Multi-Task Learning for iRAP Attribute Classification and Road Safety Assessment

Marin Kačan, Marin Oršić, Siniša Šegvić
Faculty of Electrical Engineering and Computing,
University of Zagreb

Marko Ševrović
Faculty of Transport and Traffic Sciences,
University of Zagreb

Abstract— **We address the automatic recognition of road safety attributes according to the iRAP methodology. We formulate the problem as a separate multi-class classification of each iRAP attribute in georeferenced video clips that correspond to particular road segments. We propose a solution based on an efficient multi-task model with shared features, which can recognize all attributes with a single forward pass and learn in an end-to-end fashion. We perform experiments on a novel real dataset acquired along 1850 km of public roads in Bosnia and Herzegovina, in which all iRAP attributes have been annotated by human experts. We express recognition accuracy as per-attribute macro-F1 scores due to a significant class imbalance present within most attributes. We thoroughly validate different variants of our model, analyze the contributions of several hyper-parameters, and report recognition accuracy on the independent test set.**

I. INTRODUCTION

With more than 1.35 million people killed on roads throughout the world each year, road traffic crashes are among the world's most significant public health and injury prevention problems [1]. Improvements in traffic infrastructure and road safety engineering decrease the risk and severity of crashes. Introduction of safe roadsides, sidewalks, pedestrian crossings, bicycle paths and other road safety attributes results in fewer road deaths and injuries.

The International Road Assessment Programme (iRAP) is a registered charity [2]. It aims to eliminate high-risk roads by proposing the iRAP star rating — a robust, evidence-based approach for road-safety assessment [3]. iRAP Star Rating is a simple and objective measure of the in-built safety of the road, with 5-star roads being the safest and 1-star roads the most unsafe. The rating is produced by calculating the Star Rating Score for each 100-meter road segment and then averaging over longer lengths [4]. The Star Rating Score is calculated from 52 attributes (or risk factors) for a particular road segment [5]. According to iRAP terminology, assigning values to road segment attributes is denoted as attribute *coding*.

iRAP attributes are described in detail by the iRAP Coding Manual [2]. IRAP attributes can be assessed with on-site surveys by teams of experts [6]. However, off-line operation can be advantageous due to opportunity to perform repeated assessments [7], [8]. Currently, most off-line assessments are performed manually by trained experts which code attributes by annotating georeferenced video. Though faster and cheaper than live surveys, manual off-site assessment is still expensive, time-consuming, and prone to error.

Although a substantial amount of research has been directed towards automatic inspection of traffic infrastructure, such as traffic sign [9] and road markings detection [10], there have been only a few attempts to automatically recognize road-safety features in video [6], [7], [8]. As opposed to human coders, automatic coding is objective and consistent through time. It benefits from more data, so it keeps improving as more roads are coded. It can also be used as an internal validation tool for human coders, It can also potentially speed-up manual coding by pre-coding a video that an expert coder only needs to check and verify [11].

A crucial step of the iRAP coding process is quality assurance, where 10% of the coded roads needs to be checked by accredited experts. This process, if done automatically using AI, could be extended to 100% and sped up, That could result in more accurate, dependable, and valuable assessments.

This paper provides a preliminary feasibility study for automatic iRAP attribute recognition [12]. We address this problem throughout a multi-task learning approach [13]. Our model recovers shared features by leveraging a pre-trained convolutional backbone [14], [15] and a spatial pyramid pooling module [16]. The shared features are concatenated with per-attribute attention pools and fed to per-attribute classifiers. We test various combinations of backbones, output layers, loss functions, input sequence lengths, image dimensions, and explore the impact of color jittering. We evaluate the contribution of each design decision that we vary in experiments. We show per-attribute results of our best model, and discuss class imbalance and other problems which hamper correct recognition of some attributes.

II. RELATED WORK

There has been much previous research in areas related to road safety, such as localization of control devices [9], [17], recognition of fleet-management attributes [18], or semantic segmentation [19], [20]. These works are related to our task, but they target only a subset of the attributes that we address.

