NoVA: Learning to See in Novel Viewpoints and Domains

Benjamin Coors[1,3] Alexandru Paul Condurache[2,3] Andreas Geiger[1]
[1]Autonomous Vision Group, MPI for Intelligent Systems and University of Tübingen
[2]Institute for Signal Processing, University of Lübeck [3]Robert Bosch GmbH

Abstract

Domain adaptation techniques enable the re-use and transfer of existing labeled datasets from a source to a target domain in which little or no labeled data exists. Recently, image-level domain adaptation approaches have demonstrated impressive results in adapting from synthetic to real-world environments by translating source images to the style of a target domain. However, the domain gap between source and target may not only be caused by a different style but also by a change in viewpoint. This case necessitates a semantically consistent translation of source images and labels to the style and viewpoint of the target domain. In this work, we propose the Novel Viewpoint Adaptation (NoVA) model, which enables unsupervised adaptation to a novel viewpoint in a target domain for which no labeled data is available. NoVA utilizes an explicit representation of the 3D scene geometry to translate source view images and labels to the target view. Experiments on adaptation to synthetic and real-world datasets show the benefit of NoVA compared to state-of-the-art domain adaptation approaches on the task of semantic segmentation.

1. Introduction

Deep neural networks for semantic segmentation require huge labeled datasets. However, as labeling is expensive and time-consuming, the re-use and transfer of existing labeled datasets is desirable whenever possible. Yet, in many cases available datasets do not exactly match the setup of the problem we are interested in but instead differ in style (e.g. simulation vs. reality), camera model (e.g. rectilinear vs. omnidirectional) or in camera viewpoint.

While the adaptation to a different image style has been extensively addressed in previous domain adaptation work [1, 19, 31, 36], viewpoint adaptation has not yet been widely considered. However, as demonstrated by our experiments, such a change in viewpoint can lead to a dramatic performance drop. Thus, we formally introduce the challenge of domain and viewpoint adaptation and propose the *Novel Viewpoint Adaptation* (NoVA) model. NoVA enables the adaptation of source domain data to the view and style of a target domain in which no labeled data is available.

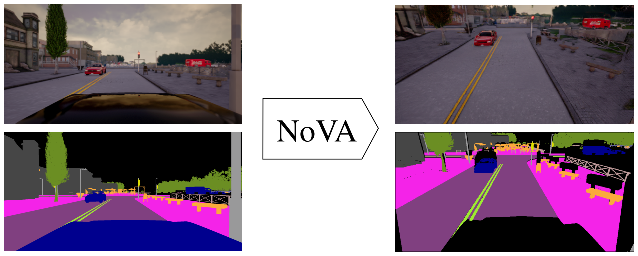


Figure 1: NoVA enables the adaptation from a source domain view to a novel viewpoint in a target domain. It performs a geometry-aware image and label translation from a source (left) to a target (right) view, in which no labels exist.

Specifically, we investigate the adaptation of a semantic segmentation model for the task of autonomous driving. In this field of autonomous driving, large labeled datasets exist [7, 15]. However, most of them are recorded from similar viewpoints, hindering their application to novel viewpoints. Here, NoVA enables the re-use and adaptation of datasets to the novel viewpoints of autonomous buses, trucks or drones.

State-of-the-art domain adaptation approaches perform an image-level adaptation to translate source images to the style of the target domain [19, 36]. Yet, these approaches struggle when faced with performing a semantically consistent translation from the source domain to the novel viewpoint of the target domain, as evidenced in our experiments. A further limitation of current domain adaptation models is that they do not take a translation of the source view segmentation labels into consideration. This is problematic as the translated source images are no longer compatible with the original source labels. In contrast, NoVA adopts ideas from the field of novel view synthesis by proposing an adaptation pipeline that features an explicit representation of the scene geometry, which enables to translate both the source view images and the source view labels to the target domain.

Our NoVA pipeline is split into four stages, which can be trained jointly or independently. First, we estimate the scene geometry by predicting a depth map from a source view image. Next, we utilize the predicted depth map as well as prior knowledge about the transformation between the two viewpoints, which we assume to be given, to forward warp the source image and source label to the target viewpoint. A refinement network then performs inpainting of occluded image areas and stylizes the warped image in the style of the target domain. Finally, we train a target segmentation network with the translated source domain data.

Thereby, NoVA effectively reduces the domain and viewpoint adaptation task to the well-studied problems of depth estimation [10, 17, 45] and image inpainting [28, 40] / stylization [19, 36], for which deep neural networks have already demonstrated remarkable performance. While NoVA builds on recent advances in supervised and self-supervised depth estimation, it utilizes a novel residual refinement network which enables the model to focus on filling in occluded image areas and updating the overall image style without having to synthesize a new image from scratch.

Compared to current image-level domain adaptation approaches, which focus on directly translating the source images to the target domain with a single generative model, NoVA uses a modular architecture that utilizes an explicit representation of the scene geometry. This enables NoVA to perform a geometry-aware translation of both source images and labels to the target domain viewpoint. In addition, it makes it possible to efficiently utilize information about how the source and target domain viewpoints are related. This prior knowledge is commonly available yet not used by current state-of-the-art domain adaptation models that are designed to mainly account for a change in image style.

We demonstrate the benefit of using NoVA over existing domain adaptation approaches for adapting to a novel viewpoint within a simulation environment as well as for adapting from simulation to a complex real-world dataset. In summary, this paper makes the following **contributions**:

- We introduce the task of domain and viewpoint adaptation, a variant of the domain adaptation task for which the domains do not only differ in style but also correspond to different viewpoints. In particular, we consider the unsupervised adaptation task, in which no labels are available in the target domain viewpoint.

