Interactive Learning of Intrinsic and Extrinsic Properties for All-Day Semantic Segmentation

Bi, Q.; You, S.; Gevers, T.

DOI
10.1109/TIP.2023.3290469

Publication date
2023

Document Version
Final published version

Published in
IEEE Transactions on Image Processing

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/in-the-netherlands/you-share-we-take-care)

Citation for published version (APA):
Interactive Learning of Intrinsic and Extrinsic Properties for All-Day Semantic Segmentation

Qi Bi, Student Member, IEEE, Shaodi You, Senior Member, IEEE, and Theo Gevers

Abstract—Scene appearance changes drastically throughout the day. Existing semantic segmentation methods mainly focus on well-lit daytime scenarios and are not well designed to cope with such great appearance changes. Naively using domain adaption does not solve this problem because it usually learns a fixed mapping between the source and target domain and thus has limited generalization capability on all-day scenarios (i.e., from dawn to night). In this paper, in contrast to existing methods, we tackle this challenge from the perspective of image formulation itself, where the image appearance is determined by both intrinsic (e.g., semantic category, structure) and extrinsic (e.g., lighting) properties. To this end, we propose a novel intrinsic-extrinsic interactive learning strategy. The key idea is to inter-calibrate between intrinsic and extrinsic representations during the learning process with spatial-wise guidance. In this way, the intrinsic representation becomes more stable and, at the same time, the extrinsic representation gets better at depicting the changes. Consequently, the refined image representation is more robust to generate pixel-wise predictions for all-day scenarios. To achieve this, we propose an All-in-One Segmentation Network (AO-SegNet) in an end-to-end manner. Large scale experiments are conducted on three real datasets (Mapillary, BDD100K and ACDC) and our proposed synthetic All-day CityScapes dataset. The proposed AO-SegNet shows a significant performance gain against the state-of-the-art under a variety of CNN and ViT backbones on all the datasets.

Index Terms—All-day scenarios, semantic segmentation, interactive learning, intrinsic and extrinsic properties.

I. INTRODUCTION

When the golden day is done, through the closing portal, child and garden, flower and sun, vanish all things mortal. This poem written by Robert Louis Stevenson reflects the large changes in appearance for all-day scenarios, i.e., appearance changes throughout a full day, from dawn to night. This large variation in appearance is especially challenging for semantic segmentation, as existing segmentation models are usually trained on consistent (not extreme) imaging conditions and therefore have limited generalization capabilities to all-day scenarios such as twilight [1], [2], [3] and night [4], [5] scenarios (Fig. 1).

Fig. 1. Semantic segmentation performance degradation is shown for current segmentation models such as DDRNet-23 [6] (second row) and domain generalization based RobustNet [2] (third row) for all-day scenarios. Errors are denoted by red bounding boxes.

To address this problem, existing methods aim to enhance the generalization capability to night-time images by exploiting a domain adaption strategy [7], [8], [9], [10], [11], [12], [13], [14]. Unfortunately, as these methods learn a fixed mapping between a source and target domain (e.g., day to night), the predictions for other conditions, included in all-day scenarios (e.g., dawn, twilight), can still be unreliable. Recently, domain generalization based segmentation [2], [3], [15], [16], [17], which learns a shared representation from multiple domains, has been exploited to deal with all-day scenarios. However, its generalization capability is still limited by not considering intra-domain variance and inter-domain diversity [18], [19], [20], [21].

In this paper, we approach the problem of all-day semantic segmentation from the perspective of image formulation itself. The image appearance is usually determined by a combination of both intrinsic (e.g., reflectance, semantics of an object) and extrinsic properties (e.g., illumination, viewpoint) [21], [22], [23]. For standard semantic segmentation, as imaging conditions are usually similar, the extrinsic properties are rather unified making it easier to learn intrinsic properties such as semantics from the observed image appearance. However, for all-day scenarios, as extrinsic properties vary significantly, learning the intrinsic properties such as semantics from varied image appearance become highly ill-posed [21]. Consequently, pixel-wise semantic prediction may become rather unstable.

To this end, we propose a novel intrinsic-extrinsic interactive learning strategy to refine the image representation for all-day semantic segmentation (Fig. 2). The general idea is to learn both intrinsic and extrinsic representations from image appearance observations by the process of interaction between the intrinsic and extrinsic representations under the guidance...
Our contribution can be summarized as follows:

- To the best of our knowledge, this is the first work to systematically study all-day semantic segmentation. We propose a novel intrinsic-extrinsic interactive learning strategy. It generates image representations robust to appearance changes for all-day scenarios.
- We propose an All-in-One Segmentation Network (AO-SegNet) consisting of the three key steps of our interactive learning in an end-to-end manner. It is versatile to a variety of CNN and ViT backbones. Moreover, a progressive multi-scale semantic mining module and a semantic align loss are proposed.
- We design an all-day semantic segmentation benchmark all-day CityScapes. It is the first semantic segmentation benchmark that contains samples from all-day scenarios, i.e., from dawn to night. Our dataset will be made publicly available at this repository.
- Extensive experiments on three real-world datasets and our All-Day CityScapes dataset demonstrate that our proposed method outperforms existing directly-supervised, all-day based, and domain generalization based segmentation methods.

The remainder of this paper is organized as follows. In Section II, we discuss the recent work related to the all-day semantic segmentation task. In Section III, a preliminary about the definition of all-day semantic segmentation and our proposed intrinsic-extrinsic interactive learning strategy are discussed. In Section IV, the proposed All-in-One Segmentation Network (AO-SegNet) guided by an interactive learning strategy is provided. In Section V, extensive experiments on three benchmarks are conducted and the a discussion is given. Finally, in Section VI, conclusions are drawn.

II. RELATED WORK

A. Semantic Segmentation Under Varied Scenarios

Deep learning models have demonstrated great success for semantic segmentation. Typical models include FPN [24], SegNet [25], UNet [26], ICNet [27]. Mask R-CNN [28], DeepLab [29], DDRNet [6] and etc. More recently, the performance of these models is further enhanced by either using multi-scale attention based fusion for convolutional networks (e.g., DDRNet [6], OCR [30]) or applying vision Transformer (e.g., Segmenter [31], Mask2Former [32]). However, these segmentation models are usually trained on consistent (no-extreme) imaging conditions, and may suffer from performance decline when the trained model is generalized on more extreme scenarios.

In the past few years, many methods focused on bridging the gap between day- and night- images, and domain adaptive segmentation methods become the dominant trend. Among the prior works, Dai and Van Gool [33] propose to adapt the style of night images during the training stage, so that the framework can show a reasonable inference on the night-time images. Similarly, Sun et al. [34] use the style transfer to improve the night-time segmentation performance. However, such methods, either supervised [7], [8], [9], [10], [11], [12], [13], [14], [35], [36], [37] or unsupervised [1], [5], [38], [39], [40], [41], [42], usually learn a fixed mapping between a source domain to a target domain, and thus are not well suited for all-day scenarios containing more varied day/night scenarios [1], [2].