Some approaches tackle iRAP attribute classification using an intermediate semantic segmentation step, from which they then extract the attributes [8]. These approaches are more related to our work. However, producing a training dataset for semantic segmentation requires dense pixel-level annotation and consequently implies significant annotation effort. That is a much slower and more tedious process than simply reusing the attributes provided by human coders as image-wide annotations. Furthermore, a semantic segmentation model only gets the learning signal from segmentation labels, and not from attribute class labels. Thus, such a model learns features that are optimal for semantic segmentation, not necessarily for our actual task. There is also a possibility of error propagation due to the isolated training of model components. For those reasons, we opt for an end-to-end system that we train to predict the classes of all considered attributes directly from video frames, without the intermediate segmentation step.

Song et al. [21] try to go even further by predicting the Star Rating Score of a road segment directly from video. We avoid directly inferring the star rating in order to preserve the explainability of our model's decisions. The Star Rating Score can be calculated directly from attribute values. We do not want to encumber our model with the additional task of learning a formula that is already known.

III. Attributes

Here we briefly list and describe the 7 attribute categories used to determine the iRAP Star Rating Score.

1) Attributes for road details and context: Most of these attributes contain metadata about the coding process (coder name, coding date, road name, etc.). The only attribute that can be coded is Divided carriageway.

2) Observed flow attributes: These attributes are obtained by counting appearances of various road users - motorcycles, bicycles, and pedestrians - in recorded segments. We do not consider attributes from this category since observing the traffic flow from a single video is bound to result in high variance estimates. That being said, we plan to address these attributes in future work.

3) Speed limit attributes: These attributes are coded according to the speed limit, which has to be deduced from traffic signs and local regulations regarding road type and post-intersection policy. In this work, we do not consider attributes from this category since the existing research [9] suggests that the problem of traffic sign detection is mostly solved and hence not a priority.

There is one attribute in this category that does not code a speed limit - Speed management. It is concerned with the presence of infrastructure features that reduce the operating speed. We chose not to address it since it only had 20 positive examples in the whole dataset.

4) Mid-block attributes: Mid-block attributes make up the bulk of all coding work. They are concerned with objects on the road median (as opposed to objects on the roadside) and within the road surface itself. We discard two attributes from this category (Service road, Centre line rumble strips) due to the scarcity of positive examples in the dataset.

5) Roadside severity: These attributes encode the most severe object on each roadside (passenger-side and driver-side), and their distances from the road. They are among the most challenging attributes for a classifier to learn because of a couple of reasons. First, the front dashboard camera captures just a fraction of each roadside. That is why, in our experiments, we also test a model that operates on sequences of image frames. These sequences are assembled to contain frames from two preceding road segments, along with the current one.

Second, various objects (houses, fences, trees) can appear on the roadside simultaneously, which complicates the estimation of the severity level for the road user. iRAP coding manual requires that these attributes are affected only by the most severe object, which is a very weak form of supervision [17]. To determine the severity of any given pair of roadside object type and its distance from the road, the coding manual defines a priority list. It contains all the possible pairs sorted with decreasing severity. Out of all the potential objects appearing on the roadside at various distances, the model has to predict only the most severe one. This is rather difficult for monocular recognition models since they need to learn to regress metric distances without being explicitly instructed to do so. A model that correctly recognizes a roadside object present in the segment may get penalized if that object turns out not to be the most severe one. In general, such kind of inference can be learned, but it requires much more training data than simple detection of objects which define the visual class.

Apart from the aforementioned roadside severity attributes, this class also includes the attribute Shoulder rumble strip. We do not address it since there is no occurrence of shoulder rumble strips in the entire dataset (all examples are negative).

6) Intersections: This category contains attributes such as Intersection quality and Intersection type, which are combined to evaluate the risk of an intersection [2]. Both of these attributes are difficult to recognize. Intersection quality includes several soft factors such as sight distance or deflection angles, which are not easy to learn. Intersection type contains 17 kinds of intersections, such as merge lane, roundabout, and railway crossing. These classes are well defined, but the many options make it easy to miss the right answer.

In this category, we discard the attributes Intersection channelization and Intersecting road volume. The former has too few positive examples in the dataset. The latter is concerned with the average daily number of vehicles passing through the segment from the intersecting road. It would not be realistic to estimate this number from a few seconds of video that captures the particular intersection. The information for this attribute is usually gathered from a different source.

7) Vulnerable road user facilities and land use: These attributes concern the presence of various facilities for pedestrians and cyclists, as well as the area type and land use at the segment.