- We improve upon current domain adaptation models by using an explicit representation of the scene geometry that enables NoVA to forward warp source view images and labels to the target domain. Thereby, the viewpoint change itself no longer has to be learned and instead the task is reduced to the well-studied problems of depth estimation and image inpainting/stylization.

- We demonstrate improved performance compared to state-of-the-art domain adaptation models on synthetic and challenging real-world datasets.

2. Related Work

This work relates to domain adaptation approaches and novel view synthesis methods. In this section, we briefly review the most related works.

Domain Adaptation. The simplest domain adaptation approach is to *fine-tune* a model, which has been pre-trained on source data, on labeled target samples [16, 26]. However, this approach is not applicable to the unsupervised scenario where no target labels are available and may result in overfitting when only limited target labels are available.

Alternatively, *feature-level* adaptation aims at learning domain invariant features by aligning the feature distributions of the source and target domain via an additional loss term such as a domain confusion loss [24, 25, 39] or via domain-adversarial training [12, 13, 37, 38]. However, enforcing domain invariance via feature-level adaptation may be detrimental with respect to the model's discriminative power [29] and may fail due to not enforcing semantic consistency between the source and the target image.

A third class of domain adaptation approaches thus consider *image-level* adaptation [1, 19, 31]. Here, the distribution alignment between the source and target domain are not performed in feature space but in image space by translating source images to the target domain, commonly with a variant of Generative Adversarial Networks (GANs) [18].

In this context, Hoffman et al. [19] proposed CyCADA, an image-to-image translation framework based on Cycle-Consistent Adversarial Networks (CycleGANs) [47] where consistency between the source and synthesized target image is enforced with a cycle-consistency loss. Recently, Tzeng et al. [36] have demonstrated that a more efficient cycle-free semantic consistency loss that exploits semantic labels from the source domain can also ensure consistency between the semantics of source images and translated source images. While current state-of-the-art image-to-image translation approaches work well for changes in image style [19, 36], they struggle with changes that require an understanding of the scene geometry such as a change of viewpoint, as evidenced in our experimental results.

Recent image-translation works that utilize a depth representation use it to simulate synthetic foggy images [30] or to preserve semantic information during image translation [4]. In contrast, in this work, we exploit depth cues for adapting to a novel viewpoint.

The adaptation to a different camera is considered in some recent works but is limited to an adaptation to a different camera style [43, 44] or to novel camera intrinsics [11]. With regards to adapting to a different viewpoint, Di Mauro et al. [8] consider the adaptation to a single novel camera view. They do not perform image translation but instead propose an encoder-decoder model in which the latent code corresponds to a semantic segmentation map of the input. It is trained with a segmentation loss (source images only),

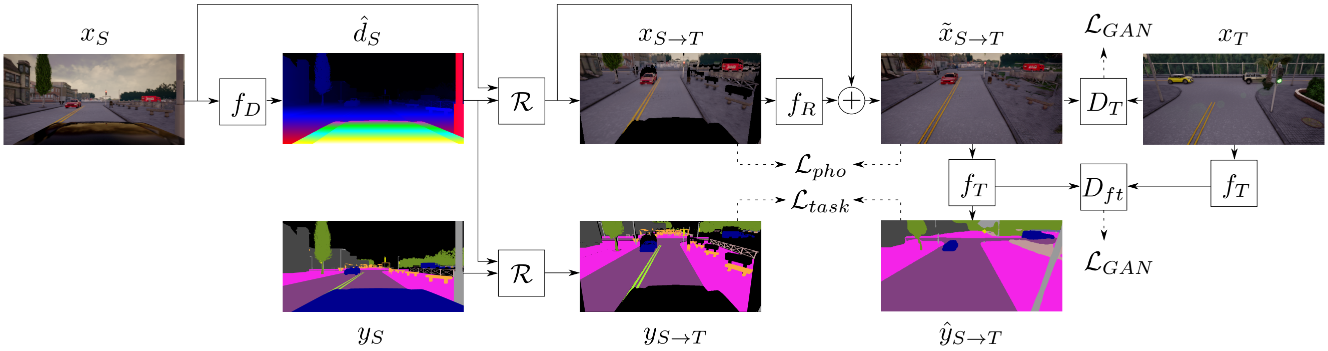


Figure 2: **NoVA Pipeline.** A depth estimation model f_D estimates $\hat{d}_S$ given a source image x_S. Based on $\hat{d}_S$, the image x_S and corresponding label y_S are warped by a differentiable rendering operator $\mathcal{R}$ to the target view. The warped image $x_{S\to T}$ is refined by a residual refinement network f_R to create $\tilde{x}_{S\to T}$. A discriminator D_T ensures the realism of refined images, while a photometric consistency loss $\mathcal{L}_{pho}$ ensures consistency between warped and refined images. A segmentation model f_T that predicts $\hat{y}_{S\to T}$ is trained with a task loss $\mathcal{L}_{task}$ on refined images $\tilde{x}_{S\to T}$ and warped labels $y_{S\to T}$. During training of f_T, a feature-level discriminator D_{ft} ensures alignment between features of refined images $\tilde{x}_{S\to T}$ and target images x_T.

a reconstruction loss and an adversarial loss. Thus, unlike NoVA, their model is not able to provide any task supervision in the target domain. Our experiments confirm that NoVA compares favorably to their SceneAdapt method.

In very recent work, Tran et al. [33] propose a domain adaptation approach which, similarly to NoVA, draws on ideas from novel view synthesis. They utilize a keypoint-based appearance flow for a perspective transformation of source images to a novel viewpoint and perform a photometric refinement using a CycleGAN. However, unlike NoVA, they do not utilize a dense depth estimation but instead use a sparse representation of a small number (i.e. 36) of 2D object keypoints. In contrast to NoVA, which is able to utilize self-supervision for its depth estimators and supports depth estimation for complex multi-object scenes, their keypoint localization network requires ground truth depth for training and only supports the localization of keypoints for a single foreground object.