Recently, domain generalized segmentation is proposed [2], [3], [15], [16], [17]. These methods usually learn a shared representation from multiple domains and are somehow robust to varied scenarios [2]. However, their generalization capability is still limited by the intra-domain variance (multiple samples) and inter-domain diversity (varied real-world scenarios).

Another bottleneck for all-day semantic segmentation is the dataset. As discussed in Sec. V-A, despite a limited number of datasets, including samples from night or twilight scenarios [1], [4], [33], [43], [44], [45], [46], the diversity for all-day scenarios is limited [1], [4], [33], [43], [45], [47].

B. All-Day Representation Learning

In the past years, representation learning under all-day scenarios is studied. For segmentation, Zhang et al. [18] combine images from multiple day/night scenarios and fuse them into a semantic complementary image for all-day road segmentation.

For other visual tasks, Hu et al. propose an all-season depth estimation benchmark [48]. To find feature correspondences for all-day and all-season images, Spencer et al. [20] propose a weakly supervised learning framework and Larsson et al. [49] enhance this topic by creating a cross-season correspondence...
dataset. More recently, Liu et al. [19] propose an all-day depth estimation framework but still use day and night image pairs as input. Sakaridis et al. propose a real-captured segmentation dataset named Adverse Conditions Dataset with Correspondences [50]. It provides images under the rain, snow, fog and night-time conditions.

In conclusion, most of the existing work is not specifically designed for semantic segmentation and the computational cost of point-wise feature correspondence [20], [49] is too burdensome to be transferred to segmentation.

III. PRELIMINARY

A. INTRINSIC & EXTRINSIC PROPERTY FORMULATION

While the image appearance \( I \) is highly dynamic in all-day scenarios, without loss of generality, the underlying properties, which contribute to appearance, are separated as intrinsic and extrinsic properties [21], [22], [23]. As illustrated in Fig. 2, the intrinsic properties \( \sigma_{I_{in}} \) are independent of the environment, e.g., the semantics of an object, the structure, and albedo. The extrinsic properties \( \sigma_{E_{ex}} \) are scene dependent, e.g., lighting, shading, spatial position. For all-day segmentation, we focus on semantics as intrinsic properties and illumination conditions as extrinsic properties because semantic and illumination account for the majority of appearance changes. Accordingly, we can formulate the concept by

\[
I = f(\sigma_{I_{in}}, \sigma_{E_{ex}}),
\]

where \( f \) denotes the image formulation process.

For all-day scenarios (e.g., dawn, dusk, night, and etc.), a collection of extrinsic properties \( \{\sigma_{E_{ex}}\} \) and a collection of intrinsic properties \( \{\sigma_{I_{in}}\} \) are given. The varied image appearance under all-day scenarios \( \{I_{j,i}\} \) can be generated by

\[
\{I_{j,i}\} = \{f(\sigma_{I_{in}}, \sigma_{E_{ex}})\}.
\]

For all-day semantic segmentation, we first need to consider the inverse operation of the image formulation process \( f \) in Eq. 1, i.e., \((\sigma_{I_{in}}, \sigma_{E_{ex}}) = f^{-1}(I)\). It is ill-posed and can not be solved directly. To obtain the semantic prediction \( S \), a quasi-inverse \( F^{-1} \) is defined by \( S = F^{-1}(I) \).

Now consider Eq. 2. Although \( \sigma_{E_{ex}} \) may vary for the image appearance \( \{I_{j,i}\} \), the semantic \( \{S_{j}\} \) should be invariant:

\[
\{S_{j}\} = F^{-1}(\{f(\sigma_{I_{in}}, \sigma_{E_{ex}})\}).
\]

Unfortunately, it is unfeasible to explicitly model (e.g. by a reflection model) the relation between \( I \), \( \sigma_{I_{in}} \) and \( \sigma_{E_{ex}} \) for complex real-world scenarios due to its highly ill-posed nature. Hence, in this paper, we focus on a data-driven approach to implicitly learn this relation. Following existing representation learning strategies in deep learning [8], [23], [51], the appearance representation \( R_{I} \) of a deep learning model is a combination of the intrinsic representation \( R_{I_{in}} \) from \( \sigma_{I_{in}} \) and the extrinsic representation \( R_{E_{ex}} \) from \( \sigma_{E_{ex}} \). It is defined by

\[
R_{I} = R_{I_{in}} + R_{E_{ex}}.
\]

Consequently, the segmentation prediction \( S \) can be inferred from \( R_{I} \), \( R_{I_{in}} \) and \( R_{E_{ex}} \). Following this idea, a novel intrinsic-extrinsic interactive learning strategy is proposed (Fig. 2) and is presented in Sec. IV.

\[\begin{align*}
(I_{j,i}) & \rightarrow \{S_{j}\} \\
(I_{j,i}) & \rightarrow \{S_{j}\} \\
& \cdots
\end{align*}\]

(a) domain adaptation segmentation

\[\begin{align*}
(I_{j,i}) & \rightarrow \{S_{j}\} \\
(I_{j,i}) & \rightarrow \{S_{j}\} \\
& \cdots
\end{align*}\]

(b) domain generalization segmentation

\[\begin{align*}
(I_{j,i}) & \rightarrow \{S_{j}\} \\
(I_{j,i}) & \rightarrow \{S_{j}\} \\
& \cdots
\end{align*}\]

(c) all-day segmentation

Fig. 3. Problem definition and pipeline of (a) domain adaptive, (b) domain generalization and (c) all-day semantic segmentation. For domain generalization segmentation (b), only samples from a specific day/night domain are utilized for training. On the other hand, all-day segmentation (c) leverages samples from all-day/night domains for training.

B. DIFFERENCE FROM DOMAIN ADOPTION & GENERALIZATION

Fig. 3 illustrates the difference between all-day semantic segmentation, domain adaption and domain generalization segmentation methods.

Domain adaptive segmentation models \( F^{-1}(\cdot, \cdot) \) considers appearance \( \{I_{j,i}\} \) as a domain, and requires a source and target domain as input. The model learn a fixed mapping between the source domain (e.g., \( \{I_{j,1}\}\)) and target domain (e.g., \( \{I_{j,i}\}\)). Note that its generalization capability to cover all-day scenarios is limited as there is (theoretically) an infinite number of scenarios.