We do not address the attribute School zone warning since its class distribution is nearly identical to the attribute School zone crossing supervisor. We also discard Pedestrian fencing, Facilities for motorcycles, and Facilities for bicycles due to the scarcity of positive examples in our dataset.

Some attributes have fine-grained classes of very low frequency. In such cases, we group those rare fine-grained classes into coarser ones. For example, the attribute "Intersection type" has 4 different classes for variants of a 3-leg intersection. They are the following: *3-leg*, *3-leg with a protected turn lane*, *3-leg signalized*, and *3-leg signalized with a protected turn lane*. We group these classes into a coarser class containing all 3-leg intersections variants. Since we use one shared model for all attributes, it will be easy to extend it to predict the remaining attributes that are not yet covered when more training data becomes available.

IV. Dataset

We perform experiments on a novel road-safety corpus acquired along 1850 km of public roads in Bosnia and Herzegovina. Even though the iRAP Star Rating Score considers 100-meter road segments, the corpus is coded over 10-meter segments to get better estimates by averaging. There are about 185,000 10-meter segments in the dataset. Human experts have annotated each of them with all 52 iRAP attributes as part of a regular iRAP coding campaign. Most of these segments span about 30-40 frames, although this depends on road and traffic conditions during acquisition. All videos are recorded in 2704x2028 RGB format at 25 frames per second.

We create the dataset for our recognition experiments by exploiting geographical information from the iRAP coding database. Each iRAP record corresponds to a 10-meter road segment. It contains the values of the 52 iRAP attributes and the GPS coordinates of the segment's two endpoints. We leverage these GPS coordinates to find video frames that correspond to each of the two endpoints. We first find the two closest GPS references in the corresponding georeferenced road video and then interpolate the GPS location using the estimated vehicle speed. Finally, we assign all intermediate frames to that particular segment.

In our experiments, we work with image frames resized to 384×228 and 768×576. Images were resized without cropping or changing the aspect ratio. We make our experiments feasible by using only a subset of all videos containing about 40,000 segments. About 28,500 segments are in the training set, while the validation set and the test set each contain about 5,500 segments. Our single-frame models operate on the middle frame of the corresponding segment. Our multi-frame models operate on sequences of 3 middle frames from the current and the two previous segments. All our models predict the values (labels) of the considered 33 attributes.

Some attributes suffer from an extreme class imbalance in the sense that nearly all the segments in the dataset belong to one class, while all other classes have much fewer segments. For example, the attribute *Pedestrian crossing - side road* is dominated by examples of the class *No crossing* that make up 99,6% of the training set. The two remaining classes - *Unsignalised crossing* and *Signalised crossing* - account for only 0,25% and 0,15% of training set examples.

V. Models

Each segment in our dataset is annotated with all 33 attributes. Hence, we can train a single multi-task model with a shared backbone [13] to recognize all attributes with a single forward pass, as illustrated in Figure 1. Note that the backbone has to learn features that are good for multiple different tasks. This may regularize the learning algorithm by favorizing features that are good generally [22]. Deep convolutional models are known for having enough capacity to learn features for multiple different tasks [23]. Also, the train time is significantly reduced since we do not need to train a different model for each attribute.

We test two standard convolutional backbones: ResNet-18 [14] and DenseNet-121 [15] (CNN in Figure 1). After the backbone, we perform spatial pyramid pooling [16] (SPP in Figure 1) with grid dimensions (6, 3, 2, 1). Our SPP module starts with a 1×1 convolution that reduces the number of the feature maps to 128. It proceeds by pooling and completes with another 1×1 convolution for each grid size, which reduces the depth from 128 to 42. The SPP layer produces a fixed-size output regardless of the input image size and captures information at different scales. We flatten and concatenate SPP outputs for all grid dimensions into a single vector, concluding the shared part of the network.

The attribute-specific parts of our network consist of soft spatial attention pooling [24], [25] (ATT$_i$ in Figure 1) and a fully-connected classifier with softmax activation (FC$_i$ in Figure 1). The attention module receives shared features $\mathbf{F}$ and a learnable query vector $\mathbf{q}_i$ and produces the attention pool $\mathbf{a}_i$. The query vector and the attention pool have the same shape: their size is equal to the depth of shared features $\mathbf{F}$. The attention vector is determined as a weighted spatial pool of shared features $\mathbf{F}$. The weights are determined by a softmax-activated similarity map between the query vector and the shared features [26]. Finally, the attention pool is concatenated with SPP output and classified into per-attribute posteriors $P(A_i|\mathbf{x})$.

