# ADeLA: Automatic Dense Labeling with Attention for Viewpoint Shift in Semantic Segmentation

Hanxiang Ren[1]* Yanchao Yang[2]†* He Wang[3]‡ Bokui Shen[2]
Qingnan Fan[4]‡ Youyi Zheng[1]† C. Karen Liu[2] Leonidas Guibas[2]
[1]Zhejiang University [2]Stanford University [3]Peking University [4]Tencent AI Lab
{hanxiang.ren,youyizheng}@zju.edu.cn {yanchaoy,karenliu,guibas}@cs.stanford.edu
hewang@pku.edu.cn bshen88@stanford.edu fqnchina@gmail.com

## Abstract

*We describe a method to deal with performance drop in semantic segmentation caused by viewpoint changes within multi-camera systems, where temporally paired images are readily available, but the annotations may only be abundant for a few typical views. Existing methods alleviate performance drop via domain alignment in a shared space and assume that the mapping from the aligned space to the output is transferable. However, the novel content induced by viewpoint changes may nullify such a space for effective alignments, thus resulting in negative adaptation. Our method works without aligning any statistics of the images between the two domains. Instead, it utilizes a novel attention-based view transformation network trained only on color images to hallucinate the semantic images for the target. Despite the lack of supervision, the view transformation network can still generalize to semantic images thanks to the induced "information transport" bias. Furthermore, to resolve ambiguities in converting the semantic images to semantic labels, we treat the view transformation network as a functional representation of an unknown mapping implied by the color images and propose functional label hallucination to generate pseudo-labels with uncertainties in the target domains. Our method surpasses baselines built on state-of-the-art correspondence estimation and view synthesis methods. Moreover, it outperforms the state-of-the-art unsupervised domain adaptation methods that utilize self-training and adversarial domain alignments. Our code and dataset will be made publicly available.*

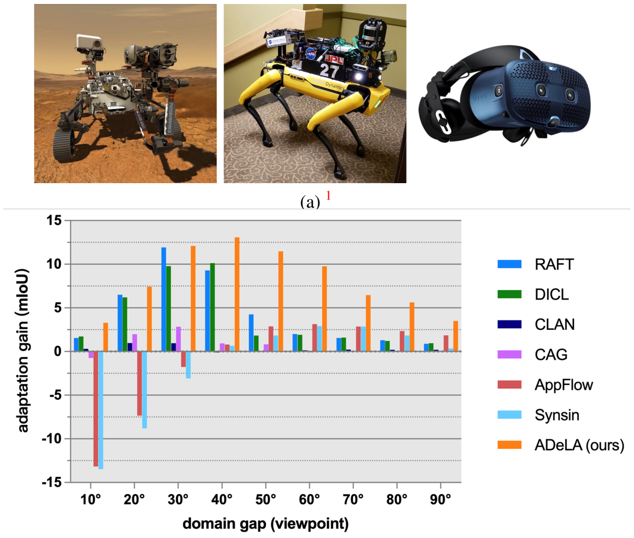

Figure 1. (a): Multiple cameras towards different viewpoints can help autonomous or assistive agents to better understand the scene. However, the performance of the semantic segmentation network trained on the forward view (typical view of existing datasets) drops sharply when tested with viewpoint shifts (Tab. 5). (b): Adaptation gain obtained by state-of-the-art methods across different viewpoints. Our method consistently shows positive gains and works robustly towards substantial viewpoint change,

## 1. Introduction

Parsing the environment from multiple viewing angles to arrive at a comprehensive understanding of the surroundings is critical for autonomous agents, assistive robots, and AR/VR equipment (Fig. 1a). These multi-camera systems can capture temporally paired data in practice from different viewpoints, and the need to train a scene parsing network that performs well at multiple viewpoints is key to estimating traversable surfaces and preventing accidents. However, viewpoint changes across cameras induce significant domain gaps – a scene parsing network trained with annotations in one view often encounters a large performance drop on another (Tab. 5). We aim to reduce this performance drop by transporting semantic information from the

*equal contribution. †corresponding author
‡work done while at Stanford

[1] image credits: NASA, Boston Dynamics and HTC

views with rich annotations (source) to views with no available annotation (target) utilizing temporally paired images readily available from multi-camera systems.

Most methods dealing with domain gaps build on the idea that an alignment in a shared latent space helps the task-specific network trained in the source domain generalize to the target. Despite its effectiveness, domain alignment generally assumes (sufficient) invariance exists for the task, which can be computed through the alignment so that the mapping from the aligned space to the output is transferable across domains (Fig. 2a & 2b). However, the domain discrepancy we consider here is mainly the content shift caused by the viewpoint change (Fig. 2c). As dense scene parsing (semantic segmentation) is viewpoint elevation-dependent, any alignment that learns away viewpoint will result in (insufficient) invariances which are not adequate or suitable for the task, thus inducing negative adaptation (Fig. 1b).

We break this conundrum by hallucinating the target semantic images using their source counterparts. Our method employs a view transformation network that outputs the target semantic image, conditioned on a source semantic image and a pair of temporally aligned regular color images. The hallucinated semantic images are then converted to semantic labels to adapt the task network.

However, without a proper inductive bias, the view transformation network would completely fail on semantic images due to their different structures. We propose that the right inductive bias is to encourage learning *spatial transportation* instead of transformation in color space. Further, we introduce a novel architecture for view transformation where the desired inductive bias is injected via an attention mechanism. To combat noise in the hallucination and better decode the semantic labels, we treat the view transformation network as a functional representation of an unknown mapping signified by the color images. Accordingly, we propose a functional label hallucination strategy that generates the soft target labels by taking in the indicator functions of each class. The proposed decoding strategy improves the label accuracy by a large margin and makes the labels more suitable for adaptation by incorporating uncertainties.

Due to the lack of datasets in semantic segmentation whose domain gaps are mainly from viewpoint change, we also propose a new dataset where the viewpoint is varied to simulate different levels of content shift (Fig. 7). To our best knowledge, the problem we study here is largely under-explored. To validate, we perform a comprehensive study of various state-of-the-art methods, including dense correspondence estimation, novel view synthesis, and unsupervised domain adaptation (UDA) methods. We demonstrate the effectiveness of our method by showing the best adaptation gains across different target domains, even for perpendicular viewing angles. Our contributions are:

- A benchmarking of state-of-the-art UDA methods for semantic segmentation on viewpoint shifts.
- A novel architecture for semantic information hallucination trained with only RGB images and an uncertainty-aware functional decoding scheme.
- A state-of-the-art method that deals with performance drops in semantic segmentation caused by viewpoint shifts for multi-camera systems.

