

SEMANTIC IMAGE SEGMENTATION METHODS AND SYSTEMS

TECHNICAL FIELD

5

The present disclosure relates to machine learning and image processing. In particular, the present disclosure relates to semantic image segmentation.

BACKGROUND

10

Semantic segmentation has been a longstanding challenge in computer vision, which aims to assign class labels to every pixel of an image. Deep-learning based approaches have achieved great successes at the price of large-scale densely-annotated datasets which are prohibitively expensive to collect. One way of circumventing this constraint is to use synthesized images with automatically generated labels (e.g. synthesized or game-engine produced data) in network training. Unfortunately, such models usually undergo a drastic performance drop when applied to real-world images due to the domain bias and shift.

15

20

Unsupervised domain adaptation (UDA) has been introduced to address the domain bias/shift issue. To reduce the cross-domain discrepancy, most state-of-the-art UDA methods exploit adversarial learning for distribution alignment in the intermediate feature, output or latent space. Among this cohort of adversarial-based methods, a common and pivotal step is the employment of a discriminator that predicts a domain label for data being either source or target domain. However, the discriminator works only on image-level and merely achieves global consistency (i.e. locational/spatial distributions consistency), where local contextual consistency (i.e. region-wise contextual-relationships) is largely neglected.

25

30

Local contextual-relationships are ubiquitous and provide important cues for scene segmentation. They can be formulated in terms of semantic compatibility/incompatibility relations between one object and its neighbouring objects. Under this formulation, a compatibility relation is an indication of visual patterns with high co-occurrence frequency, e.g. a pole beside a side-walk, and an incompatibility

relation is an indication of visual patterns with low co-occurrence frequency, e.g. a person above a driving car. The contextual information has been extensively explored in supervised semantic segmentation, whereas the local contextual-relationships is largely neglected in unsupervised domain adaptive semantic segmentation though they're beneficial in addressing local contextual consistency and inconsistency in the target domain.

SUMMARY

According to a first aspect of the present disclosure, a semantic image segmentation method of segmenting a target image is provided. The method comprises: receiving a source image and pixel labels for the source image, the pixel labels for the source image indicating, for each respective pixel of the source image, a class from a plurality of classes; training a feature extractor to extract image features from the source image and from the target image, training a pixel scale classifier to classify pixel scale features of the source image and the target image and training a local scale contextual relation classifier to classify local scale image contextual relation features of the source image and the target image and thereby obtaining a trained feature extractor, a trained pixel scale classifier and a trained local scale contextual relation classifier, wherein training the feature extractor, training the pixel scale classifier and training the local scale classifier is carried out while aligning domain shift on a pixel scale, a local scale and a global scale between a domain of the source image and a domain of the target image; and using the trained feature extractor, the trained pixel scale classifier and the trained local scale contextual relation classifier to generate pixel labels for the target image.

The proposed unsupervised domain adaptation method for semantic segmentation explicitly models the local contextual-relationships in the feature space of source domain (with label) and then transfers this contextual information into the target domain (without label), ultimately improving target domain segmentation quality. We first establish local contextual-relationships pseudo annotations in the source domain. This can be achieved by sampling regions from pixel-level ground-truth maps of source images and clustering the sampled regions to indexed groups. With the local contextual-relationships pseudo annotations in source domain, we can train a

classifier to explicitly model and learn the local contextual-relations in the feature space of source domain, and then transfer and enforce these local contextual-relations into target domain.

5 Following current discriminator-based global alignment methods, an intuitive idea is to employ hundreds of discriminators to align hundreds of contextual-relations across domain where a single discriminator focuses on a single contextual-relation, or employ just one discriminator to align all contextual-relations across domains. Obviously, the former is cumbersome which requires much redundant computation, while the latter is
10 not aware of a variety of contextual-relations in the data distribution and may end up biasing to low-level/simple difference. Therefore, different from current discriminator-based global alignment methods, we enforce these local contextual-relations on target domain via adaptive entropy max-minimizing (AEMM) between classifier C and feature extractor that estimates prototypical feature representations of these local contextual-
15 relations and congregates neighboring target incorrect samples/contextual-relations to the approximated correct source prototypes alternatively, ultimately leading to consistent local contextual-relations across domains. In this way, our method requires no discriminator which is normally used in UDA-based semantic segmentation and introduces training instability and extra components. In addition, this AEMM learning
20 scheme can also be applied into pixel-/global-scale training.

The present disclosure provides an unsupervised domain adaptation method for semantic segmentation that explicitly models the local contextual-relations in the feature space of source domain (with label) and then transfers this contextual
25 information into target domain (without label). Thus, contextual information is used for UDA-based semantic segmentation. The present disclosure further provides a novel adaptive entropy max-minimizing adversarial learning scheme to effectively align hundreds of local contextual-relations across domain, which requires no discriminator and adds no overhead.

30

In an embodiment, training the local scale contextual relation classifier to classify local scale image contextual relation features of the source image and the target image comprises implementing an adversarial training scheme between the local scale contextual relation classifier and the feature extractor.

In an embodiment, training the pixel scale classifier to classify pixel scale features of the source image and the target image comprises implementing an adversarial training scheme between the pixel scale classifier and the feature extractor.

5

In an embodiment, the method further comprises training a domain scale classifier to predict a domain label for a layout probability map generated from probability maps generated by the pixel scale classifier and the local scale contextual relationship classifier.

10

In an embodiment, training the domain scale classifier to predict a domain label for a layout probability map generated from probability maps generated by the pixel scale classifier and the local scale contextual relationship classifier comprises implementing an adversarial training scheme between the domain scale classifier and the feature
15 extractor.