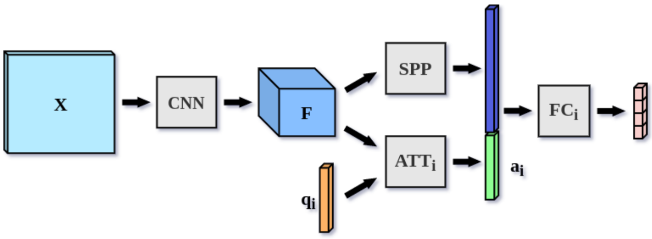

Fig. 1. The proposed recognition model feeds convolutional features ($\mathbf{F}$) into the spatial pyramid pooling module (SPP). The resulting image-wide representation SPP($\mathbf{F}$) is concatenated with per-attribute attention pools $\mathbf{a_i}(\mathbf{F}, \mathbf{q}_i)$ and processed by per-attribute classification heads FC$_i$. Note that $\mathbf{q}_i$ denotes a learnable query vector associated with attribute i. In the multi-frame case, the classification heads receive concatenations of single-image representations (SPP($\mathbf{F}_t$), $\mathbf{a}_{ti}$) where $t \in \{1..T\}$.

We train our model by averaging per-attribute classification loss. We alleviate class-imbalance by using balanced cross-entropy [27], [28]. The main idea is to promote learning of rare classes by multiplying the cross-entropy with a weight that is inversely proportional to the frequency of the correct class.

To get more spatio-temporal context for recognition, we also experiment with multi-frame models which observe images from neighboring segments as well. In this case, we detach fully-connected layers and independently apply our single-frame model to all images from the input sequence of length T. This results in T image-wide representations (SPP($\mathbf{F}_t$), $\mathbf{a}_{ti}$) for each attribute. Finally, we concatenate these representations and classify them with per-attribute fully-connected layers with softmax activation, as we did in the single-frame model.

We have validated various combinations of network architectures, loss functions, input sequence lengths, and image dimensions. Interestingly, none of these models outperformed all others on all attributes. Hence, we perform most of our experiments on model ensembles. In particular, we have applied model ensembles in our validation experiments where we want to see whether some design decision makes a consistent performance impact. We have also used model ensembles while performing error analysis on the train set to find hard examples or annotation mistakes.

VI. Experiments

We promote fast training on large quantities of training data by equipping our models with efficient convolutional backbones ResNet-18 and DenseNet-121 and by reducing the input resolution. We consider color-jittering as the only data-augmentation technique. We avoid horizontal flipping since all videos from a particular country are recorded on the same side of the road. We also avoid random cropping, since visual cues for some attributes may appear on the edges of frames, so we want to preserve them.

We use the Adam optimizer with a batch size of 10 and weight decay equal to 1e-3. We do training in two stages. In the first stage, we train the attention queries, the spatial pyramid pooling modules, and the classifiers for 4 epochs with the learning rate set to 5e-5, while keeping the backbone parameters frozen. Subsequently, we train all parameters for additional 16 epochs, with the learning rate lowered to 5e-6. In all experiments, we learn on the train split and evaluate on the test split of our dataset.

A. Methodology

Our experiments address the following 7 binary design decisions: i) single-frame vs. multi-frame recognition, ii) pre-training (ImageNet vs. Vistas), iii) backbone (rn18 vs. dn121), iv) loss balancing, v) attention, vi) input resolution (384×288 vs. 768×576), and vii) color jittering.

Each experiment addresses one of the 7 design decisions. To reduce the noise due to model variance, for each experiment, rather than just comparing two models, we compare two groups of models. Models in each group have the same value for the design decision being addressed, while they vary in the values for the remaining decisions. The two groups are symmetric: each model from the first group has its counterpart in the second group, differing only in the particular binary decision. We assess the impact of each binary decision by comparing i) average accuracy and ii) ensemble accuracy of the two model groups.

Note that the total number of trained models N is fewer than 2^7 (all possible design decision combinations). That is because parameterizations pre-trained on Vistas are available only for the ResNet-18 backbone, and we also avoided training all models on the larger input resolution due to time constraints.