## 2. Related work

We focus on unsupervised domain adaptation (UDA) methods for the pixel-level prediction task of semantic segmentation. The core ingredient of UDA is to reduce the domain gap between the source and the target data [2, 9, 14, 18, 34, 55], where the domain gap can be measured by the maximum mean discrepancy [17, 28] or central moment discrepancy [61]. Deep learning based methods resort to adversarial measurements, where discriminator networks are used to confuse the two domains [24, 31, 40, 43, 44, 51] in a shared feature space. In contrast to classification, feature space alignment is much less effective for pixel-level prediction tasks like semantic segmentation [29, 41], due to the difficulties in keeping the aligned features informative about the spatial structure of the output.

The recent success of unsupervised domain adaptation for semantic segmentation mainly relies on image-to-image translation [27, 60, 71] where the goal is to reduce the style difference between domains while preserving the underlying semantics [20, 26, 66]. Multi-level feature alignment is proposed [58] and [19] introduces intermediate styles that gradually close the gap. A disentanglement of texture and structure is also beneficial [4], and [23] performs style randomization to learn style invariance. To ease the difficulty in adversarial training, [59] proposes a style transfer via Fourier Transform while enforcing semantic consistency. On the other hand, [12, 13, 30, 56, 65] propose class-wise alignments, given that each of the semantic classes may possess a different domain gap. Similarly, [49] proposes patch-wise alignment, and [21] utilizes local contextual-relations for a consistent adaptation. [22] also performs alignment on consistently matched pixels among source and target images. The alignment can also be done in the output space [53], or in a curriculum manner. For example, [33] employs inter and intra domain adaptation with an easy-to-hard split, and [25] pre-selects source images with similar content to the target. With aligned domains, self-training using pseudo labels can be utilized to further close the gap [26, 59, 64].

Our method tackles the domain gap caused by different camera views, which renders the image space alignment ineffective as the domain gap is mainly content shift but not the style difference. Unlike cross-view image classification [1, 10, 16, 37, 63], aligning domains of different viewpoints for pixel-level prediction tasks is ill-posed, since the task is indeed view dependent [7]. The most relevant are [8, 11],

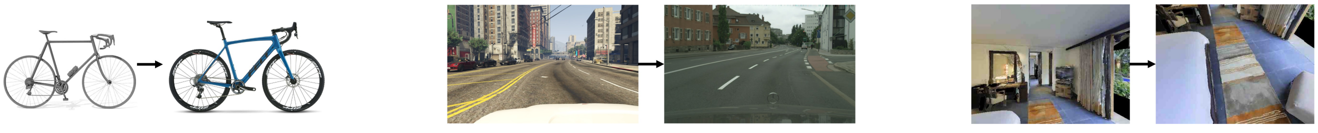
(a) (b) (c)

Figure 2. (a): image classification where image style and viewpoint are nuisances [36, 39]; (b): semantic segmentation at similar views where image style is the major nuisance for domain gaps [20, 35]; (c): semantic segmentation with changing view (*e.g.*, forward to downward), a nuisance that should not be aligned away. We focus on viewpoint shifts in semantic segmentation.

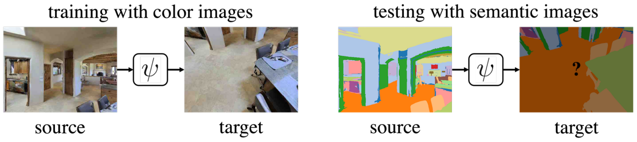

Figure 3. Left: a network $\psi$ is trained to hallucinate color images from the source to the target and is never exposed to semantic images; Right: $\psi$ is directly applied on the corresponding source semantic image to hallucinate the target semantic image to provide training labels for the target domain.

which again resort to adversarial domain alignment. Additionally, [8] requires known camera intrinsics and extrinsics. Note, both assume the viewpoint change is small or there is a large overlap between the two views, therefore the applicability to a broader setting is limited, whereas our method is not constrained by any of these assumptions. Also related is novel view synthesis [6, 15, 46, 69], particularly, single view synthesis [50, 57, 70], where multiple posed images of the same scene are needed during training. Hence, if the goal is to synthesize semantic images of a different view, the target domain's semantic images are needed, which, however, are not available in our problem setting. Another related topic is dense correspondence estimation [48,54,67], which can be used to warp labels to help adaptation between domains.

## 3. Method

Let $\mathcal{D}^s = \{(x^s_i, y^s_i)\}^n_{i=1}$ be the source dataset collected at the source viewpoint $s$, where $x^s_i \in \mathbb{R}^{h\times w\times 3}$ is an RGB image, and $y^s_i \in \mathbb{R}^{h\times w\times 3}$ is the corresponding semantic image (Fig. 3) that is usually used for visualizing the semantic labels $\bar{y}^s_i \in \mathbb{R}^{h\times w\times k}$ (we use the semantic image $y^s_i$ or labels $\bar{y}^s_i$ interchangeably depending on the context). Further, let $\mathcal{D}^\tau = \{x^\tau_i\}^n_{i=1}$ be the target dataset collected at the target viewpoint $\tau$, whose semantic label/image $y^\tau_i$ is missing. In order to make our method generally applicable, we assume no knowledge about the viewpoints $s, \tau$ except that $x^s_i$ and $x^\tau_i$ are captured simultaneously. Note, this is the *only* assumption we make since synchronization in multi-camera systems is *default*. Therefore, the domain gap between $\mathcal{D}^s$ and $\mathcal{D}^\tau$ comes from viewpoint shifts. However, the synchronized source and target view images may or may not share co-visible regions, which is unknown and determined by the actual difference between the two views. Please see Fig. 7 for examples of the source and target domains with various viewpoint shifts.

Similar to unsupervised domain adaptation, our ultimate goal is to train a semantic segmentation network $\phi : \mathbf{x} \to \mathbf{y}$ given only the annotations from the source dataset $\mathcal{D}^s$ so that $\phi$ can perform well on the target dataset $\mathcal{D}^\tau$ with the presence of viewpoint shifts. So the domain gap we are considering here is mainly the content shift induced by different viewing angles, *i.e.*, the discrepancy in the output, which violates the assumptions made by most unsupervised domain adaptation methods that rely on either image space or feature space alignment, or both [23,25,30,53,59,64,65]. Instead of aligning distributions of any kind between the two domains, which may result in negative adaptation gains (Fig. 1b), we resort to a network that can hallucinate the target view semantic images ($y^\tau$) from the source ($y^s$) guided by the color images ($x^s, x^\tau$). Specifically, we want to have a network $\psi : \mathbf{y} \times \mathbf{x} \times \mathbf{x} \to \mathbf{y}$, whose output $\psi(y^s_i; x^s_i, x^\tau_i)$ can be used as pseudo ground-truth for improving $\phi$ to make better predictions on $\mathcal{D}^\tau$.

