

# DeceptionNet: Network-Driven Domain Randomization

Sergey Zakharov <sup>\*†</sup>, Wadim Kehl <sup>✧</sup>, and Slobodan Ilic <sup>\*†</sup>

<sup>\*</sup> Technical University of Munich   <sup>✧</sup> Toyota Research Institute   <sup>†</sup> Siemens Corporate Technology  
sergey.zakharov@tum.de, wadim.kehl@tri.global, slobodan.ilic@siemens.com

## Abstract

*We present a novel approach to tackle domain adaptation between synthetic and real data. Instead of employing "blind" domain randomization, i.e., augmenting synthetic renderings with random backgrounds or changing illumination and colorization, we leverage the task network as its own adversarial guide toward useful augmentations that maximize the uncertainty of the output. To this end, we design a min-max optimization scheme where a given task competes against a special deception network to minimize the task error subject to the specific constraints enforced by the deceiver. The deception network samples from a family of differentiable pixel-level perturbations and exploits the task architecture to find the most destructive augmentations. Unlike GAN-based approaches that require unlabeled data from the target domain, our method achieves robust mappings that scale well to multiple target distributions from source data alone. We apply our framework to the tasks of digit recognition on enhanced MNIST variants, classification and object pose estimation on the Cropped LineMOD dataset as well as semantic segmentation on the Cityscapes dataset and compare it to a number of domain adaptation approaches, thereby demonstrating similar results with superior generalization capabilities.*

## 1. Introduction

The alluring possibility of training machine learning models on purely synthetic data allows for a theoretically infinite supply of both input data samples and associated label information. Unfortunately, for computer vision applications, the domain gap between synthetic renderings and real-world imagery poses serious challenges for generalization. Despite the apparent visual similarity, synthetic images structurally differ from real camera sensor data. First, synthetic image formation produces clear edges with approximate physical shading and illumination, whereas real images undergo many types of noise, such as optical aber-

rations, Bayer demosaicing, or compression artifacts. Second, the visual differences between synthetic CAD models and their actual physical counterparts can be quite significant. Apart from the visual gap, supervised approaches also require cumbersome and error-prone human labeling of real training data in the form of 2D bounding boxes, segmentation masks, or 6D poses [25, 17]. For other approaches, such as robotic control learning, solutions must be found by exploration in tight simulation-based feedback loops that require synthetic rendering [28, 41, 31].

The gap between the visual domains is nowadays mainly bridged with adaptation and/or randomization techniques. In the case of supervised domain adaptation approaches [57, 32, 3, 30], a certain amount of labeled data from the target domain exists, while in unsupervised approaches [13, 48, 43, 4] the target data are available but unlabeled. In both cases, the goal is to match the source and target distributions by finding either a direct mapping, a common latent space, or through regularization of task networks trained on the source data. Recent unsupervised approaches are mostly based on generalized adversarial networks (GANs) [4, 22, 26, 48, 51, 24, 45, 38, 1] and although these methods perform proper target domain transfers, they can overfit to the chosen target domain and exhibit a decline in performance for unfamiliar out-of-distribution samples.

Domain randomization methods [49, 21, 29, 55, 47] have no access to any target domain and employ the rather simple technique of randomly perturbing (synthetic) source data during training to make the tasks networks robust to perceptual differences. This approach can be effective, but is generally unguided, and needs an exhaustive evaluation to find meaningful augmentations that increase the target domain performance. Last but not least, results from pixel-level adversarial attacks [6, 46] suggest the existence of architecture-dependent effects that cannot be addressed by "blind" domain randomization for robust transfer.

We propose herein a general framework that performs guided randomization with the help of an auxiliary deception network trained in a similar min-max fashion as GAN

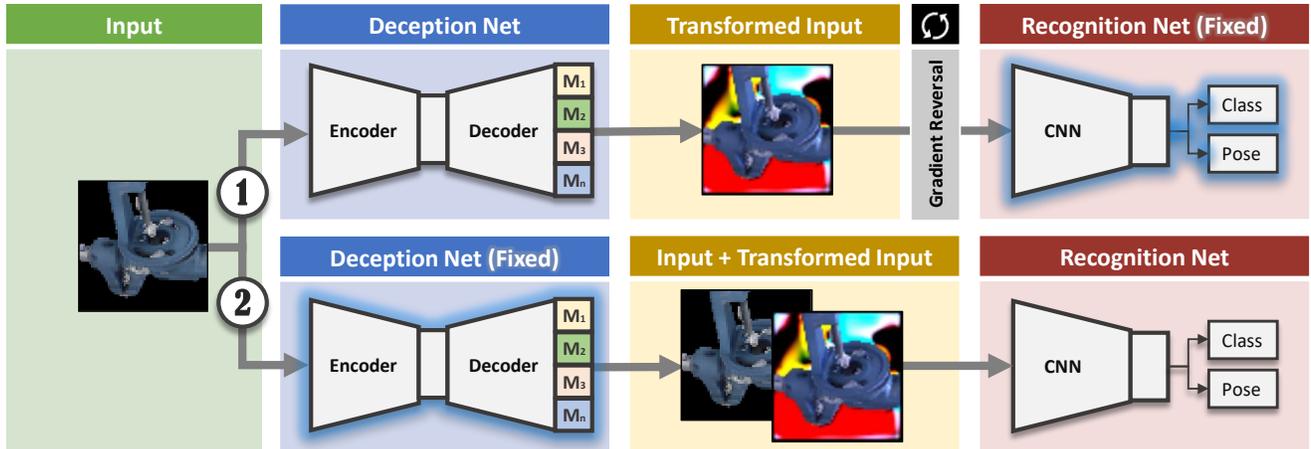


Figure 1: **Training pipeline.** Training is performed in two alternating phases. **Phase 1:** The weights of the deception network are updated, while those of the recognition network are frozen. The recognition network’s objective is maximized instead of being minimized, forcing the deception network to produce increasingly confusing images. **Phase 2:** The generated deceptive images provided by the deception network, whose weights are now frozen, are passed to the recognition network and its weights are updated such that the loss is minimized. As a result of this min-max optimization, the input images are automatically altered by the deception network, forcing the recognition network to be robust to these domain changes.

networks. This is done in two alternating phases, as illustrated in Fig. 1. In the first phase, the synthetic input is fed to our deception network responsible for producing augmented images that are then passed to a recognition network to compute the final task-specific loss with provided labels. Then, instead of minimizing the loss, we maximize it via gradient reversal [12] and only back-propagate an update to the deception network parameters. The deception network parameters are steering a set of differentiable modules  $M_1, \dots, M_N$ , from which augmentations are sampled. In the next phase, we feed the augmented images to the recognition network together with the original images to minimize the task-specific loss and update the recognition network. In this way, the deception network is encouraged to produce domain randomization by confusing the recognition network, and the recognition network is made resilient to such random changes. By adding different modules and constraints we can influence how much and which parts of the image the deception network alters. In this way, our method outputs images completely independent from the target domain and therefore generalizes much better to new unseen domains than related approaches. In summary, our contributions are:

- DeceptionNet framework that performs a min-max optimization for guided domain randomization;
- Various pixel-level perturbation modules employed in such a framework suited for synthetic data;
- Novel sequences: MNIST-COCO and Extended Cropped LineMOD that allow to demonstrate our strong generalization capabilities to unseen domains.