B. Results

Table I explores the influence of the proposed seven binary decisions on the recognition accuracy on the validation dataset. The validation performance improves by about 2.4 percentage points (pp) when the recognition is performed on sequences of three frames instead of only one frame. Pre-training the backbone parameters for semantic segmentation on the Vistas dataset increases recognition accuracy by about 1.2 pp. Using DenseNet-121 instead of ResNet-18 increases accuracy by about 0.9 pp. Loss balancing increases accuracy by 1.8 pp. Including per-attribute attention pooling did not significantly impact the results. Resizing the images to 384x228 preserves the important visual information in the image since the improvement gained from quadrupling the number of pixels is not substantial. Finally, as simple an augmentation procedure as it is, color-jittering increases accuracy by 1.3 pp.

TABLE I
Influence of the considered binary design decisions to the validation accuracy of the proposed model.

Design decisions	Mean macro-F1	
	Average	Ensemble
Single-frame	56,4	57,5
Multi-frame (T=3)	**58,6**	**59,0**
Pre-training on ImageNet	57,4	57,4
Pre-training on Vistas	**58,5**	**58,7**
Backbone: ResNet-18	56,9	56,1
Backbone: DenseNet-121	**57,7**	**57,2**
Loss: standard CE	56,0	56,9
Loss: balanced CE	**58,1**	**58,6**
No attention pooling	57,0	58,4
Per-attribute attention pooling	**57,1**	58,4
Resolution: 384x288	58,3	58,0
Resolution: 768x576	**59,3**	**58,5**
No jitter	58,0	58,1
Color jitter	**59,3**	**59,4**

The best model configuration has Vistas pre-training, multi-frame training on color-jittered 384x228 images, balanced cross-entropy loss, and attention pooling. The model gets an average macro-F1 score of 61.5%. Table II shows the performance of this configuration on each attribute. We also report the top-1 accuracy and the majority class baseline.

Numerous attributes show large discrepancies between F1 score and accuracy. This can be caused by imbalance in the number of examples for different classes of a particular attribute. Poorly balanced attributes can be reliably identified by looking for high values of the majority class baseline (MCB). This is why it is important to have a strict metric like macro-F1 (M-F1), that penalizes a classifier that ignores an infrequent class of some attribute.

TABLE II

PER-ATTRIBUTE PERFORMANCE OF OUR BEST SINGLE MODEL. M-F1 - MACRO-F1, ACC. - ACCURACY, MCB - MAJORITY CLASS BASELINE.

Attribute name	**M-F1**	**Acc.**	**MCB**
Delineation	96,3	97,2	73,9
Divided carriageway	96,2	98,1	92,5
Sidewalk - passenger-side	90,7	95,3	70,4
Sidewalk - driver-side	89,5	94,3	84,2
Area type	87,1	87,3	60,9
Street lighting	81,5	84,0	50,4
Roadworks	77,6	99,4	99,1
Land use - passenger-side	71,0	75,0	41,8
Land use - driver-side	70,2	73,4	44,9
Lane width	65,4	92,4	86,3
Sight distance	63,6	88,6	85,7
Quality of curve	61,3	75,8	59,1
Paved shoulder - passenger-side	60,4	91,7	63,6
Paved shoulder - driver-side	59,8	90,9	67,4
Median type	58,8	91,8	89,4
School zone	58,8	98,1	96,5
Grade	57,7	94,4	97,6
Vehicle parking	56,2	85,9	84,7
Roadside severity - driver-side distance	54,4	66,7	62,3
Roadside severity - passenger-side dist.	54,2	66,9	71,0
Property access points	52,9	75,8	47,2
Upgradef cost	52,9	69,3	67,0
Road condition	52,5	89,3	84,1
Curvature	51,4	69,0	59,1
Sealed road	49,9	99,5	98,2
Pedestrian crossing - inspected road	48,0	99,0	99,1
Roadside severity - passenger-side object	47,9	58,5	20,9
Number of lanes	47,2	97,3	91,4
Pedestrian crossing quality	46,3	98,8	98,9
Intersection quality	45,5	96,9	96,9
Intersection type	43,9	96,9	96,9
Roadside severity - driver-side object	39,7	53,5	21,2
Pedestrian crossing - side road	39,5	99,7	99,6

C. Discussion

Analysis of the obtained results has revealed that some attributes require a larger temporal context. For instance, the Street lighting attribute is coded as being present on stretches of road that can span more than 10 segments between two street lamps. When we set the temporal context to only three segments, the recognition head often has insufficient visual cues to deduce that street lighting is present on that stretch of the road. Thus, for future work, we plan to extend our model to significantly longer image sequences.

