Dynamic Loss Balancing and Sequential Enhancement for Road-Safety Assessment and Traffic Scene Classification

Marin Kačan, Marko Ševrović, and Siniša Šegvić, Member, IEEE

Abstract—Road-safety inspection is an indispensable instrument for reducing road-accident fatalities related to road infrastructure. Recent work formalizes the assessment procedure in terms of carefully selected risk factors that are also known as road-safety attributes. In current practice, these attributes are manually annotated in geo-referenced monocular video for each road segment. We propose to reduce dependency on tedious human labor by automating attribute collection through a two-stage deep learning approach. The first stage recognizes more than forty road-safety attributes by observing a local spatio-temporal context. Our design leverages an efficient convolutional pipeline, which benefits from pre-training on semantic segmentation of street scenes. The second stage enhances predictions through sequential integration across a larger temporal window. Our design leverages per-attribute instances of a lightweight recurrent architecture. Both stages alleviate extreme class imbalance by incorporating a multi-task variant of recall-based dynamic loss weighting. We perform experiments on the novel iRAP-BH dataset, which involves fully labeled geo-referenced video along 2,300 km of public roads in Bosnia and Herzegovina. Moreover, we evaluate our approach against the related work on three road-scene classification datasets from the literature: Honda Scenes, FM3m, and BDD100k. Experimental evaluation confirms the value of our contributions on all three datasets.

Index Terms—Image classification, road safety, iRAP attributes, deep learning, multi-task learning.

I. INTRODUCTION

ROAD accidents are a significant public health problem that causes more than 1.3 million deaths every year [1]. With its Global Plan for the Decade of Action for Road Safety [2], the UN has committed to halving the number of road traffic deaths and injuries by the end of this decade.

The holistic Safe System Approach [3], [4] identifies road infrastructure as one of the five pillars of traffic safety [5], [6]. Many earlier approaches to road-safety assessment estimate high-risk road sections (black spots) from historical data [7]. These approaches are reactive in nature since they require accidents to occur for a section to be deemed dangerous [8], [9]. In contrast, a proactive approach to road safety formalizes the safety in terms of road-infrastructure attributes. Besides being able to operate when historical data are absent, this approach also provides an insight into specific requirements to improve the safety rating. While road-safety attributes can be assessed via on-site surveys [10], off-line assessments are usually more practical and cost-effective [11], [12]. Currently, these assessments are performed manually by trained operators [13]. Automating this time-consuming process is a step toward reduced costs and increased scalability.

This work presents a two-stage visual recognition approach for assessing road-safety attributes in monocular video. Our models operate on street-level imagery by complementing local recognition and sequential enhancement with dynamic loss weighting. We implement local recognition as a multi-task [14] model with shared latent representation and per-attribute classification heads. We facilitate high-quality latent representation by pre-training on a large street-level semantic segmentation dataset [15]. Our sequential model enhances local predictions by providing access to a larger temporal context. It consists of per-attribute bidirectional LSTM models in order to account for attribute-specific temporal behavior. We address the extreme imbalance of training data by proposing a multi-task variant of dynamic loss weighting with respect to per-class recall [16]. Such practice benefits both local recognition and sequential enhancement.

We evaluate our approach on a novel dataset for road-safety assessment, which we denote as iRAP-BH. We have acquired the dataset along 2,300 km of public roads in Bosnia and Herzegovina. It covers 230,000 10-meter road segments annotated with all 52 iRAP attributes [13]. The proposed approach has no handcrafted attribute-specific rules. Hence, it could contribute to different road-safety programmes and visual event recognition problems. We demonstrate this through performance evaluation on three public road-driving classification datasets: Honda Scenes [17], FM3m [18], and BDD100k [19]. Experiments reveal that our approach outperforms the state of the art on all tasks of Honda Scenes. The greatest improvements occur on...
the Road place task which contains temporal annotations and highly imbalanced classes. Our models also perform very well on FM3m and BDD100k, even though the former is fairly well balanced and both are unsuitable for multi-frame recognition.

The contributions of this paper are as follows. Firstly, building on our preliminary research [20], we enhance attribute classification performance through pre-training on the task of semantic segmentation. This favours extraction of localized semantic features from street scenes. Empirical evidence shows a clear advantage with respect to other forms of pre-training, and highlights the importance of nuanced visual representations for road-attribute recognition. Secondly, we advance beyond our preliminary work [20] by addressing all road-safety attributes that can be visually assessed, as well as all of their fine-grained classes. Thirdly, we mitigate class imbalance through multi-task dynamic loss weighting. Our policy dynamically adjusts class weights with respect to the corresponding recall scores. Furthermore, it normalizes per-attribute loss components in order to prevent undesired cross-attribute interference. This enables effective joint learning across many attributes, even in the presence of extreme class imbalances. Fourthly, we enhance our local predictions with per-attribute sequential models. These models can learn improved attribute-specific temporal patterns by observing a wider temporal context. We provide thorough ablation studies which demonstrate the impact of our contributions.

The remainder of this paper is structured as follows. Section II surveys the related work in the areas of road-safety assessment and road-driving image classification. Section III introduces and analyzes a widely used collection of road-safety attributes. We describe our proposed approach and its technical contributions in Section IV. Section V outlines the datasets utilized in our experiments. Section VI describes the conducted experiments and analyzes the results. Finally, Section VII draws conclusions from our work, highlights the significance of our contributions and considers their implications for future research.

II. RELATED WORK

A. Road Infrastructure Safety

Traditional approaches to road-infrastructure safety assessment are based on historical accident statistics. Time, location, severity, type, and other data [21] about previous crashes can be used to identify black spots [22], create crash-risk maps [23], and design crash-prediction models [24]. These approaches can capture complex and non-trivial risk factors that history-agnostic approaches might overlook [9]. However, these approaches are reactive since they require accidents to occur for road sections to be assessed as unsafe [25]. Moreover, historical estimation tends to produce high-variance predictions due to the sparsity of accidents [26].

Proactive approaches to safe road infrastructure rely on periodic inspections of static features. The International Road Assessment Programme (iRAP) Star Rating [11] is an internationally accepted standard for assessing and improving the safety of road infrastructure. Empirical evidence shows that it can significantly reduce the number of traffic fatalities and serious injuries by identifying high-risk roads across many countries [6], [12]. The iRAP Rating assesses the in-built safety of road segments according to 52 attributes [13] that we review in Section III.