In an embodiment, the training is carried out using an adaptive entropy max-minimizing learning scheme.

20

In an embodiment, training the local scale contextual relation classifier to classify local scale image contextual relation features of the source image and the target image comprises backpropagating using a reverse gradient layer.

25

In an embodiment, training the local scale contextual relation classifier to classify local scale image contextual relation features of the source image and the target image comprises sampling regions of the source image and clustering the sampled regions of the source image to generate prototypical features in the source domain.

30

In an embodiment, clustering the sampled regions of the source image to generate prototypical features in the source domain comprises clustering using a density-based spatial clustering of applications with noise algorithm.

According to a second aspect of the present disclosure, a computer readable medium storing processor executable instructions which when executed on a processor cause the processor to carry out a method as set out above is provided.

5 According to a third aspect of the present disclosure, a semantic image segmentation system for segmenting a target image is provided. The system comprises a processor and a data storage device storing a feature extractor, a pixel scale classifier, a local scale contextual relation classifier, and computer program instructions operable to cause the processor to: receive a source image and pixel labels for the source
10 image, the pixel labels for the source image indicating, for each respective pixel of the source image, a class from a plurality of classes; train the feature extractor to extract image features from the source image and from the target image, train the pixel scale classifier to classify pixel scale features of the source image and the target image and train the local scale contextual relation classifier to classify local scale image
15 contextual relation features of the source image and the target image and thereby obtaining a trained feature extractor, a trained pixel scale classifier and a trained local scale contextual relation classifier, wherein training the feature extractor, training the pixel scale classifier and training the local scale classifier is carried out while aligning domain shift on a pixel scale, a local scale and a global scale between a domain of
20 the source image and a domain of the target image; and use the trained feature extractor, the trained pixel scale classifier and the trained local scale contextual relation classifier to generate pixel labels for the target image.

In an embodiment, the computer program instructions are further operable to cause
25 the processor to train the local scale contextual relation classifier to classify local scale image contextual relation features of the source image and the target image by implementing an adversarial training scheme between the local scale contextual relation classifier and the feature extractor.

30 In an embodiment, the computer program instructions are further operable to cause the processor to train the pixel scale classifier to classify pixel scale features of the source image and the target image comprises by implementing an adversarial training scheme between the pixel scale classifier and the feature extractor.

In an embodiment, the data storage device further stores a domain scale classifier and the computer program instructions are further operable to cause the processor to train the domain scale classifier to predict a domain label for a layout probability map generated from probability maps generated by the pixel scale classifier and the local scale contextual relationship classifier.

In an embodiment, the computer program instructions are further operable to cause the processor to train the domain scale classifier to predict a domain label for a layout probability map generated from probability maps generated by the pixel scale classifier and the local scale contextual relationship classifier by implementing an adversarial training scheme between the domain scale classifier and the feature extractor.

In an embodiment, the training is carried out using an adaptive entropy max-minimizing learning scheme.

In an embodiment, the computer program instructions are further operable to cause the processor to train the local scale contextual relation classifier to classify local scale image contextual relation features of the source image and the target image by backpropagating using a reverse gradient layer.

In an embodiment, the computer program instructions are further operable to cause the processor to train the local scale contextual relation classifier to classify local scale image contextual relation features of the source image and the target image by sampling regions of the source image and clustering the sampled regions of the source image to generate prototypical features in the source domain.

In an embodiment, the computer program instructions are further operable to cause the processor to cluster the sampled regions of the source image to generate prototypical features in the source domain by clustering using a density-based spatial clustering of applications with noise algorithm.

BRIEF DESCRIPTION OF THE DRAWINGS

In the following, embodiments of the present invention will be described as non-limiting examples with reference to the accompanying drawings in which:

FIG.1A schematically shows a prior art image segmentation process;

5

FIG.1B schematically shows an image segmentation process according to an embodiment of the present invention;

FIG.2 is a block diagram showing a semantic image segmentation system according to an embodiment of the present invention;

10

FIG.3 is a flow chart showing a semantic image segmentation method according to an embodiment of the present invention;

15

FIG.4 shows an overview of the training of the feature extractor, the pixel scale classifier, and the local scale image contextual relation classifier in an embodiment of the present invention;

20

FIG.5A and FIG.5B show an overview of local contextual-relation pseudo label establishment in an embodiment of the present invention;

FIG.6A to FIG.6E show an overview and comparison of the proposed adaptive entropy max-minimizing at different scales according to an embodiment of the present invention;

25

FIG.7 is a table showing a comparison of results of a domain adaptation task using prior art methods and a method according to an embodiment of the present invention;

30

FIG.8 is a table showing a comparison of results of a domain adaptation task using prior art methods and a method according to an embodiment of the present invention;

FIG.9 is a table showing a complementary study of the proposed contextual-relation consistent domain adaptation with local-scale to current global alignment UDA methods;

5 FIG.10 is a table showing qualitative results for a semantic image segmentation task comparing the approach of embodiments of the present invention with existing approaches; and

FIG.11 is a set of images showing a comparison of image segmentation using
10 methods according to embodiments of the present invention with existing methods.

DETAILED DESCRIPTION

The present disclosure relates to image segmentation and in particular to semantic
15 segmentation with contextual relation consistent domain adaptation. Recent advances in unsupervised domain adaptation for semantic segmentation have shown great potentials to relieve the demand of expensive per-pixel annotations. However, most existing works address the domain discrepancy by aligning the data distributions of two domains at a global image level whereas the local consistencies are largely
20 neglected.

