Semantic Nighttime Image Segmentation with Synthetic Stylized Data, Gradual Adaptation and Uncertainty-Aware Evaluation

Christos Sakaridis[1], Dengxin Dai[1], and Luc Van Gool[1,2]

[1]ETH Zürich, [2]KU Leuven

Abstract

This work addresses the problem of semantic segmentation of nighttime images. The main direction of recent progress in semantic segmentation pertains to daytime scenes with favorable illumination conditions. We focus on improving the performance of state-of-the-art methods on the nighttime domain by adapting them to nighttime data without extra annotations, and designing a new evaluation framework to address the uncertainty *of semantics in nighttime images. To this end, we make the following contributions: 1) a novel pipeline for dataset-scale guided style transfer to generate synthetic nighttime images from real daytime input; 2) a framework to gradually adapt semantic segmentation models from day to night via stylized and real images of progressively increasing darkness; 3) a novel uncertainty-aware annotation and evaluation framework and metric for semantic segmentation in adverse conditions; 4) the* Dark Zurich *dataset with 2416 nighttime and 2920 twilight unlabeled images plus 20 nighttime images with pixel-level annotations that conform to our newly-proposed evaluation. Our experiments evidence that both our stylized data per se and our gradual adaptation significantly boost performance at nighttime both for standard evaluation metrics and our metric. Moreover, our new evaluation reveals that state-of-the-art segmentation models output overly confident predictions at indiscernible regions compared to visible ones.*

1. Introduction

The last years have witnessed considerable progress in semantic segmentation and the state of the art is constantly improving. Despite the advance, most methods are designed to operate at daytime, under favorable illumination conditions. However, many outdoor applications require robust vision systems that perform well at all times of day, under challenging lighting conditions, and in bad weather [21]. In this work, we investigate semantic segmentation of nighttime scenes, both at the method and the evaluation level.

Recognition at nighttime poses further challenges compared to daytime. The features used by recognition systems at daytime are highly corrupted due to visual hazards [37] such as over-/underexposure, noise, and motion blur. Degradation of the input is often so intense that recognizing its semantic content is difficult or even impossible for humans, let alone vision algorithms developed for daytime. At the evaluation level, we recognize this challenge and design an uncertainty-aware annotation and evaluation framework for semantic segmentation that takes into account the potential existence of indiscernible regions in the input image. The highlight of this framework is the new uncertainty-aware intersection-over-union (UIoU) metric.

The current popular approach to solving high-level tasks such as semantic segmentation is to train deep neural networks [17, 34, 38] using large-scale human annotations [4, 6, 22]. This supervised scheme has achieved great success for daytime images, but it scales badly over all possible adverse conditions. At the method level, this work instead adapts semantic segmentation models trained on daytime to nighttime without annotations in the latter domain.

To this end, we first present a novel, dataset-scale style transfer pipeline to stylize labeled real images of daytime scenes as nighttime images, which are directly usable as training data for nighttime semantic segmentation. Rather than constituting a new method for style transfer, our pipeline features a principled reference selection mechanism to scale standard style transfer approaches that require paired images to entire datasets, as well as a guidance mechanism that uses an auxiliary daytime reference set, which combines the style of the input set with the content of the main reference set, to assist style transfer.

Secondly, we adapt segmentation models from daytime to nighttime gradually by using 1) labeled stylized twilight and nighttime images which are generated with our aforementioned pipeline; and 2) unlabeled real images captured at daytime and twilight and acting as an intermediate stage for supervision transfer.

Finally, we present *Dark Zurich*, a dataset of real images which contains multiple versions of the same driving scenes corresponding to daytime, twilight and nighttime. We use this dataset to apply our guided style transfer pipeline to Cityscapes [4], to feed real data to our gradual adaptation and to create a prototype test set of 20 images for our uncertainty-aware evaluation.

2. Related Work

We present the related work in four topics.

Vision at Nighttime. Due to the ubiquitous nature of this domain, nighttime scenes have attracted a lot of attention in the literature. Several works pertain to human detection at nighttime, by using FIR cameras [8, 33], visible light cameras [13], or a combination of both [2]. In driving scenarios, a few methods have been proposed to detect cars [15] and vehicles' rear lights [27]. Contrary to these domain-specific methods, previous work also includes methods designed for robustness to illumination changes, either by employing domain-invariant representations [1, 23] or by fusing information from complementary modalities and spectra [31]. Most of the aforementioned research has been conducted before the wide deployment of deep neural networks.

A very recent work [5] on semantic nighttime image segmentation shows that real images captured at twilight are helpful for supervision transfer from daytime to nighttime. This work is partially inspired by [5] and extends it by proposing a gradual adaptation framework which learns *jointly* from stylized images and unlabeled real images of increasing darkness.

Domain Adaptation and Transfer Learning. Performance of semantic segmentation on daytime scenes has undergone a rapid increase in recent years due to the learning capacity of deep neural networks and the availability of large-scale annotated datasets. How to extend the relevant methods to adverse conditions is becoming more interesting. A recent effort has been made for model adaptation from clear weather to fog [24, 25] by using both labeled partially synthetic foggy images and unlabeled real foggy images of increasing fog density. This work moves in a similar research direction but for the nighttime domain, which poses different challenges than the foggy domain of [24] for formulating our gradual adaptation framework.

Our work bears resemblance to works from the broad field of transfer learning, particularly those on semantic segmentation. Recently, domain adaptation has been widely used to adapt semantic segmentation models from synthetic images to real environments [3, 10, 25, 26, 30]. Most of these works are based on adversarial domain confusion, which is further extended in [32] by proposing incremental adversarial domain adaptation to utilize the continuity of the domain shift. Our work is complementary to this vein, as we adapt the model parameters with carefully selected data instead of modifying the overall model architecture.