B. Computer Vision for Road-Infrastructure Assessment

Computer vision has been applied to detect and recognize various elements of road infrastructure such as traffic signs [27], [28], road surface markings [29], [30], [31], fleet management attributes [18], etc. Outputs from detection and segmentation models [32], [33], [34], [35] have been utilized as intermediate results to recognize certain road-safety attributes with rule-based systems [36], [37], [38]. For instance, Yi et al. [39] leverage active learning to detect road-safety elements such as guardrails and utility poles. Ni et al. [40] use an object detection network to extract local features to augment a global feature-extracting convolutional network for road-place classification. These approaches can benefit from more advanced data augmentation procedures such as CutMix [41] or Mosaic [42], which have been beneficial for detection performance in road and railway environments [43].

In contrast, our local recognition pipeline aims to recognize road-safety attributes directly from input data, in an end-to-end image-wide manner [20]. This avoids error accumulation and produces better latent representations. In this vein, the authors of the Honda Scenes dataset [17] present a baseline approach for infrastructure-related event detection in street video. They pre-train the ResNet-50 backbone on Places365 and leverage a frozen semantic segmentation model to ignore traffic participants. The pipeline concludes by recurrent processing of frozen convolutional features and standard softmax classification. Unlike their recurrent task-agnostic temporal region proposals and subsequent classification into events, our method enhances existing convolutional predictions for specific tasks using recurrent processing. Context MTL [44] addresses recognition on Honda Scenes with a multi-task architecture that is related to our work [20]. They regularize the loss with a lower bound of mutual information between the input and the latent-space features according to the Jensen-Shannon divergence [45]. Multi-Task Attention Network (MTAN) [46] uses a shared WideResNet backbone to generate task-specific attention masks for dense prediction. It dynamically selects relevant features from the shared global feature map for each task. The network is trained using a combined loss function where task-specific losses are weighted according to their current rate of change. None of the described approaches address class imbalance with multi-task dynamic loss weighting, nor use semantic-segmentation pre-training.

C. Learning on Imbalanced Data

Class imbalance refers to non-uniform class proportions within training data [47]. Single-task setups often mitigate imbalance with data-level strategies such as oversampling rare classes and undersampling frequent ones [32], [48]. However, these methods are less feasible in multi-task scenarios [44], [46] where tasks are uncorrelated and non-uniformly balanced and both are unsuitable for multi-frame recognition.

The contributions of this paper are as follows. Firstly, building on our preliminary research [20], we enhance attribute classification performance through pre-training on the task of semantic segmentation. This favours extraction of localized semantic features from street scenes. Empirical evidence shows a clear advantage with respect to other forms of pre-training, and highlights the importance of nuanced visual representations for road-attribute recognition. Secondly, we advance beyond our preliminary work [20] by addressing all road-safety attributes that can be visually assessed, as well as all of their fine-grained classes. Thirdly, we mitigate class imbalance through multi-task dynamic loss weighting. Our policy dynamically adjusts class weights with respect to the corresponding recall scores. Furthermore, it normalizes per-attribute loss components in order to prevent undesired cross-attribute interference. This enables effective joint learning across many attributes, even in the presence of extreme class imbalances. Fourthly, we enhance our local predictions with per-attribute sequential models. These models can learn improved attribute-specific temporal patterns by observing a wider temporal context. We provide thorough ablation studies which demonstrate the impact of our contributions.

The remainder of this paper is structured as follows. Section II surveys the related work in the areas of road-safety assessment and road-driving image classification. Section III introduces and analyzes a widely used collection of road-safety attributes. We describe our proposed approach and its technical contributions in Section IV. Section V outlines the datasets utilized in our experiments. Section VI describes the conducted experiments and analyzes the results. Finally, Section VII draws conclusions from our work, highlights the significance of our contributions and considers their implications for future research.

II. RELATED WORK

A. Road Infrastructure Safety

Traditional approaches to road-infrastructure safety assessment are based on historical accident statistics. Time, location, severity, type, and other data [21] about previous crashes can be used to identify black spots [22], create crash-risk maps [23], and design crash-prediction models [24]. These approaches can capture complex and non-trivial risk factors that history-agnostic approaches might overlook [9]. However, these approaches are reactive since they require accidents to occur for road sections to be assessed as unsafe [25]. Moreover, historical estimation tends to produce high-variance predictions due to the sparsity of accidents [26].

Proactive approaches to safe road infrastructure rely on periodic inspections of static features. The International Road Assessment Programme (iRAP) Star Rating [11] is an internationally accepted standard for assessing and improving the safety of road infrastructure. Empirical evidence shows that it can significantly reduce the number of traffic fatalities and serious injuries by identifying high-risk roads across many countries [6], [12]. The iRAP Rating assesses the in-built safety of road segments according to 52 attributes [13] that we review in Section III.

B. Computer Vision for Road-Infrastructure Assessment

Computer vision has been applied to detect and recognize various elements of road infrastructure such as traffic signs [27], [28], road surface markings [29], [30], [31], fleet management attributes [18], etc. Outputs from detection and segmentation models [32], [33], [34], [35] have been utilized as intermediate results to recognize certain road-safety attributes with rule-based systems [36], [37], [38]. For instance, Yi et al. [39] leverage active learning to detect road-safety elements such as guardrails and utility poles. Ni et al. [40] use an object detection network to extract local features to augment a global feature-extracting convolutional network for road-place classification. These approaches can benefit from more advanced data augmentation procedures such as CutMix [41] or Mosaic [42], which have been beneficial for detection performance in road and railway environments [43].

In contrast, our local recognition pipeline aims to recognize road-safety attributes directly from input data, in an end-to-end image-wide manner [20]. This avoids error accumulation and produces better latent representations. In this vein, the authors of the Honda Scenes dataset [17] present a baseline approach for infrastructure-related event detection in street video. They pre-train the ResNet-50 backbone on Places365 and leverage a frozen semantic segmentation model to ignore traffic participants. The pipeline concludes by recurrent processing of frozen convolutional features and standard softmax classification. Unlike their recurrent task-agnostic temporal region proposals and subsequent classification into events, our method enhances existing convolutional predictions for specific tasks using recurrent processing. Context MTL [44] addresses recognition on Honda Scenes with a multi-task architecture that is related to our work [20]. They regularize the loss with a lower bound of mutual information between the input and the latent-space features according to the Jensen-Shannon divergence [45]. Multi-Task Attention Network (MTAN) [46] uses a shared WideResNet backbone to generate task-specific attention masks for dense prediction. It dynamically selects relevant features from the shared global feature map for each task. The network is trained using a combined loss function where task-specific losses are weighted according to their current rate of change. None of the described approaches address class imbalance with multi-task dynamic loss weighting, nor use semantic-segmentation pre-training.