FIG.1A schematically shows a prior art image segmentation process. The process takes a source image 10 and a set of source image pixel labels which indicate image segments 20 as inputs. The aim of the process is to generate a set of target image
25 pixel labels 22A for a target image 12. In the prior art process, global alignment 30A between the source domain and the target domain is carried out, however, this can lead to incompatibility relations inconsistencies 40A between image features 42A of the source domain and image features 44A of the target domain. It can also lead to compatibility relation inconsistencies 50A between image features 52A of the source
30 domain and image features 54A of the target domain. Here, “compatibility relations” refer to visual patterns with high co-occurrence frequency (e.g., “pole” should be beside the “sidewalk”), and “incompatibility relations” refer to visual patterns with low co-occurrence frequency (e.g., “sky” should not in the “building”).

FIG.1B schematically shows an image segmentation process according to an embodiment of the present invention. As with the process described with reference to FIG.1A above, the process takes a source image 10 and a set of source image pixel labels 20 as input and generates a set of pixel labels 22B which indicate different semantic segments of a target image 12. The image segmentation process of the present disclosure provides an innovative local contextual-relation consistent domain adaptation (CrCDA) technique that aims to achieve local-level consistencies during the global-level alignment. The idea is to take a closer look at region-wise feature representations and align them for local-level consistencies. Specifically, the CrCDA technique learns and enforces the prototypical local contextual-relations explicitly in the feature space of a labelled source domain while transferring them to an unlabeled target domain via backpropagation-based adversarial learning. An adaptive entropy max-min adversarial learning scheme is designed to optimally align these hundreds of local contextual-relations across domain without requiring discriminator or extra computation overhead. Thus, the process shown in FIG.1B provides local alignment 30B between the source domain and the target domain ensures incompatibility relations consistency 40B between image features 42B of the source domain and image features 44B of the target domain. It also ensures compatibility relation consistencies 50B between image features 52B of the source domain and image features 54B of the target domain.

FIG.2 is a block diagram showing a semantic image segmentation system according to an embodiment of the present invention. The semantic image segmentation system 100 is a computer system with memory that stores computer program modules which implement semantic image segmentation methods according to embodiments of the present invention.

The semantic image segmentation system 100 comprises a processor 110, a working memory 120, an input interface 130, an output interface 140, and program storage 150. The processor 110 may be implemented as one or more central processing unit (CPU) chips. The program storage 150 is a non-volatile storage device such as a hard disk drive which stores computer program modules. The computer program modules are loaded into the working memory 120 for execution by the processor 110. The input interface 130 is an interface which allows data including the source image, pixel label

data for the source image, and the target image to be received by the semantic image segmentation system 100. The output interface 140 is coupled to an output device which allows the labelled target image to be output. The output interface 140 may be coupled to display device or a printer, or to a system for processing and analyzing the labelled target image.

The program storage 150 stores a clustering segmentation module 152, a training module 154, a segmentation module 156, a feature extractor 162, a pixel scale classifier 164, a local image contextual relation classifier 166 and a domain scale classifier 168. The computer program modules cause the processor 110 to execute image segmentation processing which is described in more detail below. The program storage 150 may be referred to in some contexts as computer readable storage media and/or non-transitory computer readable media. As depicted in FIG.2, the computer program modules are distinct modules which perform respective functions implemented by the semantic image segmentation system 100. It will be appreciated that the boundaries between these modules are exemplary only, and that alternative embodiments may merge modules or impose an alternative decomposition of functionality of modules. For example, the modules discussed herein may be decomposed into sub-modules to be executed as multiple computer processes, and, optionally, on multiple computers. Moreover, alternative embodiments may combine multiple instances of a particular module or sub-module. It will also be appreciated that, while a software implementation of the computer program modules is described herein, these may alternatively be implemented as one or more hardware modules (such as field-programmable gate array(s) or application-specific integrated circuit(s)) comprising circuitry which implements equivalent functionality to that implemented in software.

The feature extractor 162, the pixel scale classifier 164, the local scale image contextual relation classifier 166 and the domain scale classifier 168 implement machine learning feature extractors and classifiers which are trained by the training module 154. The pixel scale classifier 164, the local scale image contextual relation classifier 166 and the domain scale classifier 168 are implemented as neural networks which when trained classify image features extracted by the feature extractor 162.

FIG.3 is a flow chart showing a semantic image segmentation method according to an embodiment of the present invention. The method 300 shown in FIG.3 is carried out by the semantic image segmentation system 100 shown in FIG.2.

- 5 In step 302, the input interface 130 of the semantic image segmentation system 100 receives the source image, pixel labels for the source image, and the target image.

In the present disclosure, we focus on the problem of unsupervised domain adaptation (UDA) in semantic segmentation. Given the source data $X_s \subset \mathbb{R}^{H \times W \times 3}$ with C-class pixel-scale segmentation labels $Y_s \subset (1, C)^{H \times W}$ (e.g., stimulated images from game engines) and the target data $X_t \subset \mathbb{R}^{H \times W \times 3}$ without labels (i.e., real images), our goal is to learn a semantic segmentation model G that performs well on the target dataset X_t . Current adversarial learning methods rely heavily on discriminators to align the distributions of source and target domains via two loss functions: segmentation loss
15 on source data and adversarial loss for alignment.