In the experimental section we will show that steered randomization by leveraging the network structure actually generalizes much better to new domains than unsupervised approaches with access to the target data while performing comparably well to them on known target domains.

## 2. Related Work

**Domain Adaptation.** Various domain adaptation works put their efforts to bridge the gap between the domains mostly based on unsupervised conditional generative adversarial networks (GANs) [48, 43, 4, 1] or style-transfer solutions [14]. These methods use an unlabeled subset of target data to improve the synthetic data performance. For example, the authors of [4, 1] proposed to use GANs to learn the mapping from synthetic images to real. Extending this idea, approaches of [45, 38] use GANs to tune the parameters of user-defined transformations to fit to the target distribution. As opposed to GANs, work [11] used a sequence autoencoder to extract the feature vector pairs from the available data, which are then decoded to generate new data samples.

Alternatively, domain-invariant features that work well for both real and synthetic domains can be learned. [37] mapped real image features to the feature space of synthetic images and used the mapped information as an input to a task-specific network, trained on synthetic data only.

Another example is *DSN* [5], which proposes the extraction of image representations that are partitioned into two subspaces: private to each domain and one which is shared across domains (learning domain-invariant features). The shared subspace is then used to train a classifier that per-

forms well on both domains. Similarly, *DRIT* [23] embeds images on a domain-invariant content space (capturing shared information across domains) and a domain-specific attribute space by introducing a cross-cycle consistency loss based on disentangled representations. Other approaches, such as *DANN* [13] or *ADDA* [51] instead focus on adapting the recognition methods themselves to make them more robust to the domain changes.

**Domain Randomization.** However, what if one does not have real data available? The answer for this case is domain randomization. Domain randomization is a popular approach [49, 21, 55, 36, 47, 40] that aims to randomize parts of the domain that we do not want our algorithm to be sensitive to. For example, [49] and [40] trained complex recognition methods by means of adding variability to the input render data, i.e., different illumination conditions, texture changes, scene decomposition, *etc.* This sort of parameterization allows to learn features that are invariant to the particular properties of the domain. The authors of [55] used a sophisticated depth augmentation pipeline trying to cover possible artifacts of the common commodity depth sensors. It was then used to train a network removing these artifacts from the input and generating a clean, synthetically-looking image. Building on top of this idea, the methods of [36, 47] extended this to the RGB domain.

Nevertheless, the main question remains unsolved: What is the main cause of confusion given the domain change? Domain randomization tries to target all possible scenarios, but we do not really know which of them are actually useful to bridge the domain gap. Moreover, covering all possible variations present in the real world by applying simple augmentations is almost impossible.

Our approach can be placed between domain randomization and GAN methods, however, instead of forcing randomization without any clear guidance on its usefulness, we propose to delegate this to a neural network, which we call *deception network*, which tries to alter the images in an automated fashion, such that the task network is maximally confused. Moreover, to do so, we do not require any images, labeled or unlabeled, from the target domain.

### 3. Methodology

As outlined, our approach towards steered domain randomization is essentially an extension of the task algorithm. Therefore, we have the actual task network  $T(\mathbf{x}; \theta_T) \rightarrow \hat{\mathbf{y}}$ , which, given an input image  $\mathbf{x}$ , returns an estimated label  $\hat{\mathbf{y}}$  (e.g., class, pose, segmentation mask, *etc.*), and (2) a deception network  $D$  that takes the source image  $\mathbf{x}^s$  and returns the deceptive image  $\mathbf{x}^d$ , which, when provided to the task net  $T(D(\mathbf{x}^s)) \rightarrow \hat{\mathbf{y}}^d$ , maximizes the difference between  $\hat{\mathbf{y}}^d$  and  $\mathbf{y}^s$ . While the recognition network architectures are standard and follow related work [4, 12], we will first focus

herein on our structured deception network first, and then describe the optimization objective and the training.

To formalize our pipeline, let  $X_c^s := \mathbf{x}_{c,i}^s \forall i \in N_c^s$  be a source dataset composed of  $N_c^s$  source images  $\mathbf{x}_c^s$  for an object of class  $c$ . Then,  $X^s := X_c^s \forall c \in C$  is the source dataset covering all object classes  $C$ . A dataset of real images  $X^r$  (not used by us for training) is similarly defined.

### 3.1. Deception Modules

The deception network  $D$  follows the encoder-decoder architecture where input  $\mathbf{x}^s$  is encoded to a lower-dimensional 2D latent space vector  $\mathbf{z}$  and given as an input to multiple decoding modules  $M_1, \dots, M_n$ . The final output of  $D$  is then a weighted sum of decoded outputs  $\mathbf{x}^d := \sum_i \mathbf{w}_i \cdot M_i(\mathbf{z})$  where  $\mathbf{w}_i \in [0, 1]^{H \times W}$  act as spatial masking operations. While such a formulation allows for flexibility, the decoders must follow a set of predefined constraints to create meaningful outputs and leverage an inherent image structure instead of finding trivial mappings to decrease the task performance (e.g., by decoding always to 0). Note that our proposed framework is general and, thus, requires instantiations of the deception network for specific datasets. Similar to architecture search, discovering the "best" instantiation is infeasible, but good ones can be found by analyzing the data source. After a reasonable experimentation we settled on certain configurations for MNIST (RGB) and LineMOD (RGB-D), depicted in Fig. 2. We continue by providing more detail on the used decoder modules and their constraint ranges.

#### 3.1.1 Background Module (BG)

Since our source images have black backgrounds, they hardly transfer over to the real world with infinite background variety, resulting in a significant accuracy drop. [21, 29] tackle this problem by rendering objects on top of images from large-scale datasets (e.g., MS COCO [25]).

Instead, our background module produces its output by chaining multiple upsampling and convolution operations. While the output is rather simple at start, the module regresses very complex and visually confusing structures in the advanced stages of training.