C. Learning on Imbalanced Data

Class imbalance refers to non-uniform class proportions within training data [47]. Single-task setups often mitigate imbalance with data-level strategies such as oversampling rare classes and undersampling frequent ones [32], [48]. However, these methods are less feasible in multi-task scenarios [44], [46] where tasks are uncorrelated and non-uniformly balanced and both are unsuitable for multi-frame recognition.
imbanced. In such cases, an image marked as rare in one task might simultaneously be frequent in another, which complicates the direct application of oversampling. A possible alternative would be to construct dedicated oversampled datasets for each task and then train the multi-task model by cyclically optimizing tasks in a round-robin fashion. However, our preliminary experiments indicate that round-robin training underperforms with respect to standard training by a wide margin. Consequently, we choose to address multi-task class imbalance by a custom loss weighting approach.

Several algorithmic-level and cost-based approaches try to adapt learning algorithms by assigning larger loss weights to misclassified examples of underrepresented classes [49], [50], [51], [52]. However, simple implementations of inverse-frequency loss-weighting improve recall at the expense of precision [16]. Tian et al. [16] address this issue by introducing dynamic assignment of class weights. During training, the class weights are dynamically set according to the current false negative rate of the corresponding class. This prevents the classes that achieve high recall from suffering excessive false positives. Unfortunately, this approach is not directly applicable in multi-task setups as we explain in Section IV-B.

D. Domain Shift

Robustness to domain shifts with regards to environment or driving conditions is important for real-world visual recognition systems [53], [54], [55]. Domain adaptation approaches have successfully been employed for cross-domain traffic scene recognition. Saffari et al. [56] perform traffic scene classification under varying weather conditions by combining a generative domain-invariant feature extractor, with nonlinear task-relevant dictionary learning. Di et al. [57] tackle semantic segmentation of rainy night-time scenes by introducing a near-scene semantic adaptation approach that leverages the corresponding daytime images. They minimize domain shift on the feature representation level and subsequently align the segmentation output space of the pre-trained daytime model with the rainy night-time domain. They evaluate the approach on a dataset with paired daytime and rainy night-time images.

Our road-safety attribute dataset iRAP-BH is not applicable for domain adaptation experiments, since it does not contain driving conditions annotations nor images from different countries. We aim to improve within-domain robustness through semantic segmentation pre-training, loss balancing and sequential enhancement.

E. Recurrent Models for Video Recognition

Long Short-Term Memory (LSTM) networks [58] have been used for video classification and action recognition [59]. They have also been utilized to enhance existing sequential predictions in speech recognition [60] and rainfall regression [61].

Inspired by temporal region proposal approaches [62], [63], Narayanan et al. [17] use an LSTM-based architecture to perform decoupled event proposal and traffic scene classification. Their two-stage approach generates task-agnostic event proposals as video intervals and subsequently classifies those into events from spatio-temporally pooled features. Since these two stages are decoupled, the second stage is treated as a single-task multi-class learning problem. Trabelsi et al. [64] extend the LSTM network with multi-head attention and combine it with a convolutional neural network to effectively capture and interpret complex dynamics of driver behaviour from traffic scenes. In contrast, we use recurrent processing to enhance existing convolutional predictions for a specific task, rather than for task-agnostic feature aggregation.

III. Road-Safety Attributes

The iRAP Star Rating quantifies the overall protection that road infrastructure provides to the four most common road user types [13]. The assessment procedure targets categorical values of a carefully selected set of 52 attributes related to road-infrastructure elements and roadside objects within the corresponding 100-meter or 10-meter road segment. Each attribute assumes a class from the attribute-specific taxonomy. The number of classes varies across attributes. Attributes with the most classes are those that encode the speed limit (21 classes), roadside severity (17 classes), and intersection type (16 classes). On the other hand, there are 11 binary attributes. Hence, we formulate attribute recognition as separate multi-class classification problems. We had to discard four attributes that assume only one class throughout our whole dataset: Shoulder rumble strips, Centre line rumble strips, Motorcycle facility, and Pedestrian fencing. We also had to discard five attributes that are not suitable for a visual recognition setup. These include the four speed limit attributes and Intersecting road volume. The latter captures the average daily number of vehicles that pass through a road segment from an intersecting road. This information can be estimated from road counters or aerial imagery. Consequently, our experiments address 43 iRAP attributes.

A. iRAP Attribute Groups

We provide a short overview of the seven attribute groups defined by the iRAP standard [13]. More details can be found in our preliminary work [20].

Road and context attributes (1 attribute) contain the attribute Carriageway label along with twelve attributes related to data acquisition and annotation metadata.

Observed flow attributes (5 attributes) record the flow of motorcycles, bicycles, and pedestrians through the segment.

Speed limit attributes (5 attributes) record the speed limits (4 attributes) and speed-reducing infrastructure such as speed bumps.

Mid-block attributes (16 attributes) focus on the road’s intrinsic features rather than its surroundings. The attribute Median type stands out as particularly challenging, as it requires distinguishing among 15 kinds of physical separators or median markings.

Roadside attributes (7 attributes) involve dual-side (passenger and driver) attributes that assess the risk of roadside features. The attribute Roadside Severity captures the most hazardous roadside object based on its type and proximity to the road. The ground truths for this attribute are assigned
according to the priority table [13] that ranks object and distance combinations according to the risk level.

Intersection attributes (5 attributes) capture various intersection characteristics. Most notably, the attribute Intersection type has 16 classes that cover different combinations of intersecting roads, signalization, and special features like roundabouts and railway crossings.

Vulnerable road-user facilities and land use attributes (13 attributes) detail the presence of pedestrian, cyclist, and motorcyclist amenities, as well as capturing the characteristics of the surrounding area (e.g. Area type, Land use, School zone).

B. Attribute Analysis

We provide a conceptual and empirical analysis of various aspects of iRAP attributes in our dataset.

1) Class Imbalance: Many attributes in our dataset suffer from class imbalance, which occurs when there is a significant disproportion among the number of examples belonging to different classes. Class imbalance can hamper the performance of accuracy-oriented classifiers, resulting in the minority classes being ignored [47]. We thoroughly study and address this issue through improved loss functions, training procedures and macro-F1 evaluation, as described in IV-B and VI-A.