Analysis of the hardest train errors (the lowest probability of the correct class) of ensemble models indicates the opportunity to detect inconsistent annotations. This opens up the possibility of using the proposed automatic coding system as a tool to discover the divergent practices of different annotators.

D. Explaining the model decisions

Figure 2 shows the performance of our best model on six input frames from our dataset. Alongside each input image, we visualize where the model is "looking" for the particular attribute. The visualizations display the strongest gradient of the correct logit with respect to the input image [29].

The top row shows recognition of three attributes for which the model performs well in Table II: Sidewalk - passenger-side (left), Delineation (middle), Divided carriageway (right). In all three cases, the model makes a correct decision while "observing" feasible image locations: right sidewalk, middle of the road, and the median strip, respectively.

The bottom row shows the recognition of three attributes from the bottom of Table II: Pedestrian crossing - side road (left), Roadside severity - driver-side object (middle), Number of lanes (right). The model incorrectly predicts that the sidewise pedestrian crossing is absent, despite "observing" the correct image location (left). The model also incorrectly predicts that roadside severity is defined by Safety barrier, while the correct class is Semi-rigid object (the fence). This shows why Roadside severity is a difficult attribute: many classes are hard to differentiate. Note that the street lighting pole on the left is also a potential roadside severity object, but the fence overrides it. It may be the case that multiple roadside objects confuse the model, as suggested by the dispersed attention. Finally, the model also incorrectly predicts that there is only one lane in the direction of travel, which, looking at the image, seems correct. However, that particular road is a one-way road, meaning that the model should have predicted 2 lanes. The only indication that this

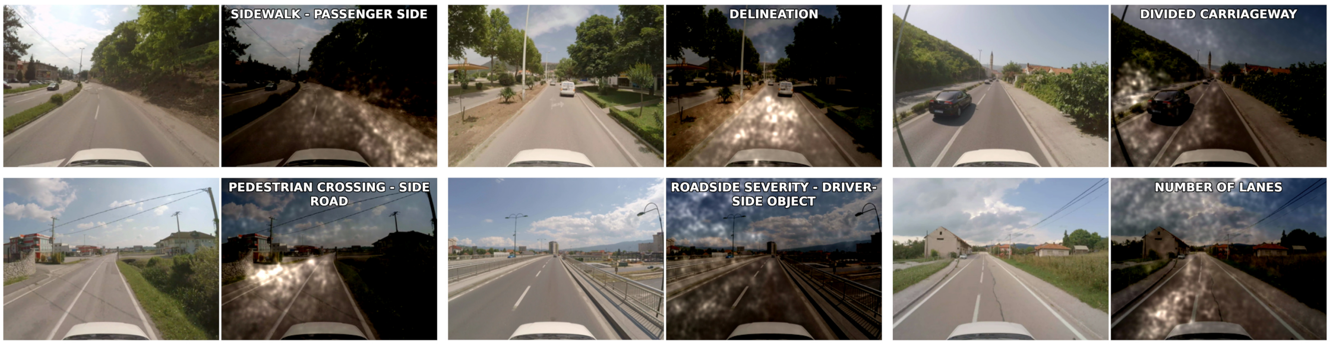
SIDEWALK - PASSENGER SIDE
DELINEATION
DIVIDED CARRIAGEWAY
PEDESTRIAN CROSSING - SIDE ROAD
ROADSIDE SEVERITY - DRIVER-SIDE OBJECT
NUMBER OF LANES

Fig. 2. Each image pair highlights regions which are responsible for the model decision. As shown in the images, the chosen attributes are, respectively: Sidewalk - passenger-side, Delineation, Divided carriageway (top), and Pedestrian crossing - side road, Roadside severity - driver-side object, and Number of lanes (bottom).

is a one-way road is a traffic sign which appears more than 400m before the segment.

VII. Conclusion

We have presented a preliminary feasibility study on automatic recognition of iRAP road-safety attributes by relying on monocular video as the only input modality. We have addressed a subset of 33 iRAP attributes with sufficient coverage in our dataset. We have devised a custom convolutional model specifically for this purpose. Our model transforms the input video clip into a shared convolutional representation, which is subsequently classified in a per-attribute multi-task manner. Thus, inference produces 33 categorical distributions for each 10-meter segment of the input video. The model is trained in an end-to-end fashion on actual annotations by iRAP experts.