The whole pipeline can be summarized as 1) train the view transformation network $\psi$ using temporally aligned source and target view color images to learn *information transport* between the two domains; 2) once $\psi$ is trained, we use it to hallucinate target view semantic images/labels; 3) the hallucinated target labels are then used to train the semantic segmentation network $\phi$ to adapt to the target views. During testing, i.e., semantic inference, $\psi$ is not in operation since we can apply the adapted semantic segmentation network $\phi$ directly on the test images from the target domain to make predictions. Thus the source images are not required. Please refer to Fig. 4 for an overview.

### 3.1. Auto-labeling with attention

Looking at a pair of color images $x^s_i, x^\tau_i$ shown in Fig. 3, one could hallucinate to some extent the target semantic image $y^\tau_i$ associated with $x^\tau_i$ given the source semantic image $y^s_i$. On the other hand, if a network learns how to hallucinate the target image $x^\tau_i$ from the source image $x^s_i$, we would expect it to make a reasonable hallucination of the target semantic image $y^\tau_i$ from the source semantic image $y^s_i$, since $x^s_i$ and $y^s_i$ are simply two different appearances of the same geometry. However, without a proper inductive bias, a network trained to hallucinate color images between

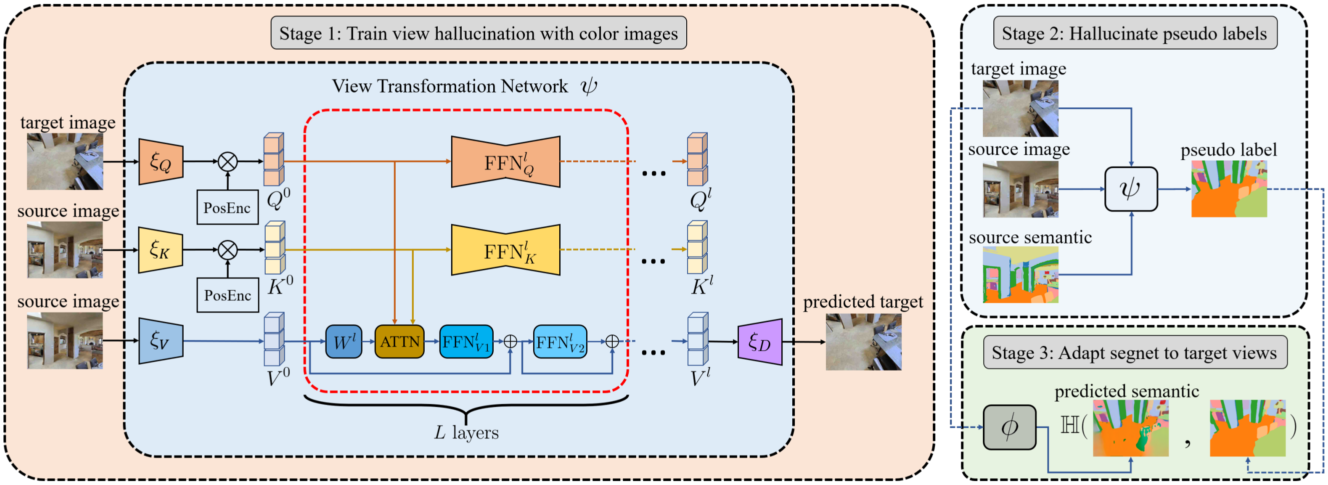

Figure 4. The proposed architecture for hallucinating arbitrary target views together with the whole pipeline for adapting the semantic segmentation network to target domains where labeling is missing. In stage 1, the view transformation network $\psi$ is only trained on color images, and is used for generating pseudo labels in the target domain in stage 2. The semantic segmentation network $\phi$ is then adapted to the target view in stage 3 using the target pseudo labels.

different views may fail completely when tested on semantic images due to their statistical difference.

To validate, we train a UNet [38] $\psi^{unet}$ to hallucinate $x_i^\mathcal{T}$ from $x_i^s$ using an L1 reconstruction objective at fixed $\tau$. After training, we check if $\psi^{unet}(y_i^s)$ is similar to $y_i^\mathcal{T}$. As shown in Fig. 5b (2nd column), the UNet trained on color images does not generalize well on semantic images, which confirms the difficulties of hallucinating the novel appearance of a seen view, even if the geometry is unchanged.

We propose that the key to generalizing to novel appearance is to bias the view transformation network towards learning spatial transportation instead of color transformation. For example, the network needs incentives to learn where the color should be copied to in the target view instead of how the color should change to form the target view. If so, the view hallucination should generalize to any novel appearance since the color transformation may depend on domains while the transport conditioned on the same views and geometry is invariant.

**Biasing towards information transport with attention.** The self-attention mechanism proposed in [52] represents a layer that processes the input by first predicting a set of keys ($K$) and a set of queries ($Q$), whose dot-products are then used to update a set of values ($V$) to get the output (updated values):

$$\text{ATTN}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

By examining how a single output value $v_i'$ is computed, we can see why attention helps to bias towards transport that facilitates the generalization of the hallucination. Let $q_i$ be the corresponding query for $v_i'$, and $[k_1, k_2, ..., k_m]$ be the keys, then $v_i' = \sum_{j=1}^m \alpha_j \cdot v_j$, with $\alpha_j$'s the elements of $\text{softmax}([k_1 q_i^T, k_2 q_i^T, ..., k_m q_i^T])$ (scaling factor omitted for simplicity). Note if $q_i$ is extremely similar to a certain key, *e.g.*, $k_{j^*}$, but dis-similar to the other keys, we may write $v_i' \approx v_{j^*}$. This signals that the attention is transporting values from different locations to $i$ through the weighted summation. In the extreme case, it can even stimulate point-wise transportation of the values.

To verify the effectiveness of attention in hallucinating labels (novel appearance), we simply reorganize the tunable parameters in the UNet $\psi^{unet}$ such that the convolutional layers near the bottleneck are now replaced by attention layers of the same capacity. We term it $\psi^{attn}$ and train it to hallucinate the target color images from the source color images, *i.e.*, $\hat{x}_i^\mathcal{T} = \psi^{attn}(x_i^s)$, and test it on the semantic images. As shown in Fig. 5b (3rd column), $\psi^{attn}$ can hallucinate reasonable target semantic images even it is only trained on color images. Given the effectiveness of the inductive bias introduced by the attention mechanism in semantic information hallucination for a single target view, we now detail our view hallucination network for multiple target views and the technique that we propose to generate soft labels for adaptation to different target domains.