For MNIST, we used a simpler variant that outputs a single RGB background color  $\in [0, 1]$  and an RGB foreground bias  $\in [0.1, 0.9]$  (restricted not to intersect with the background color). To form the output, we first apply the background color and then add the foreground bias using the mask. We ensure that the final values are in the range  $[0, 1]$ .

#### 3.1.2 Distortion Module (DS)

The module is based on the idea of the elastic distortions first presented in [44]. Essentially, a 2D deformation field is randomly initialized from  $[-1, 1]$  and then convolved with

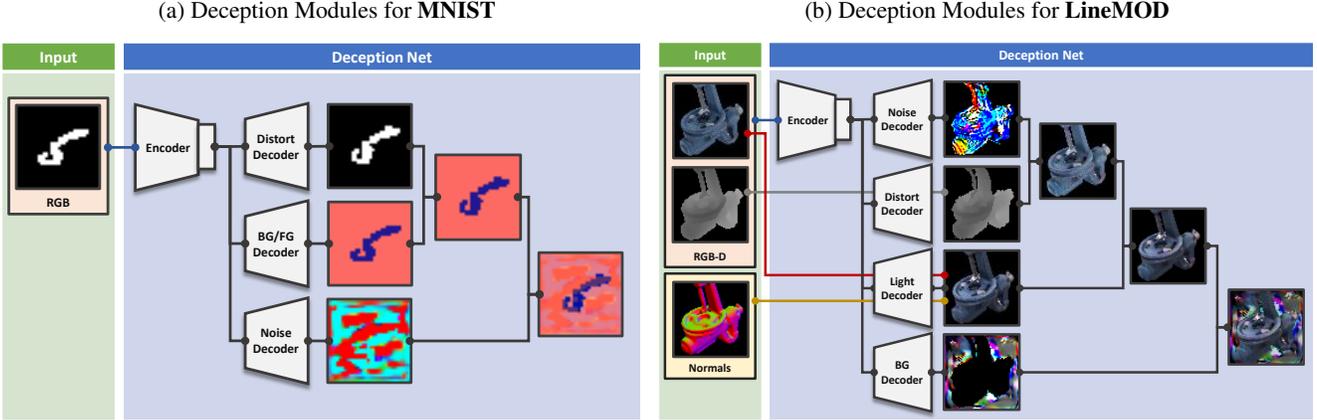


Figure 2: **Architecture of the deception networks used for the presented experiments.** For the case of **MNIST** classification, three deception modules are used: the *distortion module* applying elastic deformations on the image, the *BG/FG module* responsible for generating background and foreground colors, and the *noise module* additionally distorting the image by applying slight noise. The **LineMOD** dataset requires a more sophisticated treatment and features four deception modules: noise and distortion (applied on depth channel only), modules similar to the previous case, pixel-wise BG module and light module generating different illumination conditions based on the Phong model.

a Gaussian filter of standard deviation  $\sigma$ . For large values of  $\sigma$ , the resulting field approaches 0, whereas smaller values of  $\sigma$  keep the field mostly random. However, the moderate values of  $\sigma$  make the resulting field perform elastic deformations, where  $\sigma$  defines the elasticity coefficient. The resulting field is then multiplied by a scaling factor  $\alpha$ , which controls the deformation intensity.

Our implementation closely follows the described approach but we use the decoder output as the distortion field and apply resampling, similar to spatial transformer networks [20]. We fix  $\sigma = 4$ , but learn both  $\alpha \in (0, 5]$  and the general decoder parameters. This means that the network itself controls where and how much to deform the object.

### 3.1.3 Noise Module (NS)

Applying slight random noise augmentation to the network input during training is common practice. In a similar fashion, we use the noise decoder to add generated values to the input. The noise decoder regresses a tensor of the input size with values in the range  $[-0.01, 0.01]$ , which are then added to the input of the module.

### 3.1.4 Light Module (L)

Another feature not well covered by synthetic data is proper illumination. Recent methods [21, 29, 16, 56] prerender a number of synthetic images featuring different light conditions. Here, we instead implement differentiable lighting based on the simple Phong model [35], which is fully operated by the network. While more complex parametric and

differentiable illumination models do exist, we found this basic approach to already work quite well.

The module requires surface information which is provided in form of normal maps. From this, we generate three different types of illumination, namely ambient, diffusive, and specular. The light decoder outputs a block of 9 parameters that are used to define the final light properties, i.e., a 3D light direction, an RGB light color (restricted to the range of  $[0.8, 1]$ ), and a weight for each of the three illumination types ( $\mathbf{w}_a \in [0.6, 1]$ ,  $\mathbf{w}_d \in [0, 1]$ ,  $\mathbf{w}_s \in [0, 1]$ ).

## 3.2. Optimization Objective

The optimization objective of the deception network is essentially the loss of the recognition network; however, instead of minimizing it, we maximize it by updating the parameters in the direction of the positive gradient. This is achieved by adding a gradient reversal layer [12] between the deception and recognition nets as shown in Fig. 1. The layer only negates the gradient when back-propagating, thereby resulting in the reversed optimization objective for a given loss. Therefore, the general optimization objective can be written as follows:

$$\min_{\theta_T} \max_{\theta_D} \mathcal{L}_t(T(D(\mathbf{x}; \theta_D)), \mathbf{y}; \theta_T) \quad (1)$$

$$\text{subject to } C_{M_n} \quad \text{for } \mathbf{n} = 1, \dots, N_m \quad (2)$$

where  $\mathbf{x}$  is the input image,  $\mathbf{y}$  is the ground truth label,  $T$  is the task network,  $\mathcal{L}_t$  is the task loss,  $D$  is the deception network, and  $C_m$  denotes the hard constraints defined by the deception modules enforced by projection after a gradient step. In this framework, the deception network's objective



Figure 3: **Example samples of the MNIST modalities:** MNIST (Source), MNIST-M (Target), and MNIST-COCO (Generalization) on the left; and example augmentation images generated by PixelDA and our method respectively.

only depends on the objective of the recognition task and can, therefore, be easily applied to any other task.

### 3.3. Training Procedure

We use two different SGD solvers, where the actual task network has a learning rate of 0.001 with a decaying factor of 0.95 every 20000<sup>th</sup> iteration. The learning rate of the deception network was found to work well with a constant value of 0.01. We train with a batch size of 64 for all the experiments and we stop training after 500 epochs. During the experimentation, we also discovered that concatenating real and perturbed images led to a consistent improvement in numbers.

## 4. Evaluation

In this section, we conduct a series of experiments to compare the capabilities of our pipeline with the state-of-the-art domain adaptation methods. We first compare ourselves against these baselines for the problem of adaptation and will then compare in terms of generalization. We will conclude with an ablative analysis to measure the impact of each module and modality on the final performance.

