

# Visualizing Adapted Knowledge in Domain Transfer

Yunzhong Hou      Liang Zheng  
Australian National University  
{firstname.lastname}@anu.edu.au

## Abstract

A source model trained on source data and a target model learned through unsupervised domain adaptation (UDA) usually encode different knowledge. To understand the adaptation process, we portray their knowledge difference with image translation. Specifically, we feed a translated image and its original version to the two models respectively, formulating two branches. Through updating the translated image, we force similar outputs from the two branches. When such requirements are met, differences between the two images can compensate for and hence represent the knowledge difference between models. To enforce similar outputs from the two branches and depict the adapted knowledge, we propose a source-free image translation method that generates source-style images using only target images and the two models. We visualize the adapted knowledge on several datasets with different UDA methods and find that generated images successfully capture the style difference between the two domains. For application, we show that generated images enable further tuning of the target model without accessing source data. Code available at [https://github.com/hou-yz/DA\\_visualization](https://github.com/hou-yz/DA_visualization).

## 1. Introduction

Domain transfer or domain adaptation aims to bridge the distribution gap between source and target domains. Many existing works study the unsupervised domain adaptation (UDA) problem, where the target domain is unlabeled [27, 6, 46, 1, 11]. In this process, we are interested in what knowledge neural networks learn and adapt.

Essentially, we should visualize the knowledge difference between models: a source model trained on the source domain, and a target model learned through UDA for the target domain. We aim to portray the knowledge difference with image generation. Given a translated image and its original version, we feed the two images to the source and the target model, respectively. It is desired that differences between image pairs can compensate for the knowledge dif-

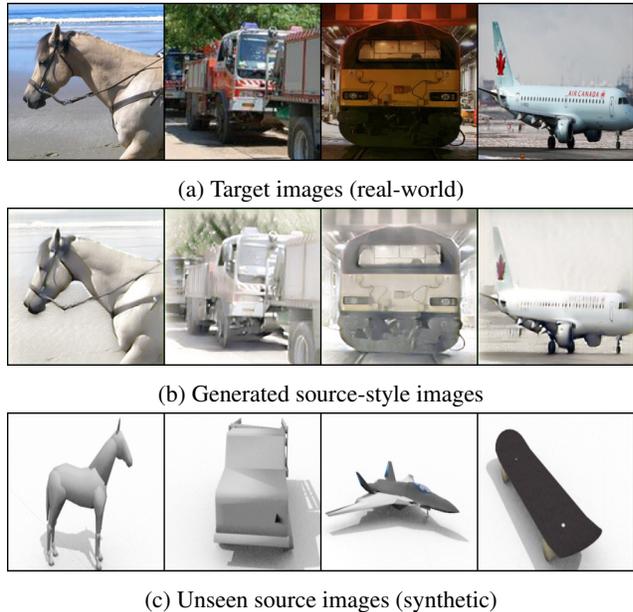


Figure 1: Visualization of adapted knowledge in unsupervised domain adaptation (UDA) on the VisDA dataset [38]. To depict the knowledge difference, in our source-free image translation (SFIT) approach, we generate source-style images (b) from target images (a). Instead of accessing source images (c), the training process is guided entirely by the source and target models, so as to faithfully portray the knowledge difference between them.

ference between models, leading to similar outputs from the two branches (two images fed to two different models). Achieving this, we could also say that the image pair represent the knowledge difference.

This visualization problem is very challenging and heretofore yet to be studied in the literature. It focuses on a relatively understudied field in transfer learning, where we distill knowledge differences from *models* and embed it in generated *images*. A related line of works, traditional image translation, generates images in the desired style utilizing content images and style images [7, 13, 48], and is applied

in pixel-level alignment methods for UDA [26, 2, 44, 11]. However, relying on *images from both domains* to indicate the style difference, such works cannot faithfully portray the knowledge difference between source and target *models*, and are unable to help us understand the adaptation process.

In this paper, we propose a source-free image translation (SFIT) approach, where we translate target images to the source style without using source images. The exclusion of source images prevents the system from relying on *image pairs* for style difference indication, and ensures that the system only learns from the two *models*. Specifically, we feed translated source-style images to the source model and original target images to the target model, and force similar outputs from these two branches by updating the generator network. To this end, we use the traditional knowledge distillation loss and a novel relationship preserving loss, which maintains relative channel-wise relationships between feature maps. We show that the proposed relationship preserving loss also helps to bridge the domain gap while changing the image style, further explaining the proposed method from a domain adaptation point of view. Some results of our method are shown in Fig. 1. We observe that even under the source-free setting, knowledge from the two models can still power the style transfer from the target style to the source style (SFIT decreases color saturation and whitens background to mimic the *unseen* source style).

On several benchmarks [19, 36, 39, 38], we show that generated images from the proposed SFIT approach significantly decrease the performance gap between the two models, suggesting a successful distillation of adapted knowledge. Moreover, we find SFIT transfers the image style at varying degrees, when we use different UDA methods on the same dataset. This further verifies that the SFIT visualizations are faithful to the models and that different UDA methods can address varying degrees of style differences. For applications, we show that generated images can serve as an additional cue and enable further tuning of target models. This also falls into a demanding setting of UDA, source-free domain adaptation (SFDA) [17, 20, 24], where the system has no access to source images.