## 3.2. Labeling multiple target domains

Here we specify the proposed network architecture that can seamlessly work for different target views, *e.g.*, the target domain is a mixture of views, which eliminates the need to train separate networks. Again, the view transformation network $\psi(x_V; x_K, x_Q)$ takes in a pair of color images, which guide $\psi$ to predict the target view from the source whose appearance is determined by either the source color image or the source semantic image, *i.e.*, $\hat{x}_i^\mathcal{T} =$

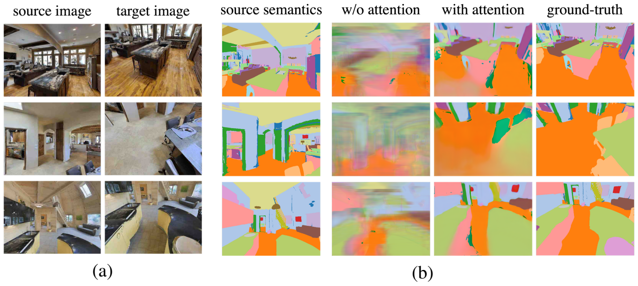

Figure 5. (a): source and target color images for training the view transformation network $\psi$; (b): applying $\psi$ on source semantic images. The hallucinated semantic images by the network without attention ("w/o attention") are inaccurate and not consistent with the target images; however, the hallucinations from the network with attention ("with attention") are sharp and more precise.

$\psi(x^s_i; x^s_i, x^\tau_i)$ (stage 1) or $\hat{y}^\tau_i = \psi(y^s_i; x^s_i, x^\tau_i)$ (stage 2). As illustrated in Fig. 4 (stage 1), we let $x_Q = x^\tau_i$, $x_K = x^s_i$ and $x_V = x^s_i$, which are lifted to query, key and value features through the following procedure:

$$Q^0 = \xi_Q(x_Q)[\mathbf{1}; u_{pos}; v_{pos}]$$
$$K^0 = \xi_K(x_K)[\mathbf{1}; u_{pos}; v_{pos}]$$
$$V^0 = \xi_V(x_V)$$

here $\xi_Q, \xi_K, \xi_V$ are separate encoders with strided convolutions to reduce the spatial dimensions of the features, and $u_{pos}, v_{pos}$ are fixed positional encodings that represent the normalized image grids (horizontal and vertical), i.e., for coordinate $(m, n)$ we have $u_{pos}(m, n) = n/w, v_{pos}(m, n) = m/h$, with $h$ and $w$ the height and width of the image. These lifted features are processed by $L$ of the proposed *information transport* layer:

$$Q^l = \text{FFN}^l_Q(Q^{l-1}) \tag{1}$$
$$K^l = \text{FFN}^l_K(K^{l-1}) \tag{2}$$
$$\hat{V}^l = \text{ATTN}(Q^{l-1}, K^{l-1}, V^{l-1}W^l) \tag{3}$$
$$\bar{V}^l = \text{FFN}^l_{V1}(\hat{V}^l) + V^{l-1} \tag{4}$$
$$V^l = \text{FFN}^l_{V2}(\bar{V}^l) + \bar{V}^l \tag{5}$$
$$x^l_Q = \xi_D(V^l) \tag{6}$$

where $\text{FFN}^l_Q, \text{FFN}^l_K$ are two feed-forward networks of downsampling and upsampling convolutional layers with layernorm to maintain the size of the updated keys and queries. And the feed-forward networks $\text{FFN}^l_{V1}, \text{FFN}^l_{V2}$ are simply convolutional layers whose stride is equal to one. Using two $\text{FFN}_V$ with skip connections can improve the convergence speed with a light network capcity increase. We ablate on update schemes of $K, Q$ in the experiments. Note, for each $V^l$, we apply the shared decoder $\xi_D$ to map it to the image space, and $x^L_Q$ will be the final output of the proposed view transformation network $\psi$.

**Training loss and data augmentation.** For training the network $\psi(x_V; x_K, x_Q)$ in Fig. 4 (stage 1), we apply color jittering to the input images. Specifically, the hue of $x^s_i, x^\tau_i$ are perturbed by a random factor to get $x_Q$ and $x_K$, and by a different factor to get $\bar{x}_Q$ and $x_V$, where $\bar{x}_Q$ is the expected output of $\psi(x_V; x_K, x_Q)$. Different hue perturbations can help prevent information leakage, since now $x_Q$ (input) and $\bar{x}_Q$ (expected output) are not identical, yet the consistency between $x_V$ and $\bar{x}_Q$ is preserved to enable meaningful hallucination. In addition, we apply the same color permutation to $x_V$ and $\bar{x}_Q$, to further prevent information leakage from $x_Q$ to the output. More details can be found in the appendix. The training loss for $\psi$ is:

$$\mathcal{L}^\psi = \sum_{x_Q \in \{\mathcal{D}^\tau\}} \sum_{l=1}^{L} \lambda_l \|x^l_Q - \bar{x}_Q\|_1 \tag{7}$$

with $\lambda_l$ the weighting coefficient for the $l$-th layer's output $x^l_Q$, which is decoded from $V^l$, and we set $\lambda_l = 2^{-(L-l)}$ so that early predictions are weighted less.

### 3.3. Functional label hallucination

Given the trained $\psi$, we can hallucinate the target semantic images for $x^\tau_i$'s, *i.e.*, $\hat{y}^\tau_i = \psi(x_V; x_K, x_Q)$, by setting $x_Q = x^\tau_i$, $x_K = x^s_i$ and $x_V = y^s_i$. We can then convert the hallucinated semantic images into semantic labels (integers) via nearest neighbor search in the color space to adapt the semantic segmentation network ($\phi$) to the target domains. However, the converted labels sometimes could be wrong due to noise in the predicted color ( Fig. 6 (3rd, 4th columns)).