Novel View Synthesis. Several models have been proposed for generating novel views from a single input image.

While some directly predict a novel view using an encoder-decoder architecture [21, 32, 41] or generative adversarial network [34, 42], others utilize the appearance flow, a dense flow field that specifies how to warp the input to the target view [27, 46]. However, flow-based warping can lead to the distortion of local structures in the output.

Liu et al. [22] demonstrated that an explicit representation of the 3D geometry of a scene improves upon flow-based view synthesis. The benefit of an explicit depth representation has also been confirmed in other recent novel view synthesis works [3, 5]. NoVA builds on this insight but extends the geometry-aware image warping to unsupervised depth estimation models [17, 45] and integrates it into a framework for domain and viewpoint adaptation.

3. Novel View Adaptation

This section introduces our Novel Viewpoint Adaptation (NoVA) framework. NoVA performs a geometry-aware translation of the source domain data to the target domain. It is an unsupervised and unpaired method, requiring no annotations in the target domain nor corresponding image pairs.

The NoVA pipeline is split into four stages (see Fig. 2). In a first step, a depth map is estimated from a given source domain image. Next, the source image and label are forward warped to the viewpoint of the target domain. Afterwards, occluded areas are inpainted and the style of the warped image is adapted to the style of the target domain by a refinement network. Finally, the translated images and labels are utilized to train a target segmentation network.

Problem Setup. We consider the challenging problem of domain and viewpoint adaptation for the task of semantic segmentation. More precisely, we consider unsupervised adaptation, where we are provided with a set of images X_S and labels Y_S in the source domain and with unpaired images X_T and no labels in the target domain. In addition, we assume to know the transformation between the source and target viewpoint $V_{S\to T} = (K_S, K_T, R_{S\to T}, t_{S\to T})$, where K_S, K_T are the camera intrinsics and $R_{S\to T}, t_{S\to T}$ is the transformation between source and target view. Based on the source dataset (X_S, Y_S) we can train a source segmentation model f_S, parameterized by a CNN with weights θ_s, for K-way classification using a cross-entropy loss:

$$\begin{aligned}&\mathcal{L}_{task}(f_S, X_S, Y_S) = \\ &\qquad -\mathbb{E}_{(x_S,y_S)\sim(X_S,Y_S)} \sum_{k=1}^{K} \mathbb{1}_{k=y_S} \log\left(\sigma(f_S^{(k)}(x_S|\theta_s))\right)\end{aligned} \tag{1}$$

where σ is the softmax function.

However, the source model will not perform well on images from the target viewpoint, as evidenced in our experiments (see Section 4). Thus, we aim to train a target network f_T that is optimized for target domain images.

Depth Estimation. In order to utilize the knowledge about the transformation between the source and target viewpoints, which is encapsulated in $V_{S\to T}$, we utilize an explicit depth representation of the scene. In addition, this enables us to not only translate the source images but also the source labels. Given a source view image $x_S \sim X_S$, a depth estimation network f_D, parameterized by a CNN with weights θ_d, estimates a depth map $\hat{d}_S = f_D(x_S|\theta_d)$.

Rendering. Given a predicted source view depth map $\hat{d}_S$, we warp the source view image x_S as well as the corresponding source view label $y_S \sim Y_S$ to the target view. Using a differentiable rendering operator $\mathcal{R}$ we generate a target view image $x_{S\to T} = \mathcal{R}(x_S, \hat{d}_S, V_{S\to T})$ and target view label $y_{S\to T} = \mathcal{R}(y_S, \hat{d}_S, V_{S\to T})$ according to $V_{S\to T}$. While we allow backpropagation through $\mathcal{R}$ for warping the source images x_S , we stop the gradients from backpropgating through $\mathcal{R}$ when warping the source labels y_S.

While self-supervised monocular depth estimation methods utilize inverse warping to provide supervision via view synthesis [14, 45], our problem setup, which considers unpaired images, necessitates the use of forward warping. As in [35], we utilize a forward-splatting approach. We first perform a forward projection of each pixel p_S in the source image to the pixels in the target frame p_T using the inverse predicted depth $\hat{d}_S^{-1}$, the camera intrinsics K_S, K_T and the transformation between the views $R_{S\to T}, t_{S\to T}$:

$$\begin{bmatrix} p_T^x \\ p_T^y \\ 1 \\ \hat{d}_T^{-1} \end{bmatrix} \sim \begin{bmatrix} K_T & \hat{0} \\ \hat{0} & 1 \end{bmatrix} \begin{bmatrix} R_{S\to T} & t_{S\to T} \\ \hat{0} & 1 \end{bmatrix} \begin{bmatrix} K_S^{-1} & \hat{0} \\ \hat{0} & 1 \end{bmatrix} \begin{bmatrix} p_S^x \\ p_S^y \\ 1 \\ \hat{d}_S^{-1} \end{bmatrix} \tag{2}$$

The target image is initialized with an empty canvas onto which the projected source pixels are splatted. Several source pixels may map to the same target pixel, thus we require the use of z-buffering to deal with occlusions. For this, we use a differentiable *soft* z-buffer where the contribution of each source pixel to a target pixel is weighted according to its inverse depth in the target view $\hat{d}_T^{-1}$. Finally, the image is normalized by a weighted average of the contributions of points which splat to a given target pixel. For more details, we refer the reader to [35].