As the first dataset, we used the popular handwritten digits dataset MNIST as well as MNIST-M, introduced in [13] for unsupervised domain adaptation (depicted in Figs. 3a 3b). MNIST-M blends digits from the original monochrome set with random color patches from BSDS500 [2] by simply inverting the color values for the pixels belonging to the digit. The training split containing 59001 target images is then used for domain adaptation. The remaining 9001 target images are used for evaluation. That means that around 86% of the target data is used for training. Note that while MNIST is not technically synthetic, its clean and homogeneous appearance is typical for synthetic data.

The second dataset is the Cropped LineMOD dataset [53] consisting of small centered, cropped 64×64 patches of 11 different objects in cluttered indoor settings displayed in various of poses. It is based on the LineMOD dataset [15] featuring a collection of annotated RGB-D sequences

recorded using the Primesense Carmine sensor and associated 3D object reconstructions. The dataset also features a synthetic set of crops of the same objects in various poses on a black background. We will treat this Synthetic Cropped LineMOD as the source dataset and the Real Cropped LineMOD as the target dataset. Domain adaptation methods use a split of 109208 rendered source images and 9673 real-world target images, 1000 real images for validation, and a target domain test set of 2655 images for testing. We show examples in Figs. 4a and 4b.

The last dataset pair we used for the experiments is SYNTHIA [39] and Cityscapes [10]. SYNTHIA is a collection of pixel-annotated road scene frames rendered from a virtual city. Cityscapes is its real counterpart acquired in the street scenes of 50 different actual cities. Following a common evaluation protocol, we used a subset of 9400 SYNTHIA images, also known as SYNTHIA-RAND-CITYSCAPES, as the source data and 500 Cityscapes validation images as the target data.

### 4.1. Adaptation Tests

All domain adaptation methods use a significant portion of the target data for training, making the resulting mapped source images very similar to the target images (e.g., Fig. 3b vs 3d and Fig. 4b vs 4d). A common benchmark for domain adaptation is then to compare the performance of a classifier trained on the mapped data against a classifier trained on the source data only (lower baseline) and against a classifier trained directly on the target data (upper baseline).

Our approach is generally disadvantaged since we can structure our domain mapping only through the source data and the deception architecture. To show that our learned randomization is indeed guided, we additionally implement an unguided randomization variant that applies train time augmentation similar to the related work. It employs the same modules and constraints as our deception network, but its perturbations are conditioned on random values in each forward pass instead of latent codes from the input.

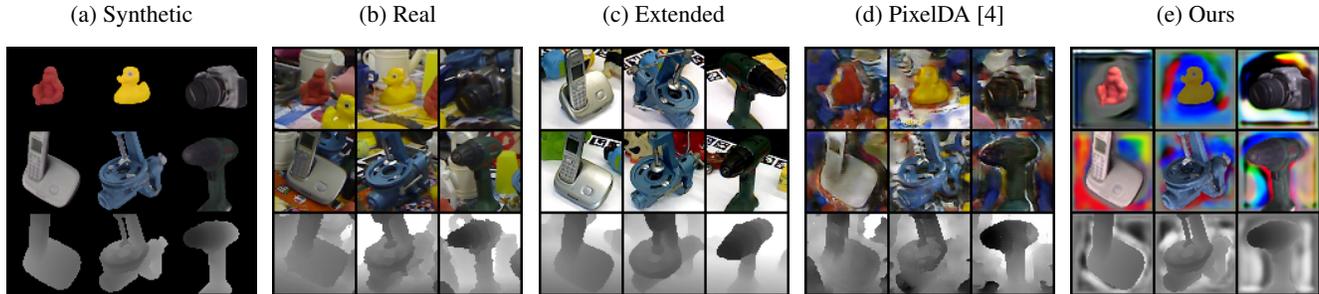


Figure 4: **Example samples of the LineMOD modalities:** Synthetic (Source), Real (Target), and Extended (Generalization) on the left; and example augmentation images generated by PixelDA and our method respectively.

#### 4.1.1 Classification on MNIST

In Table 1 we collect the results of the most relevant methods tested on the MNIST  $\rightarrow$  MNIST-M scenario and split them according to the type of data used. Since domain adaptation methods use both source and target data for training, they are allocated to a separate group ( $S + T$ ). Both our method and the unguided randomization variant only have access to the source data and are therefore grouped in  $S$ . The task network follows the architecture presented in [12], which is also used by the other methods. The task’s objective  $\mathcal{L}_t$  is a simple cross entropy loss between the predicted and the ground truth label distributions.

We can identify three key observations: (1) our method shows very competitive results (90.4% classification) and is on par with the latest domain adaptation pipelines: DSN – 83.2%, DRIT – 91.5% and PixelDA – 95.9%. Moreover, we outperform most of the methods by a significant margin despite the fact that they had access to a large portion of the target data to minimize the domain shift. (2) Guiding the randomization leads to 7% higher accuracy which supports our claim convincingly. (3) Surprisingly, unguided randomization (with appropriate modules) alone is in fact enough to outperform most methods on MNIST.

#### 4.1.2 Classification and Pose Estimation on LineMOD

As before, the domain adaptation methods are trained on a mix of source (Synthetic Cropped LineMOD) and target (Real Cropped LineMOD) data and we compare to the pre-defined baselines. We use the common task network for this benchmark from [12] and the associated task loss:

$$\mathcal{L}_t(G) = \mathbb{E}_{\mathbf{x}^s, \mathbf{y}^s} \left[ -\mathbf{y}^s \top \log \hat{\mathbf{y}}^d + \xi \log(1 - \mathbf{q}^s \top \hat{\mathbf{q}}^d) \right] \quad (3)$$

where the first term is the classification loss and the second term is the log of a quaternion rotation metric [19].  $\xi$  weighs both terms whereas  $\mathbf{q}^s$  and  $\hat{\mathbf{q}}^d$  are the ground truth and predicted quaternions, respectively.

The results in Table 1 present a more nuanced case. On this visually complex dataset, unguided randomization per-

forms only above the lower baseline and is far behind any other method. Our guided randomization, on the other hand, with – 95.8% classification and 51.9° angle error is competitive with those of the latest domain adaptation methods using target data: DSN – 100% & 53.3°, DRIT – 98.1% & 34.4°, and PixelDA – 99.9% & 23.5°. Nonetheless, we believe that both DRIT and PixelDA are not fully reachable by target-agnostic methods like ours since the space of all needed adaptations (e.g., aberrations or JPEG artifacts) has to be spanned by our deception modules. The augmentation differences between PixelDA and our method (Figs. 4d and 4e) suggest the existence of some visual phenomena we are still not accounting for with our deception network.