## 2. Related Work

**Domain adaptation** aims to reduce the domain gap between source and target domains. Feature-level distribution alignment is a popular strategy [27, 6, 46, 40]. Long *et al.* [27] use the maximum mean discrepancy (MMD) loss for this purpose. Tzeng *et al.* [46] propose an adversarial method, ADDA, with a loss function based on the generative adversarial network (GAN). Pixel-level alignment with image translation is another popular choice in UDA [26, 2, 44, 42, 1, 11]. Hoffman *et al.* propose the CyCADA [11] method based on CycleGAN [48] image translation. Other options are also investigated. Saito *et al.* [40]

align the task-specific decision boundaries of two classifiers. Source-free domain adaptation (SFDA) does *not* use the source data and therefore greatly alleviates the privacy concerns in releasing the source dataset. As an early attempt, AdaBN [22] adapts the statistics of the batch normalization layers in the source CNN to the target domain. Li *et al.* [20] generate images with the same distribution of the target images and use them to fine-tune the classifier. Liang *et al.* [24] fine-tune a label smoothed [34] source model on the target images. To the authors' knowledge, there is still yet to be any visualization that can indicate what models learn during adaptation.

**Knowledge distillation** transfers knowledge from a pre-trained teacher model to a student model [10], by maximizing the mutual information between teacher outputs and student outputs. Some existing works consider the relationship between instance or pixels for better distillation performance [45, 23, 37]. Instead of distilling teacher knowledge on a given training dataset, data-free knowledge distillation (DFKD) [30, 35, 3, 33, 8, 47] first *generates* training data and then *learns* a student network on this generated dataset. Training data can be generated by aligning feature statistics [30, 8, 47], enforcing high teacher confidence [30, 35, 3, 8, 47], and adversarial generation of hard examples for the student [33, 47]. In [8, 47], batch normalization statistics are matched as regularization. Our work, while also assuming no access to source images, differs significantly from these works in that our image translation has to portray the transferred knowledge, whereas data-free knowledge distillation just generates whatever images that satisfy the teacher networks.

**Image translation** renders the same content in a different artistic style. Some existing works adopt a GAN-based system for this task [26, 44, 14, 48, 11], while others use a pre-trained feature extractor for style transfer [7, 15, 32, 13]. Zhu *et al.* adopt a cycle consistency loss in the image translation loop to train the CycleGAN system [48]. Gatys *et al.* consider a content loss on high-level feature maps, and a style loss on feature map statistics for style transfer [7]. Huang and Belongie [13] propose a real-time AdaIN style transfer method by changing the statistics in instance normalization layers. Based on AdaIN, Karras *et al.* propose StyleGAN for state-of-the-art image generation [16]. Our work differs from traditional image translations in that rather than *images* from the two domains, only *models* from two domains are used to guide the image update.

## 3. Problem Formulation

To achieve our goal, *i.e.*, visualizing adapted knowledge in UDA, we translate a image  $x$  from a certain domain to a new image  $\tilde{x}$ . It is hoped that feeding the original image to its corresponding model (trained for that certain domain) and the generated image to the other model can minimize

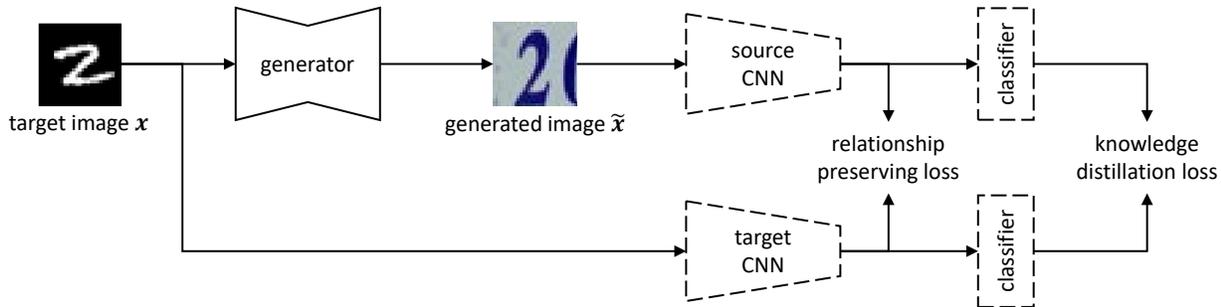


Figure 2: The proposed source-free image translation (SFIT) method for visualizing the adapted knowledge in UDA. The system includes two branches: original target images are fed to the target CNN, whereas generated source-style images are fed to the source CNN. We minimize the knowledge distillation loss and the relationship preserving loss, and update the generator network accordingly. If the two branches get similar results while adopting different models, then the difference between the original target image  $x$  and the generated source-style image  $\tilde{x}$  should be able to *mitigate* and therefore *exhibit* the knowledge difference between models. Dashed lines indicate fixed network parameters.

the output difference between these two branches. The update process is directed only by the source model  $f_S(\cdot)$  and the target model  $f_T(\cdot)$ , and we prevent access to the images from the other domain to avoid distractions. We formulate the task of visualizing adapted knowledge as a function of the source model, the target model, and the image from a certain domain,

$$\mathcal{G}(f_S, f_T, x) \rightarrow \tilde{x}. \quad (1)$$

In contrast, traditional image translation needs access to images from both domains for content and style specification. In addition to the source image  $x_S$  and the target image  $x_T$ , traditional image translation also relies on certain neural network  $d(\cdot)$  as the criterion. Instead of the source and target models, ImageNet [4] pre-trained VGG [43] and adversarially trained discriminator networks are used for this task in style transfer [7, 13] and GAN-based methods [48, 11], respectively. Traditional image translation task can thus be formulated as,

$$\mathcal{G}(d, x_S, x_T) \rightarrow \tilde{x}. \quad (2)$$

Comparing our goal in Eq. 1 and traditional image translation in Eq. 2, we can see a clear gap between them. Traditional image translation learns the style difference indicated by *images* from both domains, whereas our goal is to learn to visualize the knowledge difference between the source and target *models*  $f_S(\cdot), f_T(\cdot)$ .