2) Non-Orthogonal Design: Some iRAP attributes capture multiple features that seem orthogonal to each other, resulting in classes that are (nearly) cartesian products of different values of those features. For example, the attribute Skid resistance is supposed to capture the skidding resistance and the texture depth of the road surface. It covers two dimensions: whether the surface grip is low ("poor"), medium, or adequate; and whether the road is sealed or unsealed. This results in the attribute having the following 5 classes: Unsealed - poor, Unsealed - adequate, Sealed - poor, Sealed - medium, Sealed - adequate. The attribute might have been divided into two attributes: Sealed road (true/false) and Surface grip (poor/medium/adequate). An orthogonal formulation of these concepts would mitigate the issue of class imbalance since there would be fewer classes with more examples.

3) Fine-Grained and Visually Similar Classes: Some attributes have very fine-grained classes that can be visually very similar. For instance, the attribute Roadside severity contains classes that cover different types of safety barriers (metal, concrete, wire, motorcycle friendly) as well as a separate class for semi-rigid structures, such as various fences. These classes are well defined, but the many options make it easy to miss the correct answer and also exacerbate class imbalance by spreading the already infrequent examples over many classes.

In general, it might be advantageous to design attribute sets with fewer and more general classes. Such a decision would trade off some precision and specificity for improved recognition quality.

C. Temporal Behaviour

The videos in our dataset cover long road sections. A road section is composed of a sequence of successive 10-meter segments. We identify several distinct patterns of temporal attribute behavior along sequences. According to the iRAP standard, some attributes have a default “negative” class [13]. These attributes are usually concerned with capturing countable occurrences of various infrastructure elements, such as intersections or pedestrian crossings. The default class in these attributes is None, while the “positive” classes correspond to concrete realizations of that attribute (e.g. 3-leg intersection).

The iRAP standard mandates that any occurrence of a positive class should only be annotated in the segment closest to its occurrence. All other neighboring segments are annotated as the negative class. We call such attributes “single-peak” attributes.

Let us consider a segment that contains an occurrence of a positive class (a peak) and a few neighboring segments that appear immediately before it. The visual features that a model might use to recognize such an attribute will typically be present in the neighboring segments as well. For example, an intersection gradually becomes more and more visible in the segments leading up to the peak segment. A model that predicts an intersection in a neighboring segment is not entirely wrong, and it might be hard for it to discern which exact segment is the peak one.

The attribute Street lighting has an especially peculiar temporal behaviour. It is treated as a single-peak attribute when a single light post appears in isolation. On the other hand, for a sequence of light posts, street lighting should be recognized as present in all segments from the first to the last light post. Since the light posts in such sequences can be up to 100 meters apart, it may be difficult for a vision-based model to differentiate between single occurrences and sequences of street lights.

There is another subset of attributes that is opposite in nature to single-peak attributes. We call these attributes “smooth” since their classes rarely change and generally do not oscillate. They describe larger areas, environments, zones, or infrastructure features that are likely to remain unchanged in consecutive segments. Examples of such attributes include Area type, Road delineation, Carriageway label, etc.

Figure 3 shows examples of four iRAP attributes. It presents sequences of five frames from consecutive 10-meter segments, along with the ground truth labels and model predictions for the corresponding attribute. Only the third segment of row 1 is annotated with the positive class of the single-peak attribute Intersection type. Conversely, the two smooth attributes maintain constant ground truth label throughout rows 3 and 4. The ground truth label of Street lighting also does not change even though the discriminative visual features are not visible in certain segments.

D. Motivation for Sequential Enhancement

This subsection analyzes class co-occurrence along consecutive segments of iRAP-BH. For a given attribute A and a given pair of segments t and (t+1), the pair of corresponding classes (c_{A,t}, c_{A,t+1}) constitutes a single co-occurrence. For an attribute with n classes, we can construct an n x n co-occurrence matrix where the element (i, j) corresponds to the number of occurrences where c_{A,t} = i and c_{A,t+1} = j.
A. Recognition in the Local Spatio-Temporal Context

Figure 1 presents our convolutional architecture for multi-task visual recognition of road-safety attributes in street-level imagery. The architecture consists of a shared front-end and attribute-specific back-ends. The front-end starts with the ResNet-18 backbone which we pre-train for semantic segmentation on the Vistas dataset [15], [65]. The resulting features are subjected to Spatial Pyramid Pooling (SPP) with grid dimensions 6, 3, 2 and 1 [66]. The SPP module captures information at different scales and produces a shared fixed-size image-wide representation.

Each of the 43 attribute-specific back-ends starts with attention pooling $ATT_i$ [20] with respect to the learned attribute-specific query $q_i$. The resulting representation is concatenated with the shared SPP features into the single-frame descriptor that is fed to the corresponding prediction head $P(A_i|x)$.

This architecture is easily extended for multi-frame input. In that case, several single-frame attribute descriptors are concatenated into the multi-frame attribute descriptor. Per-attribute back-ends remain the same as in the single-frame case. Note that the number of input frames can not be arbitrarily large in order to avoid memory exhaustion during training. All our multi-frame models produce predictions for segment $T$ by observing the middle frames of segments $T$, $T-1$, and $T-4$. Each per-attribute prediction head is subject to the corresponding cross-entropy loss. Following the multi-task learning paradigm [14], the total loss is the mean of all per-attribute losses.

B. Dynamic Loss Weighting for Multi-Task Learning

We wish to alleviate multi-task class imbalance (cf. III-B) by increasing the influence of rare classes on the training objective. We denote our training set with $\{x_n, y_n\}$, where $x_n \in \mathbb{R}^d$, $y_n \in \{1, \ldots, C\}$, and $n \in \{1, \ldots, N\}$. Let $P^c_n = P(Y = c | x_n)$ denote the predicted posterior of class $c$ for input $x_n$. The standard cross-entropy can be interpreted as negative logarithm of the geometric mean of the correct class posterior

$$\overline{P} = \left(\prod_{n=1}^N P^c_n\right)^{1/N}$$

everall samples [16]:

$$\text{CE} = -\frac{1}{N} \sum_{n=1}^{N} \ln P^c_n = -\frac{1}{N} \ln \left(\prod_{n=1}^{N} P^c_n\right)$$

$$\quad = -\ln \left(\prod_{n=1}^{N} P^c_n\right)^{1/N} = -\ln \overline{P}$$

(1)

