS

Table 5. **Data Augmentation.** Baselines for object detection and pose estimation; p denotes augmentation probability

Object Detection	Pose Estimation		
Augmentation	p	Augmentation	p
Coarse Dropout	0.5	Coarse Dropout	0.5
Gaussian Blur	0.4	Gaussian Blur	0.5
Enhance Sharpness	0.3	Add	0.5
Enhance Contrast	0.3	Invert	0.3
Enhance Brightness	0.5	Multiply	0.5
Enhance Color	0.3	Linear Contrast	0.5
Add	0.5		
Invert	0.3		
Multiply	0.5		
Gaussian Noise	0.1		
Linear Contrast	0.5		

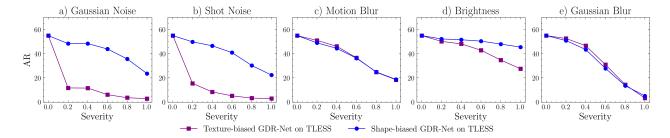


Figure 6. Shape Bias Robustness to Image Perturbations Influence of various image perturbations on AR scores with varying severity; GDR-Net with one pose estimation model trained per object; ground truth detections used for pose estimation

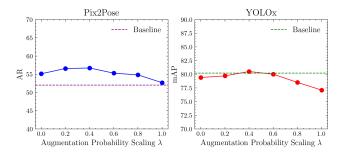


Figure 7. **Online Data Augmentation Severity.** Ablation study on the severity of online augmentation, for Pix2Pose and YOLOx on TLESS; ground truth detection priors for Pix2Pose.

mation with Pix2Pose, the right part shows that for detection with YOLOx on TLESS. For pose estimation a λ of $0.2 \le \lambda \le 0.4$ results in the most accurate estimates. For detection a λ of 0.4 achieves the best results. Generally, while shape-biased pose estimation improves with any set of online data augmentations, the data augmentation for object detection requires tuning to result in the highest mAP when training for a shape bias.

5.6.3 Influence of the Mesh Origin

This sections ablates the influence of the mesh origin and the type connected bias. Table 6 reports the AR when training Pix2Pose for TLESS. Ground truth detections are used as image location priors. The results show that inducing a shape bias with reconstructed objects improves over using the object prior and its reconstructed texture. The accurate geometry prior of CAD models is in general preferable over the reconstruction, yet again, training for a shape bias improves over training for a color bias; color bias refers to sampling object colors in grayscale, according to the parameters in Table 1 and [30].

6. Conclusion

This paper presents a simple and easily applicable strategy for improving pose estimation of textureless and metal-

Table 6. **Mesh Origin.** Pose estimation accuracy using Pix2Pose with diverse combination of object geometry and texture, on TLESS. The ground truth detections and an augmentation probability scaling factor of $\lambda=0.2$ are used.

Mesh Origin	Bias	AR
Reconstructed	texture	42.64
Reconstructed	none (ours)	46.57
CAD	color	59.24
CAD	none (ours)	62.95

lic objects. Inducing a texture bias to estimators by randomizing object textures during data rendering results in negligible computational overhead to the standard strategy. The improvements, however, are manifold. Object detection improves in most of the tested cases, particularly for the most recent detector, pose estimation is on par or improves for both tested object detectors, the pose estimators are more robust to image perturbations.

Future studies will investigate adversarial learning for biasing estimators for shape, in order to further improve detection and pose estimation accuracy and robustness. The presented study does not consider Vision Transformers and Diffusion models. Future work will account for them. Furthermore, since ITODD does not provide ground truth annotations, future work will provide more extensive evaluations for metallic objects.

Acknowledgement

We gratefully acknowledge the support of the EU-program EC Horizon 2020 for Research and Innovation under grant agreement No. 101017089, project TraceBot, the Austrian Science Fund (FWF), under project No. I 6114, project iChores, and the city of Vienna (MA23 – Economic Affairs, Labour and Statistics) through the research project AIAV (MA23 project 26-04).

References

- [1] Dominik Bauer, Timothy Patten, and Markus Vincze. VeRE-FINE: Integrating object pose verification with physics-guided iterative refinement. *IEEE Robotics and Automation Letters*, 5(3):4289–4296, 2020. 1
- [2] Sima Behpour, Kris M. Kitani, and Brian D. Ziebart. Ada: Adversarial data augmentation for object detection. In *Preceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1243–1252, 2019.
- [3] Maximilian Denninger et al. Blenderproc. *CoRR*, abs/1911.01911, 2019. 3, 4
- [4] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Hartinger, and Carsten Steger. Introducing MVTec ITODD