Target View Refinement. As the target view may contain new scene content, which was occluded or outside of the camera frame in the source view image, we refine the warped image $x_{S\to T}$ by inpainting blank image areas and stylizing the image in the target domain style. This task is performed by a refinement network f_R, which is parameterized by a CNN with weights θ_r. This network is not tasked with synthesizing an image from scratch but instead only needs to output a residual r which is added to the warped source image $x_{S\to T}$ before the hyperbolic tangent activation in the network's last layer to create the refined image $\tilde{x}_{S\to T} = \tanh(x_{S\to T} + r)$. By modeling the refinement with a residual connection, we make the inpainting task easier to learn and encourage the network to keep the overall image structure of the warped image in the refined image.

The supervision for training the refinement network is provided by a discriminator network D_T that is trained with an adversarial loss $\mathcal{L}_{GAN}$:

$$\begin{aligned} &\mathcal{L}_{GAN}(\tilde{G}_{S\to T}, D_T, X_T, X_S) \\ &\quad = \mathbb{E}_{x_T \sim X_T}[\log D_T(x_T)] \\ &\quad + \mathbb{E}_{x_S \sim X_S}[\log(1 - D_T(\tilde{G}_{S\to T}(x_S)))] \end{aligned} \tag{3}$$

where the depth estimation, forward rendering and refinement steps are encapsulated into a single virtual generator step $\tilde{G}_{S\to T} = f_R(\mathcal{R}(x_S, f_D(x_S|\theta_d), V_{S\to T})|\theta_r)$.

As we do not have any target labels available, we cannot apply the same unsupervised refinement approach to the warped source labels $\hat{y}_{S\to T}$. Instead, we use the warped source labels without refinement to provide a sparse supervision signal for training the target network f_T, in which the task loss is only applied in regions with label information.

Enforcing Photometric Refinement Consistency. An important aspect of the refinement step is that the original scene structure and content of the warped source image should be preserved. In order to enforce consistency between the warped and the refined image, we propose a lightweight photometric refinement loss that penalizes differences between the warped and the refined source pixels:

$$\begin{aligned} &\mathcal{L}_{pho}(G_{S\to T}, \tilde{G}_{S\to T}, X_S) = \\ &\quad w_{pho}\, \mathbb{E}_{x_S \sim X_S}[||G_{S\to T}(x_S) - \tilde{G}_{S\to T}(x_S)||_1] \end{aligned} \tag{4}$$

where $G_{S\to T} = \mathcal{R}(x_S, f_D(x_S|\theta_d), V_{S\to T})$ and w_{pho} is a binary pixel-wise mask. The weight w_{pho} is 0 for empty pixels in the warped image $x_{S\to T}$, onto which no source pixel was mapped, and 1 for all non-empty pixels.

While the refinement network has the freedom to change any warped source pixel if desired, $\mathcal{L}_{pho}$ encourages the refinement network to effectively act as an inpainting model. However, even when the source and target domains do not only differ in viewpoint but also in their overall image style, we find the photometric refinement loss to be an effective alternative to a semantic [36] or cycle consistency loss [19].

Target Network Training. The target task network f_T, which is parameterized by a CNN with weights θ_t, is trained with the translated source domain data. Let us denote the translated source view dataset as $(\tilde{X}_{S\to T}, Y_{S\to T})$ where $\tilde{X}_{S\to T} = \{f_R(\mathcal{R}(x_S, f_D(x_S|\theta_d), V_{S\to T})|\theta_r)|x_S \in X_S\}$, $Y_{S\to T} = \{\mathcal{R}(y_S, f_D(x_S|\theta_d), V_{S\to T})|x_S \in X_S, y_S \in Y_S\}$.

For training f_T we again utilize the cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{task}(f_T, \tilde{X}_{S\to T}, Y_{S\to T}) = \\ -\mathbb{E}_{(\tilde{x}_{S\to T}, y_{S\to T})\sim(\tilde{X}_{S\to T}, Y_{S\to T})} \\ \sum_{k=1}^{K} \mathbb{1}_{k=y_{S\to T}} \log\left(\sigma(f_T^{(k)}(\tilde{x}_{S\to T}|\theta_t))\right)\end{aligned} \tag{5}$$

In addition, we perform feature-level alignment of f_T between the the target images X_T and the refined images $\tilde{X}_{S\to T}$. For this, we add a discriminator D_{ft} to distinguish between features of target images and refined images:

$$\begin{aligned}\mathcal{L}_{GAN}(f_T, D_{ft}, f_T(\tilde{X}_{S\to T}|\theta_t), X_T) = \\ \mathbb{E}_{\tilde{x}_{S\to T}\sim\tilde{X}_{S\to T}}[\log D_{ft}(f_T(\tilde{x}_{S\to T}|\theta_t))] + \\ \mathbb{E}_{x_T\sim X_T}[\log(1 - D_{ft}(f_T(x_T|\theta_t)))]\end{aligned} \tag{6}$$

Overall Learning Objective. Our complete learning objective encapsulates the above losses, which optimize for target view segmentation ($\mathcal{L}_{task}$, see Eq. (5)), image refinement ($\mathcal{L}_{GAN}$, see Eq. (3)), photometric consistency ($\mathcal{L}_{pho}$, see Eq. (4)) and feature alignment ($\mathcal{L}_{GAN}$, see Eq. (6)):

$$\begin{aligned}\mathcal{L}_{NoVA}(f_T, \tilde{G}_{S\to T}, D_T, D_{ft}, X_S, X_T, Y_S) \\ = \mathcal{L}_{task}(f_T, \tilde{X}_{S\to T}, Y_{S\to T}) \\ + \mathcal{L}_{GAN}(\tilde{G}_{S\to T}, D_T, X_T, X_S) \\ + \mathcal{L}_{pho}(G_{S\to T}, \tilde{G}_{S\to T}, X_S) \\ + \mathcal{L}_{GAN}(f_T, D_{ft}, f_T(\tilde{X}_{S\to T}|\theta_t), X_T)\end{aligned} \tag{7}$$

4. Experiments

In order to demonstrate the effectiveness of NoVA, we perform experiments to adapt to a novel viewpoint within a simulation environment (see Section 4.2) as well as from simulation to a complex real environment (see Section 4.3). Section 4.1 gives an overview of our experimental setup.

