

Seeing Beyond Appearance – Mapping Real Images into Geometrical Domains for Unsupervised CAD-based Recognition

Benjamin Planche^{CO,1,2}, Sergey Zakharov^{CO,1,3}, Ziyang Wu¹,
Andreas Hutter¹, Harald Kosch², Slobodan Ilic^{1,3}

¹Siemens Corporate Technology

{benjamin.planche, andreas.hutter, slobodan.ilic, ziyang.wu}@siemens.com

²University of Passau

harald.kosch@uni-passau.de

³Technical University of Munich

sergey.zakharov@tum.de

Abstract

While convolutional neural networks are dominating the field of computer vision, one usually does not have access to the large amount of domain-relevant data needed for their training. It thus became common to use available synthetic samples along domain adaptation schemes to prepare algorithms for the target domain. Tackling this problem from a different angle, we introduce a pipeline to map unseen target samples into the synthetic domain used to train task-specific methods. Denoising the data and retaining only the features these recognition algorithms are familiar with, our solution greatly improves their performance. As this mapping is easier to learn than the opposite one (i.e. to learn to generate realistic features to augment the source samples), we demonstrate how our whole solution can be trained purely on augmented synthetic data, and still perform better than methods trained with domain-relevant information (e.g. real images or realistic textures for the 3D models). Applying our approach to object recognition from texture-less CAD data, we present a custom generative network which fully utilizes the purely geometrical information to learn robust features and achieve a more refined mapping for unseen color images.

1. Introduction

The ever-increasing popularity of deep convolutional neural networks seems well-deserved, as they are adopted for more and more complex applications. This success has to be slightly nuanced though, as these methods usually rely on large annotated datasets for their training. In many cases still (e.g. for scalable industrial applications), it would be extremely costly, if not impossible, to gather the required

data. For such use-cases and many others, synthetic models representing the target elements are however usually pre-available (industrial 3D CAD blueprints, simulation models, etc.). It thus became common to leverage such data to train recognition methods e.g. by rendering huge datasets of relevant synthetic images and their annotations.

However, the development of exhaustive, precise models behaving like their real counterparts is often as costly as gathering annotated data (e.g. acquiring precise texture information, to render proper images from CAD data, actually imply capturing and processing images of target objects). As a result, the salient discrepancies between model-based samples and target real ones (known as *realism gap*) still heavily impairs the application of synthetically-trained algorithms to real data. Research in *domain adaptation* thus gained impetus the last years. Several solutions have been proposed, but most of them require access to real relevant data (even if unlabeled) or access to synthetic models too precise for scalable real-world use-cases (e.g. access to realistic textures for 3D models).

In our work, we introduce a novel approach, *SynDA*, tackling domain adaptation and realism gap from a different angle. *SynDA* is composed of a custom generative network to map unseen real samples toward a relevant, easily-available synthetic domain (e.g. normal maps), in order to improve recognition for methods themselves trained on this noiseless synthetic modality. Along this paper, we demonstrate that it is sensible to train task-specific networks on noiseless information so they learn clean discriminative features, and then develop a mapping function from real to synthetic data; rather than to focus on developing or learning pseudo-realistic noise models to train against (though the two can be complementary to bridge the gap both ways).

Applied in this paper to CAD-based recognition in color pictures, our approach is based on the assumptions that real-world images can be mapped to the synthetic domain; and

^{CO}These authors contributed equally to the work.

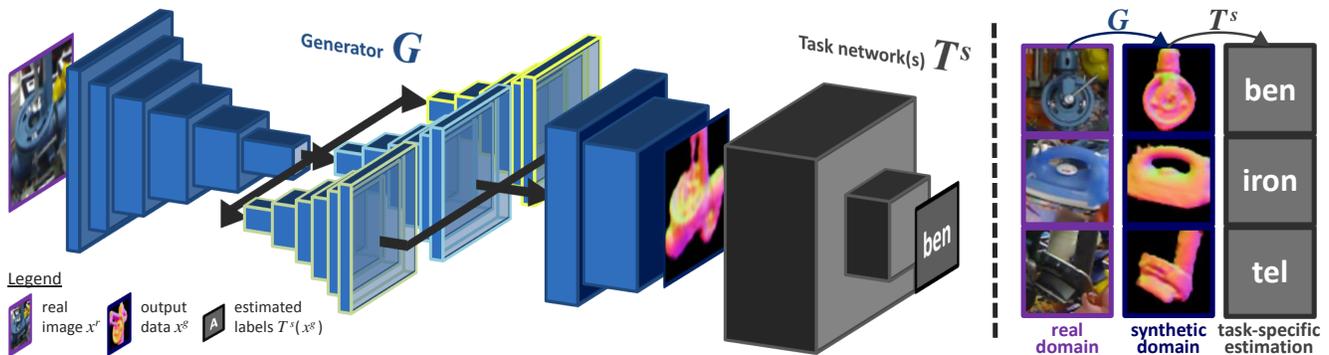


Figure 1: **Pipeline Usage.** We present G , a custom generative network which maps real unseen data into discriminative synthetic domains X^s available for training (*e.g.* normal maps rendered from provided CAD models). The pre-processed data can then be handed to any recognition methods T^s , themselves simply trained on X^s , to achieve high performance despite the lack of any domain-relevant training information.

that, in absence of any real training data, this mapping can be learned by recovering the synthetic samples altered by a stochastic noise source. Since our method only needs to eliminate noise and retain features, it performs better than usual generative solutions for domain adaptation, which learns the more difficult task of generating complex features to mimic the target data. As long as the synthetic domains contain all relevant features and as long as those features are contained in real images, our approach successfully enhances recognition, as demonstrated through our empirical evaluation. In summary, we are making the following contributions:

(a) Synthetic modality regression for domain adaptation

– We propose a novel framework to learn a mapping from unseen real data to relevant synthetic domains, denoising and recovering the information needed for further recognition. Our solution thus not only covers the real-synthetic gap, but also takes care of cross-modality mapping. More specifically, we present how color images can be mapped to normal maps, to help pose-regression and classification in absence of reliable texture information for training.

(b) Decoupling domain adaptation from recognition

– Most domain adaptation schemes constrain the methods training, by adding pseudo-realistic or noisy features to their training set, editing their architecture or losses, *etc.* In our framework, task-specific algorithms simply learn on available, relevant synthetic data (clean normal maps from CAD models in our case), while separately our network G is trained on noisy data to map them into the selected synthetic domain. This decoupling makes training more straightforward, and allows G to be used along any number of recognition methods. We furthermore observe better results compared to recognition methods directly trained on augmented data. Results even compare to solutions using real information for training.

(c) Performance in complete absence of real training

data – Domain adaptation approaches usually assume the realism gap to be already partially bridged, requiring access to some target domain images or realistic synthetic models (*e.g.* textured 3D data). Opting for recognition tasks with texture-less CAD data for only prior, we demonstrate how our pipeline can be trained on purely synthetic data and still generalize well to real situations. For that we leverage an extensive augmentation pipeline, used as a noise source applied to training samples so our solution learns to denoise and retain the relevant features.

(d) Multi-task network with self-attentive distillation

– The one advantage of synthetic models such as CAD data is the possibility to easily extract various precise modalities and ground-truths to learn from. We thus consolidate several state-of-the-art works [22, 18, 45, 48] to develop a custom generator with multiple convolutional decoders for each relevant synthetic modality (*e.g.* normal maps, semantic masks, *etc.*), and a distillation module on top making use of self-attention maps to refine the final outputs.

After providing a pertinent survey in Section 2 and describing our methodology in Section 3, we evaluate our solution in Section 4 over a set of different recognition methods and datasets to support our claims.

2. Related Work

Domain adaptation became an increasingly present challenge with the rise of deep-learning methods. We thus dedicate most of our literature review to listing main solutions developed to bridge the gap between real and synthetic data. In a second time, we present convolutional neural network (CNN) methods for shape regression, as we put emphasis on the mapping from real color images to synthetic geometrical domains in this paper.

Bridging the realism gap: The realism gap is a very well known problem for computer vision methods that rely on synthetic data, as the knowledge acquired on these modal-

ities usually poorly translates to the more complex real domain, resulting in a dramatic accuracy drop. Several ways to tackle this issue have been investigated so far. A first obvious solution is to improve the quality and realism of the synthetic models. Several works try to push forward simulation tools for sensing devices and environmental phenomena. State-of-the-art depth sensor simulators work fairly well for instance, as the mechanisms impairing depth scans have been well studied and can be rather well reproduced [24, 35]. In case of color data however, the problem lies not in the sensor simulation but in the actual complexity and variability of the color domain (*e.g.* sensibility to lighting conditions, texture changes with wear-and-tear, *etc.*). This makes it extremely arduous to come up with a satisfactory mapping, unless precise, exhaustive synthetic models are provided (*e.g.* by capturing realistic textures). Proper modeling of target classes is however often not enough, as recognition methods would also need information on their environment (background, occlusions, *etc.*) to be applied to real-life scenarios. For this reason, and in complement of simulation tools, recent CNN-based methods are trying to further bridge the realism gap by learning a mapping from rendered to real data, directly in the image domain. Mostly based on unsupervised conditional generative adversarial networks (GANs) [39, 38, 4] or style-transfer solutions [11], these methods still need a set of real samples to learn their mapping.