 A dataset for 3D object recognition in industry. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 2200–2208, 2017. 2, 4
- [5] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 3
- [6] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding YOLO series in 2021. *arXiv* preprint arXiv:2107.08430, 2021. 2, 4, 5
- [7] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In Proceedings of the International Conference on Learning Representations, 2019. 1, 2, 3, 5
- [8] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. Advances in Neural Information Processing Systems, 31, 2018. 2, 4, 5
- [9] 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.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4, 5
- [11] Zaixing He, Yue Chao, Mengtian Wu, Yilong Hu, and Xinyue Zhao. G-GOP: Generative pose estimation of reflective texture-less metal parts with global-observation-point priors. *IEEE/ASME Transactions on Mechatronics*, 29(1):154–165, 2023. 1, 3
- [12] Zaixing He et al. ContourPose: Monocular 6-D pose estimation method for reflective textureless metal parts. *IEEE Transactions on Robotics*, 39(5):4037–4050, 2023. 1, 3
- [13] Tomáš Hodaň et al. BOP: Benchmark for 6D object pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 19–34, 2018. 5
- [14] Tomáš Hodaň et al. Photorealistic image synthesis for object instance detection. In *Proceedings of the IEEE International Conference on Image Processing*, pages 66–70. IEEE, 2019. 1, 4

- [15] Tomáš Hodaň et al. BOP challenge 2020 on 6D object localization. Proceedings of the European Conference on Computater Vision Workshops, 2020. 5
- [16] Tomáš Hodaň, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pages 880–888, 2017. 2, 4
- [17] Yinlin Hu, Pascal Fua, and Mathieu Salzmann. Perspective flow aggregation for data-limited 6d object pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 89–106. Springer, 2022. 1
- [18] Yinlin Hu, Sebastien Speierer, Wenzel Jakob, Pascal Fua, and Mathieu Salzmann. Wide-depth-range 6D object pose estimation in space. In *Proceedings of the IEEE/CVF Con*ference on Computer Vision and Pattern Recognition, pages 15870–15879, 2021. 7
- [19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 5
- [20] Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. HomebrewedDB: RGB-D dataset for 6D pose estimation of 3D objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019. 3
- [21] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 574–591. Springer, 2020. 1
- [22] Tsung-Yi Lin et al. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755, 2014. 5
- [23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 2980–2988, 2017. 2, 4, 5
- [24] Xingyu Liu, Ruida Zhang, Chenyangguang Zhang, Bowen Fu, Jiwen Tang, Xiquan Liang, Jingyi Tang, Xiaotian Cheng, Yukang Zhang, Gu Wang, and Xiangyang Ji. Gdrnpp. https://github.com/shanice-l/gdrnpp_bop2022, 2022. 5, 6, 7
- [25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [26] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3DUnderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF* Conference on Computer Vision and Pattern Recognition, pages 55–64, 2020.
- [27] 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*, Oct 2019. 1, 2, 4, 6

- [28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 28, 2015. 2, 4, 5
- [29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1, 3
- [30] Martin Sundermeyer et al. Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 1, 2, 4, 8
- [31] Stefan Thalhammer, Dominik Bauer, Peter Hönig, Jean-Baptiste Weibel, José García-Rodríguez, and Markus Vincze. Challenges for monocular 6D object pose estimation in robotics. arXiv preprint arXiv:2307.12172, 2023. 1
- [32] Stefan Thalhammer, Markus Leitner, Timothy Patten, and Markus Vincze. PyraPose: Feature pyramids for fast and accurate object pose estimation under domain shift. In *IEEE In*ternational Conference on Robotics and Automation, pages 13909–13915, 2021. 2
- [33] Stefan Thalhammer, Timothy Patten, and Markus Vincze. COPE: End-to-end trainable constant runtime object pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2859–2869, 2023. 1, 2
- [34] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 23–30, 2017. 2, 3
- [35] 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. 1, 2, 4, 6, 7
- [36] Jun Yang, Wenjie Xue, Sahar Ghavidel, and Steven L. Waslander. 6D pose estimation for textureless objects on RGB frames using multi-view optimization. In *Proceedings* of the IEEE International Conference on Robotics and Automation, pages 2905–2912, 2023. 1, 3
- [37] Long Yang, Qingan Yan, Yanping Fu, and Chunxia Xiao. Surface reconstruction via fusing sparse-sequence of depth images. *IEEE Transactions on Visualization and Computer Graphics*, 24(2):1190–1203, 2018.
- [38] Haixin Yu, Shoujie Li, Houde Liu, Chongkun Xia, Wenbo Ding, and Bin Liang. TGF-Net: Sim2Real transparent object 6D pose estimation based on geometric fusion. *IEEE Robotics and Automation Letters*, 8(6):3868–3875, 2023. 1,
- [39] Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. Learning data augmentation strategies for object detection. In *Proceedings of the*

European Conference on Computer Vision, pages 566–583. Springer, 2020. 2