## 4. Method

To investigate what neural networks learn in domain adaptation, we propose source-free image translation (SFIT), a novel method that generates source-style images from original target images, so as to mitigate and represent the knowledge difference between models.

### 4.1. Overview

Following many previous UDA works [6, 27, 46, 24], we assume that only the feature extractor CNN in the source model is adapted to the target domain. Given a source CNN  $f_S(\cdot)$  and a target CNN  $f_T(\cdot)$  sharing the same classifier  $p(\cdot)$ , we train a generator  $g(\cdot)$  for the SFIT task. We discuss why we choose this translation direction in Section 4.3. As the training process is source-free, for simplicity, we refer to the target image as  $x$  instead of  $x_T$  in what follows.

As shown in Fig. 2, given a generated image  $\tilde{x} = g(x)$ , the source model outputs a feature map  $f_S(\tilde{x})$  and a probability distribution  $p(f_S(\tilde{x}))$  over all  $C$  classes. To depict the adapted knowledge in the generated image, in addition to the traditional knowledge distillation loss, we introduce a novel relationship preserving loss, which maintains relative channel-wise relationships between the target-image-target-model feature map  $f_T^N(x)$  and the generated-image-source-model feature map  $f_S^N(\tilde{x})$ .

### 4.2. Loss Functions

With a knowledge distillation loss  $\mathcal{L}_{KD}$  and a relationship preserving loss  $\mathcal{L}_{RP}$ , we have the overall loss function,

$$\mathcal{L} = \mathcal{L}_{KD} + \mathcal{L}_{RP}. \quad (3)$$

In the following sections, we detail the loss terms.

**Knowledge distillation loss.** In the proposed source-free image translation method, portraying the adapted knowledge in the target model  $f_T(\cdot)$  with source model and generator combined  $f_S(g(\cdot))$  can be regarded as a special case of knowledge distillation, where we aim to distill the adapted knowledge to the generator. In this case, we include a knowledge distillation loss between generated-image-source-model output  $p(f_S(\tilde{x}))$  and target-image-

target-model output  $p(f_T(\mathbf{x}))$ ,

$$\mathcal{L}_{\text{KD}} = \mathcal{D}_{\text{KL}}(p(f_T(\mathbf{x})), p(f_S(\tilde{\mathbf{x}}))), \quad (4)$$

where  $\mathcal{D}_{\text{KL}}(\cdot, \cdot)$  denotes the Kullback-Leibler divergence.

**Relationship preserving loss.** Similar classification outputs indicate a successful depiction of the target model knowledge on the generated images. As we assume a fixed classifier for UDA, the global feature vectors from the target image target CNN and the generated image source CNN should be similar after a successful knowledge distillation. Promoting similar channel-wise relationships between feature maps  $f_T^N(\mathbf{x})$  and  $f_S^N(\tilde{\mathbf{x}})$  helps to achieve this goal.

Previous knowledge distillation works preserve relative batch-wise or pixel-wise relationships [45, 23]. However, they are not suitable here for the following reasons. Relative batch-wise relationships can not effectively supervise the per-image generation task. Besides, the efficacy of pixel-wise relationship preservation can be overshadowed by the global pooling before the classifier. By contrast, channel-wise relationships are computed on a per-image basis, and are effective even after global pooling. As such, we choose the channel-wise relationship preserving loss that is computed in the following manner.

Given feature maps  $f_T^N(\mathbf{x}), f_S^N(\tilde{\mathbf{x}})$ , we first reshape them into feature vectors  $\mathcal{F}_S$  and  $\mathcal{F}_T$ ,

$$\begin{aligned} f_S^N(\tilde{\mathbf{x}}) \in \mathbb{R}^{D \times H \times W} &\rightarrow \mathcal{F}_S \in \mathbb{R}^{D \times HW}, \\ f_T^N(\mathbf{x}) \in \mathbb{R}^{D \times H \times W} &\rightarrow \mathcal{F}_T \in \mathbb{R}^{D \times HW}, \end{aligned} \quad (5)$$

where  $D, H$ , and  $W$  are the feature map depth (number of channels), height, and width, respectively. Next, we calculate their channel-wise self correlations, or Gram matrices,

$$G_S = \mathcal{F}_S \cdot \mathcal{F}_S^T, \quad G_T = \mathcal{F}_T \cdot \mathcal{F}_T^T, \quad (6)$$

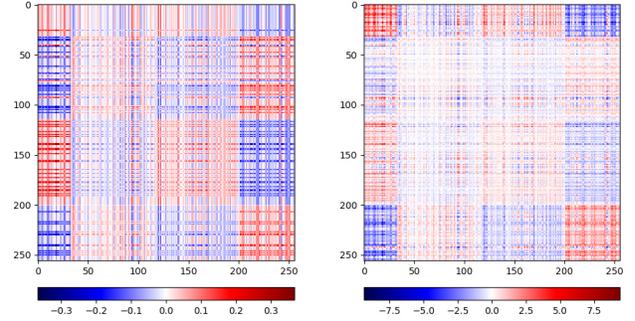
where  $G_S, G_T \in \mathbb{R}^{D \times D}$ . Like other similarity preserving losses for knowledge distillation [45, 23], we then apply the row-wise  $\mathcal{L}_2$  normalization,