To resolve the ambiguities, we propose the following functional label hallucination by treating $\psi(\cdot; x^s_i, x^\tau_i)$ as the functional representation of an unknown mapping $T(x^s_i, x^\tau_i) : \Omega_s \to \Omega_\tau$ conditioned on the color images $x^s_i, x^\tau_i$. Here $\Omega_s, \Omega_\tau$ represent the source and target image domains/grids. According to [32], if $T$ is a bijective mapping between $\Omega_s$ and $\Omega_\tau$, the actual mapping $T$ can then be recovered from $\psi(\cdot; x^s_i, x^\tau_i)$ by checking its output of indicator functions of the elements in $\Omega_s$. However, recovering the underlying unknown mapping $T$ is unnecessary in our scenario, and, indeed, we do not have any constraints that $T$ is bijective. Instead, we utilize the functional representation $\psi(\cdot; x^s_i, x^\tau_i)$ to find regions in $\Omega_\tau$ that share the same label with those in $\Omega_s$. Let $\mathbf{1}_{y^s_i=c}$ be the indicator function of the regions that are classified as class $c$, then $\hat{y}^\tau_{ic} = \psi(\mathbf{1}_{y^s_i=c}; x^s_i, x^\tau_i)$ indicates the regions of class $c$ in $\Omega_\tau$. And the hallucinated labels can be written as:

$$\hat{y}^\tau_i = \text{softmax}(\psi(\mathbf{1}_{y^s_i=1}; x^s_i, x^\tau_i), ..., \psi(\mathbf{1}_{y^s_i=C}; x^s_i, x^\tau_i)) \tag{8}$$

with $C$ the number of semantic classes, and now the hallucinated target view labels $\hat{y}^\tau_i$ represent the probabilistic distributions over the $C$ classes for each pixel.

**Adapting to target domains.** With the functional hallucination strategy, we can avoid performing a nearest neighbor search in the color space, which improves the accuracy

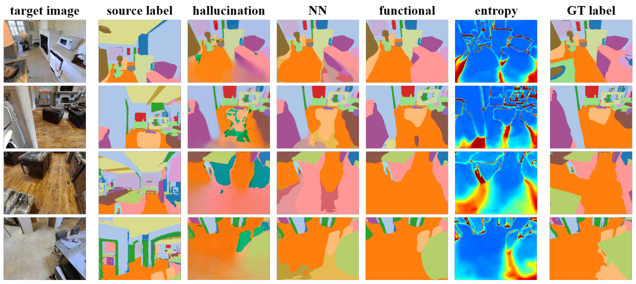

Figure 6. Effectiveness of the proposed functional hallucination strategy. The target semantic images (3rd) are hallucinated from the source counterparts (2nd), which are decoded into semantic labels using nearest neighbor search (4th) or the proposed functional strategy (5th) with uncertainties (6th).

of the generated labels even when the hallucinated color is noisy (Fig. 6 (5th column)). Moreover, the soft probabilistic labels (Fig. 6 (6th column)) are more suitable for adapting the semantic segmentation network $\phi$ to the target domains, avoiding errors of hard labels when the hallucination is of low confidence. We then finetune $\phi$ for each target domain using the target dataset $\mathcal{D}^{\tau} = \{(x_i^{\tau}, \hat{y}_i^{\tau})\}$ augmented with the soft labels:

$$\mathcal{L}^{\phi} = \sum_i \mathbb{H}(\hat{y}_i^{\tau}, \phi(x_i^{\tau})) \tag{9}$$

where $\mathbb{H}$ is the cross-entropy between the network prediction and target pseudo labels.

# 4. Experiments

## 4.1. Data generation

Due to the lack of benchmarks for evaluating UDA methods under viewpoint shifts, we propose a new dataset whose source and target domains are generated by varying camera elevation and viewing angles. Moreover, we explicitly control the viewpoint shifts, such that we can quantitatively assess the adaptation performance with respect to the degree of domain gaps. We resort to simulation for data collection since 1) it is much easier to obtain semantic segmentation ground-truth in simulation; 2) the degree of the domain gap caused by viewpoint change is more controllable; and 3) it is more friendly to the personnel who is in charge of the data collection given the pandemic.

Furthermore, we maximize the realism of the generated data by employing the Matterport3D dataset [3], which contains 90 building-scale real-world scenes with pixel-wise semantic annotations[2]. The scenes from Matterport3D are then imported into the Habitat simulation platform [42] for our data generation. Specifically, we first randomly sample two states in the scene, with one state (the position and yaw angle of a virtual camera) representing the starting state, and the other the end state. We then perform collision-free path planning between these two states. The resulted path is accepted if it has a length larger than 15 path points, and images at each point along the path are collected. To synthesize the domain gaps, we set the pitch angle of the virtual camera to 0° for collecting the source domain videos (annotations), which resembles the working viewpoint for semantic segmentation networks trained on existing scene parsing datasets [45, 47, 68]. Moreover, we increase the pitch angle of the virtual camera by 10° (up to 90°) for collecting target domain videos (no annotations), which results in 9 different target domains. For each domain, we collect 13,500 training images and 2,700 test images with resolution 384×512. Please see Fig. 7 for samples from the collected datasets.

## 4.2. Implementation details

We adapt the UNet structure [38] with reduced capacity and layernorm activation to construct the feed-forward networks $\text{FFN}_Q$ and $\text{FFN}_K$. Similar to [62], $W$ is a convolutional layer with kernel size $1\times1$, $\text{FFN}_{V1}, \text{FFN}_{V2}$ consist of one and two convolutional layers respectively. Both $\text{FFN}_{V1}$ and $\text{FFN}_{V2}$ use leakyrelu activation function. Network $\psi$ contains $L = 8$ attention modules. Training of $\psi$ is carried out on eight Nvidia V100 GPUs, with batch size 16. We use the Adam optimizer with an initial learning rate of 1e-4 and momentums of 0.9 and 0.999. The training converges after 10 epochs. We use the DeepLabv2 [5] with the ResNet101 backbone as the semantic segmentation network $\phi$, which is initialized with the pre-trained weights on ImageNet [25, 30, 53, 59, 65]. Soft labels for each target view $\tau$ are hallucinated using Eq. (8). The semantic segmentation networks $\phi^{\tau}$ for each target domain are trained using Eq. (9) with the Adam optimizer, a batch size of 6 and an initial learning rate of 7.5e-5. The learning rate is then halved after 10 and 15 epochs. The training converges at 25 epochs. To have a fair comparison with the state-of-the-art domain adaptation methods that adapt from a single source domain to a single target domain, we also train the segmentation network for each target domain separately. We use mean intersection-over-union (mIoU) as the metric. [3]

## 4.3. Ablation study

**Effectiveness of the proposed inductive bias.** Qualitative comparisons in Fig. 5 show that the proposed *spatial transport* inductive bias and the architecture facilitate zero-shot semantic image hallucination. In Tab. 1 we quantitatively confirm its effectiveness and check how it extends across different levels of viewpoint shift. Besides the *color transformation bias* ("UNet"), we also test the inductive biases introduced by explicitly modeling the dense 2D correspondence ("Flow") and by explicitly modeling the image

[2] we completed the Terms of Use agreement form and obtained consent from the creators, and the data does not contain any privacy information