This equation can also be expressed as a weighted sum of per-class geometric means $\overline{P}^c$:

$$\text{CE} = -\frac{1}{N} \sum_{c=1}^{C} \sum_{n : y_n = c} \ln P^c_n = -\frac{1}{N} \sum_{c=1}^{C} \frac{1}{N_c} \ln \left(\prod_{n : y_n = c} P^c_n\right)$$

$$\quad = -\sum_{c=1}^{C} \frac{N_c}{N} \ln \left(\prod_{n : y_n = c} P^c_n\right) = -\sum_{c=1}^{C} \frac{N_c}{N} \ln \overline{P}^c$$

(2)

The range $\{n : y_n = c\}$ denotes examples of class $c$. The symbol $\overline{P}^c = \left(\prod_{n : y_n = c} P^c_n\right)^{1/N_c}$ denotes the geometric mean

IV. MULTI-TASK RECOGNITION OF ROAD-SAFETY ATTRIBUTES IN VIDEO

Our approach performs multi-task recognition of road-safety attributes through two stages: local recognition (section IV-A) and sequential enhancement (section IV-C). Both stages alleviate class imbalance by dynamic loss weighting according to our multi-task recall analysis (section IV-B).

We build two such matrices for each attribute: one using the ground truth labels and the other using predictions produced by our local recognition pipeline from Figure 1. For single-peak attributes, the only possible ground truth transitions go from the default class to a positive class and back. Thus, their ground truth matrices will have non-zero diagonal elements only in the row and the column of the default class. For smooth attributes, most consecutive segments belong to the same class. Thus, the diagonal elements of their ground truth matrices will be significantly larger than off-diagonal elements.

Our analysis reveals consistent discrepancies between ground truth and local prediction matrices for these two groups of attributes. In the local prediction matrices, single-peak attributes have many non-zero diagonal values. These discrepancies occur when the model assigns the same positive class to two consecutive segments that are visually very similar (e.g. two segments inside an intersection). This is a reasonable error, but it shows that the convolutional recognition model fails to learn the single-peak annotation convention. For smooth attributes, significantly larger values of off-diagonal elements were observed in prediction matrices compared to ground truth matrices. This reveals the presence of spurious class transitions in local predictions for consecutive segments, which implies that our local model fails to learn the inertia of smooth attributes.

This result motivates us to extend the local recognition pipeline with per-attribute sequential enhancement models that learn temporal behavior patterns without requiring costly back-propagation through hundreds of video frames. The technical details of this component are described in IV-C.
posterior of the correct class in samples that belong to class c. The equation shows that standard cross-entropy maximizes the weighted arithmetic mean of per-class geometric mean posteriors with the weights being the relative class frequencies \(N_c/N\).

If we want each class to have the same contribution to the loss, we can assign a weight to each class that is the inverse of its relative frequency: \(w_c = N/N_c\). This yields the inverse-frequency-weighted cross-entropy loss [50], [51]:

\[
\text{CE}^{\text{IFW}} = -\frac{1}{N} \sum_{n=1}^{N} w_{y_n} \ln P_{y_n}^{c} = -\frac{1}{N} \sum_{c=1}^{C} w_c \sum_{y_n=c} \ln P_{y_n}^{c}
\]

The standard cross-entropy does not take into account the distribution of the posterior over incorrect classes. In that sense, cross-entropy can be viewed as measuring the extent to which a particular prediction is a false negative while ignoring the false positives. Thus, assigning a large weight to a particular class might also increase the incidence of false positives for that class. This analysis has been confirmed empirically [16]: increasing the class weight indeed decreases the precision of predictions for that class.

These observations suggest that placing a large weight on a rare class that already achieves high recall is likely to decrease precision with little gain in recall. This can be prevented by adapting the loss weights with respect to the model performance in terms of per-class recall [16]. Let \(R_{c,t}\) denote the validation recall of class c after epoch \(t - 1\). Then, recall-balanced class weights \(w_{c,t}^{R}\) can be expressed as [16]:

\[
w_{c,t}^{R} = \frac{N}{N_c} (1 - R_{c,t}) + \epsilon
\]

When recall is close to zero, the weight (4) approaches the inverse relative frequency. As the recall of a class increases, its weight diminishes. We add \(\epsilon = 10^{-4}\) to prevent the weight going to zero in the unlikely event of perfect recall.

We note that batches with examples from extremely rare classes will have a much larger loss magnitude than batches with no such examples. Hence, learning with small batches and large class imbalances may lead to drastic changes in loss magnitude across training iterations.

In multi-task learning, the total loss is calculated as the arithmetic mean over all tasks. Thus, a task with examples from rare classes may suffocate other tasks due to larger loss. If we have many tasks that suffer from class imbalance, it is not unlikely that for any given batch, there will be one task that impedes the progress of other tasks. This may prevent the model to learn any of the tasks, since they all intermittently hamper each other. We address this problem by favoring a stable loss magnitude of individual tasks. We achieve this by normalizing the loss with the sum of the weights of individual examples. If we reuse the weights \(w_{c,t}^{R}\) from equation 4, then the loss for each individual task can be expressed as follows:

\[
\text{CE}^{\text{MT}} = -\frac{\sum_{n=1}^{N} w_{y_n,t}^{R} \ln P_{y_n}^{c}}{\sum_{n=1}^{N} w_{y_n,t}^{R}}.
\]

C. Sequential Enhancement

The second stage of our recognition approach enhances local predictions by aggregating evidence across a large temporal window. We form temporal inputs as sequences of \(T = 21\) vectors for all consecutive segments from \((t-10)\) to \((t+10)\). Instead of hand-crafted post-processing rules, we propose to classify prediction sequences with deep recurrent models [67]. While the first stage of our approach predicted all attributes in a single forward pass, sequential enhancement involves per-attribute recurrent models. Thus, these models can learn attribute-specific temporal behavior patterns from III-C.

Our recurrent models consist of four layers with bidirectional long short-term memory cells (Bi-LSTM). Each layer processes the sequence in two directions using two separate unidirectional LSTM modules. The inputs are a concatenation of the local logits \(s^a_t\) in segment \(t\) by observing \(T = 21\) vectors that correspond to segments from \((t-10)\) to \((t+10)\). Each of these vectors is a concatenation of the logits \(s^a_t\) and the jointly learned embedding \(e^c_t\) of the the most probable class according to the local model.
in a vector of size 1280, which is then fed to a fully-connected softmax classifier. The output is the predicted posterior distribution over classes $P(A_i = c_j | x_{i-T:T+T})$. The whole model is trained with dynamically weighted cross-entropy as presented in IV-B.