#### 4.2. Generalization Tests

For the second set of experiments, we test the generalization capabilities of our method as well as the competing approaches. The major advantage of our pipeline is its independence from any target domain by design. To support our case we designed two new datasets:

		MNIST $\rightarrow$ MNIST-M	Synthetic Cropped LineMOD $\rightarrow$ Real Cropped LineMOD	
Model		Classification Accuracy (%)	Classification Accuracy (%)	Mean Angle Error (°)
S	Source (S)	56.6	42.9	73.7
	Unguided	83.1	53.1	52.6
	Ours	90.4	95.8	51.9
S + T	CycleGAN [59]	74.5	68.2	47.5
	MMD [52, 27]	76.9	72.4	70.6
	DANN [13]	77.4	99.9	56.6
	DSN [5]	83.2	100	53.3
	DRIT [23]	91.5	98.1	34.4
	PixelDA [4]	95.9	99.9	23.5
Target (T)		96.5	100	12.3

Table 1: **Baseline tests:** While performing slightly worse than the leading state-of-the-art domain adaptation methods using target data, we still manage to achieve very competitive performance without access to target data.

- **MNIST-COCO** The data collection follows the exact same generation procedure of MNIST-M and has the same exact number of images for both training and testing. The only difference here is that instead of the BSDS500 dataset, we use crops from MS COCO. Fig. 3e demonstrates some of the newly generated images.
- **Extended Real Cropped LineMOD** Thanks to the help of the authors of the original LineMOD dataset [15], we were able to get some of the original LineMOD objects, namely "phone", "benchvise", and "driller". We repeated the physical acquisition setup and generated an annotated scene for each object. Each scene depicts a specific object placed on a white markerboard atop a turntable and coarsely surrounded by a small number of cluttered objects, slightly occluding the object at times. Each sequence contains 130 RGB-D images covering the full 360° rotation at an elevation angle of approximately 60°. Given the acquired and refined poses, we again crop the images in the same fashion as in the Cropped LineMOD dataset [53]. All 390 images are used for evaluation, with some examples shown in Fig. 4c.

For a comparison with the strongest related methods, i.e., DSN, DRIT, and PixelDA, we used open source implementations and diligently ensured that we are able to properly train and reproduce the reported numbers from Table 1. While the DRIT implementation worked well for the adaptation experiments, we failed to produce reasonably high numbers for the generalization experiment and chose to exclude it from the comparison.

Similar to before, we train them using the target data from MNIST-M and Real Cropped LineMOD. After the training is finished and the corresponding accuracies on the target test splits are achieved, we test them on the newly acquired dataset. While different, these extended datasets still bear a certain resemblance to the target dataset and we could expect to see a certain amount of generalization. For our randomization methods, we can immediately test on the

		MNIST → MNIST-COCO	Synthetic Cropped LineMOD → Extended Real Cropped LineMOD	
	Model	Classification Accuracy (%)	Classification Accuracy (%)	Mean Angle Error (°)
	Source (S)	57.2	63.1	78.3
S	Unguided	85.8	77.2	48.5
	Ours	89.4	99.0	46.5
S+	DSN [5]	73.2	45.7	76.3
	PixelDA [4]	72.5	76.0	84.2
	Target (T)	96.1	100	14.7

Table 2: **Generalization tests:** Our method generalizes well to the extended datasets, while the adaptation methods underperform due to overfitting.

new data, since retraining is not necessary.

As is evident from Table 2, the accuracy of our method on MNIST-COCO is very close to the MNIST-M number (90.4% and 89.4% respectively). For the case of Extended Real Cropped LineMOD, we get even better results than for the Real Cropped LineMOD for both accuracy and angle error: We only need to classify 3 objects instead of 11 with a much smaller pose space, and the scenes are in general cleaner and less occluded. These results underline our claim with respect to generalization. This is, however, not the case for the domain adaptation methods showing drastically worse results. Interestingly, we observe an inverse trend where better results on the original target data lead to a more significant drop. Despite of having a very high accuracy on the target data and the ability to generate additional samples that do not exist in the dataset, these methods present typical signs of overfit mappings that cannot generalize well to the extensions of the same data acquired in a similar manner. The simple reason for this might be the nature of these methods: they do not generalize to the features that matter the most for the recognition task, but to simply replicate the target distribution as close as possible. As a result, it is not clear what the classifier exactly focuses on during inference; however, it could very likely be the particular type of images (e.g., in case of MNIST-COCO) or a specific type of backgrounds and illumination (e.g., in case of Extended Real Cropped LineMOD). In contrast to domain adaptation methods, our pipeline is designed not to replicate the target distribution, but to make the classifier invariant to the changes that should not affect classification, which is the reason why our results remain stable.

### 4.3. Ablation Studies

In this section, we perform a set of ablation studies to gain more insight into the impact of each module inside the deception network. Obviously, our modules model only a fraction of possible perturbations and it is important to understand the individual contributions. Moreover, we demonstrate how well we perform provided different types of input modalities for the LineMOD datasets.

		MNIST → MNIST-M	Synthetic Cropped LineMOD → Real Cropped LineMOD	
	Modules	Classification Accuracy (%)	Classification Accuracy (%)	Mean Angle Error (°)
	None	56.6	42.9	73.7
	BG	82.4	74.8	50.4
	BG + NS	86.5	77.6	52.8
	BG + NS + DS	90.4	78.7	48.2
	BG + NS + DS + L	-	95.8	51.9

Table 3: **Module ablation:** Evaluation of the importance of the deception network’s modules. BG – background, NS – noise, DS – distortion, L – light.