4.1. Experimental Setup

Datasets We utilize synthetic data generated in CARLA [9] as well as the real-world dataset CityScapes [7].

In the CARLA simulation framework, we generate data from a car and a truck viewpoint. For both views we generate 30 train, 15 test and 5 validation sequences of $1,000$ frames each, where every frame consists of a stereo RGB image pair, a semantic segmentation label and a depth map of resolution 2048 × 1024. In CityScapes, we use 2975 train and 500 test frames with fine annotations. Please see the supplementary material for more details, including the camera intrinsic and extrinsic parameters of both datasets.

Baselines. The naïve baseline for all of our experiments is to train a segmentation model f_S on source data only.

In addition, we use two state-of-the-art image-level domain adaptation models, CyCADA [19] and SPLAT [36], as well as the SceneAdapt model by Maura et al. [8]. It should be noted that CyCADA and SPLAT have been proposed for general domain adaptation and not viewpoint adaptation.

For CyCADA, we follow the generator and discriminator architectures of [47]. The input to both networks is resized to 512×256. We train CyCADA with the Adam optimizer, single image batches and a learning rate of 0.0002 for 20 epochs after which the learning rate is linearly decayed to zero over the course of the next 20 epochs. We adopt the same training scheme for SPLAT but replace the cycle-consistency with a semantic-consistency loss that uses a segmentation network pre-trained on the source dataset.

The SceneAdapt model is constructed from an encoder based on our base segmentation network (see below) and a decoder that uses the architecture of the generator models of CyCADA and SPLAT. The discriminator architecture and training scheme are consistent with CyCADA and SPLAT.

NoVA. While our differentiable rendering formulation enables a joint end-to-end training of the complete NoVA pipeline, we found it beneficial to train NoVA in stages.

For depth estimation, we evaluate three classes of estimators: A self-supervised, monocular approach [17], a supervised monocular approach [20] and a supervised stereo approach [2]. Each approach is trained on source images of resolution 512×256 as outlined in the respective paper.

For rendering, we bilinearly upsample the predicted depth maps to match the source resolution of 2048×1024. In order to avoid empty pixels in the warped output, the rendering operator $\mathcal{R}$ outputs images and labels which are downscaled by a factor of 4 to a resolution of 512×256.

The forward warped images are refined by a residual refinement network f_R that is based on the CyCADA generator architecture which is modified to include a residual connection. The overall training scheme remains unchanged.

Segmentation Model. For the segmentation model we use a VGG16-FCN8s [23]. We train it for $100,000$ steps with batches of size 4 using a learning rate of $1e-3$ with stochastic gradient descent and momentum of 0.9. Feature-level adaptation is performed as described in [19].

Evaluation Metrics. We consider the metrics of mean intersection-over-union (mIoU), frequency-weighted intersection-over-union (fwIoU) and pixel accuracy:

$$\text{mIoU} = \frac{1}{N} \cdot \frac{\sum_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \tag{8}$$

$$\text{fwIoU} = \frac{1}{\sum_k t_k} \cdot \frac{\sum_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \tag{9}$$

$$\text{pixAcc} = \frac{\sum_i n_{ii}}{\sum_i t_i} \tag{10}$$

where N is the number of segmentation classes, n_{ij} is the number of pixels of class i predicted as class j and $t_i = \sum_j n_{ij}$ is the total number of pixels of class i.

4.2. Sim2Sim

For a first set of experiments, we aim to evaluate the task of viewpoint adaptation only. To this effect, we select two domains which only differ in viewpoint but not in style.

Table 1 presents the results for adapting from a car to a truck viewpoint in the CARLA simulation [9]. It shows that NoVA outperforms the scene and domain adaptation baselines by a large margin on the task of viewpoint adaptation.

In fact, given ground truth depth, NoVA is able to get very close to reaching the performance of a target oracle model, which is trained on the target images and labels. For predicted depth estimation, we find that all variants of NoVA still perform significantly better than the baselines. Here, the supervised stereo model (*stereo-sup*) outperforms the monocular self-supervised (*mono-self*) and the monocular supervised (*mono-sup*) approach.

For SceneAdapt, we find that it improves over the source segmentation model for the viewpoint adaptation task. As for CyCADA and SPLAT, we find that both struggle with the task. Because they are designed for domain adaptation and not for viewpoint adaptation, they do not take a translation of the source labels to the target viewpoint into consideration. This leads to a mismatch between the translated source images and the original source labels. Indeed, we see that the performances of CyCADA and SPLAT improve when we combine their translated images with the actual target labels. However, even when using target view labels, their performances do not reach the level of NoVA. This indicates that NoVA's image translation pipeline is overall better adapted to the task of viewpoint adaptation.

When inspecting the translated images of CyCADA and SPLAT (see Fig. 5), we find that a shortcoming for both models is that the semantics of translated images are not always consistent with the semantics of the source images. Interestingly, we find that CyCADA is nonetheless able to reconstruct the source from the translated image well, which suggests that it has learned to encode some of the source semantics in the noise of the translated image (see Fig. 9) [6].