[3] orientation within a horizontal plane, similarly, the pitch angle is the orientation within a vertical plane

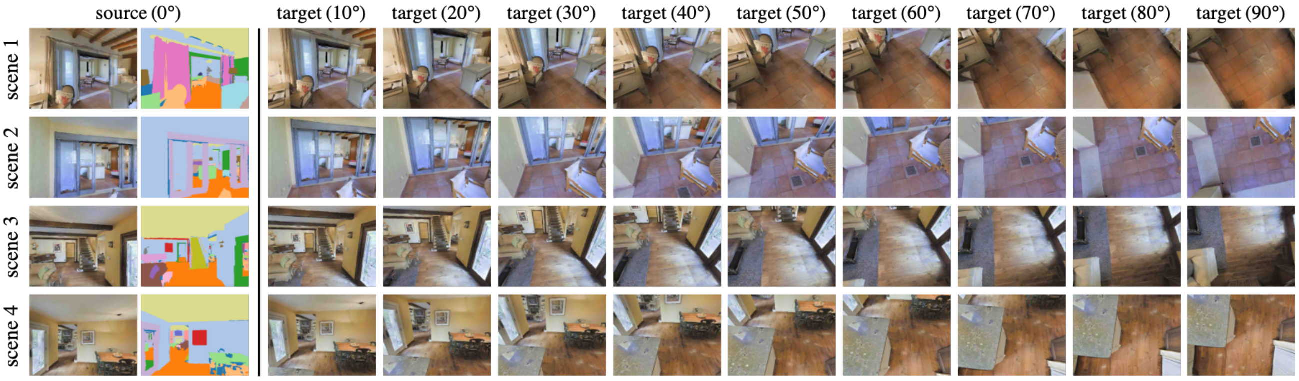

Figure 7. Samples from the proposed dataset (one source and nine target domains) for benchmarking unsupervised domain adaptation methods under viewpoint shifts in semantic segmentation.

| | Target Domains | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | 10° | 20° | 30° | 40° | 50° | 60° | 70° | 80° | 90° |
| UNet [38] | 49.76 | 28.19 | 13.69 | 9.26 | 6.56 | 4.71 | 2.59 | 1.63 | 1.28 |
| Flow [48] | 33.04 | 27.59 | 22.72 | 19.36 | 17.02 | 14.21 | 11.55 | 9.67 | 8.34 |
| RAFT [48] | 70.62 | 61.25 | 53.92 | 42.54 | 18.17 | 9.36 | 7.57 | 6.24 | 5.58 |
| 3D [57] | 28.16 | 22.12 | 18.35 | 15.80 | 13.14 | 11.22 | 9.20 | 6.61 | 2.86 |
| ADeLA(S) | 54.85 | 46.29 | 42.66 | 37.75 | 27.71 | 21.33 | 14.18 | 8.69 | 4.17 |
| ADeLA(M) | 48.42 | 41.87 | 35.73 | 30.39 | 24.11 | 17.40 | 11.79 | 8.82 | 7.34 |
| UNet+F [38] | 73.62 | 49.07 | 27.12 | 20.08 | 16.48 | 13.68 | 11.61 | 9.79 | 8.53 |
| ADeLA(S)+F | 70.07 | **67.63** | **58.62** | **54.33** | **47.45** | **37.81** | **28.39** | **19.78** | **15.17** |
| ADeLA(M)+F | **75.75** | 66.29 | 57.45 | 49.57 | 40.38 | 30.00 | 20.96 | 15.44 | 12.60 |

Table 1. Ablation study on different inductive biases for zero-shot semantic image hallucination. Numbers are the mIoUs of the hallucinated semantic labels on the training set of each target domain.

formation process in 3D ("3D"). For "Flow," we adapt the architecture of RAFT [48] and train it to estimate the flow that reconstructs the target color image from the source, and use the flow for warping the semantic labels. For "3D", we adapt the state-of-the-art single view synthesis framework [57], and supply it with ground-truth camera poses for semantic image synthesis. We report the performance of our method under two settings: the single source to single target setting ("ADeLA(S)"), and the the single source to multiple targets setting ("ADeLA(M)"). The labels for "ADeLA(S)" and "ADeLA(M)" are generated using the nearest neighbor search. We also report the score of the warped labels using the fully supervised RAFT model for reference.

We can make the following observations: 1) "UNet" (color transformation) does not work at large viewpoint shifts. 2) the 2D dense correspondence inductive bias ("Flow") works better for large viewpoint shifts, which verifies our proposal for biasing towards transportation. However, the comparison between "Flow" and "RAFT" shows that the spatial correspondence learned from color images can be erroneous, so "Flow" is much worse than "RAFT" at small viewpoint changes. Moreover, "RAFT" is worse than "Flow" at large viewpoint shifts, which indicates that the exact dense correspondence may not be suitable for semantic label hallucination. 3) The 3D inductive bias ("3D") does not perform well since the learned 3D representation from color images does not generalize to semantic images. 4) Our model performs well across all target domains, due to the proposed spatial transportation bias, and the capability to hallucinate beyond exact correspondence.

Moreover, we show the quality of the semantic labels hallucinated using the proposed functional label hallucination strategy ("UNet+F," "ADeLA(S)+F," "ADeLA(M)+F"). As seen in Tab. 1 (bottom), functional hallucination significantly improves the performance of UNet and our models, demonstrating its effectiveness in resolving the ambiguities in the hallucinated semantic images. Note, "Flow" and RAFT warp labels with explicit dense correspondence, thus they are unable to take advantage of the functional strategy. The same observation holds for "3D", whose 3D representation learned with color images does not generalize even with ground-truth camera poses.

**Effectiveness of the update scheme for $K, Q$.** We conduct experiments to investigate different $K, Q$ update schemes. The *information transport* layer uses UNet structures $\text{FFN}_K$ and $\text{FFN}_Q$ to update $K, Q$. To check the effectiveness of the UNet structure, which performs spatial downsampling and upsampling (feature resolution preserved), we replace them with several 1x1 convolutions with the same capacity to maintain the spatial resolution and update $K, Q$. Also, to confirm the need for updates in $K, Q$, we remove all $\text{FFN}_K$ and $\text{FFN}_Q$ modules in our model so that $K, Q$ do not change across different layers. As shown in Tab. 2, there is a significant performance drop if we replace the proposed UNet structure with other options. These experiments confirm that updating $K, Q$ is necessary, and introducing spatial downsampling and upsampling while updating $K, Q$ is not only more computationally efficient but can also improve the accuracy.