Style Transfer and Image Translation. Automatic data-driven manipulation of the apparent time of day had been addressed before the advent of deep neural networks [28]. However, it was the seminal works of Gatys *et al*. [7] on style transfer with deep features and Isola *et al*. [11] on image translation with GAN that initiated intense research on these two topics, both of which are applied to transfer of the time of day. Recent style transfer methods including [16, 19] enjoy improved photorealism and preserve the content of the input image in the output, but they require a reference image as additional input to dictate the style. As a consequence, these works only present results on a small number of selected image pairs and have not been scaled to large datasets yet. On the other hand, state-of-the-art image translation techniques such as [18, 39] lift the requirement of paired samples from the two domains and have thus been applied at dataset scale, enabling evaluation of their output with respect to its utility for domain adaptation. However, the lack of paired samples combined with the generative structure of these models often leads to hallucinated content in the translated image. In this paper, we focus on the traditional style transfer setting with paired images and propose a novel pipeline to *scale* it to large datasets.

Semantic Segmentation Evaluation. Semantic segmentation evaluation is commonly performed with the IoU metric [6]. Cityscapes [4] introduced an instance-level IoU (iIoU) metric to remove the large-instance bias, as well as mean average precision for the task of semantic instance segmentation. The two tasks have recently been unified into panoptic segmentation [14], with a respective panoptic quality metric. The most closely related work to ours in this regard is WildDash [36], which uses standard IoU together with a fine-grained evaluation to measure the impact of visual hazards on performance. On the contrary, we introduce UIoU, a new metric that handles images with regions of uncertain semantic content and is suited for benchmarks characterized by adverse conditions.

3. Semantic Nighttime Image Segmentation

Our proposed approach for semantic segmentation of nighttime images comprises two main components: a guided style transfer pipeline for generation of partially synthetic nighttime images from real daytime inputs and a gradual adaptation framework that jointly leverages these synthetic images as well as real images of increasing darkness.

3.1. Nighttime Image Synthesis from Real Daytime Scenes with Guided Style Transfer

We consider the case of two image datasets, an input dataset $\mathcal{I} = \{I_i : i = 1, \dots, M\}$ and a reference dataset

$\mathcal{R} = \{R_j : j = 1, \ldots, N\}$. Contrary to the usual style transfer application where we are given a fixed pair of images serving as the content image and the reference style image, we aim to perform style transfer at a dataset scale, so that we generate a stylized version $\mathcal{I}^s$ of the input dataset $\mathcal{I}$ that is characterized by the style of the reference dataset $\mathcal{R}$. However, due to the requirement of the standard style transfer setting for paired data, $\mathcal{I}^s$ is not uniquely defined from $\mathcal{I}$ and $\mathcal{R}$. In particular, denoting style transfer as a transformation T, its output for the input image I_i depends on the image R_j that is selected as reference:

$$I_i^{(j)} = T(I_i, R_j). \tag{1}$$

3.1.1 Feature-Based Reference Selection

In order to define $\mathcal{I}^s$ uniquely, we need to determine an assignment $A : \{1, \ldots, M\} \to \{1, \ldots, N\}$ which links each input image I_i to a unique reference image $R_{A(i)}$ that will dictate the style of the respective output. We name this process *reference selection*. A simple, data-agnostic baseline for reference selection is random assignment, e.g. via random permutation of the reference set, which ensures that every reference image is represented equally in the output. We instead propose to select the reference image by comparing its high-level features with those of the input. As style needs to be transferred across regions of the two images with similar semantic content for the output to be realistic, we expect that using pairs of images with similar high-level features will result in better outputs. More formally, denote by F an extractor of dense high-level features and by $\rho(x, y) \in [0, 1]$ a similarity measure. Feature-based reference selection is defined through the assignment

$$A(i) = \arg \max_{j \in \{1, \ldots, N\}} \{\rho(F(I_i), F(R_j))\}. \tag{2}$$

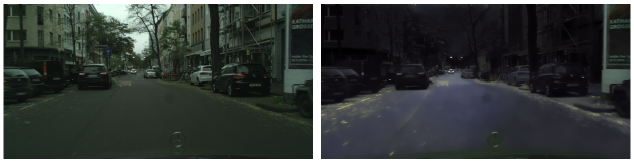
We instantiate the similarity measure ρ with the contextual similarity (CX) proposed in [20], since it can match non-aligned feature maps, which fits to our setting where $\mathcal{I}$ and $\mathcal{R}$ depict different scenes.

3.1.2 Source Style Guidance

In practice, extracting the features $F(R_j)$ from the reference image is challenging due to the adversity of the domain that corresponds to the reference style, e.g. in the nighttime case, and the features may not be representative. The same problem applies to the case of segmentation-assisted style transfer, where the semantic segmentations $H(I_i)$ and $H(R_j)$ of the images serve as additional inputs:

$$I_i^{(j)} = T(I_i, H(I_i), R_j, H(R_j)). \tag{3}$$

In this case, ground truth for $H(R_j)$ is generally not available and predicting it is difficult, again due to the lack of suitable features.



(a) Cityscapes (b) Dark Cityscapes

Figure 1. Example synthetic nighttime image from Dark Cityscapes, generated using daytime guidance and contextual similarity for reference selection.