V. DATASETS

A. iRAP-BH

This paper introduces a novel corpus of georeferenced video that we have acquired along 194 public road sections (2300 km total) in Bosnia and Herzegovina for off-line road-safety assessment. All videos were recorded in $2704 \times 2028$ RGB format at 25 frames per second with a GoPro HERO4 Black camera. An average road section comprises 1175 10-meter segments, while an average segment spans 18 frames.

Our corpus was annotated with all iRAP attributes by trained human annotators. Even though the iRAP Star Rating Score requires 100-meter granularity, we have annotated iRAP-BH over 10-meter segments in order to provide better supervision for learning algorithms. We split the dataset into 214,073 training, 5,813 validation, and 6,563 testing segments. The three splits ensure that any two segments that belong to the same road section also belong to the same split. This enables training sequential and multi-frame models without data leakage. We represent each road segment with its middle frame resized to $384 \times 288$. This results in a multi-task multi-class video recognition dataset with 226,449 images. Qualitative examples are provided in Figure 3 and the appendix.

iRAP-BH is not related to any existing public dataset. We will make it publicly available upon acceptance in order to promote future research on road-safety and related tasks.

B. Honda Scenes

The Honda Scenes dataset [17] contains 80 training and 20 evaluation videos. Each frame of each video is annotated for the following four traffic scene classification problems: Road place, Road environment, Road surface, Weather.

Images for the Road place and Road environment problems have been obtained by subsampling at 3Hz. This results in 760,000 training and 160,000 evaluation frames. We evaluate our methods by treating consecutive frames as consecutive road segments. Images for the Road surface and Weather problems have been sampled from Honda Scenes and BDD100k. The Road surface dataset consists of 2,676 Honda and 7,463 BDD100k images. The training and evaluation splits contain 9,150 and 898 images. The Weather dataset is a subset of BDD100k with 11,781 training and 1,255 test images. The following paragraphs briefly describe the four problems.

1) Road Place: This is the only problem that considers multiple multi-class classification tasks. Each of those tasks has fine-grained temporal labels with classes such as Approaching (A), Entering (E), and Passing (P) that depend on the relative position of the car to the place of interest in a given frame. The tasks are: Construction zone, Intersection (3 way), Intersection (4 way), Intersection (5 way & more), Overhead bridge, Rail crossing, Merge - Gore On Left, Merge - Gore On Right, Branch - Gore On Left, Branch - Gore On Right.

2) Road Environment: This problem involves recognition of the following classes: Local, Highway, Ramp, Urban. The problem does not involve temporal labels, but it consists entirely of frames from the Honda Scenes dataset. That means we can still use our multi-frame and sequential models.

3) Road Surface: In this multi-class problem, each image is annotated with one of three road surface classes: Wet, Dry, and Snowy. We evaluate only the single-frame version of our model since the dataset includes only non-sequential images. The Road surface classes are fairly balanced, so weighted losses do not improve performance.

4) Weather: This multi-class problem involves image classification into four classes: Clear, Overcast, Rainy, and Snowy. We evaluate only our unbalanced single-frame model for the same reasons as in the case of Road surface.

C. FM3m

The third iteration of the Fleet Management Dataset (FM3) [18] consists of 11,448 images of traffic scenes from Croatian roads. The main subset of the dataset (FM3m) consists of 6,413 images. The training, validation and test splits contain 1,607, 1,600, and 3,206 images, respectively. Each image is labeled with one binary label (true/false) for each classification attribute. There are 8 classification binary attributes: highway, road, tunnel, exit, settlement, overpass, booth, traffic.

The frames were assigned to training, validation, and test splits in a uniform random fashion. This means that, in general, consecutive frames were assigned to different splits. This prevents us to take multiple segments on input, so we evaluate only single-frame models in these experiments.

D. BDD100k

The Berkeley Deep Drive (BDD100k) dataset [19] has been designed for heterogeneous multi-task visual recognition in road driving scenes. It contains 100,000 40-second video clips from a wide range of environments, including urban and rural areas, different weather conditions and times of day. From each sequence, a single keyframe was annotated with object bounding boxes, drivable areas, lane markings and full-frame panoptic labels. There are 80,000 training, 10,000 validation, and 20,000 test images, respectively. Additionally, there are three image-wide tasks: Scene, Weather, and Time of day. The Scene classes include Tunnel, Residential, Parking Lot, City Street, Countryside, Gas Station, and Highway. Our experiments focus on the Scene task since the other two tasks have not been addressed by relevant previous work. Moreover, there is an intersection between the Weather task and the homonymous task of Honda Scenes.

BDD100k does not allow sequential processing since each image comes from a different video sequence. Thus, our experiments involve single-frame models, as in FM3m, and the Road surface and Weather problems of Honda Scenes.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

We evaluate visual recognition of road-safety attributes with different variants of our approach on the iRAP-BH dataset.
Moreover, we compare our approach with previous work on three related datasets: Honda Scenes, FM3m, and BDD100k. Though not annotated for iRAP attributes, these three datasets involve classification of road and traffic-related classes, some of which are quite similar to certain iRAP attributes.

A. Evaluation Metrics

We evaluate our approaches on iRAP-BH and Honda Scenes according to the mean macro-averaged F1 score [68], [69]. This is a suitable metric due to our multi-task multi-class setup and class imbalance. Creators of Honda Scenes also use macro-F1 in their experiments. We evaluate FM3m performance according to the mean average precision (mAP) across all tasks. This metric is suitable since all tasks involve binary classification [18]. We evaluate BDD100k classification performance according to accuracy [40], [56]. We present all performance metrics as percentage points (pp).