**Number of information transport layers.** We experiment with the number $L$ of information transport layers in the view transformation network on source domain 0° and target domain 30°. The results are reported in Tab. 3. Users

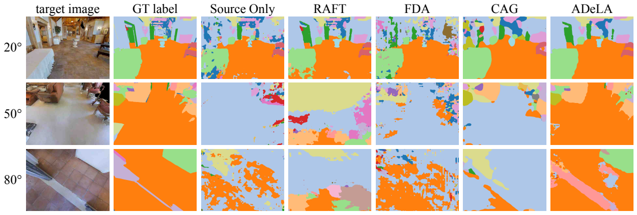

Figure 8. Qualitative comparison with competing methods on different target domains. RAFT [48], FDA [59], and CAG [65].

| $K, Q$ update scheme | UNet | 1x1 convs | no update |
|---|---|---|---|
| mIoU | 42.7 | 38.4 | 33.4 |

Table 2. Comparison between different $K, Q$ update schemes with source domain 0° and target domain 30°.

| $L$ | 1 | 3 | 5 | 7 | 8 |
|---|---|---|---|---|---|
| mIoU | 35.24 | 39.73 | 40.27 | 41.93 | 42.66 |
| #Params | 15.8M | 34.8M | 53.8M | 72.8M | 82.3M |
| FPS | 20.71 | 16.88 | 13.05 | 10.64 | 9.68 |

Table 3. Effects of the number of information transport layers.

| | Target Domains | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Type | 10° | 20° | 30° | 40° | 50° | 60° | 70° | 80° | 90° |
| Soft | 31.91 | 28.49 | 24.31 | 21.34 | 16.42 | 12.92 | 9.74 | 7.92 | 5.37 |
| Hard | 30.34 | 26.63 | 23.45 | 20.23 | 16.19 | 12.74 | 9.69 | 7.68 | 5.99 |

Table 4. Effects of using soft and hard labels for training $\phi$.

can adjust $L$ for a trade-off between the label quality and label hallucination speed based on their actual budget.

**Effectiveness of soft labels for training $\phi$.** We experiment with both soft and hard labels for training the segmentation network $\phi$. The results are shown in Tab. 4. Soft labels outperform hard labels across different viewpoint shifts, which demonstrates the effectiveness of the estimated uncertainties as analyzed in Sec.3.3.

### 4.4. Benchmarking

We carry out an extensive study of state-of-the-art methods in reducing performance drops caused by viewpoint shifts on semantic segmentation [22, 23, 25, 30, 33, 53, 59, 64, 65]. The benchmarking is reported in Tab. 5. Among those methods, [25, 64] focus on self-training, [30, 33, 65] perform class-wise and curriculum domain alignment, and [23, 53, 59] align domains in the image/output space. We also experiment with three best performing dense correspondence estimation methods [48, 54, 67], and two single view synthesis methods [57, 70] to generate target view labels to help adapt the segmentation networks. All methods are re-trained on the training sets of the proposed benchmark, and tested on the test sets of the nine target domains. Our method consistently achieves positive adaptation gain and performs much better than the other methods at large viewpoint shifts. Note that FDA [59] performs better on the target domain of 10° (small gap) due to its strong style randomization mechanism. However, our method surpasses FDA on the remaining target domains without any data augmentation in adapting the segmentation network. In order to verify if the temporally aligned data is beneficial for other methods, we select the top UDA methods and train them also on paired source and target images. The results in Tab. 5 show that the performance of these UDA methods even degrades compared to their unpaired counterparts. This concludes that the comparison is fair and comprehensive. Please see Fig. 8 for visual results.

| | | Target Domains | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Type | Method | 10° | 20° | 30° | 40° | 50° | 60° | 70° | 80° | 90° |
| Baseline | Target Only | 33.22 | 31.34 | 30.35 | 29.45 | 27.18 | 25.70 | 25.85 | 24.93 | 24.12 |
| | Source Only | 27.90 | 20.24 | 12.22 | 7.63 | 4.41 | 2.53 | 1.80 | 1.53 | 1.53 |
| | UNet [38] | 30.51 | 22.36 | 12.26 | 8.44 | 6.58 | 4.80 | 3.81 | 2.98 | 2.22 |
| Dense Corresp. Est. | RAFT [48] | 29.46 | 26.74 | 24.15 | 16.92 | 8.66 | 4.54 | 3.35 | 2.82 | 2.41 |
| | MFNet [67] | 29.20 | 24.53 | 13.16 | 6.65 | 4.76 | 3.95 | 3.00 | 2.81 | 2.48 |
| | DICL [54] | 29.62 | 26.45 | 22.01 | 17.75 | 6.25 | 4.44 | 3.40 | 2.75 | 2.49 |
| UDA (unpaired) | ProDA [64] | 25.26 | 19.35 | 12.03 | 7.46 | 4.39 | 1.74 | 1.12 | 0.83 | 0.77 |
| | CLAN [30] | 28.22 | 21.21 | 13.17 | 7.53 | 4.37 | 2.66 | 2.02 | 1.74 | 1.73 |
| | CAG [65] | 27.17 | 22.22 | 15.05 | 8.57 | 5.24 | 2.43 | 1.83 | 1.50 | 1.54 |
| | FDA [59] | **37.80** | 23.34 | 12.33 | 6.69 | 3.58 | 2.15 | 1.56 | 1.67 | 1.32 |
| | PLCA [22] | 26.83 | 19.04 | 12.42 | 7.79 | 5.37 | 3.38 | 2.49 | 2.21 | 1.93 |
| | LTIR [23] | 26.22 | 20.50 | 13.43 | 6.16 | 3.90 | 2.09 | 1.82 | 1.65 | 1.66 |
| | CCM [25] | 28.26 | 19.48 | 10.56 | 4.92 | 2.78 | 1.50 | 1.14 | 0.95 | 0.90 |
| | Advent [53] | 11.38 | 7.93 | 4.98 | 3.28 | 2.54 | 2.16 | 1.60 | 1.52 | 1.49 |
| | Intrada [33] | 10.16 | 7.84 | 6.13 | 4.08 | 2.67 | 1.98 | 1.58 | 1.67 | 0.93 |
| UDA (paired) | ProDA [64] | 20.61 | 17.82 | 10.38 | 6.71 | 4.11 | 1.85 | 1.11 | 0.91 | 0.90 |
| | CLAN [30] | 25.41 | 18.33 | 10.61 | 5.91 | 3.37 | 2.19 | 1.71 | 1.57 | 1.58 |
| | CAG [65] | 23.48 | 18.55 | 12.34 | 8.06 | 4.33 | 1.83 | 1.59 | 1.47 | 1.57 |
| | FDA [59] | 30.83 | 16.46 | 11.39 | 6.9 | 3.69 | 2.17 | 1.74 | 1.84 | 1.69 |
| | PLCA [22] | 24.86 | 19.49 | 12.36 | 8.63 | 5.52 | 3.70 | 2.84 | 2.26 | 2.04 |
| Novel View Syn. | AppFlow [70] | 14.73 | 12.91 | 10.46 | 8.43 | 7.30 | 5.68 | 4.66 | 3.87 | 3.39 |
| | Synsin [57] | 14.43 | 11.44 | 9.15 | 8.29 | 6.24 | 5.43 | 4.67 | 3.36 | 1.88 |
| Info. Trans. | ADeLA | 31.91 | **28.49** | **24.31** | **21.34** | **16.42** | **12.92** | **9.74** | **7.92** | **5.37** |