Domain generalization segmentation models also consider the appearance of each scenario \( I_{i} \) as a domain, and learn a shared representation from multiple domains, i.e., \( F^{-1}(\{I_{j,1}\}, \ldots, \{I_{j,i}\}, \ldots) \). However, its generalization capability is negatively influenced by the intra-domain variance and inter-domain diversity [2], [16]. Moreover, as the all-day scenario is an infinite set, it may suffer from the curse of dimensionality.

C. LIMITATION OF INTRINSIC & EXTRINSIC FORMULATION

Based on the intrinsic and extrinsic formulation, there are still several limitations that need to be addressed for all-day semantic segmentation.

- Different from the work on low-level vision such as intrinsic decomposition [22], [23], where the intrinsic representation (i.e., albedo) and extrinsic representation (i.e., shading) is at the level of objects or scenes, in our all-day semantic segmentation task, the image appearance is pixel-level. In other words, the intrinsic decomposition [22], [23] has the capability to split the object appearance into two parts. In contrast, our intrinsic-extrinsic formulation assumes that each pixel is determined by both intrinsic and extrinsic properties, and is not directly related to the object-level representation.

- Although the semantic of an object, e.g., car, building, belongs to the intrinsic properties, in the semantic segmentation task, some categories also need multiple cues containing extrinsic properties to cover their semantics. Therefore, in our IEM step, the intrinsic and extrinsic...
representations are fused for a final semantic prediction to enforce a better understanding of outdoor scenarios.

IV. METHODOLOGY

A. Framework Overview

Fig. 4 illustrates the overall framework of our All-in-One Segmentation Network (AO-SegNet), which is an end-to-end realization of the proposed intrinsic-extrinsic interactive learning strategy (Fig. 2).

After convolutional feature extraction by the backbone (e.g., DDRNet-23 [6]), the intrinsic-extrinsic enhancing (IEE) step consists of the intrinsic learner (I-learner) and extrinsic learner (E-learner) (Sec. IV-B). The I-learner generates the intrinsic representation \( R_{\text{In}} \), and the E-learner generates the extrinsic representation \( R_{\text{Ex}} \).

Then, the intrinsic-extrinsic interacting (IEI) step (Sec. IV-C) interacts with both representations for refinement under spatial-wise guidance, where the key positions are highlighted by given them higher weights. The extrinsic features contributing positively to the intrinsic features are selected (denoted by E2I), and vice versa (denoted by I2E). This is computed by the proposed interaction attention mechanism, where the spatial weight matrix of one branch is fused into another branch, and the key positions in the intrinsic/extrinsic branch are jointly learnt.

Finally, the intrinsic-extrinsic merging (IEM) step merges both refined intrinsic representation \( R'_{\text{In}} \) and extrinsic representation \( R'_{\text{Ex}} \) for semantic prediction, and provides a semantic latent space \( \mathcal{O} \) to constraint the intrinsic representations before and after refinement (Sec. IV-D & IV-E).

B. Intrinsic-Extrinsic Enhancing

Based on the intrinsic-extrinsic formulation (Eq. 4), we first need to initialize both the intrinsic representation and extrinsic representation. For a dual-branch structure such as DDRNet-23 backbone [6], it computes both low-resolution output \( C \) and a high-resolution output \( A \) from the image appearance \( I \). Both branches consist of multiple residual blocks but of different spatial resolution. The low-resolution output \( C \) has a resolution of 1/64, while the high-resolution output \( A \) has a resolution of 1/8. On the other hand, when scaled to one-branch structure such as ResNet, the 1/64-resolution feature serves as \( C \) and the 1/8-resolution feature serves as \( A \).

In the proposed AO-SegNet, the low-resolution output \( C \) and high-resolution output \( A \) are inherited to compute \( R_{\text{In}} \), extrinsic representation \( R_{\text{Ex}} \), respectively. The low/high-resolution output for computing the intrinsic/extrinsic representation is intuitive, because for an all-day segmentation the low-resolution features contain more semantic information for intrinsic representation, while the extrinsic variation is reflected by high-resolution features.

In the proposed AO-SegNet, the low-resolution output \( C \) and high-resolution output \( A \) are inherited to compute \( R_{\text{In}} \) and \( R_{\text{Ex}} \), respectively. The low/high-resolution output for computing the intrinsic/extrinsic representation is intuitive, because for an all-day segmentation the low-resolution features contain more semantic information for intrinsic representation, while the extrinsic variation is reflected by high-resolution features.

The implementation of our intrinsic learner (I-learner) and extrinsic learner (E-learner) is given in Algorithm 1. As the low-resolution convolutional features tend to convey more high-level information such as semantics [6], for the I-learner, intrinsic information especially related to the semantics [21] needs to be extensively mined from the low-resolution branch output \( C \) to compute \( R_{\text{In}} \). Thus, a progressive multi-scale

---

### Algorithm 1 Implementation of I-Learner & E-Learner

**Input:** low-resolution feature \( C \), high-resolution feature \( A \)

**Output:** intrinsic representation \( R_{\text{In}} \), extrinsic representation \( R_{\text{Ex}} \)

1. \( \% n: \) scale number; \( s: \) stride of convolution; \( Up: \) up-sampling; \( \text{conv}1: 1 \times 1 \) convolution filter; \( \text{diConv}3: 3 \times 3 \) dilated convolution; \( r: \) dilated rate.
2. for \( i = 1 \rightarrow n \)
3. \( C^i \leftarrow Up^{s=2^{i-1}}(\text{conv}1=C^{i=2^{i-1}}(C)) \)
4. end for
5. for \( i = 2 \rightarrow n \)
6. \( C^i \leftarrow \text{diConv}3^{s=2^{i-1}}(C^i+C^{i-1}) \)
7. end for
8. \( R_{\text{In}} \leftarrow Up^{s=8}(\text{conv}1([C^1, \cdots, C^n]) + \text{conv}1(C)) \)
9. \( R_{\text{Ex}} \leftarrow \text{conv}1(A) \)
semantic mining (PMSM) module is designed (shown in Fig. 5). Assume there are $n$ scales, for each scale $i$, $C$ is processed by a $1 \times 1$ convolution $Conv_1$ with a stride of $s = 2^{i-1}$ and then up-sampled by $2^{i-1}$ times, so that the output $C^i$ from the $i^{th}$ scale is calculated.

Then, the representation $C^i$ from the $i^{th}$ scale is progressively fused with $C^{i-1}$ from the former $(i - 1)^{th}$ scale via firstly adding by $C^{i-1}$ and then processed by a $3 \times 3$ dilated convolution $dConv_3$ with a dilated rate $r = 2i - 1$. Finally, $C^i$ are concatenated and compressed by a $1 \times 1$ convolutional layer, and then fused with $C$ and processed by another $1 \times 1$ convolutional layer. Then, it is up-sampled 8 times to compute the intrinsic representation $R_{In}$.