B. iRAP-BH

We augment input images through color jittering by varying brightness, contrast, saturation, and hue in relative ratios of 0.6, 0.3, 0.2, and 0.02. We do not employ horizontal flipping as it would disturb the detection of attributes that are specific to right-hand traffic scenarios (e.g. Roadside severity - passenger side). We also opt not to apply random cropping as it omits visual information from the peripheral parts of input images since this information is critical for identifying roadside attributes such as Street lighting. Both stages of our method are trained with the Adam optimizer. For local recognition, the learning rate is set to 1e-5, weight decay to 1e-3 and batch size to 12. We train for 15 epochs according to a multiplicative learning rate scheduler with the annealing factor of 0.88 per epoch. For sequential enhancement we set the learning rate to 5e-4, weight decay to 1e-4, and train for 10 epochs with batch size 32. We have systematically explored multiple values for each hyperparameter involved in our model architecture and the optimization process. The optimal hyperparameter values were selected by conducting two separate grid searches on the validation split of iRAP-BH. The two grid searches optimized hyperparameters of the two stages of our method. The optimal values are consistently applied in all subsequent experiments, across all datasets.
Table I explores the impact of sequential post-processing and loss weighting on the overall performance. We observe that semantic segmentation pre-training contributes 1.2 percentage points (pp) mF1. Inverse-frequency weighting delivers additional 1.5 pp. Our multi-task weighting improves upon that by 1.3 pp. Finally, sequential post-processing further increases the performance by about 5.1 pp.

Table II shows the impact of sequential enhancement and dynamic loss weighting on each attribute. The attributes which get the largest relative improvement from loss weighting are Pedestrian crossing - inspected road, Median Type, Pedestrian observed flow along the road passenger-side, Roadside severity - driver-side object, and Bicycle facility. Relative improvements compared to standard cross-entropy for those attributes range from 19% to 26.5%. All of these attributes suffer from class imbalance.

Single-peak attributes which benefit most from sequential enhancement are Speed management / traffic calming and Intersection type, with relative improvements of 44.7% and 22.3%. Smooth attributes with the largest relative improvements are Bicycle facility, Number of lanes, and Skid resistance / grip. Figure 3 shows four examples of successful sequential enhancement. In the case of Intersection type, the correction accommodated the single-peak annotation convention. For smooth attributes Bicycle facility and Number of lanes, the sequential model corrected the spurious class transition made by the local model by considering a larger context. The last example is Street lighting, which should be annotated continuously through all segments between two nearby lighting poles. In this particular example, the upcoming lighting poles are obscured by the road curvature and overgrown roadside bushes. Consequently, the local model classifies the in-between segments incorrectly. The sequential model corrects these mistakes by leveraging a larger context window.

Table III validates several pre-training strategies for the backbone of our model. We observe that dense semantic pre-training outperforms classification pre-training. This suggests that detailed spatial understanding of visual concepts benefits the recognition of road-safety attributes.

C. Honda Scenes

This subsection compares our method with prominent previous work on Honda Scenes. We include several methods from the original paper [17], Context MTL [44], MTAN [46] and two of our ablations that show the impact of our contributions.
TABLE V
EXPERIMENTAL EVALUATION ON ALL ROAD-PLACE TASKS OF HONDA SCENES (MACRO-F1, PERCENTAGE POINTS). LEGEND: BB - BACKBONE; B - Background, A - Approaching, E - Entering, P - Passing

| Model                        | BB | B | Intersection 3-way | Railway Crossing | Construction | Left Merge | Right Merge |
|------------------------------|----|---|--------------------|------------------|-------------|------------|-------------|
| Honda BiLSTM [17]           | m50| 88| 0 0 9 3            | 24 14 46 28      | 2 5 29 12   | 9 28 19    | 16 23 20    |
| Honda Event [17]            | m50| 92| 0 0 0 0            | 23 47 46 39      | 2 6 38 15   | 5.6 8 7    | 13 16 15    |
| Context MTL [44]            | m50| - | 0 6 0 2           | 1 35 52 32       | 0 4 38 14   | 4 6 5      | 26 18 22    |
| MTAN [46]                   | wn28| 92| 1 2 5 3           | 19 27 42 29      | 3 9 24 12   | 11 17 14   | 19 12 16    |
| ConvCE (ours)               | m18| 90| 19 0 9 8       | 13 49 52 38      | 2 11 56 23  | 22 29 26   | 29 33 31    |
| ConvCE\_EXP (ours)          | m18| 91| 27 0 10 12      | 15 56 59 43      | 3 12 63 26  | 27 36 32   | 31 35 33    |
| ConvCE\_EXP\_SE (ours)      | m18| 91| 29 0 9 13       | 28 55 71 51      | 11 22 64 32 | 29 43 36   | 34 45 40    |

1) Road Place: Table IV evaluates the overall performance. We denote the original sequential baseline as Honda BiLSTM and their two-stage sequential approach as Honda Event [17]. Table V focuses on individual tasks. In both tables, our baseline (multi-frame model, standard loss, no sequential enhancement) outperforms all previous approaches in spite of a weaker backbone. This makes the improvements in Tables I – III even more convincing since now we know that they start from a very strong baseline. Multi-task dynamic loss weighting increases our performance by 2.9 pp. This improvement is due to class imbalance [17]. We observe the greatest relative improvement on Intersection (5-way or more), Railway, Left Merge, and Left Branch subtasks. Sequential enhancement brings a further improvement of 3.9 pp.

2) Road Environment: Table VI compares our multi-frame model with Context MTL, MTAN, and the two original frame-based approaches [17]. These approaches involve a ResNet-50 backbone pre-trained on Places365 and leverage the DeepLabV2 semantic segmentation model to either mask out traffic participants (Honda Frame - Mask) or enrich the input image with its segmentation map (Honda Frame - SemSeg). Our model prevails on most classes. The largest improvement occurs on the most challenging class (Ramp). That is the least frequent class in the dataset, and the class that benefits the most from multi-task loss weighting. Sequential enhancement brings a further improvement of 1.2 pp.

3) Road Surface: Table VI shows that our single-frame model outperforms previous approaches. Loss weighting does not improve performance since the classes are fairly balanced. We could not use our multi-frame and sequential enhancements since the task only allows single-frame prediction.

Table VII shows that pre-training the ResNet-18 for semantic segmentation on Vistas works better than pre-training the ResNet-50 for classification on ImageNet-1k.

4) Weather: Table VI shows that our single-frame model outperforms previous approaches on classes Overcast, Snow and overall, while underperforming on Clear and Rain. As in the Road Surface task, loss weighting does not contribute since the classes are balanced, while sequential approaches are not applicable to the single-frame prediction task.