However, there exists a crucial limitation for these approaches: even if perfect adaptation is achieved through a discriminator, the alignment is implemented on global level (i.e., image-level), where local contextual information may be lost/deconstructed.
20 The reason lies in that the discriminator can only implement alignment at global level, which inputs the whole map but outputs a digit to represent domain labels (e.g., 0 or 1). In some cases, parts of local regions (e.g., local contextual relations) have been well aligned across domains. However, the discriminator might deconstruct this existing local alignment during implementing the global marginal distribution alignment.
25 In this disclosure, we denote phenomenon as lack of local consistency (i.e., local contextual inconsistency)", which is important to semantic segmentation in dense pixel-scale prediction.

In step 304, the training module 154 is executed by the processor to train the feature
30 extractor 162, the pixel scale classifier 164, the local scale image contextual relation classifier 166 and the domain scale classifier 168.

FIG.4 shows an overview of the training of the feature extractor, the pixel scale classifier, and the local scale image contextual relation classifier in an embodiment of the present invention.

- 5 As shown in FIG.4, the semantic segmentation model G consists of a feature extractor E 162 and two classifiers (i.e., C_{seg} 164 and C_{cr} 166) where C_{seg} is for pixel-scale segmentation and C_{cr} is for local-scale contextual-relations learning/classification.

10 The inputs to the semantic segmentation model are a source image 10, a target image 12 and a set of pixel labels 20 for the source image. Given images from source and target domains, the feature extractor extracts E features and feeds them to classifier C_{seg} and C_{cr} for classification at pixel and region scales. In the source flow 402, $\mathcal{L}_{seg}$ is computed based on the segmentation probability map from C_{seg} , $\mathcal{L}_{cr}$ is computed based on the classification probability maps from C_{cr} . In the target flow 404, $\mathcal{L}_{ent_{cr}}$ is
15 computed based on the classification probability maps from C_{cr} .

Backpropagation 406 is applied in the pixel level training. The local-scale alignment is implemented in back-propagation 408 by training the parts before and after the gradient reverse layer 410 in adversarial scheme w.r.t $\mathcal{L}_{ent_{cr}}$.

20

The feature extractor E extracts features from input images. C_{seg} and C_{cr} classify features generated by E into pre-defined semantic classes. Specifically, C_{seg} processes features at pixel-scale, which aims to predict pixel-scale labels. The pre-defined semantic class domain for, C_{seg} is the pixel-scale ground-truth, so there
25 is no difference between C_{seg} and traditional segmentation classifier. C_{cr} processes features at local scales, which aims to predict region-scale/contextual-relations labels. The pre-defined semantic class domain for C_{cr} is the clustered contextual-relations ground-truth.

30 The feature extractor E and the classifiers (i.e., C_{seg} and C_{cr}) are trained in an adversarial scheme to reduce domain shifts at local scales to achieve local contextual-relation consistency.

In order to implement local-scale task, we sample regions on the feature space and implement domain alignment at local scales to achieve local contextual-relation consistent domain adaptation, as shown in FIG.5A and FIG.5B. Different from system which implement mode-agnostic patches alignment or that align patch-indexed representation of the whole image only at global scales by a discriminator (i.e., the probability distributions of patch index prediction of the whole images.), we aim to align inter-class relations within each single patch, i.e., the probability distributions of pixel class prediction within each patch, w.r.t its mode via a classifier.

As shown in FIG.5A, the preliminary is to establish the region-scale label, where we first crop the pixel-scale ground-truth which corresponds to the pixel labels for the source image into many larger regions 504 and then use density-based spatial clustering of applications with noise (DBSCAN) to cluster them to assign each region a certain index label (i.e., contextual-relation pseudo label) 506.

As shown in FIG.5B, we assign the index label to regions according to the clustering results based on the histogram of gradient. For region-scale label (i.e., contextual-relation pseudo label), we cluster regions into different groups based on the histogram of gradient and assign the index label. These region-scale/contextual-relation pseudo labels can assist our network to implement alignment at local scales.

The clustering shown in FIG.5A is carried out by the clustering module 152 of the semantic image segmentation system shown in FIG.2.

In local-scale adaptation, C_{cr} aims to approximate the prototypical feature representations for each contextual-relation (e.g., road-sidewalk, sky-building, pole-sidewalk, etc.) by implementing entropy maximization in target domain according to the source prototypical feature representations found via supervised learning in source domain. E focuses on extracting discriminative feature representations (near the approximated prototypical feature representations) by implementing entropy minimization. Specifically, the prototypical feature representations of source domain found with supervision are first utilized to estimate the prototypical feature representations for target data by entropy maximizing w.r.t C_{cr} . E then adapts the

extracted feature representations to the corresponding prototypical feature representations by minimizing the entropy. The overall unsupervised domain adaptation at local scales is achieved by the adversarial training between C_{cr} and E as illustrated in FIG.6A to FIG.6E.

5

FIG.6A shows the mechanism of traditional global-scale domain adaptation, where some samples are adapted into the wrong area due to the lack of local consistency (i.e. local contextual-relation consistency).

10

Different from that applied in semi-supervised learning, our unsupervised domain adaptation training method, referred as adaptive entropy max-minimizing (AEMM) implements entropy max-min with a regularizer $\mathcal{R}(P) = \text{ave}\{P \log P\} \times \lambda_R$ (λ_R decreases with training iteration) for better estimating the prototypes in the target domain where no labels are available.

15

FIG.6B shows the various symbols used in FIG.6A and FIG.6C to FIG.6E

FIG.6C shows alignment on the pixel scale. FIG.6D shows alignment on the local scale and FIG.6E shows alignment on the global scale.