		Road	SW	BLDG	Wall	Fence	Pole	TL	TS	VEG	Sky	PRSN	Rider	Car	Bus	Mbike	Bike	mIoU	mIoU*
	Source (S)	3.8	10.2	46.3	1.8	0.3	19.1	4.0	7.5	71.8	72.2	44.6	3.4	24.9	5.2	0.0	2.5	19.8	22.8
S	Unguided	17.9	8.8	59.2	0.8	0.4	22.1	3.5	6.1	71.4	70.4	40.3	7.3	37.9	3.3	0.2	7.3	22.3	25.7
	Ours	51.4	17.8	62.5	1.6	0.4	22.6	6.0	11.9	70.9	73.5	42.1	8.2	40.9	8.1	3.9	18.4	27.5	32.0
S+T	FCNs Wld [18]	11.5	19.6	30.8	4.4	0.0	20.3	0.1	11.7	42.3	68.7	51.2	3.8	54.0	3.2	0.2	0.6	20.1	22.9
	CDA [58]	65.2	26.1	74.9	0.1	0.5	10.7	3.7	3.0	76.1	70.6	47.1	8.2	43.2	20.7	0.7	13.1	29.0	34.8
	Cross-City [9]	62.7	25.6	78.3	-	-	-	1.2	5.4	81.3	81.0	37.4	6.4	63.5	16.1	1.2	4.6	-	35.7
	Tsai et al. [50]	78.9	29.2	75.5	-	-	-	0.1	4.8	72.6	76.7	43.4	8.8	71.1	16.0	3.6	8.4	-	37.6
	ROAD-Net [8]	77.7	30.0	77.5	9.6	0.3	25.8	10.3	15.6	77.6	79.8	44.5	16.6	67.8	14.5	7.0	23.8	36.1	41.7
	LSD-seg [42]	80.1	29.1	77.5	2.8	0.4	26.8	11.1	18.0	78.1	76.7	48.2	15.2	70.5	17.4	8.7	16.7	36.1	42.1
	Chen et al. [7]	78.3	29.2	76.9	11.4	0.3	26.5	10.8	17.2	81.7	81.9	45.8	15.4	68.0	15.9	7.5	30.4	37.3	43.0
	Target (T)	96.5	74.6	86.1	37.1	33.2	30.2	39.7	51.6	87.3	90.4	60.1	31.7	88.4	52.3	33.6	59.1	59.5	65.5

Table 4: **Real-world application:** Segmentation performance on SYNTHIA  $\rightarrow$  Cityscapes benchmark based on Intersection over Union (IoU) tested on 16 (mIoU) and 13 (mIoU\*) classes of the Cityscapes dataset. Our method outperforms *source* and *unguided* by a significant margin and remains competitive to the methods relying on the target data.

### 4.3.1 Deception Modules

We tested 4 different variations of the deception net that use varying combinations of the deception modules: *background (BG)*, *noise (NS)*, *distortion (DS)*, and *light (L)*. The exact combinations and the results on both datasets are listed in Table 3.

It can be clearly seen that each additional module in the deception network adds to the discriminative power of the final task network. The most important modules can also be easily distinguished based on the results. Apparently, the background module always makes a significant difference: the purely black backgrounds of the source data are drastically different from the real imagery. Another interesting observation is the strong impact the lighting perturbation has in the case of the Cropped LineMOD dataset. This enforces the notion that real sequences undergo many kinds of lighting changes that are not well-represented by synthetic renderings without any additional relighting. Note that the MNIST deception network does not employ lighting.

### 4.3.2 Input Modalities

For the task of simultaneous instance classification and pose estimation, we (as well as the other methods) always use the full RGB-D information. This ablation aims to show how well we fare provided only a certain type of data and the impact on the final results. Table 5 shows that RGB allows for better classification, whereas depth provides better pose estimates. We can further boost the classification enormously and reduce the pose error by combining both modalities.

Input	Synthetic Cropped LineMOD $\rightarrow$ Real Cropped LineMOD		Synthetic Cropped LineMOD $\rightarrow$ Extended Real Cropped LineMOD	
	Classification Accuracy (%)	Mean Angle Error ( $^{\circ}$ )	Classification Accuracy (%)	Mean Angle Error ( $^{\circ}$ )
D	73.3	36.6	78.7	34.9
RGB	84.8	57.4	85.9	49.4
RGB-D	95.8	51.9	99.0	46.5

Table 5: **Input modality ablation:** Performance evaluation based on the input data type used: depth, RGB, or RGB-D.

### 4.4. Real-world Scenario

We demonstrate a real-world application of our approach on a more practical problem of semantic segmentation using the common SYNTHIA  $\rightarrow$  Cityscapes benchmark. Having only synthetic SYNTHIA renderings, we try to generalize to the real Cityscapes data by evaluating our method on 13 and 16 classes using the Intersection over Union (IoU) metric. This setup is particularly difficult since the domain gap problem here is intensified by a completely different set of segmentation instances and camera views. For a fair comparison, all methods use a VGG-16 base (FCN-8s) recognition network. The deception modules used in this case are as follows: 2D noise (NS), elastic distortion (DS), and light (L). Normal maps for the light module are generated from the available synthetic depth data.

Table 4 shows that even without access to target domain data, our pipeline remains competitive with the methods relying on target data, showing mIoU of 27.5% and mIoU\* of 32% (16 and 13 classes) – well above *source* and *unguided*. The results also confirm the generality of the approach with respect to the different task architectures and datasets.

## 5. Conclusion

In this paper we presented a new framework to tackle the domain gap problem when no target data is available. Using a task network and its objective, we show how to extend it with a simple encoder-decoder deception network and bind both in a min-max game in order to achieve guided domain randomization. As a result, we obtain increasingly more robust task networks. We demonstrate a comparable performance to domain adaptation methods on two datasets and, most importantly, show superior generalization capabilities where the domain adaptation methods tend to drop in performance due to overfitting to the target distribution. Our results suggest that guided randomization, because of its simple but effective nature, should become a standard procedure to define baselines for domain transfer and adaptation techniques.

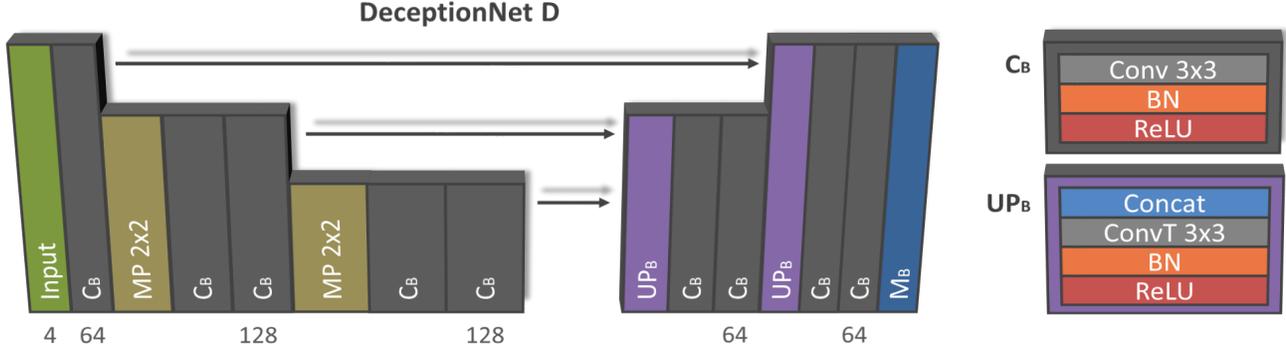


Figure 5: **DeceptionNet architecture.** Our network features a typical encoder-decoder architecture. The encoder part consists of 2 consecutive downsamplings followed by a sequence of convolutional blocks  $C_B$ . The decoder part shares a similar architecture for all presented augmentation modules. Arrows show the skip connections between blocks.