We propose to use an auxiliary reference set $\mathcal{R}' = \{R'_k : k = 1, \ldots, K\}$ which depicts the same set of scenes as $\mathcal{R}$ (with some variation due to dynamic content) but is characterized by the same source style as the input set $\mathcal{I}$. This auxiliary set acts as a bridge between $\mathcal{I}$ and $\mathcal{R}$ as we detail subsequently. A known assignment $B : \{1, \ldots, K\} \to \{1, \ldots, N\}$ of images in $\mathcal{R}'$ to images in $\mathcal{R}$ is assumed, e.g. through localization ground truth, so that R'_k depicts roughly the same scene as $R_{B(k)}$. We use feature-based reference selection between $\mathcal{I}$ and $\mathcal{R}'$ similarly to (2) to define the assignment $A' : \{1, \ldots, M\} \to \{1, \ldots, K\}$. In this way, the features $F(R'_k)$ which are used as surrogate for $F\left(R_{B(k)}\right)$ are more suitable for comparison to $F(I_i)$. The assignment A is then simply the composition $A = B \circ A'$.

In addition, for the segmentation-assisted case, we propose to predict the semantic segmentations of the auxiliary set $\mathcal{R}'$ and use them in style transfer as surrogate for the respective segmentations of $\mathcal{R}$, so that (3) is modified to

$$I_i^{(A(i))} = T(I_i, H(I_i), R_{A(i)}, H(R'_{A'(i)})). \tag{4}$$

3.1.3 Dark Cityscapes

We apply our dataset-scale style transfer pipeline using the Cityscapes dataset [4] as the input set $\mathcal{I}$ and generate Dark Cityscapes. In particular, we use the training split of our *Dark Zurich* dataset (cf. Sec. 5 for details) to form the reference sets $\mathcal{R}$ and $\mathcal{R}'$. In this setting, $\mathcal{R}'$ contains daytime images to match the source style of Cityscapes. We select the state-of-the-art closed-form solution to photorealistic stylization [16] as the style transfer method of choice and use VGG-19 [29] conv4_2 features for reference selection. We employ the segmentation-assisted variant (3) of [16] and use RefineNet [17] to predict the segmentations for *Dark Zurich* as well as ground-truth annotations for Cityscapes. Style transfer is performed at full resolution. Dark Cityscapes inherits the semantic annotations of the original Cityscapes due to identical image content and can thus be leveraged to train semantic segmentation models for nighttime. The style of the reference set $\mathcal{R}$ determines the style of Dark Cityscapes, so we can generate multiple versions of the latter with different times of day apart from night (e.g. twilight), which is essential for gradual adaptation to nighttime as presented in Sec. 3.2. An example image from Dark Cityscapes with nighttime style is shown in Fig. 1.

3.2. Gradual Adaptation with Synthetic and Real Data

Using solely stylized data for adapting segmentation models to nighttime is prone to the discrepancy between the appearance of the synthesized images and the target domain of real nighttime images. Instead, we propose to learn jointly from our stylized data and unlabeled real data corresponding to a brighter time of day than the target time. We are given a semantic segmentation model ϕ_1 tailored for the time of day T_1 and aim to adapt it to T_2, which is darker than T_1. Our approach uses an unlabeled set of real images $\mathcal{D}_1^r$ captured at T_1 and a labeled set of synthetic images $\mathcal{D}_2^s$ with the style of T_2. We first use ϕ_1 to estimate the labels of $\mathcal{D}_1^r$ and use these labels as surrogate for the ground-truth annotations, resulting in a weakly labeled dataset $\bar{\mathcal{D}}_1^r$. We then learn a semantic segmentation model ϕ_2 for T_2 by initializing it with ϕ_1 and optimizing jointly on $\mathcal{D}_2^s$ and $\bar{\mathcal{D}}_1^r$. We use a hyper-parameter μ to weigh the contribution of $\bar{\mathcal{D}}_1^r$ to the training loss compared to that of $\mathcal{D}_2^s$. The motivation for our joint adaptation is twofold. First, T_1 is an easier domain for inferring the semantics of unlabeled images than T_2 because of its more favorable illumination and the de facto larger availability of annotated data for brighter times of day. Second, synthetic images from $\mathcal{D}_2^s$ have similar style to real images captured at T_2 but contain unrealistic artifacts, whereas real images from $\mathcal{D}_1^r$ have a different style than T_2 but are characterized by real, artifact-free textures. We iterate the above learning process for increasingly darker times of day to gradually adapt the input model to nighttime.

4. Uncertainty-Aware Evaluation

4.1. Motivation

When an image is captured under adverse conditions, such as at nighttime, there often exist large regions of the image on which recognition of the semantic class of the corresponding scene content is actually impossible, even for an experienced human subject. This is caused by visual hazards [37] that render the content of the affected regions *indiscernible*. We term such regions as *invalid* for the task of semantic segmentation. A robust model should generate predictions that are characterized by high *uncertainty* for invalid regions. Invalid regions are closely related to the concept of negative test cases that was considered in [36]. However, invalid regions constitute intra-image entities and can co-exist with valid regions in the same image, whereas a negative test case in [36] refers to an entire image as a sample that should be treated as invalid. We design an evaluation framework which incorporates invalid regions in the ground truth and uses the uncertainty (equivalently confidence) related to the evaluated predictions to identify the pixels which are predicted as *invalid*.