20

In pixel-scale alignment, C_{seg} firstly approximates the target prototypical features by maximizing entropy on target data and then E aims to congregate the features to the approximated prototypical features by minimizing entropy.

25

Local-scale alignment works in the same scheme of pixel-scale adaptation while the only difference is the processing unit size (the former adapts a larger group of features; the latter adapts single pixel-scale features). In our local-scale adaptation setting, the source data contributes to L_{seg} and L_{cr} . Given a source image $x_s \in X_s$, its corresponding segmentation label $y_s \in Y_s$ and contextual-relation pseudo-label $y_{s_cr} \in$

30

Y_{s_cr} , $P_s^{(h,w,c)} = C_{seg}(E(x_s))$ is the predicted probability map w.r.t each pixel over C classes; $P_{s_cr}^{(i,j,n)} = C_{cr}(E(x_s))$ is the predicted probability map w.r.t each region over N pre-defined contextual-relations classes. Therefore, it is a simple supervised learning objective to minimize L_{seg} and L_{cr} which are expressed as:

$$\mathcal{L}_{seg}(E, C_{seg}) = \sum_{h,w} \sum_c -y_s^{(h,w,c)} \log P_s^{(h,w,c)}$$

$$\mathcal{L}_{cr}(E, C_{cr}) = \sum_{i,j} \sum_n -y_{s_cr}^{(i,j,n)} \log P_{s_cr}^{(i,j,n)}$$

5

As the target label is not accessible, we introduce the adversarial training scheme between feature extractor E and classifier C_{cr} to extract discriminative features for target data via adaptively max-minimizing entropy in target domain. Given a target image $x_t \in X_t$, $P_{t_cr}^{(i,j,n)} = C_{cr}(E(x_t))$ is the predicted probability map w.r.t each region

10 over N pre-defined contextual-relations classes. The entropy loss $\mathcal{L}_{ent_cr}$ is expressed as:

$$\mathcal{L}_{ent_cr}(E, C_{cr}) = -\frac{1}{C} \sum_{i,j} \sum_n \max\{P_{t_cr}^{(i,j,n)} \log P_{t_cr}^{(i,j,n)} - \mathcal{R}(P_{t_cr}^{(i,j,n)}), 0\}$$

For local-scale adaptation, we use a back-propagation optimizing scheme with the gradient reverse layer which is described in [1]. The training objective can be

15 express as:

$$\begin{aligned} & \min_{\theta_E} \mathcal{L}_{seg} + \lambda_{cr} \mathcal{L}_{cr} + \lambda_{ent} \mathcal{L}_{ent_cr} \\ & \min_{\theta_{C_{seg}}} \mathcal{L}_{seg} \\ & \min_{\theta_{C_{cr}}} \mathcal{L}_{cr} - \lambda_{ent} \mathcal{L}_{ent_cr} \end{aligned}$$

20

where λ_{ent} is a weight factor to control the balance of unsupervised adaptation on target data and supervised learning on source data.

As shown in FIG.6E, the global alignment is implemented by a domain classifier.

25

In multi-scale adaptation, for $\mathcal{L}_{seg}$, $\mathcal{L}_{cr}$ and $\mathcal{L}_{ent_cr}$, the objectives are the same as that in local-scale adaptation. We extend the AEMM adversarial training scheme described above into pixel-scale and global-scale adaptation. For pixel-scale adaptation, we implement pixel-scale entropy loss $\mathcal{L}_{ent}$ on target data to E and C_{seg} . For global-scale

adaptation, we implement global-scale entropy loss $\mathcal{L}_{ent_D}$ on target data to E and C_D , where C_D is a domain classifier.

- 5 C_D takes the layout probability map concatenated by the two probability maps generated from C_{seg} and C_{cr} as input, and predicts domain label for it (e.g. 0 for source domain, 1 for target domain). The global-alignment is achieved by the adversarial training between C_D and (E, C_{seg}, C_{cr}) . Finally, our multi-scale consistent domain adaptation network is able to align domain shift at global scales, local-scale
10 and pixel-scale simultaneously.

Similar to local-scale adaptation, we formulate the pixel-scale entropy loss as:

$$\mathcal{L}_{ent_pix}(E, C_{seg}) = -\frac{1}{C} \sum_{h,w} \sum_c \max\{P_{t_pix}^{(h,w,c)} \log P_{t_pix}^{(h,w,c)} - \mathcal{R}(P_{t_pix}^{(h,w,c)}), 0\}$$

15

For multi-scale adaptation, we also use the same back-propagation optimizing scheme with the gradient reverse layer mentioned in [13] and [14]. The training objective can be express as:

$$\begin{aligned} 20 \quad & \min_{\theta_E} \mathcal{L}_{seg} + \lambda_{cr} \mathcal{L}_{C_{cr}} + \lambda_{ent} (\mathcal{L}_{ent_pix} + \mathcal{L}_{ent_cr}) + \lambda_D \mathcal{L}_D \\ & \min_{\theta_{C_{seg}}} \mathcal{L}_{seg} - \lambda_{ent} \mathcal{L}_{ent_pix} + \lambda_D \mathcal{L}_D \\ & \min_{\theta_{C_{cr}}} \mathcal{L}_{C_{cr}} - \lambda_{ent} \mathcal{L}_{ent_cr} + \lambda_D \mathcal{L}_D \\ & \max_{\theta_{C_D}} \lambda_D \mathcal{L}_D \end{aligned}$$

- 25 where $\mathcal{L}_D$ is given by the equation below:

$$\mathcal{L}_D(E, C_{seg}, C_{cr}, C_D) = \sum_{h,w} E[\log(C_D(P_{t_layout}^{(h,w,c+n)}))] + \sum_{h,w} E[\log(1 - C_D(P_{t_layout}^{(h,w,c+n)}))]$$

in which $P_{t_layout}^{(h,w,c+n)}$ is the predicted layout probability map and λ_{cr} , λ_{ent} and λ_D are the weight factors to balance the unsupervised adaptation on target data and the task-specific objectives on source data.