D. FM3m
We compare our method with two of the best approaches provided by the authors of the FM3m dataset. They train SVM classifiers with RBF kernels on image descriptors obtained by ResNet-50 and DenseNet-121 backbones pre-trained on ImageNet-1k. It should be noted that the two approaches do not fine-tune their feature extractors on FM3m, while our approach has a weaker backbone. We also compare against other concurrent approaches for traffic scene classification,
TABLE VIII
AP PERFORMANCE ON THE FM3M DATASET: H-Highway, R-Road, Tu-Tunnel, E-Exit, S-Settlement, O-Overpass, B-Booth, Tr-Traffic

| Model               | H   | R   | E   | O   | B   | Tr  | Mean |
|---------------------|-----|-----|-----|-----|-----|-----|------|
| RNS-SVM [11]        | 89  | 87  | 82  | 83  | 86  | 88  | 90   |
| DIN Li-SVM [18]     | 91  | 89  | 87  | 88  | 89  | 92  | 91   |
| Honda Frame - Mask  [17] | 95  | 90  | 88  | 90  | 91  | 96  | 94   |
| Honda Frame - SemSeg [17] | 90  | 91  | 90  | 92  | 93  | 95  | 93   |
| MTAN [46]           | 92  | 93  | 91  | 94  | 95  | 97  | 95   |
| Conv single, CE     | 100 | 94  | 93  | 96  | 95  | 94  | 95   |
| Conv single, CE seq | 100 | 94  | 91  | 94  | 95  | 94  | 94   |

TABLE IX
ABLATION OF PRE-TRAINING ON FM3M (AP):H-Highway, R-Road, Tu-Tunnel, E-Exit, S-Settlement, O-Overpass, B-Booth, Tr-Traffic

| Model     | BB | H | R | Tn | E | S | O | B | D | Mean |
|-----------|----|---|---|----|---|---|---|---|---|------|
| IN (ours) | m18| 79 | 91 | 90 | 87 | 90 | 92 | 90 | 94 | 89   |
| IN (ours) | m30| 90 | 95 | 92 | 89 | 92 | 87 | 90 | 96 | 95   |
| Vistas (ours) | m18| 100 | 96 | 90 | 96 | 96 | 94 | 95 | 92 | 97   |

TABLE X
COMPARISON WITH PRIOR WORK ON THE SCENE TASK OF BDD100k

| Model               | Accuracy |
|---------------------|----------|
| Honda Frame - Mask  | 76.8     |
| Honda Frame - SemSeg| 76.0     |
| MTAN [46]           | 73.9     |
| Local-Global FCRNN [40] | 76.0     |
| Conv single, CE     | 78.4     |
| Conv single, CE seq | 78.7     |

| Model               | Cloudy | Rainy | Snowy |
|---------------------|--------|-------|-------|
| Honda Frame - Mask  | 70.9   | 63.8  | 62.2  |
| Honda Frame - SemSeg| 72.3   | 65.1  | 62.3  |
| MTAN [46]           | 71.1   | 66.2  | 60.7  |
| SADA [56]           | 70.5   | 62.7  | 59.1  |
| Conv single, CE     | 75.9   | 71.4  | 70.0  |
| Conv single, CE seq | 76.6   | 71.5  | 70.9  |

namely the two variants of the Honda Frame [17] frame-based model and the multi-task attention network (MTAN) [46]. Table VIII shows that our single-frame models perform competitively, and that our multi-task loss-weighting improves performance on nearly all classes.

We also ablate the impact of pre-training by comparing three different variants of our single-level model. Table IX shows that Vistas pre-training contributes more than using the larger ResNet-50 backbone. Sequential enhancement is not applicable since this is a single-frame prediction task.

E. BDD100k

Table X evaluates our method on the default and the cross-domain setup of the Scene task of BDD100k. Both setups include comparisons with Honda Frame - Mask, Honda Frame - SemSeg [17], and MTAN [46]. The default setup also includes Local-Global FCRNN [40], which is a two-stream model that relies on local and global features of Faster RCNN and InceptionV2. The cross-domain setup also includes the “source-only” baseline from the Sparse Adversarial Domain Adaptation (SADA) paper [56].

The table shows that our single-frame models outperform all competing approaches. Our multi-task loss-weighting contributes in all experiments, while sequential enhancement is not applicable due to the single-frame prediction setup. Note that all experiments in the cross-domain section train only on sunny images and report on cloudy, rainy, and snowy images.

VII. CONCLUSION

We have presented a two-stage approach for automatic collection of road-safety attributes in monocular video. Our approach complements the baseline convolutional recognition in a local spatio-temporal context with three contributions: semantic segmentation pre-training, multi-task recall-based loss weighting, and sequential enhancement.

Our multi-task convolutional classifier consists of a shared backbone and per-attribute back-ends. We pre-train the shared backbone for semantic segmentation of street-level images as a part of an efficient dense prediction pipeline. Our baseline outperforms the state of the art on Honda scenes, which emphasizes the value of our subsequent two contributions. We address the extreme class imbalance with a multi-task variant of recall-based loss weighting. In this setup, the magnitudes of individual task losses are normalized in order to encourage commensurate contributions to the total loss. A closer look at iRAP attributes reveals that different attribute types exhibit very different temporal behavior patterns. For instance, intersections occur at discrete moments in time, while the number of lanes remains constant through many frames. These considerations led us to correct the local predictions with attribute-specific recurrent models that learn the temporal behavior over a larger temporal context.

We have experimentally demonstrated a substantial impact of our contributions over the strong baseline. Each of our three contributions improve the performance across the board for all attributes. The combined approach outperforms the previous work on all tasks of the Honda Scenes dataset. The greatest improvements occur on the Road place task due to significant class imbalance and the availability of multi-frame input. Our approach also delivers competitive performance on FM3m and BDD100k, in spite of a weaker backbone and inability to leverage sequential post-processing.

We emphasize that our approach delivers competitive performance on a very low computational budget. The shared convolutional features are extracted by an efficient backbone only once per frame. Per-attribute recurrent models are extremely efficient due to low-dimensional inputs and outputs. Preliminary experiments have shown that a straightforward causal adaptation can deliver road-safety assessments in real-time even on a mobile device.

To summarize, our main contributions are semantic pre-training, multi-task recall-based loss weighting, and sequential enhancement. In addition, we introduce iRAP-BH as a novel dataset for visual road-safety assessment that specifically focuses on iRAP attributes. Future work should explore the sensitivity of the proposed approach to various kinds of domain shift through evaluation on novel datasets. The new datasets should include multiple countries, different types of cameras, various seasons, and challenging meteorological conditions. Moreover, it would be interesting to attempt further improvement by leveraging panoptic predictions and monocular reconstruction.