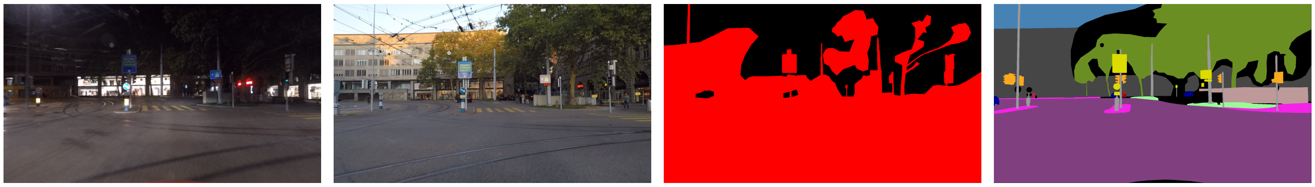
To formulate our novel uncertainty-aware evaluation protocol and metric for semantic segmentation, we build upon the evaluation of negative test cases that is proposed in [36] and generalize it so that it can be applied uniformly to all images in the evaluation set, whether they contain invalid regions or not. More specifically, we adopt the principle from [36] that for invalid regions of an image that are assigned a legitimate semantic label in the ground-truth annotation (e.g. with the aid of a corresponding hazard-free image of the same scene), both this label and the *invalid* label are considered as correct predictions. However, this principle only refers to true invalid regions and does not specify how evaluation is performed on valid (discernible) regions. We introduce the novel concept of false invalid predictions for valid regions which are incorrectly assigned the *invalid* label, and incorporate it naturally in our new uncertainty-aware evaluation metric. In the following, we elaborate on the generation of suitable ground-truth annotations and model predictions for our evaluation framework and present our proposed uncertainty-aware IoU metric.

4.2. Annotation Protocol

For each image I, the annotation process involves two main steps: 1) creation of the ground-truth invalid mask J, and 2) creation of the ground-truth semantic labeling H.

For the semantic labels, we consider a predefined set $\mathcal{C}$ of C classes, which in our case is equal to the set of Cityscapes [4] evaluation classes ($C = 19$). The annotator is first presented only with I and is asked to mark the invalid regions in it as the regions which she cannot unquestionably assign to one of the C classes nor declare as not belonging to any of the C classes. The result of this annotation is a binary mask J, which is set to 1 at invalid pixels and 0 everywhere else.

Secondly, the annotator is asked to mark the semantic labels of I, only that this time she also has access to an *auxiliary* image I'. This latter image has been captured with roughly the same 6D camera pose as I but under more favorable conditions. In our dataset, I' is captured at daytime whereas I is captured at nighttime. The large overlap of static scene content between the two images allows the annotator to label certain regions in H with a legitimate semantic label from $\mathcal{C}$, even though the same regions have been annotated as invalid (and are kept as such) in J. This allows joint evaluation on valid and invalid regions, as it creates regions which can accept both the *invalid* label and the ground-truth semantic label as correct predictions. Due to the imperfect match in the camera poses between I and I', the labeling of invalid regions is done conservatively, marking a coarse boundary which may leave unlabeled zones



(a) Input image I (b) Auxiliary image I' (c) GT invalid mask J (d) GT semantic labeling H

Figure 2. Example input images from *Dark Zurich-test* and output annotations with our protocol. Invalid pixels in J are marked black.

around the true semantic boundaries in I, so that no pixel is assigned a wrong label. The parts of I which remain indiscernible even after inspection of I' are left unlabeled in H. These parts as well as instances of classes outside $\mathcal{C}$ are not considered during evaluation. We illustrate a visual example of our annotation inputs and outputs in Fig. 2.

4.3. Uncertainty-Aware Predictions

The semantic segmentation output that is fed to our evaluation is expected to include pixels labeled as *invalid*. Instead of defining a separate, explicit *invalid* class, which would potentially require the creation of new training data to incorporate this class, we take a more flexible approach, using a *confidence threshold* to invalidate predictions.

In particular, we assume that the evaluated method outputs an intermediate soft prediction $\mathbf{s}(\mathbf{p})$ at each pixel $\mathbf{p}$ as a probability distribution among the C classes, which is subsequently converted to a hard assignment for standard evaluation by outputting the class $\tilde{H}(\mathbf{p}) = \arg\max_{c\in\mathcal{C}}\{s_c(\mathbf{p})\}$ with the highest probability. In this case, $s_{\tilde{H}(\mathbf{p})}(\mathbf{p}) \in [1/C,\, 1]$ is the effective confidence associated with the prediction. This assumption is not very restrictive, as most recent semantic segmentation methods are based on CNNs with a softmax layer that outputs such soft predictions.

The final evaluated output $\hat{H}$ is computed based on a free parameter $\theta \in [1/C,\, 1]$ which acts as a confidence threshold by invalidating those pixels where the confidence of the prediction is lower than θ:

$$\hat{H}(\mathbf{p}) = \begin{cases} \tilde{H}(\mathbf{p}) & \text{if } s_{\tilde{H}(\mathbf{p})}(\mathbf{p}) \geq \theta, \\ \textit{invalid} & \text{otherwise.} \end{cases} \tag{5}$$

Increasing θ results in more pixels being predicted as *invalid*. This approach relies on the fact that ground-truth invalid regions are identified by the uncertainty of their semantic content even for experienced human annotators, which implies that a model should place low confidence (equivalently high uncertainty) in predictions on such regions. If the model correctly does so, it is rewarded for the respective true invalid predictions by our evaluation metric.

4.4. Uncertainty-Aware IoU

We propose uncertainty-aware IoU (UIoU) as a generalization of the standard IoU metric for evaluation of semantic segmentation results which may contain pixels labeled as invalid. UIoU reduces to standard IoU if no pixel is predicted to be invalid.