Other approaches are instead focusing on adapting the recognition methods themselves, to make them more robust to domain changes. For instance, solutions like *DANN* [10] or *ADDA* [41] are also using unlabeled samples from the target domain along the source data to teach the task-specific method domain-invariant features. Considering real-world and industrial use-cases when only texture-less CAD models are provided, some researchers [37, 40] are compensating the lack of target domain information by training their recognition algorithms on heavy image augmentations or on a randomized rendering engine. The claim is that with enough variability in the simulator, real data may appear just as another variation to the model. Considering similar applications (when no real samples nor texture information are available), our method follows the same principle, but applies it to the training of a domain-mapping function instead of the recognition networks. We demonstrate how this different approach not only improves the end accuracy, but also makes the overall solution more modular.

Regression of geometrical information: As no textural information is provided for training, we apply our domain adaptation method to the mapping of real cluttered color images into the only prior domain: the geometrical representation of target objects, extracted from their CAD data. The regression of such view-based shape information (*e.g.* normal or depth maps) from monocular color images is not

a new task in the field of computer vision, and it has been already explored by several works. The pioneer approaches tackled this complex mapping either by using probabilistic graphical models relying on hand-crafted features [16, 26], or by using feature matching between an RGB image and a set of RGB-D samples to find the nearest neighbors and warp them into a final result [17, 28]. Unsurprisingly, the latest works employ CNNs as a basis for their algorithms [8, 36, 23]. Eigen *et al.* [9] are the first ones to apply a CNN (the popular AlexNet [21]) to this problem, making predictions in a two-stage fashion: coarse prediction and refinement. This approach was further improved by additionally regressing labels and normals, with a refinement step for the final estimation [8].

Another way of improving the quality of predicted depth or normal data is to use neural networks together with graph probabilistic models. Liu *et al.* [27] use a unified Deep Convolutional Neural Fields (DCNF) framework based on the combination of a CNN and conditional random field (CRF) to regress depth from monocular color images of various scenes. Their pipeline consists of two sub-CNNs with a common CRF loss layer, and yields detailed depth maps. Building on the previous framework, Cao *et al.* [6] train the DCNF model jointly for depth regression and semantic segmentation, demonstrating how joint training can improve the overall results. Similarly, Kendall *et al.* [18] proposed a multi-task Bayesian network approach (including depth regression) which weighs multiple loss functions by considering the uncertainty of each task. Another way of efficiently combining a multi-task output was presented in [45], which uses so-called distillation modules to supervise and improve the output result. Unfortunately, all aforementioned methods require real labeled images from the target domain for their training, which is too strong a constraint for real-life scalable applications. Our own method does build upon their conclusions [8, 42, 12, 22, 18, 45], making use of a custom cross-modality network with advanced distillation to learn a robust mapping from noisy domains to synthetic ones.

3. Methodology

Driven by the necessity of learning only from synthetic data for scalable recognition processes, we developed a method to map unseen real samples (*e.g.* color images) into the noiseless synthetic domain (*e.g.* normal maps rendered from CAD models) the task-specific solutions were trained on (*c.f.* Figure 1), to enable recognition.

Following the same formalization as in [47], let $X_c^s = \{x_{c,i}^s \mid \forall i \in N_c^s\}$ be a dataset made of a number N_c^s of uncluttered, noiseless training samples x_c^s of class c . Let $X^s = \{X_c^s \mid \forall c \in C\}$ be the complete clean training dataset. We similarly define X^r the set of target C -related real data, completely unavailable for training. Note

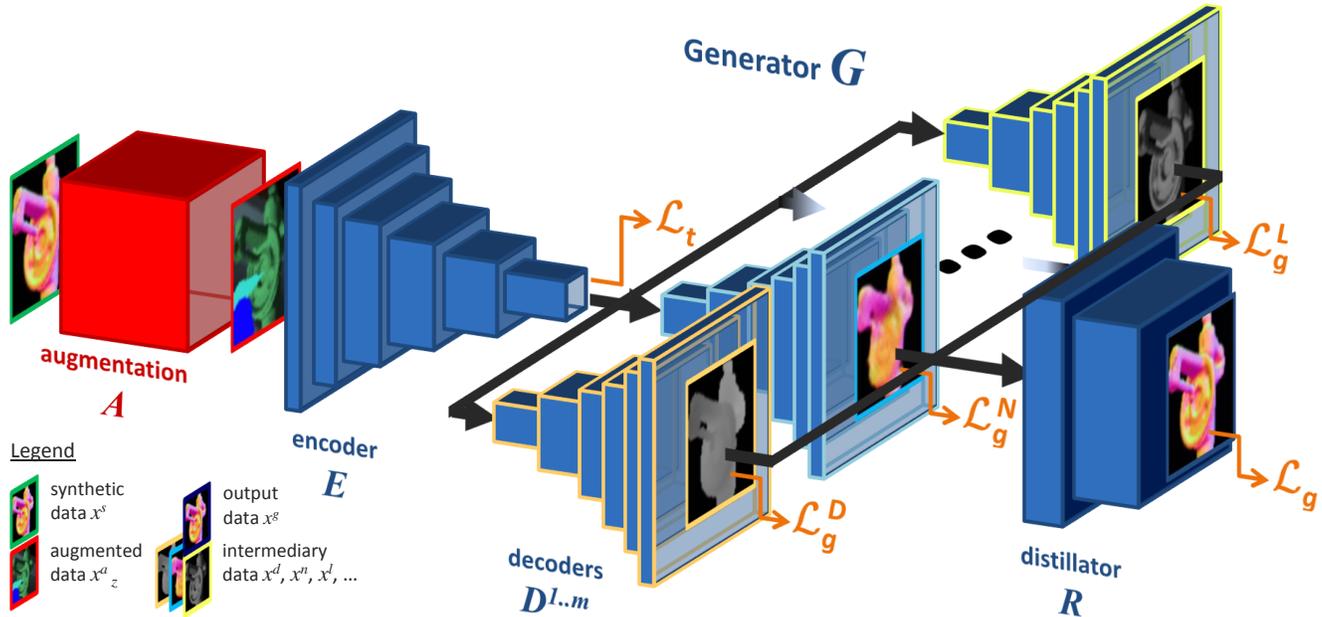


Figure 2: **Training of our network G .** Taking full advantage of available synthetic data *e.g.* texture-less CAD models, G consists of a custom multi-modal pipeline with self-attentive distillation, trained to recover noiseless geometrical and semantic modalities from randomly augmented synthetic samples (detailed architecture in Figure 3).

that samples x^r can also be of a different modality than x^s (*e.g.* x^r being color images while x^s being normal maps, when no texture were available to render synthetic color images). Finally, let $T(x; \theta_T) \rightarrow \tilde{y}$ be any recognition algorithm which given a sample x returns an estimate \tilde{y} of a task-specific label or feature y (*e.g.* class, pose, mask image, hash vector, *etc.*). We define as T^s the method trained on noiseless X^s .

Given this setup, our pipeline trains a function G purely on synthetic data (and thus in an unsupervised manner), to learn a mapping from complex C -related instances to their corresponding clean signal (*c.f.* Figure 2). To achieve this when no domain-relevant data is available for training, we describe in this section how G is trained against a data augmentation pipeline $A(x^s, z) \rightarrow x^a_z$, with z a noise vector randomly defined at every training iteration and x^a_z the resulting noisy data. Our training approach assumes that G removes the artificially introduced noise z such that only the original synthetic signals x^s are retained. Thus, G can be seen as a noise filter that removes unneeded elements in input data, and can be also applied over the domain X^r of real samples as long as synthetic information can be extracted from them. We demonstrate that, in the case of CAD-based visual recognition, we can indeed define a new generative method G fully utilizing the synthetic modalities, and a complex and stochastic augmentation pipeline A to train G against, such that G maps real images into the learnt synthetic domain with high accuracy. We even demonstrate how this process increases the probability that

$T^s(G(x^r)) = \tilde{y}^g$ is accurate compared to $T^a(x^r) = \tilde{y}^r$, with T^a the task-specific algorithm directly trained on data augmented by A . Though we focus the rest of the paper on CAD-based visual recognition for clarity, the principles behind our solution can be directly applied to other use-cases.

3.1. Real-to-Synthetic Mapping through Multi-Modal Distillation

3.1.1 Multi-Modal U-Net Architecture

As demonstrated by previous works in multi-task learning [8, 42, 12, 22, 18, 45], it is often advantageous to train a network on several tasks (even when considering only one), as the synergy between the tasks can make each of them perform better, and make the common layers for feature extraction more robust and focused on abstract, cross-modality information.

We thus adopt such a scheme to guide our generator in its main task of extracting the chosen synthetic features from noisy samples. Not limited to the scarce, sometimes imprecise, annotations of real training datasets, we can rely on a multitude of different synthetically-rendered modalities. For industrial CAD-based recognition, G would learn to map real images into a geometrical domain (normal and/or depth maps), using for sub-tasks the regression of depth and normal maps, semantic or contour mask, *etc.* (*c.f.* Section 3.2).