5

In some embodiments, multiple local scale classifiers may be implemented which perform adaptation on multiple image patch sizes. For the source domain, the sizes of the input image for datasets GTA5 and SYNTHIA are 720×1280 and 760×1280 ,
 10 respectively. In an example, two types of regions with two different sizes are used. The first sizes of regions for datasets GTA5 and SYNTHIA are 18×32 and 19×32 , respectively. The second sizes of regions for datasets GTA5 and SYNTHIA are 36×64 and 38×64 , respectively. For the target domain dataset Cityscapes, the size of input image is 512×1024 . The sizes of regions are 16×32 and 32×64 , respectively.
 15 Two independent contextual-relations (CR) classifiers are used to deal with these two types of regions with two different sizes. Here, two “contextual-relations (CR) classifiers” are used as the examples. Multiple “contextual-relations (CR) classifiers” could be used to cover multiple region/patch sizes ($1/40$, $1/20$, $1/10$, $1/5$, of the image size etc).

20

In the following description, a comparison of results obtained using prior art methods and results obtained using methods according to embodiments of the present invention are described.

25 We evaluate our unsupervised domain adaptation networks for semantic segmentation on two challenging synthesized-to-real tasks: GTA5 [2] $\rightarrow$ Cityscapes [3] and SYNTHIA [4] $\rightarrow$ Cityscapes. GTA5 contains 24,966 synthesized images with high-resolution and we use the 19 common categories between GTA5 and Cityscapes. SYNTHIA contains 9,400 synthetic images with 16 common categories in Cityscapes.
 30 We use either GTA5 or SYNTHIA as source domain. We use the unlabeled training set of Cityscapes as target domain, which includes 2975 real-world images.

For a fair comparison, we utilize Deeplab-V2 architecture [5] with ResNet-101 pretrained on ImageNet [6] as our single-scale semantic segmentation network ($E + C_{seg}$). To extend our model to multi-scale network, we simply copy and modify C_{seg} to create C_{cr} and C_D with different output channels (e.g. N and 1) and different output sizes due to various scales (i.e. region-size and global-size). We also apply our methods on VGG-16 [7] in the same way as employing ResNet-101. A gradient reverse layer is employed to reverse the entropy loss between E and (C_{seg}, C_{cr}) during pixel-/region-scale adaptation to achieve adversarial training. The domain classifier C_D works similar to a discriminator for global-scale alignment. During training, we utilize stochastic gradient descent to optimize our networks with a momentum of 0.9 and a weight decay of $1e - 4$. The initial learning rate is set as $2.5e - 4$ and decayed by a polynomial policy with a power of 0.9 . For all experiments, the hyper-parameters λ_{ent} , λ_D , λ_{cr} and N are set as $2.5e - 5$, $2.5e - 5$, $5e - 3$ and 100 , respectively.

FIG.7 is a table showing results of domain adaptation task $GTA5 \rightarrow Cityscapes$. “V” means the VGG16-based model and “R” means the ResNet101-based model. We compare the experimental results of our method and state-of-the-art algorithms in two “Synthetic-to-real” UDA tasks with two different architectures: VGG-16 and ResNet-101. For “ $GTA5 \rightarrow Cityscapes$ ”, we present the results in FIG.7 with comparisons to the state-of-the-art domain adaptation methods. The following domain adaptation methods are compared with the proposed method: Fully convolutional networks (FCNs Wild) – [8], Curriculum domain adaptation (CDA) – [9], Adapted Structured Output Space for Semantic Segmentation (AdaptSeg) – [10], class-balanced self-training (CBST) – [11], Category-level Adversaries for Semantics Consistent Domain Adaptation (CLAN) – [12], Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation (AdvEnt) – [13], Domain adaptation for structured output via discriminative patch representations (PatAlign) – [14], and Maximum Squares Loss (MaxSquare) – [15].

Our contextual-relation consistent domain adaptation, expressed as CrCDA, achieves comparable performance to other state-of-the-art approaches on both architectures. Compared to Adapt-SegMap (output space global alignment), category-level adversarial network (output space class-wise alignment) and patch-represented global

alignment (patch-indexed latent space alignment), CrCDA consistently brings over +2.1% mIoU improvements on ResNet-101. It is believed that this gain is from our end-to-end/concurrent multi-scale alignment, which indicates that local consistency (i.e., local contextual-relation consistency) is very important as well as global consistency and they are complementary to each other.

FIG.8 is a table showing results of results of domain adaptation task SYNTHIA $\rightarrow$ Cityscapes. “V” means the VGG16-based model and “R” means the ResNet101-based model. “mIoU” and “mIoU*” are calculated over 16 and 13 classes, respectively. As shown in FIG.8, the adaptation result for the task “SYNTHIA $\rightarrow$ Cityscapes” show consistent improvements w.r.t existing methods. Detailed analysis is included the following paragraphs.

We analyze our proposed CrCDA with several state-of-the-art baselines. In general, both single-scale form (CrCDA*) and multi-scale form (CrCDA) achieve comparable results to all the baselines in all the settings.