The calculation of UIoU for class c involves five sets of pixels, which are listed along with their symbols: true positives (TP), false positives (FP), false negatives (FN), true invalids (TI), and false invalids (FI). Based on the ground-truth invalid masks J, the ground-truth semantic labelings H and the predicted labels $\hat{H}$ for the set of evaluation images, these five sets are defined as follows:

$$\text{TP} = \{\mathbf{p} : H(\mathbf{p}) = \hat{H}(\mathbf{p}) = c\}, \tag{6}$$

$$\text{FP} = \{\mathbf{p} : H(\mathbf{p}) \neq c \text{ and } \hat{H}(\mathbf{p}) = c\}, \tag{7}$$

$$\text{FN} = \{\mathbf{p} : H(\mathbf{p}) = c \text{ and } \hat{H}(\mathbf{p}) \notin \{c,\, \textit{invalid}\}\}, \tag{8}$$

$$\text{TI} = \{\mathbf{p} : H(\mathbf{p}) = c \text{ and } \hat{H}(\mathbf{p}) = \textit{invalid} \text{ and } J(\mathbf{p}) = 1\}, \tag{9}$$

$$\text{FI} = \{\mathbf{p} : H(\mathbf{p}) = c \text{ and } \hat{H}(\mathbf{p}) = \textit{invalid} \text{ and } J(\mathbf{p}) = 0\}. \tag{10}$$

UIoU for class c is then expressed as

$$\text{UIoU} = \frac{|\text{TP}| + |\text{TI}|}{|\text{TP}| + |\text{TI}| + |\text{FP}| + |\text{FN}| + |\text{FI}|}. \tag{11}$$

Note that a true invalid prediction results in equal reward to predicting the correct semantic label of the pixel. Moreover, an invalid prediction does not come at no cost: it incurs the same penalty on valid pixels as predicting an incorrect label.

When dealing with multiple classes, we modify our notation to $\text{UIoU}^{(c)}$ (similarly for the five sets of pixels related to class c), which we avoided in the previous definitions to reduce clutter. The overall semantic segmentation performance on the evaluation set is reported as the mean UIoU over all C classes (mUIoU for short). Note that we can perform multiple evaluations by varying the confidence threshold θ and using the respective output, so that we obtain a parametric expression $\text{UIoU}(\theta)$. When $\theta = 1/C$, no pixel is predicted as invalid and thus $\text{UIoU}(1/C) = \text{IoU}$.

We include a brief theoretical analysis to motivate the usage of UIoU instead of the standard IoU. Our analysis is

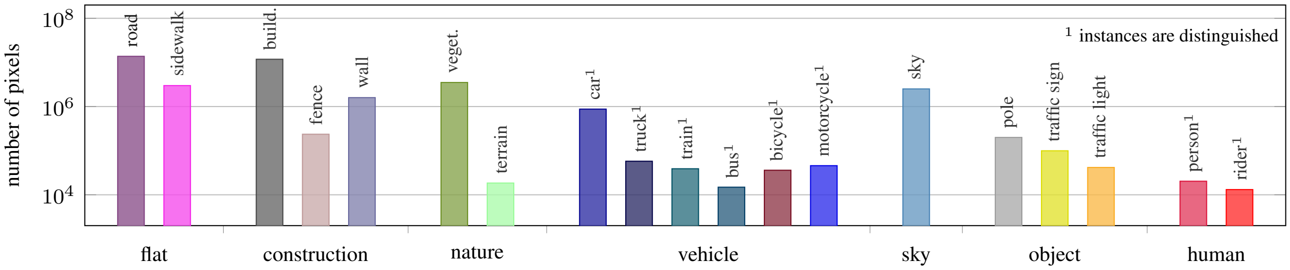


Figure 3. Number of annotated pixels per class for *Dark Zurich-test*.

driven by the expectation that predictions on invalid regions should be associated with lower confidence than those on valid regions. Based on a related assumption, we prove that UIoU becomes greater than IoU for some $\theta > 1/C$.

Theorem 1. *Assume that there exist* θ_1*,* θ_2 *such that* $\theta_1 < \theta_2$*,* $\forall p : J(p) = 1 \Rightarrow s_{\tilde{H}(p)}(p) \leq \theta_1$ *and* $J(p) = 0 \Rightarrow s_{\tilde{H}(p)}(p) \geq \theta_2$*. If we additionally assume that* $\exists p \in \mathrm{FN}^{(c)}(1/C) \cup \mathrm{FP}^{(c)}(1/C) : J(p) = 1$*, then* $\mathrm{IoU}^{(c)} < \mathrm{UIoU}^{(c)}(\theta_1)$.

Proof. For brevity, we drop the class superscript in the proof. We firstly associate the pixel sets for standard IoU with their previously defined counterparts for UIoU. It holds that

$$\begin{aligned}&|\mathrm{TP}(1/C)| + |\mathrm{FN}(1/C)| \\ &\qquad = |\mathrm{TP}(\theta)| + |\mathrm{FN}(\theta)| + |\mathrm{TI}(\theta)| + |\mathrm{FI}(\theta)|. \end{aligned} \tag{12}$$

The first assumption of Theorem 1 implies that $\mathrm{FI}(\theta_1) = \emptyset$. This leads to

$$|\mathrm{TP}(1/C)| = |\mathrm{TP}(\theta_1)| + |\mathrm{TI}(\theta_1)| + |\mathrm{FN}(\theta_1)| - |\mathrm{FN}(1/C)|. \tag{13}$$

Combining the second assumption with the first yields

$$(|\mathrm{FN}(1/C)| - |\mathrm{FN}(\theta_1)|) + (|\mathrm{FP}(1/C)| - |\mathrm{FP}(\theta_1)|) > 0. \tag{14}$$