Inspired by previous multi-modal generative pipelines [22, 18, 45], our network is composed of a

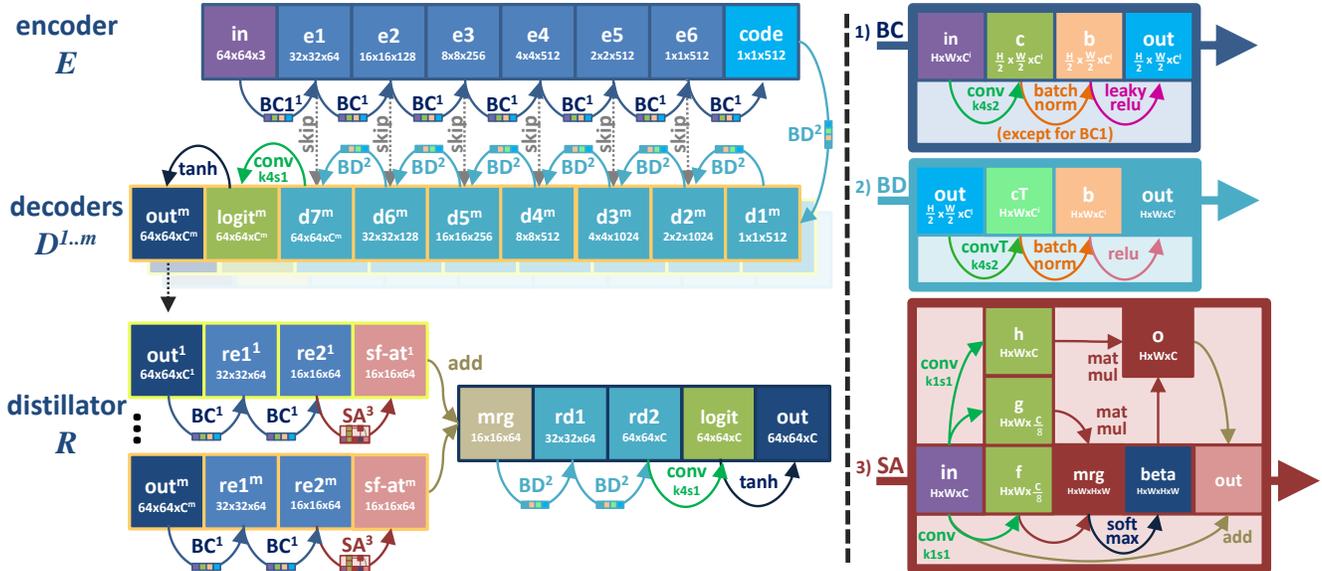


Figure 3: **Detailed architecture of our network G** , on the left. Reused layer blocks (BC for encoding, BD for decoding, SA for self-attention) are detailed on the right. $conv\ k4s2$ stands for a convolutional layer with 4×4 filters and a stride of 2; $convT$ stands for a transposed convolution.

single convolutional encoder E and m decoders D^{mod} , with m the number of sub-tasks. For the rest of the paper, we only consider up to 4 sub-tasks—normal and depth regression, semantic segmentation, foreground lightness evaluation—though it would be straightforward to add more (e.g. contour extraction as in [45]).

In our solution, each intermediary modality is fully decoded in order to be compared to its synthetic ground-truth. Each generative loss \mathcal{L}_g^{mod} (L1 distance for images, cross-entropy for binary masks) is back-propagated through its decoder, then jointly through the common encoder (c.f. Figure 2).

A triplet loss \mathcal{L}_t is optionally added at the network bottleneck to improve the feature distribution in the embedding space, using task-specific metrics to push apart encoded features of images from semantically-different images, while bringing together features of similar elements.

$$\mathcal{L}_t(E) = \sum_{(x_b, x_p, x_n) \in X} \max \left(0, 1 - \frac{\|E(x_b) - E(x_n)\|_2^2}{\|E(x_b) - E(x_p)\|_2^2 + m} \right) \quad (1)$$

with x_b the input image used as binding anchor, x_p a positive or similar sample, x_n a negative or dissimilar one, and m the task-specific margin setting the minimum ratio for the distance between similar and dissimilar pairs of samples. For instance, for the task of instance classification and pose estimation (ICPE), we set $m = 2 \arccos(|q_b \cdot q_p|)$ if x_b and x_p are images of the same class, else $m = n$ (with q_b and q_p the pose quaternions corresponding to x_b and x_p , and $n > \pi$ a fixed margin).

Further distinguishing our solution from usual multi-modal auto-encoders, we add skip connections from each encoding block to its reciprocal decoding block. As demonstrated in previous works [22, 49], passing high-resolution features from the contracting layers along the outputs of previous decoding blocks not only improves the training by avoiding vanishing gradients, but also guides the decoding blocks in upsampling and localizing the features. We observe a clear performance boost from this change, as shown in Table 3.

3.1.2 Distillation with Self-Attention

If training the target decoder along others already improves its performance by synergy, several works [45, 12, 31] demonstrated how one can further take advantage of multi-modal architectures by adding a distillation module on top of the decoders, merging their outputs to distill a final result.

In their work [45], *Pad-Net* authors present several distillation strategies, with the most efficient one making use of attention mechanism [30, 2, 29] to better weigh the cross-modality merging, bringing forward the most relevant features for the final modality.

Using this insight, we built our own module R to refine the target results from the partially re-encoded intermediary modalities by using self-attention computations [7]. This mechanism, adapted by Zhang *et al.* [48] for image generation and detailed in Figure 3, is used to efficiently model relationships between widely-separated spatial regions. Given a feature map $x \in \mathbb{R}^{C \times H \times W}$, the output of

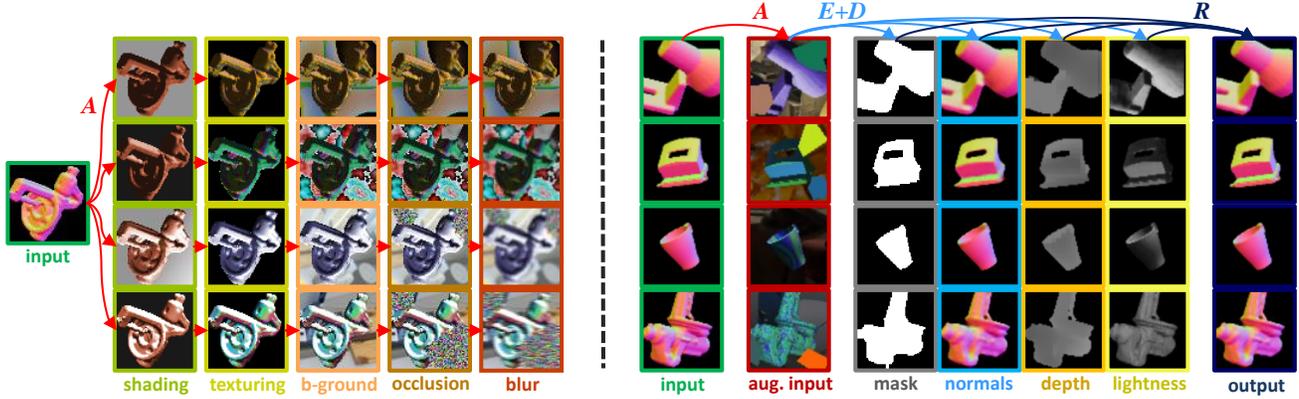


Figure 4: **Augmentation and training results.** On the *left*, we demonstrate how normal maps are step by step transformed into complex, random color images by our online augmentation pipeline. On the *right*, we present how G is trained on these images, learning to map them back to the noiseless geometrical information.

the self-attention operation is:

$$x_{sa} = x + \gamma \cdot \sigma((W_f * x)^T \cdot (W_g * x)) \cdot (W_h * x) \quad (2)$$

with σ the *softmax* activation function; $W_f \in \mathbb{R}^{\bar{C} \times C}$, $W_g \in \mathbb{R}^{\bar{C} \times C}$, $W_h \in \mathbb{R}^{C \times C}$ learned weight matrices (we opt for $\bar{C} = C/s$ as in [48]); and γ a trainable scalar weight. Instantiating and applying this process to each re-encoded modality, we sum the resulting feature maps, before decoding them to obtain the final output. This new distillation process not only allows to pass messages between the intermediary modalities, but also between distant regions in each of them.

Our distillator is trained jointly with the rest of the generator, with a final generative loss \mathcal{L}_g (L1 distance here) applied to the distillation results. Not only our whole generator can thus be efficiently trained in a single pass, but no manual weighing of the sub-task losses is needed, as the distillator implicitly covers it (this furthermore suits our use-cases, as manual fine-tuning is technically possible only when validation data from target domains are available).