For E-learner, it is straightforward to compute $R_{Ex}$ from a high-resolution output $A$. The $1 \times 1$ convolutional layer benefits the cross-channel feature interaction and it only adds little computational cost. It has been widely-used for feature enhancement of convolutional neural networks [52], [53], [54], [55]. Consequently, a $1 \times 1$ convolutional layer is used to enhance the high-resolution feature $C$ for the extrinsic representation $R_{Ex}$.

C. Intrinsic-Extrinsic Interacting

In all-day scenarios, extrinsic properties vary a lot. It is difficult to learn semantics from the varied appearance. The intrinsic-extrinsic interacting (IEI) step refines the intrinsic and extrinsic representations by the interaction attention mechanism, where the spatial weight matrix of one branch is fused into another branch for interaction. In this way, as the spatial attention mechanism emphasizes the positions of the key semantic information in an image, this spatial-wise information is jointly learnt by both the intrinsic and extrinsic branches. Consequently, under spatial-wise guidance, the intrinsic representation becomes more stable in depicting the intrinsic properties from the image appearance and the extrinsic representation becomes more powerful to depict the environmental changes.

First, both $R_{In}$ and $R_{Ex}$ are processed by an one-layer attention module to generate the spatial weight matrix $M_{In}$ and $M_{Ex}$ as follows

$$M_{In} = \text{softmax}(W_1R_{In} + b_1),$$

$$M_{Ex} = \text{softmax}(W_2R_{Ex} + b_2),$$

where $W_1$, $b_1$ and $W_2$, $b_2$ denote the weights and biases of these two $1 \times 1$ convolutional layers, and softmax denotes the softmax normalization function.

Then, for the extrinsic-to-intrinsic (E2I) branch, the highlighted spatial information from $R_{Ex}$ is embedded into the intrinsic spatial weight matrix $M_{In}$ by

$$M'_{In} = M_{In} + M_{Ex},$$

where $M'_{In}$ is the refined intrinsic spatial weight matrix.

Similarly, for the intrinsic-to-extrinsic (I2E) branch, the extrinsic spatial weight matrix $M_{Ex}$ is updated into $M'_{Ex}$ by the spatial positions with strong responses in $R_{In}$. This process is given by

$$M'_{Ex} = M'_{In} + M_{Ex},$$

Finally, the refined intrinsic representation $R'_{In}$ and extrinsic representation $R'_{Ex}$ take the skip connection into account, calculated by

$$R'_{In} = R_{In} \odot M'_{In} + R_{In},$$

$$R'_{Ex} = R_{Ex} \odot M'_{Ex} + R_{Ex},$$

where $\odot$ denotes the hadamard product operation.

D. Intrinsic-Extrinsic Merging

The intrinsic-extrinsic merging (IEM) fuses the refined $R'_{In}$ and $R'_{Ex}$ so that the refined image representation $R'$ is more suitable for segmentation of all-day scenarios. Moreover, as the semantics are stable despite the variance of the extrinsic properties, $R_{In}$ and $R'_{In}$ provide the same semantic information in latent space. It is necessary to fuse the intrinsic and extrinsic representations for the final semantic prediction, as some semantic categories (e.g., sky, street light) in outdoor scenarios are highly related to the extrinsic properties (in Sec. III-C).

1) Semantic Prediction: The final image representation $R'$ for semantic segmentation is given by

$$R' = \text{softmax}(R'_{Ex} + R'_{In}).$$

Then, $R'$ is used as input to the segmentation head, up-sampling 8 times, to generate the pixel-wise semantic prediction $S$.

2) Semantic Latent Space: The intrinsic representation $R_{In}$ provides semantic category information in semantic latent space $O$. Assume a transformation $T$ that maps $R_{In}$ to its embedding $O_{In}$ in this latent space. For each position of $R_{In}$, it is embedded by a vector $[o_1, \cdots, o_k, \cdots, o_N]$, where $N$ denotes the number of semantic categories, $o_k$ is the feature response value for the $k^{th}$ semantic category, and $k = 1, 2, \cdots, N$.

3) Latent Space Constraint: As discussed above, $R_{In}$ and $R_{In}$ convey the same semantic information. Transformation $T : R_{In} \rightarrow O_{In}$ is a $1 \times 1$ convolutional layer where the channel number equals semantic category number $N$, given by

$$O_{In} = \text{softmax}(W_3R_{In} + b_3),$$

$$O'_{In} = \text{softmax}(W_4R'_{In} + b_4),$$

where $O_{In}$ and $O'_{In}$ are the embeddings in latent space $O$ for $R_{In}$ and $R'_{In}$. $W_3$, $b_3$ and $W_4$, $b_4$ denote the parameters of transformation $T$. Finally, $O_{In}$ and $O'_{In}$ are aligned in semantic latent space $O$ by the loss term $L_{alg}$ and will be discussed in Sec. IV-E.
A BRIEF SUMMARY OF EXISTING SEGMENTATION BENCHMARKS THAT CONTAINS VARIOUS SCENARIOS OF A DAY. #NUM_EVAL: NUMBER OF NIGHT (N)/TWILIGHT (T)/ALL-DAY (AD) SAMPLES IN THE EVALUATION SET. CLASSES: SEMANTIC CATEGORY NUMBER; REAL-WORLD: WHETHER OR NOT THE SAMPLES ARE COLLECTED FROM REAL-WORLD SCENARIOS; FINE-GRAIN GT: WHETHER OR NOT BOTH TRAINING AND EVALUATION SETS HAVE FINE-GRAIN GROUND TRUTH FOR BOTH SUPERVISION AND EVALUATION. IT IS SHOWN THAT DATASET GAPS REMAIN FOR FULL ALL-DAY SCENARIOS

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Proc. &amp; Year</th>
<th>Scenarios</th>
<th># Num_eval</th>
<th># Classes</th>
<th>Real-world</th>
<th>Fine-grain GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYNTHIA [46]</td>
<td>CVPR_2018</td>
<td>day/night</td>
<td>100</td>
<td>13</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>GTAS [47]</td>
<td>ECCV_2016</td>
<td>N/A</td>
<td>120</td>
<td>13</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mapillary Vistas [43]</td>
<td>ICCV_2017</td>
<td>day/night</td>
<td>60</td>
<td>6</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>WildDash [44]</td>
<td>ECVW_2018</td>
<td>day/night</td>
<td>13</td>
<td>13</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Raincover [4]</td>
<td>ICRA_2017</td>
<td>day/night</td>
<td>95</td>
<td>3</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>BDD100K [45]</td>
<td>CoRR_2018</td>
<td>day/night</td>
<td>345</td>
<td>19</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Nighttime Driving [33]</td>
<td>ICTIS_2018</td>
<td>day/night</td>
<td>50</td>
<td>19</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Dark Zurich [1]</td>
<td>ICCV_2019</td>
<td>twilight/night</td>
<td>151</td>
<td>19</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>HyperCity 2.0 [56]</td>
<td>ICCVW_2021</td>
<td>day/night</td>
<td>68</td>
<td>19</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>All-day CityScapes (Ours)</td>
<td>ICCV_2021</td>
<td>day/night</td>
<td>500</td>
<td>19</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

### E. Loss Function

1) **Segmentation Loss:** After \( R'_i \) in Eq. 11 is computed by the aforementioned segmentation head, it is supervised by the same segmentation loss \( L_{seg} \) used by DDRNet-23 [6].