FIG.9 is a table showing results of ablation study of the proposed contextual-relation consistent domain adaptation on GTA5-to-Cityscapes using the ResNet-101 network. All settings/methods are with “ L_{seg} ” (bold texts represent our methods). CrCDA* represents the contextual-relation consistent domain adaptation with only single-scale (local scale). As shown on the first three rows in the table of FIG.9, our pixel-scale AEMM adversarial network brings +1.4% improvements in terms of mIoU over MinEnt. The reason lies in that direct entropy minimization does not take the domain gap into account while our AEMM training scheme pushes the source distribution closer to target distribution during maximizing entropy on target data.

For our CrCDA with single-scale form (CrCDA*) via AEMM, it outperforms MinEnt-based contextual-relations alignment by +1.6% on ResNet-101, as shown on the second row of block 910 in FIG.9. It is believed that these improvements are contributed by our adaptive entropy max-min training scheme which considers the domain mismatch/gap while MinEnt neglects.

Our CrCDA with multi-scale form integrating three scales' adaptation (pixel-, local- and global-scale), termed as CrCDA shown on the last row of block 920 in FIG.9, achieves state-of-the-art performances 48.6% mIoU on ResNet-101. Besides, CrCDA also outperforms all current methods by over +1.5%. Compared to "Pixel+Global", CrCDA brings +2.6% improvement in mIoU, which demonstrates that local-scale alignment is essential as well as other scales (e.g., pixel-scale and global-scale). In fact, the local contextual-relation consistent adaptation loss (i.e. L_{ent_cr}) penalize groups of pixels predictions to achieve local-scale alignment, where global-scale adaptation loss operates more on image-scale (e.g. scene layout) while that of pixel-scale works on the feature representation alignment of each independent pixels. The consistent results with different settings further confirm that complementary information has been learned in different scales' adaptation.

FIG.10 shows a complementary study of the proposed contextual-relation consistent domain adaptation with local-scale to current global alignment UDA methods on GTA5-to-Cityscapes using the ResNet-101 network. All methods are default with " L_{seg} ". As shown in FIG.9, our local contextual-relations alignment method is complementary to most existing global-scale alignment approaches.

FIG.11 is a set of images showing a comparison of image segmentation using methods according to embodiments of the present invention with existing methods. As can be seen from FIG.11, our approach (CrCDA) aligns low-level features (e.g. boundaries of sidewalk, car and person etc.) as well as high-level features by multi-scale adversarial learning. In contrast, AdvEnt ignores low-level information because global alignment focuses more on high-level information. Thus, as shown above, CrCDA achieves both local and global consistencies while AdvEnt only achieves global consistency.

As described above, the local contextual-relation consistent domain adaptation (CrCDA) to address the task of unsupervised domain adaptation for semantic segmentation. By taking a closer look at the local inconsistency (i.e. local contextual-relations inconsistency) while implementing global adaptation, CrCDA is able to align the domain shift in local and global scales at the same time, where local semantic

consistency is normally ignored by current approaches. The experimental results on the two challenging segmentation UDA tasks validate the state-of-the-art of CrCDA.

Embodiments of the present invention may be applied to image processing in a variety of scenarios such as facial segmentation, control of robots, fashion categorization, biomedical image analysis, autonomous driving, satellite image processing, geo-sensing/land usage analysis and precision agriculture.

Whilst the foregoing description has described exemplary embodiments, it will be understood by those skilled in the art that many variations of the embodiments can be made within the scope and spirit of the present invention.

REFERENCES

[1] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881-2890 (2017).

[2] Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: European conference on computer vision. pp. 102-118. Springer (2016).

[3] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213-3223 (2016).

[4] Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3234-3243 (2016).

- [5] Chen, Q., Liu, Y., Wang, Z., Wassell, I., Chetty, K.: Re-weighted adversarial adaptation network for unsupervised domain adaptation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018).
- 5 [6] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248-255. IEEE (2009).
- [7] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale
10 image recognition. arXiv preprint arXiv:1409.1556 (2014).
- [8] Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016).
- 15 [9] Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2020-2030 (2017).
- [10] Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.:
20 Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7472-7481 (2018).
- [11] Zou, Y., Yu, Z., Vijaya Kumar, B., Wang, J.: Unsupervised domain adaptation
25 for semantic segmentation via class-balanced self-training. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 289-305 (2018).
- [12] Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y.: Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In:
30 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2507-2516 (2019).

[13] Vu, T.H., Jain, H., Bucher, M., Cord, M., Perez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2517-2526 (2019).

5 [14] Tsai, Y.H., Sohn, K., Schuler, S., Chandraker, M.: Domain adaptation for structured output via discriminative patch representations. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1456-1465 (2019).

[15] Chen, M., Xue, H., Cai, D.: Domain adaptation for semantic segmentation with
10 maximum squares loss. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2090-2099 (2019).

CLAIMS

1. A semantic image segmentation method of segmenting a target image, the method comprising:

receiving a source image and pixel labels for the source image, the pixel labels for the source image indicating, for each respective pixel of the source image, a class from a plurality of classes;

training a feature extractor to extract image features from the source image and from the target image, training a pixel scale classifier to classify pixel scale features of the source image and the target image and training a local scale contextual relation classifier to classify local scale image contextual relation features of the source image and the target image and thereby obtaining a trained feature extractor, a trained pixel scale classifier and a trained local scale contextual relation classifier, wherein training the feature extractor, training the pixel scale classifier and training the local scale classifier is carried out while aligning domain shift on a pixel scale, a local scale and a global scale between a domain of the source image and a domain of the target image; and

using the trained feature extractor, the trained pixel scale classifier and the trained local scale contextual relation classifier to generate pixel labels for the target image.