On the other hand, qualitative results for NoVA (see Fig. 3 and Fig. 4) demonstrate that NoVA is effective in retaining the source image semantics in the refined images. As shown by the ablation study in Table 2, training with forward warped image-label pairs $(x_{S\rightarrow T}, y_{S\rightarrow T})$ already results in a large performance gain in comparison to the source segmentation model. NoVA's residual refinement further improve NoVA's performance over a default refinement model that synthesizes its output image from scratch.

4.3. Sim2Real

In a second set of experiments we investigate the adaptation to a novel viewpoint and domain. For this, we aim to adapt from a truck viewpoint in CARLA [9] to a car viewpoint in the complex real-world dataset of CityScapes [7]. In this setup, the domain gap is now not only caused by a change in viewpoint but also by a change in image style.

For this experiment, we restrict our evaluation to the subset of CityScapes classes which are present in CARLA.

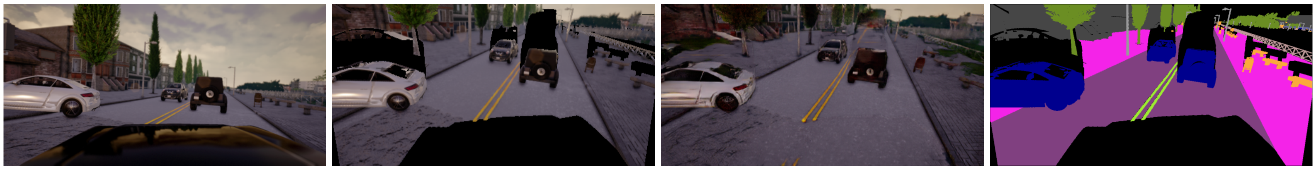
Table 3 demonstrates that NoVA also improves upon the baseline methods in closing the domain gap from a synthetic source domain to the novel viewpoint of a challenging real-world target domain. Unlike for the viewpoint adaptation experiments, SceneAdapt now yields no performance improvements over training a segmentation model with source data only. This suggests that the SceneAdapt model is better suited for the viewpoint adaptation task than for the joint domain and viewpoint adaptation task.

Method	mIoU	fwIoU	pixAcc
Source Only	26.54	43.56	55.82
SceneAdapt [8]	26.63	54.65	68.15
CyCADA [19]	10.57	21.44	30.36
CyCADA [19] + *trgt-labels*	16.31	45.89	62.55
SPLAT [36]	13.63	22.77	32.26
SPLAT [36] + *trgt-labels*	18.81	45.12	59.29
NoVA$_{mono-self}$	42.54	69.99	79.99
NoVA$_{mono-sup}$	45.27	71.20	80.49
NoVA$_{stereo-sup}$	49.67	76.44	84.97
NoVA$_{GT}$	**51.89**	**78.66**	**86.69**
Target Oracle	52.72	79.96	87.81

Table 1: **Results for Viewpoint Adaptation on Sim2Sim.** When tasked with adapting a semantic segmentation model from a car to a truck viewpoint, in which no labels are available, NoVA outperforms current state-of-the-art domain and scene adaptation baselines and closes the gap to a target oracle model, which is trained on labeled target data.

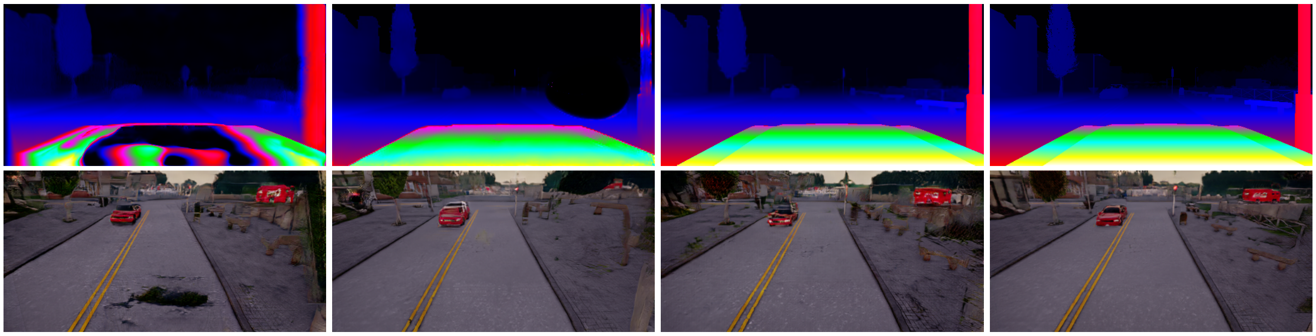
Method	mIoU	fwIoU	pixAcc
Source Only	26.54	43.56	55.82
+ Forward Warping	47.81	75.74	84.11
+ Default Refinement	49.24	76.91	85.08
+ Residual Refinement	**51.89**	**78.66**	**86.69**

Table 2: **Ablation Study for NoVA$_{GT}$ on Sim2Sim.** When the source and the target domain are separated by a change in viewpoint only, NoVA's forward warping of source images and labels to the target domain yields the largest performance improvement over training with source data only. A residual refinement, which inpaints occluded areas in the forward warped images, improves over a default refinement, that needs to synthesize a complete new image from scratch.



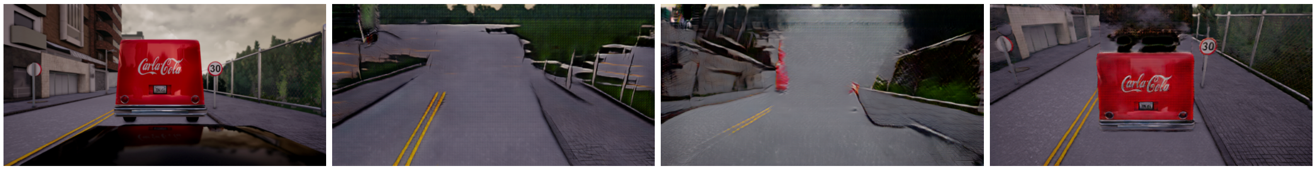
Source Image | Forward Warped Image | Refined Warped Image | Forward Warped Label

Figure 3: **NoVA$_{GT}$ Performance on Sim2Sim.** Given a source view frame, NoVA forward warps the source image and label to the target viewpoint and refines the warped image by inpainting occluded image areas with a residual refinement network.