2) **Semantic Align Loss:** For both \( O_{In} \) and \( O'_{In} \), their pixel-wise responses on semantic categories have \( N \) dimensions. Each dimension represents the feature response for a certain semantic category. Their differences are minimized by the \( l_2 \) norm. The semantic align loss \( L_{alg} \) is as follows:

\[
L_{alg} = \frac{1}{W \times H \times N} \sum_{w=1}^{W} \sum_{h=1}^{H} \sum_{i=1}^{N} |O'_{In} - O_{In}|^2 . \tag{14}
\]

where \( W \) and \( H \) denote the width and height of image \( I \) respectively, and \( 1 \leq w \leq W, 1 \leq h \leq H \), and \( | \cdot |^2 \) denotes the \( l_2 \) norm function. This loss term constrains the semantic information carried in the intrinsic representation in latent space \( \theta \) before and after the IEI steps. This allows the semantics prediction to become more stable.

3) **Overall Loss Function:** The overall loss function \( L \) is a linear combination of \( L_{seg} \) and \( L_{alg} \) and is given by

\[
L = L_{seg} + \alpha L_{alg} , \tag{15}
\]

where \( \alpha \) is a hyper-parameter to balance the impact of these two terms. Empirically, we set \( \alpha = 5 \times 10^{-4} \).

### V. EXPERIMENT AND ANALYSIS

#### A. Dataset and Evaluation Protocols

Existing segmentation datasets that contain twilight or night scenarios are listed in Tab. I. After excluding those with unreliable ground truth and maximizing the diversity in depicting all-day scenarios, BDD100K, Mapillary Vistas and ACDC are chosen for our experiments. Furthermore, to enhance the diversity for all-day scenarios, we provide a synthetic All-day CityScapes dataset for validation.

1) **BDD100K:** This dataset [45] has 19 semantic categories. The same as CityScapes. It has 7,000 training images and 1,000 validation images. In its validation set, there are 345 night-time images. Other images are from day-time, but are still, to a certain extend, varied in terms of image appearance.

2) **Mapillary Vistas [43]:** It has 66 object categories and 37 instance-specific labels. Following former works [2] for semantic segmentation under adverse conditions, in our experiments, these labels are merged into 19 categories same as CityScapes. Overall, there are 18,000 training samples and 2,000 validation samples. Its validation set has 62 twilight and night images to evaluate the performance under all-day scenarios.

3) **Adverse Conditions Dataset With Correspondences [50]:** It has 19 semantic categories, the same as CityScapes [57] and BDD100K [45]. This dataset provides 3,000 day-time images (in fog, rain and snow conditions) and 1,006 night-time images. The training, validation and test set are split by a ratio of 4:1:5. Although ACDC can be jointly used with other dataset for day-night domain adaption, these settings are beyond the scope of our problem definition in Sec. III.

4) **All-Day CityScapes:** To bridge the dataset gap of full-day scenarios, we create a dataset that contains a more complete description of all-day rather than only contains twilight and night scenarios. Hence, we use the latest SkyAR tool [58] to render the CityScapes dataset [57], so that the fine-grained ground truth of CityScapes is inherited. Details on how the SkyAR rendering package works can be found in [58]. The steps are as follows: 1) setting a certain day/night scenario and finding the sky regions as background, 2) matting the background of the target image, 3) fusing the background of the target image by rendering, and the key lighting condition is derived from the sky regions, and 4) post-processing.

Specially, from dawn to night, nine scenarios from six different time periods, i.e., dawn, morning, daytime, sunset, dusk, night, are provided (See Fig. 6 for the sample distribution of each scenario in our All-day CityScapes dataset. Note that, for daytime, sunset and night cases, two scenarios are provided to enlarge diversity and are labeled as ‘-1’ and ‘-2’ respectively). Same as CityScapes, there are 2,975 training samples and 500 testing samples with 19 semantic categories.

#### B. Implementation Details

All our experiments are implemented on a work station with an Intel® Core™ i7-10700K CPU and two GeForce RTX 2080 SUPER GPUs. For our AO-SegNet, the SGD optimizer is used with an initial learning rate of 0.003, momentum of
C. Comparing With Existing Methods

1) Compared Methods: Twelve directly-supervised segmentation models (FCN [59], SegNet [25], FPN [24], Mask RCNN [28], ICNet [27], DeepLab-V3 [29], DDRNet-23 [6], CCNet [60], DANet [61], OCR [30], Segmenter [31], Mask2Former [32]) and four domain generalization based segmentation methods (SGSN [18], Iternorm [16], SW [17], IBNet [15], RobustNet [2]) are used for comparison with our AO-SegNet. For fair comparison, the pre-trained parameters of the backbone from ImageNet are also used for initialization, and the rest components’ parameters are randomly initialized for training. All the default data augmentation on DDRNet-23 is kept for fair comparison.

2) Comparison on All-Day CityScapes: For all experiments on our All-day CityScapes benchmark, official codes with default parameter settings are used. The performance of all these methods on our synthetic All-day CityScapes dataset is reported in Tab. II. It is shown that:

(1) Our AO-SegNet outperforms all the compared methods for three categories, namely, direct supervision (denoted as DS), all-day segmentation (denoted as AD) and domain generalization (denoted as DG).

(2) For each of the backbone that AO-SegNet uses, it outperforms the other methods that use the same backbone by 9.5%, 5.8%, 1.9%, 1.2%, 1.5%, respectively.

(3) Regarding the domain generalized methods, RobustNet [2] also achieves satisfactory performance with a 72.7% mIoU. However, the rest of these methods demonstrates inferior performance.

(4) CNN based directly-supervised segmentation models generally demonstrate inferior performance compared to domain generalization based segmentation methods, but ViT based directly-supervised methods (Segmenter [31], Mask2Former [32]) generally show superior performance.