3.2. Learning from Purely Geometrical CAD Data

3.2.1 Synthetic Data Generation

The aforementioned architecture has been developed to especially shine for one particular use-case, poorly covered in the literature despite being common in industrial applications: the training of recognition methods on pure 3D CAD data, *i.e.* without any real relevant images and their annotations, nor captured textures for the 3D models to render realistic images. Despite the apparent meagerness of the available training data, covering only the geometrical aspects of target classes with no appearance information, it is still possible to render multiple synthetic modalities from the CAD models in order to build a rich annotated dataset,

to guide the training of complex generative networks such as our proposed one.

Without any relevant texture information, usual dataset rendering and training methods for the color domain cannot be directly applied. Since only geometrical information is made available, we select the surface normal and/or depth domains as target modality for the mapping performed by G . For this reason and similarly to other view-based training methods from CAD data [43, 46, 47], we use a simple 3D engine to generate noiseless normal and depth maps for each class c from a large set of relevant viewpoints (*e.g.* defined as vertices of an icosahedron centered on the target elements). This dataset of geometrical mappings is both used as ground-truth for the final outputs of G and some of its sub-tasks, and as inputs for the augmentation pipeline A deployed when no color data is available to train G .

3.2.2 Online Color Rendering and Augmentation

$A(x^s, z) \rightarrow x_z^a$ is an extensive online augmentation pipeline we designed, parametrized by a noise vector z randomly sampled at every call from a k -dimensional finite set \mathbb{Z}^k , with k the number of augmentation parameters. In order to make up for the complete lack of appearance and clutter information, A follows the principle of *domain randomization* [40], *i.e.* it is meant to add enough visual variability to the training inputs so that the trained method can generalize to real unseen samples. This means conceiving an augmentation pipeline with a large enough $|\mathbb{Z}^k|$. In our case, A first dynamically transforms the input geometrical views into color images through random shading and texturing, before applying further noise and clutter to the images, in order to prepare G for the complexity of real data.

To maximize the training variability, A is built to run in parallel of GPU-based trainings (*online*), providing new randomized samples every iteration (unlike *offline* solu-

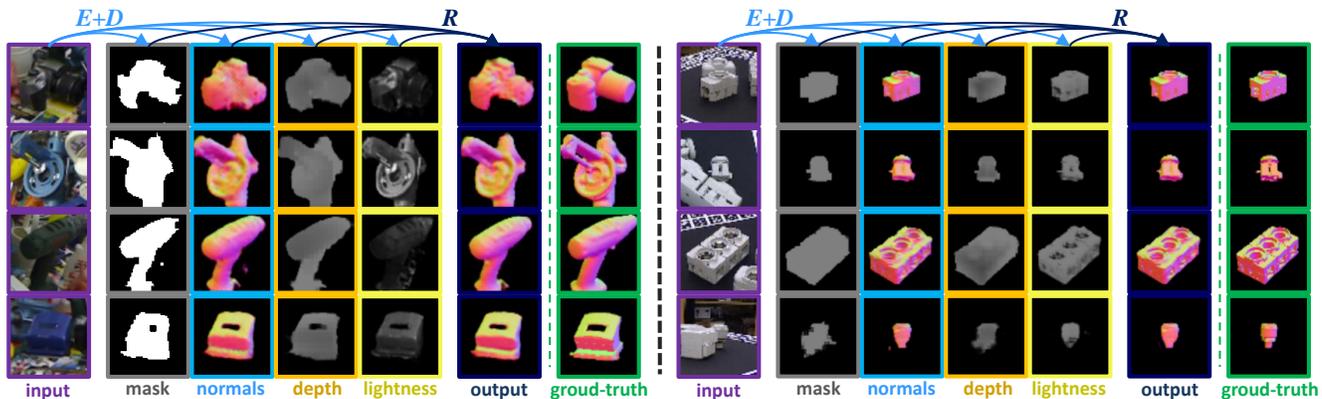


Figure 5: **Qualitative results** (intermediary and final mappings), when applying G purely trained on synthetic data to real samples, on LineMOD [14] and T-LESS [15] datasets.

tions, generating a fixed training dataset beforehand). Inspired by the literature both in computer vision [32] and computer graphics [3, 34], the following operations are thus composing A (illustrated in Figure 4):

Simple random shading: A first takes the provided normal maps and convert them into color images by applying simple Blinn-Phong shading [3]. Randomly sampling ambient and directional light sources, as well as diffusion and specular color factors for the objects, the provided surface normals are used to compute the diffuse and specular lightness maps through direct matrix products. Since distance information is lost in normal maps, this shading model is simplified by supposing the light sources at an infinite distance, hence the same light source vector for every surface point. This way, one can easily simulate an infinity of lighting conditions, returning the resulting lightness map.

Stochastic texturing: Given the lack of relevant texture information, random texture maps are procedurally generated using noise functions *e.g.* fractal Perlin noise [34] and Voronoi texturing. Downsampled to 2D vector maps, the original normals are used to index the generated textures; to achieve a more “organic” appearance, with patterns sometimes following some of the shape features.

Background addition: To simulate cluttered scenes, backgrounds are added to the rendered images, either re-using the previously-introduced noise functions, or using random patches from any publicly available image dataset (*e.g.* COCO [25]). Lightness maps from the shading step are furthermore used to homogenize the background brightness.

Random occlusion: Occlusions are introduced to further simulate clutter, but also so that G can learn to recover hidden or lost geometrical information. Based on [32], occluding polygons are generated by walking around the image, taking random angular steps and random radii at each step; then by painting it on top of the images with color noise.

Blur: To reproduce possible motion blur or unfocused images, Gaussian, uniform or median blur is applied with vari-

able intensity.

4. Evaluation

Pursuing our application of *SynDA* to CAD-based recognition tasks, we evaluate our method on two different tasks of localized object classification and pose estimation, opting for well-known algorithms on datasets commonly used in this domain [43, 4, 47]. First presenting concise qualitative observations, we then quantitatively and extensively evaluate *SynDA* through a comparison of its performance to state-of-the-art solutions depending on the available training modalities, and through an ablation study.

4.1. Experimental Setup

4.1.1 Instance Classification (IC) on T-LESS

As a preliminary experiment, we consider localized classification on T-LESS [15], a dataset of industrial objects with texture-less CAD models and RGB-D images from different complex scenes. Strong textural and geometrical similarities between the objects and heavy occlusions make it a challenging dataset for geometry-based classification. As in [47], we select the first 3 scenes and their 11 objects, building a set of 5,514 RGB patches of objects occluded up to 60%. For our task-specific method, we opt for the well-known ResNet [13], with 9 residual blocks.

4.1.2 Instance Classification and Pose Estimation (ICPE) on LineMOD

LineMOD [14] contains 15 mesh models of distinctive textured objects, along their RGB-D sequences and camera poses. We take advantage of this dataset to demonstrate how texture information is too often taken for granted in CAD-based application, and how its absence can heavily impact usual methods (*e.g.* in industrial settings). LineMOD has 4 symmetric objects. While some works simply removed

Table 1: **Quantitative comparison of recognition pipelines**, depending on the available training data, for the task of localized instance classification on T-LESS [15] with T ResNet9 network [13]. *Methods are explicitly described in annex.*

Training			Classification accuracy
Data	Method		
	$T(X_{train}^r) \rightarrow y$	(\emptyset)	99.34%
	$T^d(A(X^s); X_{train}^r) \rightarrow y, c_{dom}$	(DANN)	60.58%
	$G^{pix}(X^s, X_{train}^r) \rightarrow X^{r'}, T(X^{r'}) \rightarrow y$	(PixelDA)	63.12%
	$G^{pre}(X_{other}^r) \rightarrow X^s, T(X^s) \rightarrow y$	(Iro <i>et al.</i>)	36.03%
	$T(A(X^s)) \rightarrow y$	(\emptyset)	53.81%
	$G(A(X^s)) \rightarrow X^s, T(X^s) \rightarrow y$	(ours)	71.78%

these ambiguous elements for evaluation [4, 5], others only constrain the real views by keeping the unambiguous poses for these 4 objects [43, 46, 47]. We opt for the latter solution, to highlight the generalization capabilities of *SynDA* w.r.t. the number of objects. To further demonstrate that our method is tailored neither to a dataset nor to a recognition method, we select a different solution, the so-called *triplet* CNN [43, 46], which uses the aforementioned triplet loss \mathcal{L}_t , to map images to an embedding space which enforces separation for distinct classes and poses.

4.1.3 Tasks Preparation

For both tasks, synthetic training data X^s (normal, depth and semantic maps) are generated with a basic 3D engine (while lightness maps are obtained from A), following the procedure described in Section 3.2, taking into account in-plane rotations. For each task, the real color images are split 50/50 into a test set X_{test}^r , and a training set X_{train}^r for comparative methods which require real data. Similarly, we render $X^{s,t}$, realistic images from textured models to train some opponent methods on.

4.2. Qualitative Observations

Qualitative results can be found in Figures 1, 4, 5, as well as in the supplementary material. For both datasets, our method clearly learns to recover the clean geometrical features of target objects in unseen real images (Fig. 5), even though it has been trained with no information about the real domain (Fig. 4). The monochrome appearance of T-LESS objects may make the task easier; but as this information is not known during training, G is still trained on random noisy textures, and yet manages to map the real samples and even to recover occluded parts. As demonstrated on

LineMOD, our solution indeed learns to ignore visual properties such as textures to retain the synthetic features, using the prior CAD data.