$$\tilde{G}_{S[i,:]} = \frac{G_{S[i,:]}}{\|G_{S[i,:]}\|_2}, \quad \tilde{G}_{T[i,:]} = \frac{G_{T[i,:]}}{\|G_{T[i,:]}\|_2}, \quad (7)$$

where  $[i, :]$  indicates the  $i$ -th row in a matrix. At last, we define the relationship preserving loss as mean square error (MSE) between the normalized Gram matrices,

$$\mathcal{L}_{\text{RP}} = \frac{1}{D} \left\| \tilde{G}_S - \tilde{G}_T \right\|_F^2, \quad (8)$$

where  $\|\cdot\|_F$  denotes the Frobenius norm (entry-wise  $\mathcal{L}_2$  norm for matrix). In Section 4.3, we further discuss the relationship preserving loss from the viewpoint of style transfer and domain adaptation, and show it can align feature map distributions in a similar way as style loss [7] for style transfer and MMD loss [27] for UDA, forcing the generator to portray the knowledge difference between the two models.



(a) Relationship preserving loss (b) Traditional style loss

Figure 3: Comparison between the proposed relationship preserving loss and the traditional style loss. In (a) and (b), given 256-dimensional feature maps, we show differences of row-wise normalized Gram matrix (Eq. 8) and original Gram matrix (Eq. 9). Deeper colors indicate larger differences and therefore stronger supervision. The proposed relationship preserving loss provides evenly distributed supervision for all channels, whereas the traditional style loss focuses primarily on several channels.

### 4.3. Discussions

**Why transfer target images to the source style.** According to the problem formulation in Eq. 1, we should be able to visualize the adapted knowledge by generating either source-style images from target images, or target-style images from source images. In this paper, we select the former direction as it might be further applied in fine-tuning the target model (see Section 5.4 for application).

**Style transfer with the relationship preserving loss.** The proposed relationship preserving loss can be regarded as a normalized version of the traditional style loss introduced by Gatys *et al.* [7],

$$\mathcal{L}_{\text{style}} = \frac{1}{D^2} \|G_S - G_T\|_F^2, \quad (9)$$

which computes MSE between Gram matrices.

In the proposed relationship preserving loss (Eq. 8), instead of original Gram matrices, we use a row-wise normalized version. It focuses on *relative* relationships between channels, rather than *absolute* values of self correlations as in the traditional style loss. Preserving relative relationships provides more evenly-distributed supervision for all channels, instead of prioritizing several channels as in the traditional style loss (Fig. 3). Experiments find this evenly-distributed supervision better preserves the foreground object and allows for easier training and higher performance, while also changing the image style (see Section 5.5).

**Distribution alignment with the relationship preserving loss.** As proved by Li *et al.* [21], the traditional style

loss  $\mathcal{L}_{\text{style}}$  is equivalent to the MMD loss [27] for UDA. We can also see the relationship preserving loss as a modified version of the MMD loss, which aligns the distribution of the generated image source CNN feature map  $f_S^N(\tilde{x})$  to the target image target CNN feature map  $f_T^N(x)$ .

## 5. Experiments

### 5.1. Datasets

We visualize the knowledge difference between source and target models on the following datasets.

**Digits** is a standard UDA benchmark that focuses on 10-class digit recognition. Specifically, we experiment on MNIST [19], USPS, and SVHN [36] datasets.

**Office-31** [39] is a standard benchmark for UDA that contains 31 classes from three distinct domains: Amazon (A), Webcam (W), and DSLR (D).

**VisDA** [38] is a challenging large-scale UDA benchmark for domain adaptation from 12 classes of synthetic CAD model images to real-world images in COCO [25].

### 5.2. Implementation Details

**Source and target models.** We adopt source and target models from a recent SFDA work SHOT-IM [24] if not specified. SFDA is a special case of UDA, and it is even more interesting to see what machines learn in the absence of source data. We also include UDA methods DAN [27] and ADDA [46] for SFIT result comparisons. For network architectures, on digits dataset, following Long *et al.* [28], we choose a LeNet [18] classifier. On Office-31 and VisDA, we choose ResNet-50 and ResNet-101 [9], respectively.

**Generator for SFIT.** We use a modified CycleGAN [48] architecture with 3 residue blocks due to memory concerns.

**Training schemes.** During training, we first initialize the generator as a transparent filter, which generates images same as the original input. To this end, we use the ID loss  $\mathcal{L}_{\text{ID}} = \|\tilde{x} - x\|_1$  and the content loss  $\mathcal{L}_{\text{content}} = \|f_S^N(\tilde{x}) - f_S^N(x)\|_2$  to train the generator for initialization. The initialization performance is shown in Table 4, where we can see a mild 1.9% accuracy drop from original target images. Then, we train the generator with the overall loss function in Eq. 3 for visualizing the adapted knowledge. Specifically, we use an Adam optimizer with a cosine decaying [31] learning rate starting from  $3 \times 10^{-4}$  and a batch size of 16. All experiments are finished using one RTX-2080Ti GPU.

### 5.3. Evaluation

**Recognition accuracy on generated images.** To examine whether the proposed SFIT method can depict the knowledge difference, in Table 1-3, we report recognition results using the generated-image-source-model branch (referred as “generated images”). On the digits dataset,

Method	SVHN→MNIST	USPS→MNIST	MNIST→USPS
Source only [11]	67.1±0.6	69.6±3.8	82.2±0.8
DAN [27]	71.1	-	81.1
DANN [6]	73.8	73	85.1
CDAN+E [28]	89.2	98.0	95.6
CyCADA [11]	90.4±0.4	96.5±0.1	95.6±0.4
MCD [40]	96.2±0.4	94.1±0.3	94.2±0.7
GTA [41]	92.4±0.9	90.8±1.3	95.3±0.7
3C-GAN [20]	99.4±0.1	99.3±0.1	97.3±0.2
Source model [24]	72.3±0.5	90.5±1.6	72.7±2.3
Target model [24]	98.8±0.1	98.1±0.5	97.9±0.2
Generated images	98.6±0.1	97.4±0.3	97.6±0.3

Table 1: Classification accuracy (%) on digits datasets. In Table 1-3, “Generated images” refers to feeding images generated by SFIT to the source model.