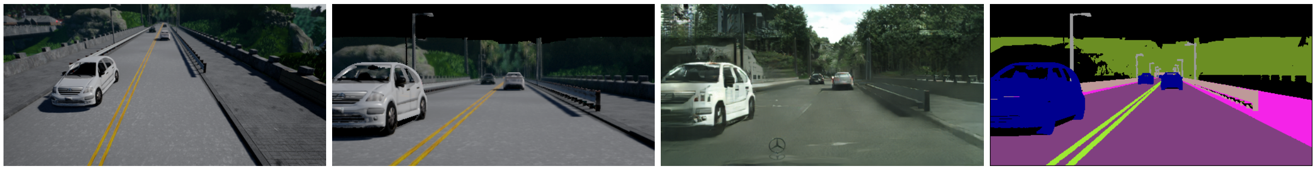
NoVA$_{mono-self}$ | NoVA$_{mono-sup}$ | NoVA$_{stereo-sup}$ | NoVA$_{GT}$

Figure 4: **NoVA Performance for Different Depth Estimators.** NoVA can utilize self-supervised and supervised monocular or stereo depth estimation models as well as ground truth depth maps (top row: predicted depth, bottom row: refined images).



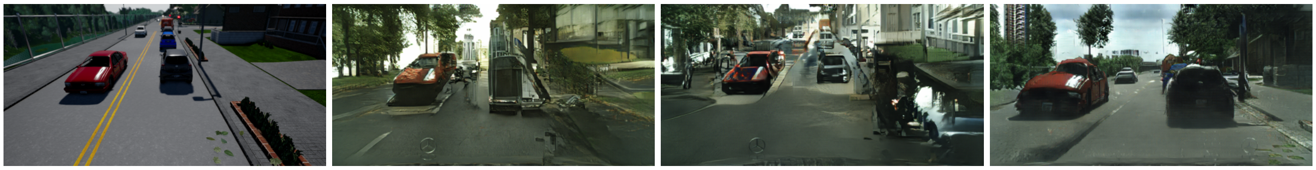
Source Image | CyCADA [19] | SPLAT [36] | NoVA$_{GT}$

Figure 5: **Qualitative Comparison of NoVA$_{GT}$ to the Baselines on Sim2Sim.** In contrast to NoVA, the CyCADA and SPLAT baseline models do not ensure a semantic consistency between the source image and the translated source image.



Source Image | Forward Warped Image | Refined Warped Image | Forward Warped Label

Figure 6: **NoVA$_{GT}$ Performance on Sim2Real.** NoVA is able to adapt to the novel viewpoint and style of the CityScapes dataset by forward warping to the target view and refining the warped images in the style of the target domain.



Source Image | CyCADA [19] | SPLAT [36] | NoVA$_{GT}$

Figure 7: **Qualitative Comparison of NoVA$_{GT}$ to the Baselines on Sim2Real.** In contrast to NoVA, the CyCADA and SPLAT baseline models fail to correctly adapt semantic objects (e.g. cars) from the source to the target domain viewpoint.

Method	mIoU	fwIoU	pixAcc
Source Only	18.84	37.34	47.59
SceneAdapt [8]	11.54	30.65	37.23
CyCADA [19]	19.26	43.90	56.43
SPLAT [36]	21.01	49.42	60.99
NoVA$_{mono-self}$	30.23	60.32	72.09
NoVA$_{mono-sup}$	34.36	66.83	78.25
NoVA$_{stereo-sup}$	32.96	63.95	75.09
NoVA$_{GT}$	**35.91**	**69.52**	**80.84**
Target Oracle	51.30	79.82	88.36

Table 3: **Results for Domain and Viewpoint Adaptation on Sim2Real.** When the source and target domain differ in both viewpoint and style, NoVA again significantly outperforms the state-of-the-art adaptation baseline models.

CyCADA and SPLAT demonstrate performance improvements with respect to the source model on the Sim2Real task. However, they still perform worse than all NoVA variants and a qualitative comparison to NoVA (see Fig. 7) reveals that they struggle to correctly warp the appearance of semantic objects (e.g. cars) to the viewpoint of the target domain. In the case of SPLAT, this can be explained by its semantic consistency loss, which encourages semantic objects to reappear in the translated image at the same spatial location as in the original source image.

Despite not using a cycle or semantic consistency loss, qualitative results in Fig. 6 and Fig. 7 confirm that NoVA preserves the scene semantics well even when adapting to a different domain. This suggests that NoVA's explicit forward warping in combination with its residual refinement and photometric refinement loss offer a lightweight but effective alternative for ensuring the semantic consistency between source images and translated source images.

In a second ablation study (see Table 4), we find NoVA's forward warping component again to be highly beneficial. However, compared to Sim2Sim, the residual refinement now improves upon the forward warping significantly, as the domains are now also separated by a change in style.

As opposed to the results on the Sim2Sim task, NoVA is now unable to fully close the gap to the target oracle model that is trained on the labeled target data. We suspect this may be due to CARLA lacking CityScapes' overall diversity. Here, semi-supervised adaptation with a limited number of labeled CityScapes examples can further improve NoVA's performance and help to close the gap to the target oracle model. As Fig. 8 visualizes, combining $N_T = 300$ labeled CityScapes frames with NoVA's $N_S = 30,000$ translated source frames can already boost NoVA's mIoU-performance by about 5% wrt. an unsupervised adaptation.