GAN-based domain adaptation methods such as *PixelDA* [4] fail to learn their opposite mapping when only geometrical properties are provided, as shown in the annex. Indeed, learning both to add clutter and assign the proper texture to each class is a much more complex task, which would require further supervision (for instance, the foreground-similarity loss of *PixelDA* cannot be used in this setting to guide the network).

Finally, we can visually observe the improvements between the intermediary normal maps (directly from decoder D^N) and the refined outputs after self-attentive distillation, both in terms of segmentation and internal details. As mentioned in Section 3.1.1, one could easily add or replace intermediary modalities (for instance, regressing the objects lightness maps may not seem fully relevant, though it can be used to provide the latest layers of the network with information from the original color domain).

4.3. Comparison with other Domain Adaptation Approaches

Given the two pre-defined evaluation tasks, we quantitatively evaluate the performance of our pipeline, and compare it with usual, state-of-the-art methods, depending on the available training data (real images, corresponding annotations, CAD models, corresponding realistic textures, or real images from a different domain). For each setup, the same task-specific network is used (ResNet for IC on T-LESS, Triplet CNN for ICPE on LineMOD), trained by itself, against our augmentation pipeline A (with texturing augmentation disabled for pre-textured data), or along some auxiliary generators or sub-networks for domain adaptation

Table 2: **Quantitative comparison of recognition pipelines**, depending on the available training data, for the task of localized instance classification and pose estimation (ICPE) on LineMOD [14] with T triplet CNN [43, 46].

Training			Angular error		Classification accuracy
Data	Method		Median	Mean	
	$T(X_{train}^r) \rightarrow y$	(\emptyset)	9.50°	12.42°	99.72%
	$T^d(A(X^{s,t}); X_{train}^r) \rightarrow y, c_{dom}$	(DANN)	14.33°	30.45°	89.84%
	$G^{pix}(X^{s,t}, X_{train}^r) \rightarrow X^{r'}, T(X^{r'}) \rightarrow y$	(PixelDA)	15.38°	35.17°	91.06%
	$T^d(A(X^s); X_{train}^r) \rightarrow y, c_{dom}$	(DANN)	43.63°	68.59°	40.13%
	$G^{pix}(X^s, X_{train}^r) \rightarrow X^{r'}, T(X^{r'}) \rightarrow y$	(PixelDA)	95.14°	97.36°	35.39%
	$T(X^{s,t}) \rightarrow y$	(\emptyset)	88.62°	92.35°	43.62%
	$T(A(X^{s,t})) \rightarrow y$	(\emptyset)	70.18°	84.22°	49.11%
	$G^{pre}(X_{other}^r) \rightarrow X^s, T(X^s) \rightarrow y$	(Iro <i>et al.</i>)	52.43°	71.69°	41.49%
	$T(A(X^s)) \rightarrow y$	(\emptyset)	41.23°	67.50°	34.38%
	$G(A(X^s)) \rightarrow X^s, T(X^s) \rightarrow y$	(ours)	13.37°	27.46°	91.28%

(e.g. for *PixelDA* [4] or DANN [10]; for T used with a pre-trained monocular-RGB-to-depth generator G^{pre} [23]; or for *SynDA*).

For both tasks, we consistently observe the positive impact of *SynDA* on recognition, as shown in Tables 1-2. Despite being trained on the scarcest data, with the largest domain gap, our generator G brings the performance of the task-specific methods T above other solutions trained on more relevant information. The accuracy improvement is even more apparent for the pose regression task, as our pipeline precisely recovers geometrical features. It also appears clear that decoupling data augmentation and recognition training is beneficial, as illustrated by the accuracy difference between the two last lines of each table. This follows our initial intuition on the logic of teaching task methods in the available clean synthetic domain, while learning in parallel a mapping to project real data into this prior domain. This separation furthermore makes it straightforward to train new task-specific methods, with G ready to be plugged on top.

4.4. Architecture Validation through Ablation

Table 3 presents the results of an extensive ablation study done on our novel network architecture. By consolidating several state-of-the-art works on generative networks [22, 18, 45, 48], we developed a robust architecture to tackle extreme domain mappings (e.g. real RGB to synthetic normals).

As mentioned in Section 4.2, we can observe how the addition of decoders for auxiliary tasks improves the final output by synergy. The inclusion of self-attention mecha-

nism (SA layers) in the distillation module further enhances this effect, weighting the contribution of features between intermediary modalities, but also between distant internal regions. Finally, the benefits of passing messages directly between each encoder block and their opposite block for each decoder $D^{1..m}$, through the use of *skip* layers (c.f. U-Net architectures [22, 49]), is clearly highlighted in the table, as well as the use of a triplet loss \mathcal{L}_t at the bottleneck to improve the quality of the embedding space.

All in all, our network relies on a powerful multi-task architecture, structured to tackle real-to-synthetic mapping challenges, by utilizing any available synthetic modalities to learn robust features. One could easily build on this solution by considering additional or more use-case relevant sub-tasks (e.g. contour regression, part segmentation, *etc.*).

5. Conclusion

We present *SynDA*, a novel strategy for complex domain adaptation scenarios. Applied to several CAD-based recognition tasks and making use of a state-of-the-art generative network, our solution outperforms other supervised or unsupervised methods. For the challenging task of localized instance recognition and pose estimation e.g. on RGB LineMOD data, *SynDA* more than doubles the angular and class accuracy compared to other methods trained on synthetic data, and even surpasses previous domain adaptation methods requiring real data.

This is made possible by tackling the domain mapping from the opposite direction, using our custom generator to denoise unseen real samples and retain only the recognition-relevant features available during training.

Table 3: Architectural ablation study, with the “IC on LineMOD” task.

Encoder	Decoders				Distill.	Layers		Losses		Angular error		Classification accuracy
	D^N	D^D	D^M	D^L		R	\overrightarrow{SA}	\overrightarrow{skip}	$\mathcal{L}_g^{1..m}$	\mathcal{L}_t	Median	
✓	✓					✓		✓		15.75°	32.80°	87.35%
✓	✓					✓		✓	✓	15.76°	33.76°	88.04%
✓	✓	✓			✓	✓	✓	✓	✓	14.32°	30.31°	89.00%
✓	✓		✓		✓	✓	✓	✓	✓	14.48°	30.71°	89.32%
✓	✓	✓	✓		✓	✓	✓	✓	✓	14.22°	29.26°	89.67%
✓	✓	✓	✓	✓		✓		✓	✓	14.66°	30.83°	88.59%
✓	✓	✓	✓	✓	✓			✓	✓	16.07°	33.22°	87.69%
✓	✓	✓	✓	✓	✓		✓	✓	✓	14.43°	29.56°	90.38%
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	13.37°	27.46°	91.28%

Supplementary Material

A. Schematic Overview of the Different Gap-Bridging Methods

Table S1 contains a schematic comparison of the training and testing solutions for recognition tasks addressed in the paper, depending on the type of data available for training (real images, corresponding annotations, CAD models, corresponding realistic textures, or real images from a different domain).

While usual domain adaptation methods such as *PixelDA* [4] or *DANN* [10] focus on use-cases when real-world data are available—in terms of unlabeled target images but also realistic textures for 3D CAD models—we develop our solution on the assumption of minimal information on the target domain. For CAD-based recognition applications, we demonstrate in the main paper how *SynDA* yields state-of-the-art results when only pure geometrical data is available. For comparison, both *PixelDA* [4] and *DANN* [10] fail to map real and synthetic domains when provided with texture-less rendered images and real pictures, as illustrated in Figure S1.

B. Implementation Details

This section contains more in depth details on network architecture and parameters, augmentation pipeline, as well as synthetic data generation.

B.1. Network Architecture and Parameters

Figure 3 of the paper already provides the readers with an exhaustive overview of our state-of-the-art architecture, layer by layer. Our solution is implemented in Python using the TensorFlow framework [1].

Layer parameterization:

- All Convolution layers have 4×4 filter kernels;
- All Dropout layers have a dropout rate of 50%;
- All LeakyReLU layers have a leakiness of 0.2;
- Input and output images are 64×64 px, normalized between -1 and 1 .

Training parameters:

- Weights are initialized from a zero-centered Gaussian distribution, with a standard deviation of 0.02 ;
- The Adam optimizer [20] is used, with $\beta_1 = 0.5$;
- The base learning rate is initialized at $2e^{-4}$.