Since both terms in (14) are nonnegative, at least one of them is strictly positive. We distinguish two cases. If the first term in (14) is strictly positive, (13) implies $|\mathrm{TP}(1/C)| < |\mathrm{TP}(\theta_1)| + |\mathrm{TI}(\theta_1)|$ and this together with (12) suffices to reach the conclusion of the theorem. If the second term is strictly positive, it follows that $|\mathrm{FP}(1/C)| > |\mathrm{FP}(\theta_1)|$ and this together with (12) and $|\mathrm{TP}(1/C)| \leq |\mathrm{TP}(\theta_1)| + |\mathrm{TI}(\theta_1)|$ suffices to reach the conclusion. □

5. The Dark Zurich Dataset

Dark Zurich was recorded with a car in the city of Zurich using a GoPro Hero 5 camera with 1080p resolution, mounted on top of the front windshield. In particular, we planned several laps in disjoint areas of the city and drove each lap multiple times on the same day, starting from daytime through twilight to nighttime. This collection protocol enables the application of the guided style transfer approach of Sec. 3.1.2, where the auxiliary reference set $\mathcal{R}'$ consists of the daytime laps. Depending on which time of day is transferred, the main reference set $\mathcal{R}$ can be defined either as the nighttime or the twilight laps. The assignment B between the two sets is established through GPS readings.

We split *Dark Zurich* and reserve one lap for testing. The rest laps are used for training and comprise 3041 daytime, 2920 twilight and 2416 nighttime unlabeled images, extracted at 1 fps and named *Dark Zurich*-{*daytime*, *twilight*, *nighttime*} respectively. We extract one image from the nighttime testing lap every 50m or 20s, whichever comes first, and assign to it a corresponding daytime image to serve as the auxiliary image I' in our annotation framework of Sec. 4.2. We annotate 20 nighttime images of the testing lap with fine pixel-level invalid masks and semantic Cityscapes labels according to our annotation protocol and name this set *Dark Zurich-test*. Detailed statistics for the annotated labels are given in Fig. 3. In total, 37.8M pixels have been annotated with semantic labels and 16.1% of these pixels are marked as invalid. We note that most large-scale datasets for semantic segmentation of road scenes, including Cityscapes [4] and Mapillary Vistas [22], contain few or no nighttime scenes. The labeled part of Nighttime Driving [5] also has a small scale and images are annotated coarsely. A notable exception is BDD100K [35], where ca. 3% of the 10000 images correspond to nighttime. However, we found that these images contain severe labeling inconsistencies, particularly in dark regions (e.g. *sky* labeled as *building*). Our uncertainty-aware annotation framework addresses these issues which naturally occur at nighttime by introducing the concept of invalid regions and utilizing auxiliary daytime images to aid annotation, and *Dark Zurich-test* is a first prototype to support our uncertainty-aware evaluation.

6. Experiments

Our architecture of choice for experiments on semantic segmentation is the modern RefineNet [17], which fea-

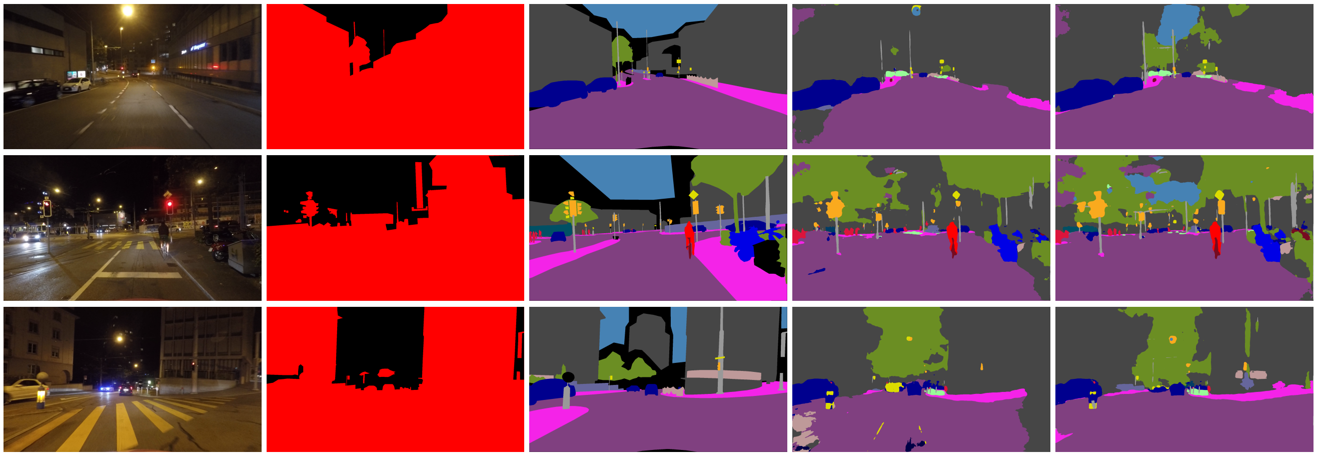

(a) Image (b) Invalid mask (c) Semantic GT (d) RefineNet [17] (e) Ours

Figure 4. Qualitative semantic segmentation results on *Dark Zurich-test*. "Ours" stands for our complete two-step gradual adaptation using Dark Cityscapes and the training split of *Dark Zurich*.

tures a ResNet [9] backbone. We use the publicly available *RefineNet-res101-Cityscapes* model, trained on the training set of the original Cityscapes, as the baseline model to be adapted to nighttime. Throughout our experiments, we train this model with a constant learning rate of 5×10^{-5} on mini-batches of size 1.