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg.
ResNet-50 [9]	68.4	96.7	99.3	68.9	62.5	60.7	76.1
DAN [27]	80.5	97.1	99.6	78.6	63.6	62.8	80.4
DANN [6]	82.6	96.9	99.3	81.5	68.4	67.5	82.7
ADDA [46]	86.2	96.2	98.4	77.8	69.5	68.9	82.9
JAN [29]	86.0	96.7	99.7	85.1	69.2	70.7	84.6
CDAN+E [28]	94.1	98.6	100.0	92.9	71.0	69.3	87.7
GTA [41]	89.5	97.9	99.8	87.7	72.8	71.4	86.5
3C-GAN [20]	93.7	98.5	99.8	92.7	75.3	77.8	89.6
Source model [24]	76.9	95.6	98.5	80.3	60.6	63.4	79.2
Target model [24]	90.8	98.4	99.9	88.8	73.6	71.7	87.2
Generated images	89.1	98.1	99.9	87.3	69.8	68.7	85.5
Fine-tuning	91.8	98.7	99.9	89.9	73.9	72.0	87.7

Table 2: Classification accuracy (%) on the Office-31 dataset. In Table 2 and Table 3, “Fine-tuning” refers to target model fine-tuning result with both generated images and target images (see Section 5.4 for more details).

Method	plane	bicycl	bus	car	horse	knife	meycl	person	plant	sktbrd	train	truck	per-class
ResNet-101 [9]	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DAN [27]	87.1	63.0	76.5	42.0	90.3	42.9	85.9	53.1	49.7	36.3	85.8	20.7	61.1
DANN [6]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
JAN [29]	75.7	18.7	82.3	86.3	70.2	56.9	80.5	53.8	92.5	32.2	84.5	54.5	65.7
ADDA [46]	88.8	65.7	85.6	53.1	74.9	96.2	83.3	70.7	75.9	26.4	83.9	32.4	69.7
MCD [40]	87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
CDAN+E [28]	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.9
SE [5]	95.9	87.4	85.2	58.6	96.2	95.7	90.6	80.0	94.8	90.8	88.4	47.9	84.3
3C-GAN [20]	94.8	73.4	68.8	74.8	93.1	95.4	88.6	84.7	89.1	84.7	83.5	48.1	81.6
Source model [24]	58.3	17.6	54.2	69.9	64.4	5.5	82.2	30.7	62.2	24.6	86.2	6.0	46.8
Target model [24]	92.5	84.7	81.3	54.6	90.5	94.7	80.9	79.1	90.8	81.5	87.9	50.1	80.7
Generated images	88.9	65.8	83.0	61.7	88.5	76.8	89.5	69.6	91.4	51.9	84.3	34.3	73.8
Fine-tuning	94.3	79.0	84.9	63.6	92.6	92.0	88.4	79.1	92.2	79.8	87.6	43.0	81.4

Table 3: Classification accuracy (%) on the VisDA dataset.

in terms of performance gaps, the knowledge differences between source and target models are 26.5% on SVHN→MNIST, 7.6% on USPS→MNIST, and 25.2% on MNIST→USPS. Generated images from SFIT bridges these differences to 0.2%, 0.7%, and 0.3%, respectively. On the Office-31 dataset, the performance gap between the two models is 8.0% on average, and the generated images shrink this down to 1.7%. Notably, the performance drops from the target-image-target-model branch to the generated-image-source-model branch are especially pronounced on D→A and W→A, two settings that transfer Amazon images with white or no background to real-world background

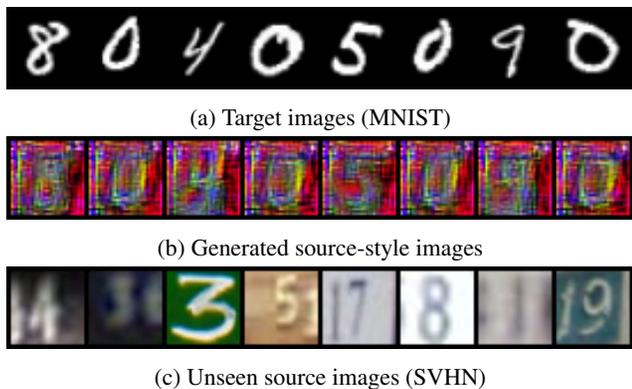


Figure 4: Results from the SFIT method on digits datasets SVHN→MNIST. In Fig. 1 and Fig. 4-6, we show in (a): target images, (b): generated source-style images, each of which corresponds to the target image above it, and (c): the unseen source images. For gray-scale target images from MNIST, our SFIT approach adds random RGB colors to mimic the full-color style in the unseen source (SVHN) without changing the content.



Figure 5: Results from the SFIT method on the Office-31 dataset Amazon→Webcam. Our translation method whitens backgrounds while increasing contrast ratios of the object (Webcam) for more appealing appearances as in the online shopping images (Amazon).

in Webcam or DSLR. In fact, in experiments we find generating an overall consistent colored background is very demanding, and the system usually generates a colored background around the outline of the object. On the VisDA

dataset, generated images bridge the performance gap from 33.9% to 6.9%, even under a more demanding setting and a larger domain gap going from real-world images to synthetic CAD model images. Overall, on all three datasets, generated images significantly mitigate the knowledge difference in terms of performance gaps, indicating that the proposed SFIT method can successfully distill the adapted knowledge from the target model to the generated images.