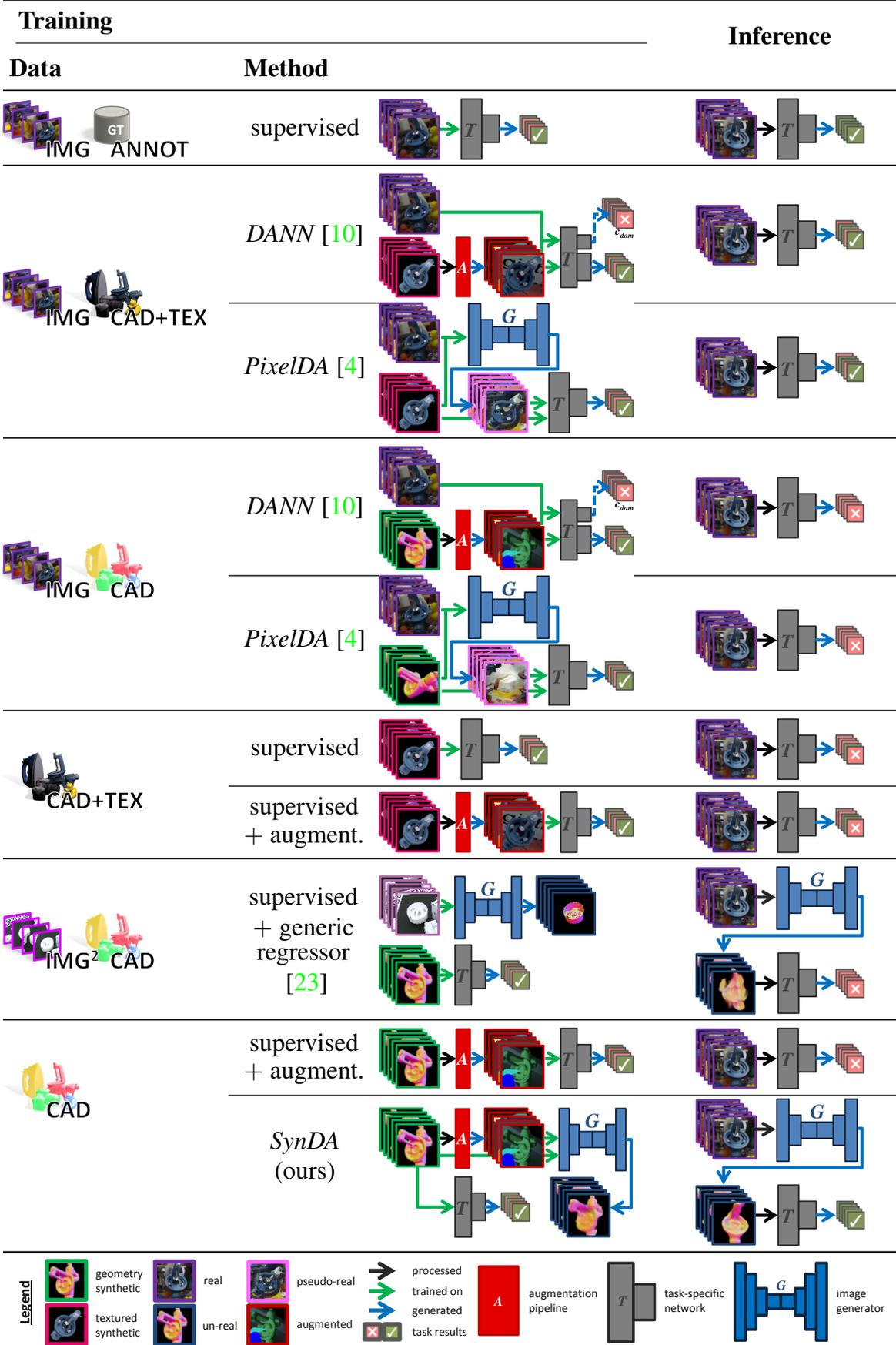
B.2. Augmentation Pipeline Details

Simple random shading: To generate a virtually infinite dataset of color images, we transform our set of geometrical maps (*i.e.* normal maps) into random images with an online augmentation procedure. Its main and first operation thus consists in shading, *i.e.* generating lightness maps from the normal maps, sampling random light conditions. It is done through a custom Blinn-Phong shading [3] method, as described in Algorithm 1.

Lighting parameters, defined through the augmentation noise vector z , are sampled using uniform distributions *e.g.* $\mathcal{U}(0.05, 0.3)$ for each color components of a , $\mathcal{U}(0.1, 0.8)$ for each component of d , $\mathcal{U}(0, 0.1)$ for each component of s , $\mathcal{U}(0.9, 1.1)$ for each component of s_p , *etc.*

Stochastic texturing: Random textures are applied to the objects every iteration. They are noise-generated, using the open-source FastNoise library [33]. In particular, Perlin noise [34], cellular noise [44], and white noise are used,

Table S1: Visual comparison of recognition schemes, depending on the available training data.



sampled from the uniform distribution $\mathcal{U}(0.0001, 0.1)$, to obtain hue and saturation maps. These maps are either directly merged to the lightness map of the objects to obtain their final HSL appearance, or either used as texture maps. In the second case, the original surface normals are projected on two dimensions (randomly dropping their X , Y or Z axis), to be used as an UV map for the texture.

Background addition: As mentioned in the paper, random backgrounds are added, either generated using the aforementioned noise sources or obtained by randomly cropping and resizing color images from public datasets (e.g. COCO [25]). The background pixel values are multiplied by the normalized foreground brightness, to get more homogenized results.

Random occlusions: To generate occlusions we use a function from [32], where they generate 2D obstacles for a drone moving planning simulation. The points p are sampled by walking around the circle taking random angular steps and random radii at each step. A polygon is then generated using the points p and filled with either random noise (c.f. main paper) or textures if provided.

The polygon’s complexity is defined by two parameters: σ (“spikeyness”), which controls how much point coordinates vary from the radius r_{ave} , and ϵ (“irregularity”), which sets an error to the default uniform angular distribution. Variables c_X and c_Y define the polygon center; r_{ave}

its average radius; $\delta\theta$ and θ are a vector of angle steps, and a vector of angles respectively; and $l_x \times l_y$ are the image dimensions (equal to 64×64 px here). The pseudocode is listed in Algorithm 2.

Occlusion parameters are also set by the noise vector z . In our experiments, the following sampling distributions are used (with \mathcal{B} – Bernoulli, \mathcal{U} – Uniform, and \mathcal{N} – Gaussian):

- $\mathcal{B}(\mathcal{U}(0, l_x/4), \mathcal{U}(l_x/4, l))$ for c_X ,
 $\mathcal{B}(\mathcal{U}(0, l_y/4), \mathcal{U}(l_y/4, l))$ for c_Y ;
- $\mathcal{U}(10, l/4)$ for r_{ave} , with $l = \min(l_x, l_y)$;
- $\mathcal{U}(3, 10)$ for N_{vert} ;
- $\mathcal{U}(0, 0.5)$ for σ .

B.3. Synthetic Data Generation

Subsection 3.2 of the main paper gives an overview of the data generation method used for our pipeline. In order to produce normal maps of objects of interest for our augmentation pipeline, we use OpenGL [19] with a custom normal shader printing its output in 3 output channels. Viewpoint are defined by the vertices of an icosahedron centered on target objects. In order to achieve a finer sampling, one needs to subsequently subdivide triangular faces of the icosahedron to smaller triangles until the desired level of detail is achieved. In-plane rotations can also be introduced by rotating the camera around the ray pointing to the object.

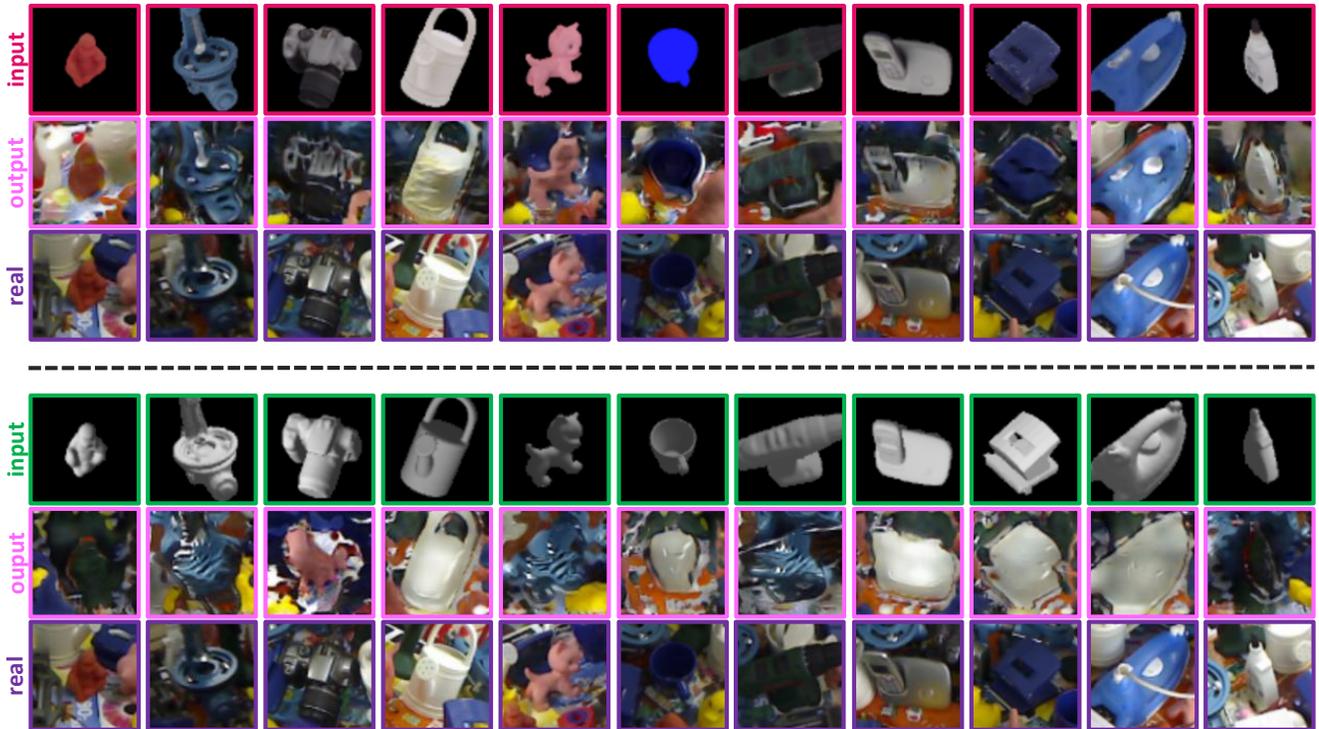


Figure S1: Qualitative comparison of results for *PixelDA* [4] trained with or without realistic texturing of the target objects. The method fails to bridge the realism gap when too wide.