6.1. Benefit of Adaptation with Our Stylized Data

Our first experiment ablates the two main components of our guided style transfer pipeline, measuring their impact on semantic segmentation performance when stylized data generated with the respective configurations of the pipeline are used to fine-tune the baseline semantic segmentation model. The input set for style transfer in this experiment is the full training set of Cityscapes comprising 2975 images, the reference set is *Dark Zurich-nighttime* and the auxiliary reference set for style guidance is *Dark Zurich-daytime*. The four compared configurations for generation of Dark Cityscapes involve random reference selection versus feature-based reference selection using contextual similarity [20] and direct style transfer from the nighttime reference set versus guided style transfer using the auxiliary daytime reference set. Each corresponding version of Dark Cityscapes is then used to fine-tune *RefineNet-res101-Cityscapes* for 4 epochs. We evaluate all models on *Dark Zurich-test* and report their standard mean IoU performance in Table 1, considering normally *invalid* pixels which are assigned a legitimate semantic label during evaluation. Fine-tuning on Dark Cityscapes consistently increases performance compared to the baseline RefineNet model. Moreover, using style guidance from the auxiliary daytime reference set results in a significant benefit. Contextual similarity brings a marginal improvement over random selection; we hypothesize that uniform selection of reference images with the random baseline leads to increased diversity in the output which counteracts the mismatch in the semantic content for the created pairs.

Mean IoU (%)		
RefineNet [17]	31.1	
Stylization | Reference selection	Random	Contextual
Direct	33.7	34.2
Daytime-guided	**37.1**	**37.3**

Table 1. Performance comparison on *Dark Zurich-test* of RefineNet and fine-tuned versions of it on Dark Cityscapes, which is generated with various style transfer configurations that all use *Dark Zurich* for the reference style. The standard mean IoU metric is used for evaluation.

Mean IoU (%)		
RefineNet [17]	31.5	
Stylization | Reference selection	Random	Contextual
Direct	39.3	37.6
Daytime-guided	**40.2**	**40.9**

Table 2. Performance comparison on Nighttime Driving [5] of RefineNet and fine-tuned versions of it on Dark Cityscapes. Details are identical to Table 1.

We repeat the same evaluation on the latest version of Nighttime Driving [5] in Table 2. The conclusions are similar; in particular, the combination of daytime style guidance and contextual similarity leads to a model with 9.4% higher performance than the baseline.

In addition, we evaluate in Fig. 5 the baseline RefineNet model and the two models corresponding to reference selec-

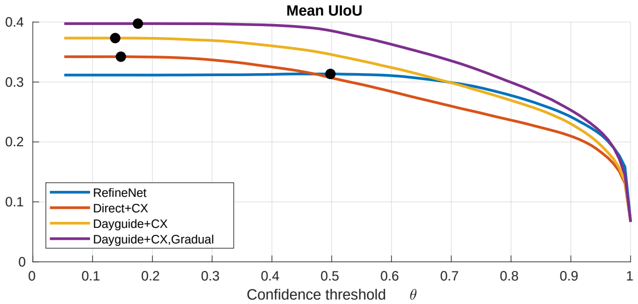


Figure 5. Performance comparison on *Dark Zurich-test* of RefineNet, fine-tuned versions of it on Dark Cityscapes and our gradually adapted model from Table 3. We evaluate mean UIoU across the entire θ range. Optimal operating points for each model are marked black. "Direct": directly transferring style from nighttime images, "Dayguide": using daytime images to guide style transfer, "CX": contextual similarity, "Gradual": two-step joint adaptation.

tion with contextual similarity on *Dark Zurich-test* with our new mean UIoU metric for variable confidence threshold θ. Contrary to our expectation based on the result of Theorem 1, the optimal mean UIoU performance of all three models is only marginally higher than their respective mean IoU performance, as mean UIoU remains virtually constant for increasing θ before starting to drop. This implies that in general correct predictions on valid regions get invalidated for lower values of the confidence threshold than incorrect predictions on invalid regions, despite the fact that a robust model should assign lower confidence to the latter, and stresses the need for robust training of semantic segmentation models [12] to avoid misinterpreting input of highly uncertain semantic content, especially in adverse conditions.

6.2. Benefit of Gradual Adaptation with Stylized and Real Data

Our second experiment demonstrates the benefit of our adaptation framework using both labeled synthetic and unlabeled real data. In this experiment, we consistently use daytime guidance and contextual similarity to generate synthetic Dark Cityscapes images. We generate a twilight and a nighttime-stylized version of Dark Cityscapes using *Dark Zurich-twilight* and *Dark Zurich-nighttime* as the reference style set respectively. These versions are used as the labeled synthetic data source of our approach. On the real data side, we use the daytime and twilight part of the training split of *Dark Zurich*. We first adapt *RefineNet-res101-Cityscapes* to the union of Dark Cityscapes-twilight and *Dark Zurich-daytime* (with noisy labels from *RefineNet-res101-Cityscapes* for the latter), then use the resulting model to label *Dark Zurich-twilight* and further adapt it to the union of Dark Cityscapes-nighttime and *Dark Zurich-twilight*. Both adaptation steps involve 30k iterations and $\mu = 1$. We report the mean IoU performance of the adapted models on *Dark Zurich-test* and Nighttime Driving in Tables 3 and 4 respectively, and compare against models adapted only to Dark Cityscapes as well as the method of [5] that only uses unlabeled real data. The gradually adapted model using both synthetic and real data ("Syn-N+Real-T") achieves the best performance on both sets (the same is true for mean UIoU as shown in Fig. 5). Both of our models that are trained jointly on synthetic and real data outperform the models that are only trained on the respective synthetic data, evidencing the benefit of using unlabeled real data in the adaptation. We provide visual results of our gradually adapted model on *Dark Zurich-test* in Fig. 4.