#### Visualization of source-free image translation results.

For digits datasets SVHN→MNIST (Fig. 4), the generator learns to add RGB colors to the gray-scale MNIST (target) images, which mimics the full-color SVHN (source) images. For Office-31 dataset Amazon→Webcam (Fig. 5), the generated images whiten the background, while having a white or no background rather than real-world background is one of the main characteristics of the Amazon (source) domain when compared to Webcam (target). Moreover, Amazon online shopping images also have higher contrast ratios for more appealing appearances, and our translated images also capture these characteristics, *e.g.*, keys in the calculator, case of the desktop computer. For VisDA dataset SYN→REAL (Fig. 1 and Fig. 6), the generator learns to decrease the overall saturation of the real-world (target) objects which makes them more similar to the synthetic (source) scenario, while at the same time whitens the background, *e.g.*, horse, truck, and plane in Fig. 1, car and skateboard in Fig. 6, and brings out the green color in the plants. Overall, image generation results exhibit minimal content changes from target images, while successfully capturing the *unseen* source style.

In terms of visual quality, it is noteworthy that generation results for digits datasets SVHN→MNIST contain colors and patterns that are not from the source domain, whereas our results on the Office-31 dataset and VisDA dataset are more consistent with the unseen source. Due to the lack of source images, rather than traditional image translation approaches [7, 13, 11, 44], SFIT only relies on source and target models, and portrays adapted knowledge according to the two models. Since a weaker LeNet classifier is used for the digits dataset, it is easier to generate images that satisfy the proposed loss terms without requiring the generated images to perfectly mimic the source style. On Office-31 and VisDA datasets, given stronger models like ResNet, it is harder to generate images that can satisfy the loss terms. Stricter restrictions and longer training time lead to generation results more coherent with *unseen* source images that also have better visual quality.

**Visualization for different UDA methods.** In Fig. 7, we show SFIT visualization results using different UDA methods. Given source and target domain, a traditional image translation method generates a certain type of images regardless of the UDA methods, indicating its incapability of presenting the knowledge difference between mod-

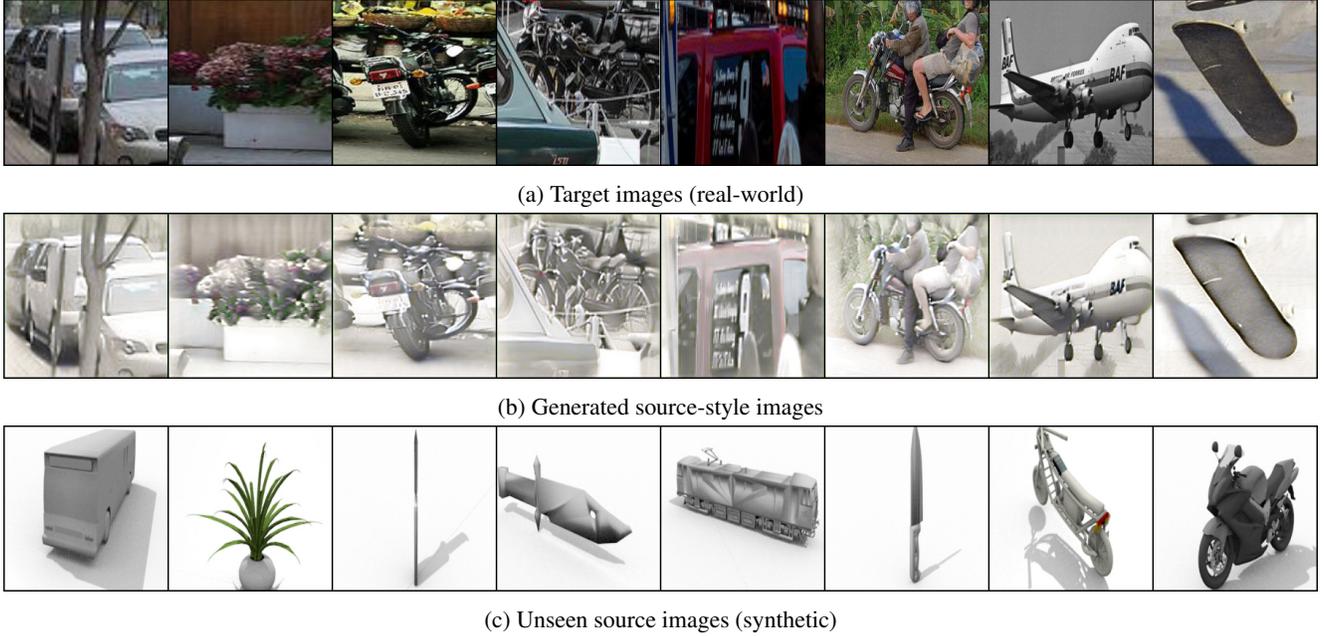


Figure 6: Results from the SFIT method on the VisDA dataset SYN→REAL. Our translation method decreases the target (real-world) image saturation and whitens the background while keeping the semantics unchanged.