Algorithm 1: Approximate Blinn-Phong shading [3]
from normal maps

Input: $N \in \mathbb{R}^{h \times w \times 3}$ normal map, $L \in \mathbb{R}^3$ directional light vector, $a \in \mathbb{R}^{\#}$ RGB ambient light coefficient, $d \in \mathbb{R}^{\#}$ RGB diffusion coefficient, $s \in \mathbb{R}^3$ RGB specular coefficient, $s_p \in \mathbb{R}$ specular hardness, $f_x \in \mathbb{R}^2$ pixel focal range used to render N

Output: $M \in \mathbb{R}^{h \times w \times 3}$ color lightness map

```

/* - Simplification #1: we recover
an approximate viewer vector  $V$ 
from  $N$  indices and  $f_x$ . */
/* - Simplification #2: we suppose
the light source at  $+\infty$  distance,
hence the same  $L$  for every
surface point. */
/* - Note: we use Einstein notation
for matrix-vector operations. */
/*
/* View vector approximation: */
1  $V \leftarrow \{(j, i, 1)\}_{j=0, i=0}^{h, w}$ ;
2  $V_j \leftarrow -\frac{(V_j - h/2)}{f_{x,j}}$ ;  $V_i \leftarrow -\frac{(V_i - w/2)}{f_{x,i}}$ ;
3  $V \leftarrow \frac{V}{\|V\|}$ ;
/* Computation of half-way vector
map: */
4  $H \leftarrow V + L$ ;
/* Diffuse shading: */
5  $D^{ij} \leftarrow N^{ij}_k L^k$ ;
/* Specular shading: */
6  $S^{ij} \leftarrow (N^{ij}_k H^k_{ij})^{s_p}$ ;
/* Adding all contributions (given
 $e_{ijc} = e_i \otimes e_j \otimes e_c$ ) : */
7  $M^{ijc} \leftarrow$ 
 $\min(\max(a \cdot e_{ijc} + d \cdot D^{ij} e_c + s \cdot S^{ij} e_c, 0), 1)$ ;
8 return  $M$ 

```

For the T-LESS dataset [15], synthetic data is rendered using the full icosahedron with a radius of 600mm and 3 subdivisions since real sequence contain objects shown from below as well. No in-plane rotations were added to the training data neither due to the same reason. This resulted in generation of 642 samples per object (given 11 objects – numbers 2, 5, 6, 7, 8, 11, 12, 18, 25, 29 and 30).

LineMOD data was generated using an icosahedron of radius 600mm with 3 consecutive subdivisions. Only the upper part of the icosahedron is used since in real sequences all objects are shot from above. In-plane rotations also added for each vertex, parametrized from -45° to 45° with a stride of 15° . Rotation invariance of four irregular LineMOD objects (*bowl*, *cup*, *eggbox*, and *glue*) was

Algorithm 2: Random polygon generation [32]

Input: $z \in \mathbb{Z}^k$ noise vector

Output: $p = \{p_i \in \mathbb{R}^2\}_{i=0}^{N_{vert}}$ polygon points

```

/* occlusion parameters sampling:
*/
1  $c_x, c_y, r_{ave}, N_{vert}, \epsilon, \sigma \leftarrow \text{sampleFromVector2}(z)$ ;
/* angle steps generation: */
2  $sum = 0$ ;
3 for  $i \in \{1, \dots, N_{vert}\}$  do
4 |  $\delta\theta_i \leftarrow \mathcal{U}(2\pi/N_{vert} - \epsilon, 2\pi/N_{vert} + \epsilon)$ ;
5 |  $sum \leftarrow sum + step$ ;
6 end
/* steps normalization: */
7  $k \leftarrow sum/(2\pi)$ ;
8 for  $i \in \{1, \dots, N_{vert}\}$  do
9 |  $\delta\theta_i \leftarrow \delta\theta_i/k$ ;
10 end
/* polygon points generation: */
11  $\theta_1 \leftarrow \mathcal{U}(0, 2\pi)$ ;
12 for  $i \in \{1, \dots, N_{vert}\}$  do
13 |  $r \leftarrow \mathcal{N}(r_{ave}, \sigma)$ ;
14 |  $p_i \leftarrow (c_x + r \cos(\theta_i), c_y + r \sin(\theta_i))$ ;
15 |  $\theta_i \leftarrow \theta_i + \delta\theta_i$ 
16 end
17 return  $p$ 

```

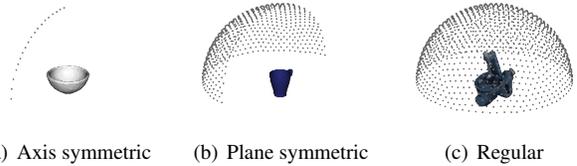


Figure S2: **Vertices sampling for different LineMOD objects** – each vertex represents a camera position from which the object is rendered.

also taken into account by limiting the amount of sampling points, such that each patch is unique. Figure S2 demonstrates the results of the output vertex sampling for different object types. We therefore generated 2,359 data points for each of the 11 regular objects, 1,239 for 3 plane symmetric objects (*cup*, *glue*, and *eggbox*) and 119 for the axis symmetric *bowl*.

C. Additional Qualitative Results

Figures S3 and S4 contain further visual results, demonstrating how our pipeline fairs on real color images when trained purely on synthetic geometrical data.

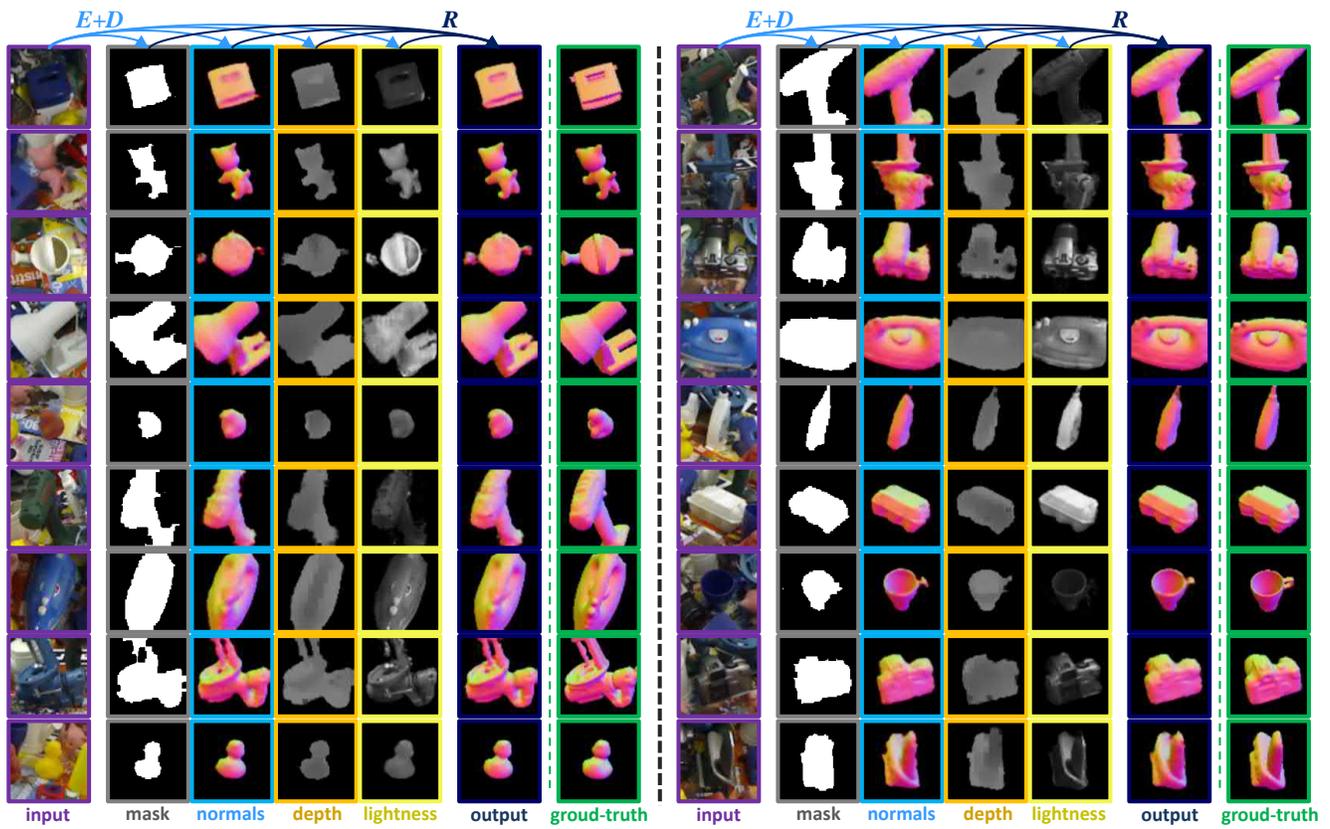


Figure S3: Qualitative results on LineMOD [14] for *SynDA* trained on texture-less CAD data.