Table 5. Quantitative comparison to state-of-the-art methods on the proposed benchmark. Numbers are mIoU scores on the test set of each target domain.

## 5. Discussion

We tackle the performance drop caused by viewpoint shifts in semantic segmentation. Experiments verify that aligning statistics between domains in a shared space could be detrimental due to the content shift across different viewing angles. Our method achieves higher adaptation gains, especially at large viewpoint shifts. However, the adaptation gain of our method also decreases towards the extreme case. In our code release, we will specify allowable uses of our system with appropriate licenses to address potential ethical and societal concerns. In the future, we would like to explore the use of temporal information to further reduce the performance drop caused by extreme viewpoint shifts.

## 6. Acknowledgement

This work was supported by the National Key R&D Program of China (No. 2018YFE0100900), a grant from the Stanford HAI Institute, a Vannevar Bush faculty fellowship, and a gift from the Adobe Corporation.

## References

[1] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 189–205, 2018. 2

[2] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 769–776, 2013. 2

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 6

[4] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1900–1909, 2019. 2

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 6

[6] Inchang Choi, Orazio Gallo, Alejandro Troccoli, Min H Kim, and Jan Kautz. Extreme view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7781–7790, 2019. 3

[7] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoff Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *arXiv preprint arXiv:2003.04475*, 2020. 2

[8] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Nova: Learning to see in novel viewpoints and domains. In *2019 International Conference on 3D Vision (3DV)*, pages 116–125. IEEE, 2019. 2, 3

[9] Gabriela Csurka. A comprehensive survey on domain adaptation for visual applications. In *Domain adaptation in computer vision applications*, pages 1–35. Springer, 2017. 2

[10] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 994–1003, 2018. 2

[11] Daniele Di Mauro, Antonino Furnari, Giuseppe Patanè, Sebastiano Battiato, and Giovanni Maria Farinella. Sceneadapt: Scene-based domain adaptation for semantic segmentation using adversarial learning. *Pattern Recognition Letters*, 2020. 2

[12] Jiahua Dong, Yang Cong, Gan Sun, Yuyang Liu, and Xiaowei Xu. Cscl: Critical semantic-consistent learning for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 745–762. Springer, 2020. 2

[13] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 982–991, 2019. 2

[14] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967, 2013. 2

[15] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2376, 2019. 3

[16] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6112–6121, 2019. 2

[17] Bo Geng, Dacheng Tao, and Chao Xu. Daml: Domain adaptation metric learning. *IEEE Transactions on Image Processing*, 20(10):2980–2989, 2011. 2

[18] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520, 2011. 2

[19] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2477–2486, 2019. 2

[20] 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 *Proceedings of the 35th International Conference on Machine Learning*, 2018. 2, 3

[21] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In *European Conference on Computer Vision*, pages 705–722. Springer, 2020. 2

[22] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander G. Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. In *Advances in neural information processing systems*, 2020. 2, 8

[23] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12975–12984, 2020. 2, 3, 8

[24] Abhishek Kumar, Prasanna Sattigeri, Kahini Wadhawan, Leonid Karlinsky, Rogerio Feris, Bill Freeman, and Gregory Wornell. Co-regularized alignment for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 9345–9356, 2018. 2

[25] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *European Conference on Computer Vision*, pages 440–456. Springer, 2020. 2, 3, 6, 8

[26] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019. 2

[27] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017. 2

[28] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015. 2

[29] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6778–6787, 2019. 2

[30] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019. 2, 3, 6, 8

[31] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5715–5725, 2017. 2

[32] Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics (TOG)*, 31(4):1–11, 2012. 5

[33] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020. 2, 8

[34] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015. 2

[35] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 3

[36] Xingchao Peng, Ben Usman, Kuniaki Saito, Neela Kaushik, Judy Hoffman, and Kate Saenko. Syn2real: A new benchmark forsynthetic-to-real visual domain adaptation. *arXiv preprint arXiv: 1806.09755*, 2018. 3

[37] Hossein Rahmani, Ajmal Mian, and Mubarak Shah. Learning a deep model for human action recognition from novel viewpoints. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):667–681, 2017. 2

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4, 6, 7, 8

[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. 3

[40] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2988–2997. JMLR. org, 2017. 2

[41] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018. 2

[42] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019. 6

[43] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 2110–2118, 2016. 2

[44] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. In *Proc. 6th International Conference on Learning Representations*, 2018. 2

[45] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *Proceedings of the International Conference on Computer Vision - Workshop on 3D Representation and Recognition*, 2011. 6

[46] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *arXiv preprint arXiv:1906.01618*, 2019. 3

[47] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 6

[48] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020. 3, 7, 8

[49] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1456–1465, 2019. 2

[50] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020. 3

[51] 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. 2

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 4

[53] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 2, 3, 6, 8

[54] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Kaihao Zhang, Pan Ji, and Hongdong Li. Displacement-invariant matching cost learning for accurate optical flow estimation. *arXiv preprint arXiv:2010.14851*, 2020. 3, 8

[55] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 2

[56] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12635–12644, 2020. 2

[57] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020. 3, 7, 8

[58] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 518–534, 2018. 2

[59] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2, 3, 6, 8

[60] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 2

[61] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*, 2017. 2

[62] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *The Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 6

[63] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2117–2126, 2017. 2

[64] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. *arXiv preprint arXiv:2101.10979*, 2, 2021. 2, 3, 8

[65] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *arXiv preprint arXiv:1910.13049*, 2019. 2, 3, 6, 8

[66] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2020–2030, 2017. 2

[67] Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I Chang, Yan Xu, et al. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6278–6287, 2020. 3, 8

[68] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6

[69] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 3

[70] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European conference on computer vision*, pages 286–301. Springer, 2016. 3, 8

[71] 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. 2