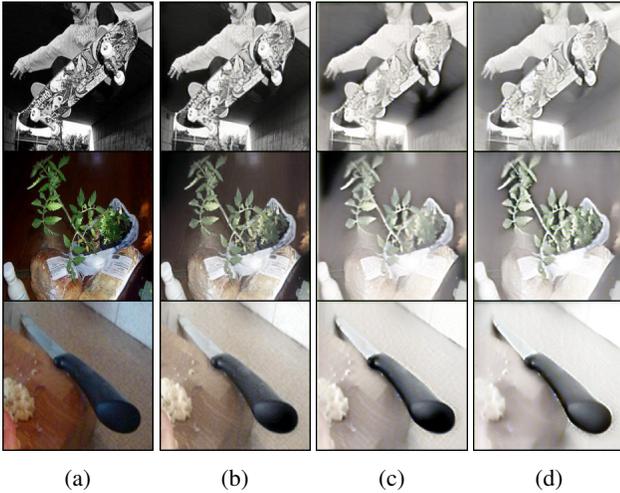


Figure 7: SFIT results on VisDA dataset with different UDA methods. (a) Target images; (b) DAN [27]; (c) ADDA [46]; (d) SHOT-IM [24].

els. In contrast, the proposed SFIT method generates different images for different UDA methods. Specifically, when comparing visualization results of the adapted knowledge in DAN [27], ADDA [46], and SHOT-IM [24], we find stronger UDA methods can better transfer the target style to the unseen source style. As shown in Fig. 7, in terms of whitening the background for style transfer, SFIT re-

sults on ADDA are less coherent than SHOT-IM but better than DAN. This further verifies that our SFIT method indeed visualizes the knowledge difference between models, and stronger adaptation methods can better endure the style difference (leading to larger knowledge difference and thus stronger style transfer results).

#### 5.4. Application

The generated images from SFIT allows for further tuning of the target model in SFDA systems, where no source image is available. We include a diversity loss on all training samples to promote even class-wise distributions,

$$\mathcal{L}_{\text{div}} = -\mathcal{H}(\mathbb{E}_{\mathbf{x} \sim P_{\text{target}}(\mathbf{x})} [p(f_{\text{T}}(\mathbf{x}))]), \quad (10)$$

where  $\mathcal{H}(\cdot)$  denotes the information entropy function. We also include a pseudo-label fine-tuning loss, if pseudo label  $\hat{y}_{\text{S}} = \arg \max p(f_{\text{S}}(\tilde{\mathbf{x}}))$  from the generated-image-source-model branch equals to the pseudo label  $\hat{y}_{\text{T}} = \arg \max p(f_{\text{T}}(\mathbf{x}))$  from the target-image-target-model branch. We then use this pseudo label  $\hat{y} = \hat{y}_{\text{S}} = \hat{y}_{\text{T}}$  to fine-tune the target model,

$$\mathcal{L}_{\text{pseudo}} = \begin{cases} \mathcal{H}(p(f_{\text{T}}(\mathbf{x})), \hat{y}), & \text{if } \hat{y} = \hat{y}_{\text{S}} = \hat{y}_{\text{T}}, \\ 0, & \text{else,} \end{cases} \quad (11)$$

where  $\mathcal{H}(\cdot, \cdot)$  denotes the cross entropy function. We combine these two loss terms in Eq. 10 and Eq. 11 to give an overall fine-tuning loss  $\mathcal{L}_{\text{FT}} = \mathcal{L}_{\text{div}} + \mathcal{L}_{\text{pseudo}}$ .

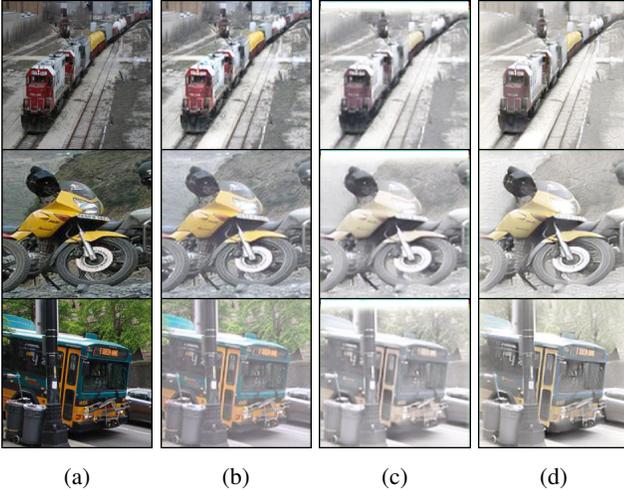


Figure 8: Visualization results on VisDA dataset with different distribution alignment methods. (a) Target images; (b) BN stats alignment [12]; (c) traditional style loss [7]; (d) relationship preserving loss.

As an additional cue, supervision from generated-image-source-model further boosts target model SFDA performance. On Office-31, fine-tuning brings a performance improvement of 0.4% according to Table 2. On VisDA, fine-tuning improves the target model accuracy by 0.7% as shown in Table 3. These improvements are statistically very significant (*i.e.*,  $p$ -value  $< 0.001$  over 5 runs), and introduce a real-world application for images generated by SFIT.

## 5.5. Comparison and Variant Study

**Comparison with the BatchNorm statistics alignment method [12].** Hou *et al.* propose to match the batch-wise feature map statistics so as to directly generate images that mimic the source style. Specifically, they explore the BatchNorm (BN) statistics stored in the BN layers in the source model for style indication, and match them against that of the generated images. Using their approach, we can mildly change the image to the unseen source style (see Fig. 8) and slightly reduce the performance difference between the two branches (see Table 4). With that said, their lack of output alignments between the *two branches* (only supervisions from the *source branch*) results in much lower quantitative performance and under-performing style transfer quality when compared to the proposed method.