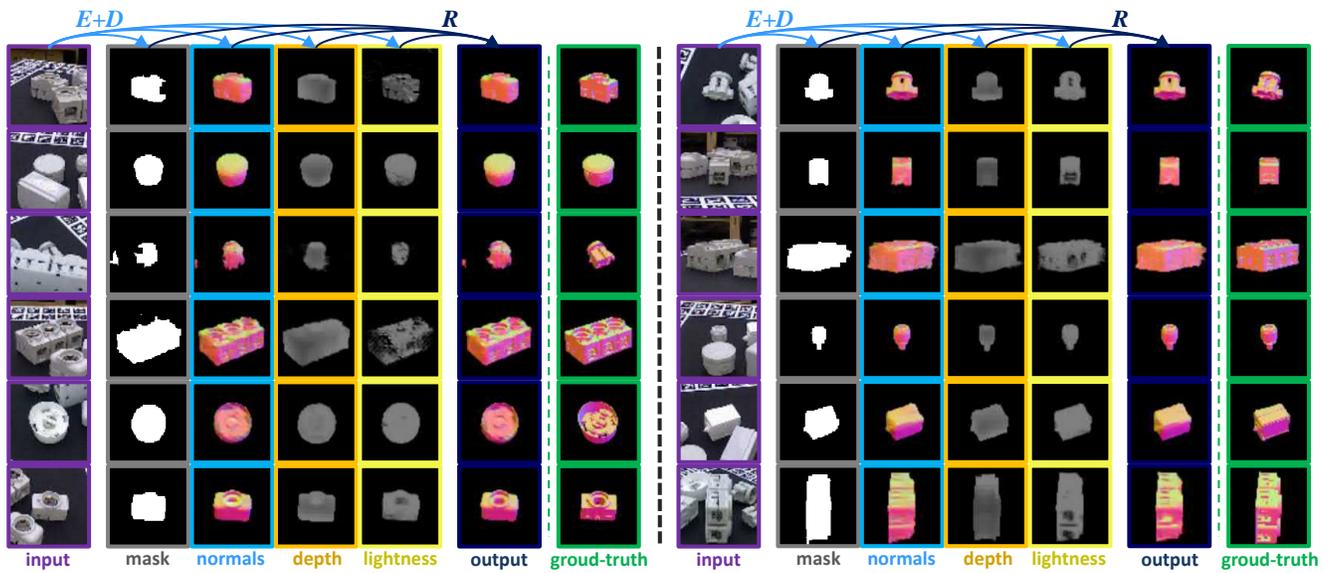


Figure S4: Qualitative results on T-LESS [15], for *SynDA* trained on texture-less CAD data.

References

- [1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. **10**
- [2] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014. **5**
- [3] J. F. Blinn. Models of light reflection for computer synthesized pictures. In *ACM SIGGRAPH computer graphics*, volume 11, pages 192–198. ACM, 1977. **7, 10, 13**
- [4] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. *arXiv preprint arXiv:1612.05424*, 2016. **3, 7, 8, 9, 10, 11, 12**
- [5] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. In *NIPS*, pages 343–351, 2016. **8**
- [6] Y. Cao, C. Shen, and H. T. Shen. Exploiting depth from single monocular images for object detection and semantic segmentation. *IEEE Transactions on Image Processing*, 26(2):836–846, 2017. **3**
- [7] J. Cheng, L. Dong, and M. Lapata. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016. **5**
- [8] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *CVPR*, pages 2650–2658, 2015. **3, 4**
- [9] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, pages 2366–2374, 2014. **3**
- [10] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. **3, 9, 10, 11**
- [11] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016. **3**
- [12] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In *CVPR*, pages 2827–2836, 2016. **3, 4, 5**
- [13] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645. Springer, 2016. **7, 8**
- [14] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *ACCV*. Springer, 2012. **7, 9, 14**
- [15] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *WACV*, 2017. **7, 8, 13, 14**
- [16] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*. IEEE, 2005. **3**
- [17] K. Karsch, C. Liu, and S. B. Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *TPAMI*, 2014. **3**
- [18] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *arXiv preprint arXiv:1705.07115*, 3, 2017. **2, 3, 4, 9**
- [19] Khronos Group. Open graphics library (opengl). <https://opengl.org>, 2006-2017. **12**
- [20] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. **10**
- [21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. **3**
- [22] R. Kuga, A. Kanezaki, M. Samejima, Y. Sugano, and Y. Matsushita. Multi-task learning using multi-modal encoder-decoder networks with shared skip connections. In *ICCV Workshop*, 2017. **2, 3, 4, 5, 9**
- [23] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*. IEEE, 2016. **3, 9, 11**
- [24] M. J. Landau, B. Y. Choo, and P. A. Beling. Simulating kinect infrared and depth images. 2015. **3**
- [25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: common objects in context. In *ECCV*, 2014. **7, 12**
- [26] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *TPAMI*, 2011. **3**
- [27] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, pages 5162–5170, 2015. **3**
- [28] M. Liu, M. Salzmann, and X. He. Discrete-continuous depth estimation from a single image. In *CVPR*, pages 716–723, 2014. **3**
- [29] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015. **5**
- [30] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *NIPS*, pages 2204–2212, 2014. **5**
- [31] N. Neverova, P. Luc, C. Couprie, J. Verbeek, and Y. LeCun. Predicting deeper into the future of semantic segmentation. In *ICCV*, pages 1–10, 2017. **5**
- [32] M. Ounsworth. *Anticipatory Movement Planning for Quadrotor Visual Servoing*. PhD thesis, McGill University Libraries, 2015. **7, 12, 13**
- [33] J. Peck. Fastnoise library. <https://github.com/Auburns/FastNoise>, 2016. **10**
- [34] K. Perlin. Improving noise. In *ACM Transactions on Graphics (TOG)*, volume 21, pages 681–682. ACM, 2002. **7, 10**
- [35] B. Planche, Z. Wu, K. Ma, S. Sun, S. Kluckner, T. Chen, A. Hutter, S. Zakharov, H. Kosch, and J. Ernst. Depthsynth: Real-time realistic synthetic data generation from cad models for 2.5 d recognition. In *3DV*. IEEE, 2017. **3**
- [36] A. Roy and S. Todorovic. Monocular depth estimation using neural regression forest. In *CVPR*, pages 5506–5514, 2016. **3**

- [37] F. Sadeghi and S. Levine. (cad)2rl: Real single-image flight without a single real image. *arXiv preprint arXiv:1611.04201*, 2016. [3](#)
- [38] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. *arXiv preprint arXiv:1612.07828*, 2016. [3](#)
- [39] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. [3](#)
- [40] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IROS*, pages 23–30. IEEE, 2017. [3](#), [6](#)
- [41] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. *arXiv preprint arXiv:1702.05464*, 2017. [3](#)
- [42] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille. Towards unified depth and semantic prediction from a single image. In *CVPR*, pages 2800–2809, 2015. [3](#), [4](#)
- [43] P. Wohlhart and V. Lepetit. Learning descriptors for object recognition and 3d pose estimation. In *CVPR*, pages 3109–3118, 2015. [6](#), [7](#), [8](#), [9](#)
- [44] S. Worley. A cellular texture basis function. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 291–294. ACM, 1996. [10](#)
- [45] D. Xu, W. Ouyang, X. Wang, and N. Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. *arXiv preprint arXiv:1805.04409*, 2018. [2](#), [3](#), [4](#), [5](#), [9](#)
- [46] S. Zakharov, W. Kehl, B. Planche, A. Hutter, and S. Ilic. 3d object instance recognition & pose estimation using triplet loss with dynamic margin. In *IROS*, 2017. [6](#), [8](#), [9](#)
- [47] S. Zakharov, B. Planche, Z. Wu, A. Hutter, H. Kosch, and S. Ilic. Keep it unreal: Bridging the realism gap for 2.5 d recognition with geometry priors only. *arXiv preprint arXiv:1804.09113*, 2018. [3](#), [6](#), [7](#), [8](#)
- [48] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018. [2](#), [5](#), [6](#), [9](#)
- [49] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. [5](#), [9](#)