From the results, the challenges for all-day scenarios can only be partially addressed by domain generalization based solutions, as the problem definition and formulation of domain generalization is actually different from the all-day definition and formulation. Specially, as discussed in Sec III-B, under domain generalization formulation, the semantics in each domain can contain rather different information from other domains. In contrast, the semantic information under all-day scenarios are more uniform and are mainly impacted by the dependency on the environment. Hence, the all-day representation learning for semantic segmentation actually demands stronger generalization capabilities than domain generalization based semantic segmentation.

On the other hand, although directly supervised segmentation can be another option towards the all-day semantic segmentation task, the varied image appearance caused by the highly dynamic environment, actually poses a great challenge.
for them to derive semantic information, as indicated by the results in Table II. This observation is particularly obvious for CNN based segmentation models. However, owing to the stronger semantic representation, ViT based segmentation methods (Segmenter [31], Mask2Former [32]) can somewhat alleviate this problem, but are still inferior to the proposed AO-SegNet by 2.7% and 1.6%, respectively.

To understand the performance of these methods, Fig. 7 (a) and (b) show the results of our AO-SegNet and four state-of-the-art methods on per-scenario from all-day and eight safety-critical semantic categories. It is shown that existing segmentation models and domain generalization methods already show some tolerance to different scenarios. However, their performance on multiple semantic categories is far from satisfactory. In contrast, our AO-SegNet improves the semantic understanding capability of these semantic categories against existing methods.

3) Comparison on Mapillary Vistas: Tab. III lists the segmentation results of our AO-SegNet and all the other compared methods on the real-world dataset Mapillary Vistas. For boarder comparison, not only the all-day semantic segmentation method, but also the directly-supervised and domain generalization segmentation methods, are all included for comparison.

Overall, our AO-SegNet outperforms all common methods. Specifically, it achieves mIoU of 72.9%, 69.5%, 73.2%, 74.8% and 76.0% when using DDRNet-23, ResNet-50, ResNet-101, ViT-B and Swin-base backbone respectively, surpassing the second best method using the same backbone by 4.3%, 5.8%, 1.9%, 2.7%, 1.6%, respectively. On the other hand, all the compared domain generalization based segmentation methods perform rather similarly on this benchmark. As discussed in Sec. III-B, the problem formulation of all-day and domain generalization for semantic segmentation is different. Despite the capability of perceiving unknown styles from other domains by domain generalization based methods, their generalization capability on all-day scenarios is still limited. This is also reflected by the performance of other all-day semantic segmentation method SCGN [18], as it also outperforms these domain generalization based methods.

4) Comparison on BDD100K: Tab. IV lists the segmentation results of our AO-SegNet and the other compared methods on another real-world dataset, BDD100K. For boarder comparison, not only the all-day semantic segmentation method, but also the directly-supervised and domain generalization segmentation methods, are all included for the comparison. Overall, our AO-SegNet outperforms all these methods with an obvious performance improvement. Specifically, it achieves mIoU of 57.3%, 55.4%, 60.1%, 61.3% and 66.2% when using DDRNet-23, ResNet-50, ResNet-101, ViT-B and Swin-base backbone respectively, surpassing the second best method using the same backbone by 4.9%, 3.9%, 1.9%, 1.6%, 2.6%, respectively.

The experimental observations on BDD100K are actually quite similar with the observations on Mapillary Vistas. For the explanations, please refer to the discussion on Mapillary Vistas for details.

5) Comparison on ACDC: Table V reports the performance of the proposed AO-SegNet and all the compared methods on ACDC dataset. Similar to the observation from the other three datasets, the proposed AO-SegNet also achieves the state-of-the-art performance. Specifically, it achieves mIoU of 69.8%, 67.2%, 70.9%, 74.5% and 77.4% when using DDRNet-23, ResNet-50, ResNet-101, ViT-B and Swin-base backbone respectively, surpassing the second best method using the same backbone by 4.3%, 7.7%, 4.6%, 2.2%, 1.2%, respectively. On the other hand, the domain generalized segmentation methods (Iternorm [16], SW [17], IBNet [15], RobustNet [2]) show great performance drop on this benchmark.

The samples in ACDC dataset are not only from nighttime, but also from adverse conditions such as rain, snow and haze. The state-of-the-art performance on ACDC dataset also indicates the strong feature representation capability of the proposed AO-SegNet and the interactive learning scheme.

**Fig. 7.** Per-scenario (a) and per-category (b) segmentation results of our AO-SegNet and four state-of-the-art methods on All-day CityScapes dataset. Evaluation metric: Mean Intersection over Union (mIoU), in percentage (%).
The framework intends to refine the intrinsic and extrinsic representation, and it may also benefit the semantic representation under other adverse conditions. However, the representation learning under adverse weather conditions is beyond the scope of this paper.

### TABLE IV

<table>
<thead>
<tr>
<th>Method</th>
<th>Category</th>
<th>Backbone</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN [59]</td>
<td></td>
<td>VGGNet-16</td>
<td>48.4</td>
</tr>
<tr>
<td>SegNet [25]</td>
<td></td>
<td>VGGNet-16</td>
<td>48.8</td>
</tr>
<tr>
<td>PPN [54]</td>
<td></td>
<td>ResNet-101</td>
<td>49.6</td>
</tr>
<tr>
<td>Mask RCNN [28]</td>
<td></td>
<td>ResNet-50</td>
<td>50.9</td>
</tr>
<tr>
<td>ICNet [27]</td>
<td></td>
<td>PSPNet-50</td>
<td>47.3</td>
</tr>
<tr>
<td>DeepLabV3 [29]</td>
<td>DS</td>
<td>ResNet-101</td>
<td>53.8</td>
</tr>
<tr>
<td>DDRNet [6]</td>
<td></td>
<td>DDRNet-23</td>
<td>52.4</td>
</tr>
<tr>
<td>CCNet [60]</td>
<td></td>
<td>ResNet-101</td>
<td>56.0</td>
</tr>
<tr>
<td>DANet [61]</td>
<td></td>
<td>ResNet-101</td>
<td>56.5</td>
</tr>
<tr>
<td>Segnetener [31]</td>
<td></td>
<td>ViT-B</td>
<td>59.7</td>
</tr>
<tr>
<td>Mask2Former [32]</td>
<td></td>
<td>Swin-base</td>
<td>63.6</td>
</tr>
<tr>
<td>SGSN [18]</td>
<td>AD</td>
<td>ResNet-50</td>
<td>51.5</td>
</tr>
<tr>
<td>Iternorm [16]</td>
<td></td>
<td>ResNet-50</td>
<td>49.2</td>
</tr>
<tr>
<td>AO-SegNet (Ours)</td>
<td>AD</td>
<td>DDRNet-23</td>
<td>57.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ResNet-50</td>
<td>55.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ResNet-101</td>
<td>60.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ViT-B</td>
<td>61.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Swin-base</td>
<td>66.2</td>
</tr>
</tbody>
</table>

The above parameter comparison demonstrates that by using each CNN or ViT based backbone, the proposed AO-SegNet framework does not result in a substantial parameter number increase.