**Effect of the knowledge distillation loss.** The knowledge distillation loss transfers the adapted knowledge to the generated images, and the removal of it results in a 1.1% performance drop.

**Effect of the relationship preserving loss.** As shown in Fig. 8, the traditional style loss can successfully transfer the target image to the source style on its own. However, using

Variant	$\mathcal{L}_{KD}$	$\mathcal{L}_{RP}$	accuracy (%)
Target image	-	-	46.8
Initialized $g(\cdot)$			44.9
BN stats alignment [12]			51.7
w/o $\mathcal{L}_{KD}$		✓	72.7
w/o $\mathcal{L}_{RP}$	✓		71.2
$\mathcal{L}_{RP} \rightarrow \mathcal{L}_{style}$	✓	$\mathcal{L}_{style}$ [7]	66.4
$\mathcal{L}_{RP} \rightarrow \mathcal{L}_{batch}$	✓	$\mathcal{L}_{batch}$ [45]	71.2
$\mathcal{L}_{RP} \rightarrow \mathcal{L}_{pixel}$	✓	$\mathcal{L}_{pixel}$ [23]	70.9
SFIT	✓	✓	73.8

Table 4: Variant study on VisDA dataset. “Initialized  $g(\cdot)$ ” refers to our transparent filter initialization in Section 5.2.

it causes a 4.8% performance drop compared to the “w/o  $\mathcal{L}_{RP}$ ” variant (see Table 4), suggesting it being unsuitable for SFIT. On the other hand, the batch-wise or pixel-wise relationship preserving variants [45, 23] are found not useful, as they fail to improve over the “w/o  $\mathcal{L}_{RP}$ ” variant.

In contrast, the proposed channel-wise relationship preserving loss  $\mathcal{L}_{RP}$  can effectively improve the recognition accuracy on the generated images, as the inclusion of it leads to a 2.6% performance increase. Moreover, as shown in Fig. 8, similar to the traditional style loss, using only the relationship preserving loss can also effectively transfer the target image to the *unseen* source style. Besides, focusing on the *relative* channel-wise relationship instead of the *absolute* correlation values, the proposed relationship preserving loss can better maintain the foreground object (less blurry and more prominent) while transferring the overall image style, leading to higher recognition accuracy.

## 6. Conclusion

In this paper, we study the scientific problem of visualizing the adapted knowledge in UDA. Specifically, we propose a source-free image translation (SFIT) approach, which generates source-style images from original target images under the guidance of source and target models. Translated images on the source model achieve similar results as target images on the target model, indicating a successful depiction of the adapted knowledge. Such images also exhibit the source style, and the extent of style transfer follows the performance of UDA methods, which further verifies that stronger UDA methods can better address the distribution difference between domains. We show that the generated images can be applied to fine-tune the target model, and might help other tasks like incremental learning.

## Acknowledgement

This work was supported by the ARC Discovery Early Career Researcher Award (DE200101283) and the ARC Discovery Project (DP210102801).

## References

- [1] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017. 1, 2
- [2] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in neural information processing systems*, pages 343–351, 2016. 2
- [3] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3514–3522, 2019. 2
- [4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 3
- [5] Geoff French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 2018. 5
- [6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 1, 2, 3, 5
- [7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 1, 2, 3, 4, 6, 8
- [8] Matan Haroush, Itay Hubara, Elad Hoffer, and Daniel Soudry. The knowledge within: Methods for data-free model compression. *arXiv preprint arXiv:1912.01274*, 2019. 2
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5
- [10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2
- [11] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 1, 2, 3, 5, 6
- [12] Yunzhong Hou and Liang Zheng. Source free domain adaptation with image translation. *arXiv preprint arXiv:2008.07514*, 2020. 8
- [13] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 1, 2, 3, 6
- [14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2
- [15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 2
- [16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2
- [17] Jogendra Nath Kundu, Naveen Venkat, and R Venkatesh Babu. Universal source-free domain adaptation. *arXiv preprint arXiv:2004.04393*, 2020. 2
- [18] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5
- [19] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, 2, 2010. 2, 5
- [20] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2020. 2, 5
- [21] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI’17*, page 2230–2236. AAAI Press, 2017. 4
- [22] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018. 2
- [23] Zeqi Li, Ruwei Jiang, and Parham Aarabi. Semantic relation preserving knowledge distillation for image-to-image translation. In *European conference on computer vision*. Springer, 2020. 2, 4, 8
- [24] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, pages xx–xx, July 2020. 2, 3, 5, 7
- [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 *European conference on computer vision*, pages 740–755. Springer, 2014. 5
- [26] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016. 2
- [27] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015. 1, 2, 3, 4, 5, 7
- [28] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1645–1655, 2018. 5

- [29] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017. 5
- [30] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*, 2017. 2
- [31] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5
- [32] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4990–4998, 2017. 2
- [33] Paul Micaelli and Amos J Storkey. Zero-shot knowledge transfer via adversarial belief matching. In *Advances in Neural Information Processing Systems*, pages 9547–9557, 2019. 2
- [34] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pages 4694–4703, 2019. 2
- [35] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4743–4751, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 2
- [36] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bisacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. 2, 5
- [37] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 2
- [38] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge, 2017. 1, 2, 5
- [39] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 2, 5
- [40] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 2, 5
- [41] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018. 5
- [42] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017. 2
- [43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3
- [44] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *ICLR*, 2017. 2, 6
- [45] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1365–1374, 2019. 2, 4, 8
- [46] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017. 1, 2, 3, 5, 7
- [47] Hongxu Yin, Pavlo Molchanov, Jose M. Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deep-inversion. In *The IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020. 2
- [48] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 2, 3, 5