## A. Supplementary Material

### A.1. Network Architecture

Let  $C_k^B$  be a convolutional block composed of the following layers:  $3 \times 3$  convolution with  $k$  filters, BatchNorm (BN), and ReLU activation function. Similarly, let  $UP_B$  be a decoding block made of: 2-factor upsampling transposed convolution, BatchNorm (BN), and ReLU.

**DeceptionNet D:** Using the defined nomenclature, the encoding part of the DeceptionNet can be described as:  $C_{64}^B - MP - C_{128}^B - C_{128}^B - MP - C_{128}^B - C_{128}^B$ ; and its decoding part as:  $UP_{64} - C_{64}^B - C_{64}^B - UP_{64} - C_{64}^B - C_{64}^B - M_B$ . Where  $M_B$  is defined by the specific module type, and  $MP$  stands for a 2-factor max-pooling layer. Encoding blocks have skip connections concatenating the channels with the opposite decoding blocks. The visual representation of the DeceptionNet’s architecture is depicted in Fig. 5.

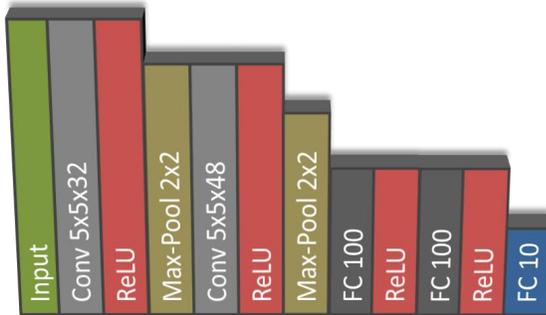


Figure 6: **MNIST Classifier:** Simple LeNet-like architecture, where 2 convolutional layers followed by ReLUs and max-poolings are finalized by 3 fully-connected layers.

**Task Network T:** In both cases, i.e., for MNIST classification and Cropped LineMOD classification and pose estimation,  $T$  follows a simple LeNet-like architecture. As for MNIST (see Fig. 6), the final layer outputs the 10D vector, whereas for Cropped LineMOD (see Fig. 7) there is a 11D classification output as well as 4D quaternion output.

### A.2. Unguided Randomization: BG Filling

One of the modalities we have compared our results with is unguided randomization that applies augmentations during the data preprocessing step. While using the same modules and constraints as our deception network, its perturbations are conditioned on random values instead of latent codes from the input.

Since our DeceptionNet is capable of generating very complex backgrounds, we have also used complex noise types for unguided randomization to make the comparison more fair (see Fig. 9). Apart from a uniform white noise, two additional noise types were used: Perlin [34] and cellular noises [54]. Sample frequencies were sampled from the



Figure 7: **Cropped LineMOD Task Network:** Simple LeNet-like architecture followed by a dropout layer with a 50% rate and outputting both a class and pose vector.

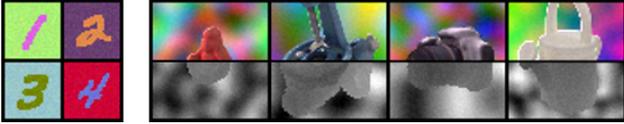


Figure 9: **Unguided samples:** We provide a sample of unguided augmentations for MNIST and LineMOD.

uniform distribution  $[0.0001, 0.1]$ . Both noise types were generated using the open source FastNoise library [33].

### A.3. Additional Qualitative Results

In this section, we present additional output examples of the deception networks for Synthetic Cropped LineMOD and SYNTHIA test cases.

The LineMOD deception network uses all of the deception modules presented in the paper, whereas the SYNTHIA deception network uses three modules: light (L), elastic distortions (DS), and foreground noise (N). The sample out-

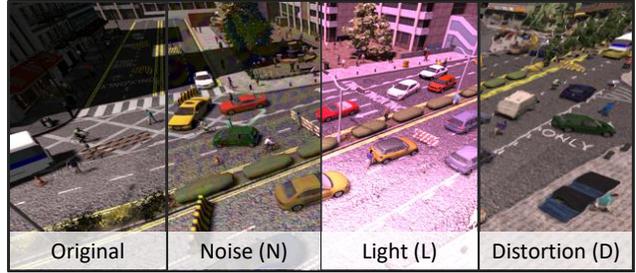


Figure 10: **Deceptive augmentations:** Augmentations applied for the SYNTHIA  $\rightarrow$  Cityscapes scenario.

puts from each of the above-mentioned modules are shown in Fig. 10. Moreover, Fig. 8 demonstrates the output of the deception network during the training process. One can see that the output becomes increasingly more sophisticated for recognition by the task network.