### TABLE V

<table>
<thead>
<tr>
<th>Method</th>
<th>Category</th>
<th>Backbone</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN [59]</td>
<td></td>
<td>VGGNet-16</td>
<td>51.3</td>
</tr>
<tr>
<td>SegNet [25]</td>
<td></td>
<td>VGGNet-16</td>
<td>54.2</td>
</tr>
<tr>
<td>PPN [54]</td>
<td></td>
<td>ResNet-101</td>
<td>58.4</td>
</tr>
<tr>
<td>Mask RCNN [28]</td>
<td></td>
<td>ResNet-50</td>
<td>59.5</td>
</tr>
<tr>
<td>ICNet [27]</td>
<td></td>
<td>PSPNet-50</td>
<td>61.8</td>
</tr>
<tr>
<td>DeepLabV3 [29]</td>
<td>DS</td>
<td>ResNet-101</td>
<td>64.5</td>
</tr>
<tr>
<td>DDRNet [6]</td>
<td></td>
<td>DDRNet-23</td>
<td>66.6</td>
</tr>
<tr>
<td>CCNet [60]</td>
<td></td>
<td>ResNet-101</td>
<td>63.2</td>
</tr>
<tr>
<td>DANet [61]</td>
<td></td>
<td>ResNet-101</td>
<td>64.1</td>
</tr>
<tr>
<td>Segnetener [31]</td>
<td></td>
<td>ViT-B</td>
<td>72.3</td>
</tr>
<tr>
<td>Mask2Former [32]</td>
<td></td>
<td>Swin-base</td>
<td>76.2</td>
</tr>
<tr>
<td>SGSN [18]</td>
<td>AD</td>
<td>ResNet-50</td>
<td>62.4</td>
</tr>
<tr>
<td>Iternorm [16]</td>
<td></td>
<td>ResNet-50</td>
<td>46.3</td>
</tr>
<tr>
<td>SW [17]</td>
<td></td>
<td>ResNet-50</td>
<td>46.6</td>
</tr>
<tr>
<td>RobustNet [2]</td>
<td></td>
<td>DDRNet-23</td>
<td>69.2</td>
</tr>
<tr>
<td>AO-SegNet (Ours)</td>
<td>AD</td>
<td>DDRNet-23</td>
<td>69.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ResNet-50</td>
<td>67.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ResNet-101</td>
<td>70.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ViT-B</td>
<td>74.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Swin-base</td>
<td>77.4</td>
</tr>
</tbody>
</table>

6) Parameter Number Comparison: Table II also compares the parameter number of each segmentation method. The proposed AO-SegNet: 1) has only 29.15M parameters when using the DDRNet-23 backbone [6], which is much less than other methods; 2) has close or even less parameter number than other segmentation methods when using ResNet-50/ResNet-101 backbone, but has a performance gain (> 5% mIoU); 3) does not result in a gain of parameter numbers when using ViT or Swin-T as the backbone. Specially, compared to the latest Mask2Former [32], a 0.46M parameter gain leads to a 1.5% mIoU performance gain.

D. Ablation Studies

1) Impact of Each Step: Table. VI reflects the impact of each component for our AO-SegNet, i.e., IEE (I-learner and E-learner), IEI and IEM step, along with two loss terms. From the experimental results, it is shown that:
Fig. 10. Visualized heatmap of the intrinsic and extrinsic representations on the All-day CityScapes dataset. Despite the day-night shift, the intrinsic representations are focusing on the salient objects, while the activations of extrinsic representations can vary and focus more on the sky, lighting sources and object surfaces.

Fig. 11. Visualized samples on All-day CityScapes dataset. Iternorm [16]; SW [17]; IBNet [15]; RobustNet [2]; SGSN [18]; Ours: AO-SegNet. From first to sixth row, sunset-2, night-1, night-2, sunset-1, daytime-2 and dawn scenario. Safety-crucial errors are labeled in brown bounding boxes.

(1) I-learner in IEE leads to an obvious performance gain, as the PMSM module uses multi-scale dilated convolutions to mine the semantic information.

(2) E-learner in IEE has a marginal performance gain of 0.2% as it uses a $1 \times 1$ convolutional layer to refine the extrinsic representation. For convolutional neural networks, the $1 \times 1$ convolutional layer is effective to enhance the feature representation with little additional computational cost [52], [53], [54], [55].

(3) IEI step improves the performance by 1.8%. It is a key step to refine the intrinsic and extrinsic representations for all-day scenarios.

(4) The IEM step with a single segmentation head leads to a performance gain of 0.6%.

(5) The full IEM with the semantic latent space for semantic alignment further results in a performance gain of 0.6%.

Each component in our framework positively contributes to the segmentation performance under all-day scenarios. Specially, the I-learner extensively mines the context information from the low-resolution branch and the E-learner refines the high-resolution branch. Hence, the performance gain by I-learner is more obvious than E-learner. On the other hand, our IEM component shows a performance gain compared to naively using the segmentation head in existing segmentation
models. It further constraints the intrinsic representation to align to the same semantic information in latent space. Finally, the key step IEI leads to a significant performance gain for all the components.

2) Impact of Each Component in IEI: A more detailed ablation study on the IEI step is provided in Table VII to investigate the impact of each component on the segmentation performance. It is shown that:

1) The dual-attention (DA) backbone leads to a 0.7% performance gain.

(2) Selecting extrinsic features contributing positively to the intrinsic (E2I) leading to a performance gain of 0.5% while propagating the intrinsic representation (I2E) to extrinsic features provides a performance gain of only 0.2%.

(3) The skip connection further enhances the feature propagation and contributes to a performance gain of about 0.2%.

The spatial-wise attention, applied on both intrinsic and extrinsic branches, improves the segmentation results. This is because the attention mechanism emphasizes on the key local information to refine the semantic representation. However, the spatial weight matrix fusion from the extrinsic branch to the intrinsic branch plays a more important role on the all-day semantic segmentation performance than fusing the spatial-wise information from the intrinsic branch to the extrinsic branch. This is because the environment dependency is not a major factor to determine the most of the pixel-wise semantic prediction in an image.

E. Understanding the Intrinsic & Extrinsic Representation

1) t-SNE Visualization: Fig. 8 (a) and (b) shows the t-SNE visualization of the intrinsic and extrinsic features before and after learnt by the interactive intrinsic-extrinsic learning. The proposed learning scheme significantly improves the separation between intrinsic and extrinsic representation.