Method	mIoU	fwIoU	pixAcc
Source Only	18.84	37.34	47.59
+ Forward Warping	26.95	55.23	69.72
+ Default Refinement	30.41	58.30	68.97
+ Residual Refinement	**35.91**	**69.52**	**80.84**

Table 4: **Ablation Study for NoVA$_{GT}$ on Sim2Real.** While we again find forward warping to be highly beneficial, residual refinement now yields a large improvement over training with the forward warped data as the refinement adapts the warped images to the style of the target domain.

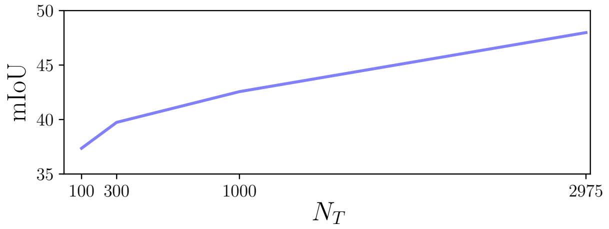
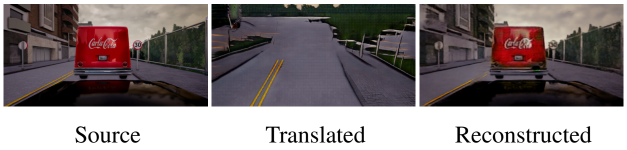


Figure 8: **Semi-Supervised Adaptation on Sim2Real.** Performance improves when we combine NoVA$_{GT}$'s translated source domain dataset of size $N_S = 30,000$ with a set of labeled target examples from CityScapes of size N_T.


Source

Translated

Reconstructed

Figure 9: **CyCADA on Sim2Sim.** CyCADA learns to encode some of the source image semantics in the noise of the translated image for the reconstruction of the source image.

5. Conclusion and Future Work

In this paper, we introduced NoVA, a new model for adapting to novel viewpoints and domains. NoVA performs a geometry-aware translation of source domain images and labels to a target domain, in which no labeled examples are available. Our experiments on the task of semantic segmentation demonstrate that NoVA significantly improves over state-of-the-art domain adaptation models for adapting to novel views in simulation and complex real world datasets.

In the future, the NoVA framework could be extended to additionally utilize temporal information to enable the model to better reason about occluded image areas.

References

[1] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[2] J.-R. Chang and Y.-S. Chen. Pyramid stereo matching network. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] X. Chen, J. Song, and O. Hilliges. NVS machines: Learning novel view synthesis with fine-grained view control. *arXiv.org*, 1901.01880, 2019.

[4] Y. Chen, W. Li, X. Chen, and L. V. Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] I. Choi, O. Gallo, A. J. Troccoli, M. H. Kim, and J. Kautz. Extreme view synthesis. *arXiv.org*, 1812.04777, 2018.

[6] C. Chu, A. Zhmoginov, and M. Sandler. Cyclegan, a master of steganography. In *Advances in Neural Information Processing Systems (NIPS) Workshops*, 2017.

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] D. Di Mauro, A. Furnari, G. Patanè, S. Battiato, and G. M. Farinella. Scene adaptation for semantic segmentation using adversarial learning. In *Proc. of International Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, 2018.

[9] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In *Proc. Conf. on Robot Learning (CoRL)*, 2017.

[10] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[11] J. M. Facil, B. Ummenhofer, H. Zhou, L. Montesano, T. Brox, and J. Civera. CAM-Convs: Camera-Aware Multi-Scale Convolutions for Single-View Depth. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] Y. Ganin and V. S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proc. of the International Conf. on Machine learning (ICML)*, 2015.

[13] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 17:2096–2030, 2016.

[14] R. Garg, B. G. V. Kumar, G. Carneiro, and I. D. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2016.

[15] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research (IJRR)*, 32(11):1231–1237, 2013.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[17] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[19] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *Proc. of the International Conf. on Machine learning (ICML)*, 2018.

[20] J. Hu, M. Ozay, Y. Zhang, and T. Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[21] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[22] M. Liu, X. He, and M. Salzmann. Geometry-aware deep network for single-image novel view synthesis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[24] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *Proc. of the International Conf. on Machine learning (ICML)*, 2015.

[25] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[26] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[27] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[28] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[29] A. Rozantsev, M. Salzmann, and P. Fua. Beyond sharing weights for deep domain adaptation. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2018.

[30] C. Sakaridis, D. Dai, S. Hecker, and L. Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2018.

[31] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised

images through adversarial training. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[32] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Multi-view 3d models from single images with a convolutional network. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2016.

[33] L. Tran, K. Sohn, X. Yu, X. Liu, and M. Chandraker. Gotta adapt 'em all: Joint pixel and feature-level domain adaptation for recognition in the wild. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[34] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[35] S. Tulsiani, R. Tucker, and N. Snavely. Layer-structured 3d scene inference via view synthesis. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2018.

[36] E. Tzeng, K. Burns, K. Saenko, and T. Darrell. SPLAT: semantic pixel-level adaptation transforms for detection. *arXiv.org*, 1812.00929, 2018.

[37] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2015.

[38] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[39] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv.org*, 1412.3474, 2014.

[40] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[41] J. Yang, S. Reed, M.-H. Yang, and H. Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[42] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2017.

[43] Z. Zhong, L. Zheng, S. Li, and Y. Yang. Generalizing a person retrieval model hetero- and homogeneously. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2018.

[44] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[45] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[46] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2016.

[47] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2017.