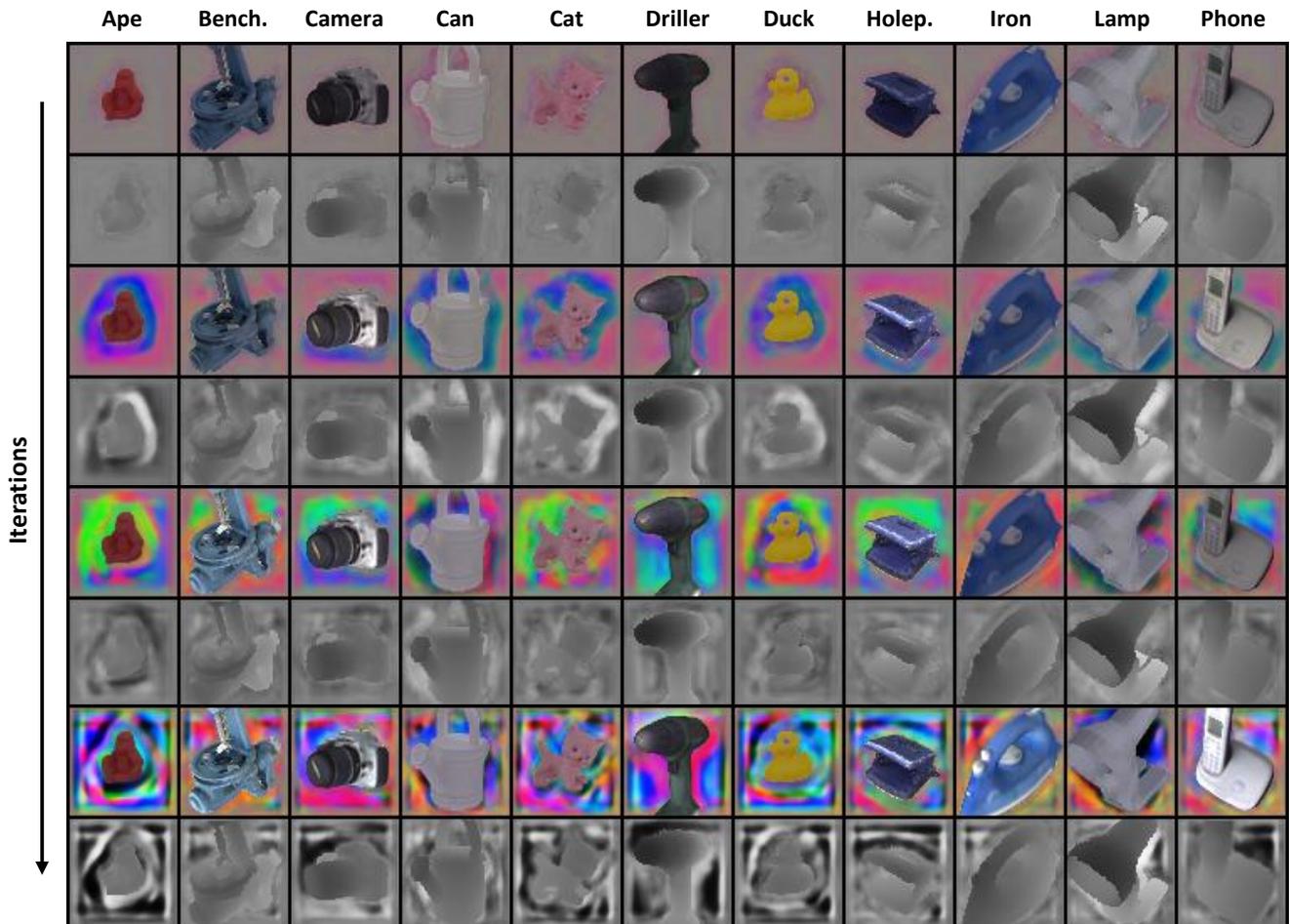


Figure 8: **Deceptive images  $x^d$  over consecutive iterations:** The output becomes increasingly more complex for  $T$ .

## References

- [1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.
- [2] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 2011.
- [3] Artem Babenko, Anton Slesarev, Alexander Chigorin, and Victor S. Lempitsky. Neural codes for image retrieval. *CoRR*, 2014.
- [4] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 2017.
- [5] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *NIPS*, 2016.
- [6] Tom B. Brown, Dandelion Man, Aurko Roy, Martn Abadi, and Justin Gilmer. Adversarial patch. In *NIPS*, 2017.
- [7] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *CVPR*, 2019.
- [8] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *CVPR*, 2018.
- [9] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *ICCV*, 2017.
- [10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.
- [11] Terrance DeVries and Graham W Taylor. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017.
- [12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.
- [13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016.
- [14] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.
- [15] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *ACCV*, 2012.
- [16] Stefan Hinterstoisser, Vincent Lepetit, Paul Wohlhart, and Kurt Konolige. On pre-trained image features and synthetic images for deep learning. 2017.
- [17] Tomáš Hodan, Pavel Haluza, Štěpán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *WACV*, 2017.
- [18] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.
- [19] Du Q Huynh. Metrics for 3d rotations: Comparison and analysis. *Journal of Mathematical Imaging and Vision*, 2009.
- [20] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*, 2015.
- [21] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *ICCV*, 2017.
- [22] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, 2016.
- [23] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018.
- [24] Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. Spigan: Privileged adversarial learning from simulation. *ICLR*, 2019.
- [25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: common objects in context. In *ECCV*, 2014.
- [26] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *CoRR*, 2016.
- [27] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *ICML*, 2015.
- [28] Jeffrey Mahler and Ken Goldberg. Learning deep policies for robot bin picking by simulating robust grasping sequences. In *CoRL*, 2017.
- [29] Fabian Manhardt, Wadim Kehl, Nassir Navab, and Federico Tombari. Deep model-based 6d pose refinement in rgb. In *ECCV*, 2018.
- [30] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. *ICCV*, 2017.
- [31] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Józefowicz, Bob McGrew, Jakub W. Pachocki, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *CoRR*, 2018.
- [32] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.
- [33] Jordan Peck. Fastnoise library. <https://github.com/Auburns/FastNoise>, 2016.

- [34] Ken Perlin. Improving noise. In *ACM Transactions on Graphics (TOG)*, 2002.
- [35] Bui Tuong Phong. Illumination for computer generated pictures. *Communications of the ACM*, 1975.
- [36] Benjamin Planche, Sergey Zakharov, Ziyang Wu, Andreas Hutter, Harald Kosch, and Slobodan Ilic. Seeing beyond appearance-mapping real images into geometrical domains for unsupervised cad-based recognition. *IROS*, 2019.
- [37] Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Feature mapping for learning fast and accurate 3d pose inference from synthetic images. In *CVPR*, 2018.
- [38] Alexander J Ratner, Henry Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. Learning to compose domain-specific transformations for data augmentation. In *Advances in neural information processing systems*, 2017.
- [39] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016.
- [40] Fereshteh Sadeghi and Sergey Levine. Cad2rl: Real single-image flight without a single real image. *arXiv preprint arXiv:1611.04201*, 2016.
- [41] Fereshteh Sadeghi and Sergey Levine. CAD2RL: Real single-image flight without a single real image. In *Robotics: Science and Systems (RSS)*, 2017.
- [42] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *CVPR*, 2018.
- [43] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2017.
- [44] Patrice Y Simard, David Steinkraus, John C Platt, et al. Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR*, 2003.
- [45] Leon Sixt, Benjamin Wild, and Tim Landgraf. Rendergan: Generating realistic labeled data. *Frontiers in Robotics and AI*, 2018.
- [46] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 2019.
- [47] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3D Orientation Learning for 6D Object Detection from RGB Images. In *ECCV*, 2018.
- [48] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *ICLR*, 2017.
- [49] 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. *IROS*, 2017.
- [50] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018.
- [51] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *CVPR*, 2017.
- [52] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, 2014.
- [53] Paul Wohlhart and Vincent Lepetit. Learning descriptors for object recognition and 3d pose estimation. In *CVPR*, 2015.
- [54] Steven Worley. A cellular texture basis function. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 291–294. ACM, 1996.
- [55] Sergey Zakharov, Benjamin Planche, Ziyang Wu, Andreas Hutter, Harald Kosch, and Slobodan Ilic. Keep it unreal: Bridging the realism gap for 2.5d recognition with geometry priors only. *3DV*, 2018.
- [56] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *ICCV*, 2019.
- [57] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, 2013.
- [58] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*, 2017.
- [59] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.