2. The semantic image segmentation method of claim 1, wherein training the local scale contextual relation classifier to classify local scale image contextual relation features of the source image and the target image comprises implementing an adversarial training scheme between the local scale contextual relation classifier and the feature extractor.

3. The semantic image segmentation method of claim 1 or claim 2, wherein training the pixel scale classifier to classify pixel scale features of the source image and the target image comprises implementing an adversarial training scheme between the pixel scale classifier and the feature extractor.

4. The semantic image segmentation method of any one of the preceding claims, further comprising training a domain scale classifier to predict a domain label for a

layout probability map generated from probability maps generated by the pixel scale classifier and the local scale contextual relationship classifier.

5. The semantic image segmentation method of claim 4, wherein training the domain scale classifier to predict a domain label for a layout probability map generated from probability maps generated by the pixel scale classifier and the local scale contextual relationship classifier comprises implementing an adversarial training scheme between the domain scale classifier and the feature extractor.

6. The semantic image segmentation method of any one of claims 2 to 5, wherein the adversarial training scheme between the local scale contextual relation classifier and the feature extractor, and / or the adversarial training scheme between the pixel scale classifier and the feature extractor, and / or the adversarial training scheme between the domain scale classifier and the feature extractor are adaptive entropy max-minimizing learning schemes.

7. The semantic image segmentation method of any one of the preceding claims, wherein training the local scale contextual relation classifier to classify local scale image contextual relation features of the source image and the target image further comprises backpropagating using a reverse gradient layer.

8. The semantic image segmentation method of any one of the preceding claims, wherein training the local scale contextual relation classifier to classify local scale image contextual relation features of the source image and the target image further comprises sampling regions of the source image and clustering the sampled regions of the source image to generate prototypical features in the source domain.

9. The semantic image segmentation method of claim 8, wherein clustering the sampled regions of the source image to generate prototypical features in the source domain comprises clustering using a density-based spatial clustering of applications with noise algorithm.

10. A computer readable medium storing processor executable instructions which when executed on a processor cause the processor to carry out a method according to any one of claims 1 to 9.

11. A semantic image segmentation system for segmenting a target image, the system comprising a processor and a data storage device storing a feature extractor, a pixel scale classifier, a local scale contextual relation classifier, and computer program instructions operable to cause the processor to:

receive a source image and pixel labels for the source image, the pixel labels for the source image indicating, for each respective pixel of the source image, a class from a plurality of classes;

train the feature extractor to extract image features from the source image and from the target image, train the pixel scale classifier to classify pixel scale features of the source image and the target image and train the local scale contextual relation classifier to classify local scale image contextual relation features of the source image and the target image and thereby obtaining a trained feature extractor, a trained pixel scale classifier and a trained local scale contextual relation classifier, wherein training the feature extractor, training the pixel scale classifier and training the local scale classifier is carried out while aligning domain shift on a pixel scale, a local scale and a global scale between a domain of the source image and a domain of the target image; and

use the trained feature extractor, the trained pixel scale classifier and the trained local scale contextual relation classifier to generate pixel labels for the target image.

12. The semantic image segmentation system of claim 11, wherein the computer program instructions are further operable to cause the processor to train the local scale contextual relation classifier to classify local scale image contextual relation features of the source image and the target image by implementing an adversarial training scheme between the local scale contextual relation classifier and the feature extractor.

13. The semantic image segmentation system of claim 11 or claim 12, wherein the computer program instructions are further operable to cause the processor to train the pixel scale classifier to classify pixel scale features of the source image and the target

image comprises by implementing an adversarial training scheme between the pixel scale classifier and the feature extractor.

14. The semantic image segmentation system of any one of claims 11 to 13, wherein the data storage device further stores a domain scale classifier and the computer program instructions are further operable to cause the processor to train the domain scale classifier to predict a domain label for a layout probability map generated from probability maps generated by the pixel scale classifier and the local scale contextual relationship classifier.

15. The semantic image segmentation system of claim 14, wherein the computer program instructions are further operable to cause the processor to train the domain scale classifier to predict a domain label for a layout probability map generated from probability maps generated by the pixel scale classifier and the local scale contextual relationship classifier by implementing an adversarial training scheme between the domain scale classifier and the feature extractor.

16. The semantic image segmentation system of any one of claims 12 to 15, wherein the adversarial training scheme between the local scale contextual relation classifier and the feature extractor, and / or the adversarial training scheme between the pixel scale classifier and the feature extractor, and / or the adversarial training scheme between the domain scale classifier and the feature extractor are adaptive entropy max-minimizing learning schemes.

17. The semantic image segmentation system of any one of claims 11 to 16, wherein the computer program instructions are further operable to cause the processor to train the local scale contextual relation classifier to classify local scale image contextual relation features of the source image and the target image by backpropagating using a reverse gradient layer.

18. The semantic image segmentation system of any one of claims 11 to 17, wherein the computer program instructions are further operable to cause the processor to train the local scale contextual relation classifier to classify local scale image contextual relation features of the source image and the target image by

sampling regions of the source image and clustering the sampled regions of the source image to generate prototypical features in the source domain.

19. The semantic image segmentation system of claim 18, wherein the computer program instructions are further operable to cause the processor to cluster the sampled regions of the source image to generate prototypical features in the source domain by clustering using a density-based spatial clustering of applications with noise algorithm.