Mean IoU (%)

DMADA	Syn-T	Syn-N	Syn-T+Real-D	Syn-N+Real-T
33.8	37.3	37.3	38.8	**39.9**

Table 3. Performance comparison on *Dark Zurich-test* of our gradual adaptation using synthetic and real data against [5] ("DMADA") and synthetic-only adaptation. "Syn": synthetic data, "Real": real data, "D": daytime, "T": twilight, "N": nighttime.

Mean IoU (%)

DMADA	Syn-T	Syn-N	Syn-T+Real-D	Syn-N+Real-T
36.1	41.3	40.9	41.8	**43.6**

Table 4. Performance comparison on Nighttime Driving [5] of our gradual adaptation using synthetic and real data against [5] ("DMADA") and synthetic-only adaptation. The rest abbreviations are identical to Table 3.

7. Conclusion

In this paper, we have presented a framework to improve semantic segmentation performance at nighttime with stylized data and unlabeled real data of increasing darkness, as well as a novel uncertainty-aware evaluation for semantic segmentation designed specifically to handle images with indiscernible content. In this context, we have introduced a novel guided pipeline for style transfer at dataset scale and used it to transfer the time of day of Cityscapes. We have also presented *Dark Zurich*, a large-scale dataset of real scenes captured at multiple times of day, and annotated 20 nighttime scenes of it with a new protocol which enables our uncertainty-aware evaluation. Detailed evaluation with the standard IoU metric on two real nighttime sets (including ours) evidences the benefit of both our style transfer pipeline for creating synthetic nighttime training data and our gradual adaptation with synthetic and unlabeled real data. Finally, evaluation on our set with the novel UIoU exposes the overly confident predictions of the examined models for semantically uncertain parts of nighttime images.

References

[1] J. M. A. Alvarez and A. M. Lopez. Road detection based on illuminant invariance. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):184–193, 2011.

[2] Y. Chen and C. Han. Night-time pedestrian detection by visual-infrared video fusion. In *World Congress on Intelligent Control and Automation*, 2008.

[3] Y. Chen, W. Li, and L. Van Gool. ROAD: Reality oriented adaptation for semantic segmentation of urban scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[5] D. Dai and L. Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *IEEE International Conference on Intelligent Transportation Systems*, 2018.

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[7] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] J. Ge, Y. Luo, and G. Tei. Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. *IEEE Transactions on Intelligent Transportation Systems*, 10(2):283–298, 2009.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[10] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, 2018.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[12] A. Kendall, V. Badrinarayanan, and R. Cipolla. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In *Proc. British Machine Vision Conference*, 2017.

[13] J. H. Kim, H. G. Hong, and K. R. Park. Convolutional neural network-based human detection in nighttime images using visible light camera sensors. *Sensors*, 17(5), 2017.

[14] A. Kirillov, K. He, R. B. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. *CoRR*, abs/1801.00868, 2018.

[15] H. Kuang, K. Yang, L. Chen, Y. Li, L. L. H. Chan, and H. Yan. Bayes saliency-based object proposal generator for nighttime traffic images. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):814–825, 2018.

[16] Y. Li, M.-Y. Liu, X. Li, M.-H. Yang, and J. Kautz. A closed-form solution to photorealistic image stylization. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[17] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, 2017.

[19] F. Luan, S. Paris, E. Shechtman, and K. Bala. Deep photo style transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[20] R. Mechrez, I. Talmi, and L. Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[21] S. G. Narasimhan and S. K. Nayar. Vision and the atmosphere. *Int. J. Comput. Vision*, 48(3):233–254, July 2002.

[22] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kontschieder. The Mapillary Vistas dataset for semantic understanding of street scenes. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[23] G. Ros and J. M. Alvarez. Unsupervised image transformation for outdoor semantic labelling. In *IEEE Intelligent Vehicles Symposium (IV)*, 2015.

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

[25] C. Sakaridis, D. Dai, and L. Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, Sep 2018.

[26] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[27] R. K. Satzoda and M. M. Trivedi. Looking at vehicles in the night: Detection and dynamics of rear lights. *IEEE Transactions on Intelligent Transportation Systems*, 2016.

[28] Y. Shih, S. Paris, F. Durand, and W. T. Freeman. Data-driven hallucination of different times of day from a single outdoor photo. *ACM Trans. Graph.*, 32(6):200:1–200:11, November 2013.

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[30] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018.

[31] A. Valada, J. Vertens, A. Dhall, and W. Burgard. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In *ICRA*, 2017.

[32] M. Wulfmeier, A. Bewley, and I. Posner. Incremental adversarial domain adaptation for continually changing environments. *ICRA*, 2018.

[33] F. Xu, X. Liu, and K. Fujimura. Pedestrian detection and tracking with night vision. *IEEE Transactions on Intelligent Transportation Systems*, 6(1):63–71, 2005.

[34] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations*, 2016.

[35] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell. BDD100K: A diverse driving video database with scalable annotation tooling. *CoRR*, 2018.

[36] O. Zendel, K. Honauer, M. Murschitz, D. Steininger, and G. Fernandez Dominguez. Wilddash - creating hazard-aware benchmarks. In *European Conference on Computer Vision (ECCV)*, September 2018.

[37] O. Zendel, M. Murschitz, M. Humenberger, and W. Herzner. How good is my test data? Introducing safety analysis for computer vision. *International Journal of Computer Vision*, 125(1):95–109, Dec 2017.

[38] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.