2) Heatmap from Intrinsic-Extrinsic Representation: To better understand the separated intrinsic and extrinsic representation, they are resized and displayed on the original image for visualization. Fig. 9 shows the visualization on some samples from the All-day CityScapes dataset. It can be seen that the feature activation from intrinsic representation is very similar to the spatial attention mechanism. It highlights the key objects in an image, which strongly relates to the semantic representation. In contrast, the feature activation of extrinsic representation is very distributed and scattered.
Fig. 13. More visualized samples on All-day CityScapes dataset. Iternorm [16]; SW [17]; IBNet [15]; RobustNet [2]; SGSN [18]; Ours: AO-SegNet. From first to sixth row, sunset-2, night-1, night-2, sunset-1, daytime-2 and dawn scenario. Safety-crucial errors are labeled in brown bounding boxes.

It usually highlights the regions such as sky and the object surface, which may be strongly correlated to the extrinsic properties such as lighting.

3) Intrinsic-Extrinsic Representation from All-Day: A qualitative analysis is investigated. For a certain sample, we use the SkyAR tool [58] to render its appearance under different times of a day. Then, the pre-trained AO-SegNet is directly used to infer the intrinsic representation and extrinsic representation. The visualization is shown in Fig. 10. Under the all-day scenarios, the activation of the intrinsic representation is stable, focusing on the key objects in an image. In contrast, under the all-day shift, the activation of extrinsic representation can be rather inflexible on the background regions of an image. Based on the above analysis, the intrinsic representation highlights the key semantic information in an image, while the extrinsic representation is more sensitive to the background and environment. These observations are consistent to the intrinsic and extrinsic property formulation in Sec. III.

F. Qualitative Segmentation Prediction

1) Results on All-Day CityScapes: Fig. 11 shows qualitative segmentation results on All-day CityScapes for a variety of scenarios. Overall, our AO-SegNet demonstrates a consistent segmentation prediction for dawn, day-time, sunset and night scenarios. Also, the semantic prediction of small objects, such as traffic lights and traffic signs, are also improved by our AO-SegNet even for day-time scenarios.
Existing domain generalization based methods still demonstrate some qualitative flaws on sunset, dawn or night scenarios, e.g., omitting the semantic prediction for e.g. traffic signs and traffic lights. Among these methods, SGSN and RobustNet show relatively stable predictions.

2) Results on Mapillary Vistas and BDD100K: Fig. 12 shows qualitative predictions on twilight, night and day-time scenarios from Mapillary Vistas and BDD100K respectively. As discussed in Tab. I, most samples in Mapillary Vistas and BDD100K are still day-time samples. The twilight and night samples from these two datasets are rare. Such typical scenarios have been deliberately selected in both figures.

Overall, our AO-SegNet provides the best visual prediction quality on these real-world scenarios. Similar to the results for All-day CityScapes, domain generalization based segmentation methods all demonstrate, to a certain extent, inaccurate segmentation prediction.

3) More Segmentation Results: In Appendix A, more segmentation predictions on All-day CityScapes and BDD100K datasets are provided in Fig. 13 and 14, respectively. The proposed AO-SegNet shows a more precise prediction on a variety of all-day scenarios on all these benchmarks, when compared with the existing directly-supervised and domain generalized semantic segmentation methods.

VI. Conclusion

In this paper, all-day semantic segmentation is considered. The definition of this task and its difference with
domain generalization is discussed. To address all-day semantic segmentation, we formulated that each pixel in image is determined by both intrinsic and extrinsic properties. We designed an intrinsic-extrinsic interactive learning strategy to generate semantic predictions on all-day scenarios.

The proposed intrinsic-extrinsic interactive learning strategy is included in an end-to-end by our new All-in-One Segmentation Network (AO-SegNet). To bridge the dataset gap for all-day semantic segmentation, we created an All-Day CityScapes dataset on top of the CityScapes dataset.

Future works include: 1) designing a larger all-day benchmark for all-day semantic segmentation, which contains both photographed and rendered real-world scenarios; 2) investigating robust semantic segmentation predictions under not only all-day, but also more diverse all-weather scenarios, e.g., snow, haze, rain.

APPENDIX A
MORE VISUALIZED SEGMENTATION PREDICTIONS

More segmentation predictions on All-Day CityScapes and BDD100K datasets are provided in Fig. 13 and 14, respectively. The proposed AO-SegNet shows a more precise prediction on a variety of scenarios in a day on all these benchmarks, when compared with the existing directly-supervised and domain generalized semantic segmentation methods.

ACKNOWLEDGMENT

The authors would like to thank the editors and the anonymous three reviewers, whose insightful suggestions and comments significantly improved their manuscript.

REFERENCES


[44] Shaodi You (Senior Member, IEEE) received the bachelor’s degree from Tsinghua University, China, in 2009, and the M.E. and Ph.D. degrees from the University of Tokyo, Japan, in 2012 and 2014, respectively. He is an Assistant Professor with the University of Amsterdam (UvA), The Netherlands. He is also with the Computer Vision Research Group, Faculty of Science, Institute of Informatics. His research interests include physics based vision, perception based vision and learning, and 3D geometry. He is best known for his work on visual enhancement for rainy days. He is the Area Chair of CVPR2023, CVPR2022, Neurips2022, ECCV2022, and ECCV2020; and the Program Chair of ICCV2017, ICCV2019, and ICCV2021 Workshop on Physics Based Vision meets Deep Learning. He is an Associate Editor of IJCV.


Qi Bi (Student Member, IEEE) received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree with the Computer Vision Research Group, University of Amsterdam, Amsterdam, The Netherlands. His research interests include robust vision in bad weather, domain generalization, and self-supervised learning. He was awarded as an Outstanding Reviewer of CVPR 2023. His works were shortlisted for CVPR 2021 Best Paper Candidates and MICCVAI 2021 Travel Award. He serves as a regular reviewer for some top-tier journals, including IJCV, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING; and some top-tier conferences, including CVPR, ICCV, NeurIPS, and ECCV. He is an organizer, the general chair, and a program committee member of a various number of conferences; and an invited speaker at major conferences.

Theo Gevers is a Professor of computer vision with the University of Amsterdam. He is also the Director of the Computer Vision Laboratory and the Co-Director of the Atlas Lab (UvA-TomTom) and Delta Lab (UvA-Bosch), Amsterdam. He has published over 250 papers and three books. His research interests include artificial intelligence, with a focus on computer vision and deep learning, in particular image processing, 3D (object) understanding, and human-behavior analysis with industrial and societal applications. He is the Co-Founder of 3Duniversum.

He is an organizer, the general chair, and a program committee member of a various number of conferences; and an invited speaker at major conferences.