Our study shows that the described setup suffers from many problems that are not present in academic datasets. Numerous attributes suffer from extreme class imbalance. Because of that, some very simple models that ignore infrequent classes might seem deceptively good. It is easy to envision a scenario where a rare but dangerous attribute is missed because the model ignores the rare classes to maintain high accuracy. We address this problem by learning with the balanced cross-entropy loss and evaluating accuracy with the per-attribute macro-F1 metric.

The recognition success wildly varies across the attributes. Some attributes such as Delineation or Divided carriageway are ready for full automation, even with our simple model and a limited quantity of training data. On some other attributes, our model is not ready for industrial exploitation, either because of too few training data (Pedestrian crossing, Sealed road) or because of the sheer difficulty of the recognition problem (Intersection quality, Roadside severity).

Suitable directions for future work include improving the recognition model, better exploitation of knowledge transfer, semi-supervised learning, and increasing the quantity of training data.

ACKNOWLEDGMENT

This work has been has been made possible under the SLAIN project (Saving Lives Assessing and Improving TEN-T Road Network Safety) co-financed by the Connecting Europe Facility of the European Union. We also thank the European Regional Development Fund for support within the frame of the project DATACROSS (KK.01.1.1.01.0009).

References

[1] World Health Organization, "Global status report on road safety 2018: summary," 2019.

[2] iRAP - International Road Assesment Programme, "iRAP Coding Manual Version 5.0 – Drive on Right Edition," 2019.

[3] Road Safety Foundation, "Engineering Safer Roads: Star Rating roads for in-built safety," 2015.

[4] iRAP - International Road Assesment Programme, "iRAP Star Rating and Investment Plan Implementation Support Guide," 2017.

[5] Steve Lawson, "iRAP methodology and safer roadsides," 2017.

[6] S. Segvic, K. Brkic, Z. Kalafatic, V. Stanisavljevic, M. Sevrovic, D. Budimir, and I. Dadic, "A computer vision assisted geoinformation inventory for traffic infrastructure," in *Proc. ITSC*, 2010, pp. 66–73.

[7] Z. M. Jan, B. K. Verma, J. Affum, S. Atabak, and L. Moir, "A convolutional neural network based deep learning technique for identifying road attributes," in *Proc. IVCNZ*. IEEE, 2018, pp. 1–6.

[8] T. G. P. Sanjeewani and B. K. Verma, "Learning and analysis of ausrap attributes from digital video recording for road safety," in *Proc. IVCNZ*. IEEE, 2019, pp. 1–6.

[9] S. Segvic, K. Brkic, Z. Kalafatic, and A. Pinz, "Exploiting temporal and spatial constraints in traffic sign detection from a moving vehicle," *Mach. Vis. Appl.*, vol. 25, no. 3, pp. 649–665, 2014.

[10] P. Foucher, Y. Sebsadji, J. Tarel, P. Charbonnier, and P. Nicolle, "Detection and recognition of urban road markings using images," in *Proc. ITSC*. IEEE, 2011, pp. 1747–1752.

[11] R. Yan, A. Natsev, and M. Campbell, "Formal models and hybrid approaches for efficient manual image annotation and retrieval," in *Semantic Mining Technologies for Multimedia Databases*, 2009.

[12] I. J. Goodfellow, Y. Bengio, and A. C. Courville, *Deep Learning*, ser. Adaptive computation and machine learning. MIT Press, 2016.

[13] Y. Bengio, A. C. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[15] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*. IEEE Computer Society, 2017, pp. 2261–2269.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015.

[17] V. Zadrija, J. Krapac, S. Segvic, and J. Verbeek, "Sparse weakly supervised models for object localization in road environment," *Comput. Vis. Image Underst.*, vol. 176-177, pp. 9–21, 2018.

[18] I. Sikiric, K. Brkic, P. Bevandic, I. Kreso, J. Krapac, and S. Segvic, "Traffic scene classification on a representation budget," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 1, pp. 336–345, 2020.

[19] I. Krešo, J. Krapac, and S. Segvic, "Efficient ladder-style densenets for semantic segmentation of large images," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–11, 2020.

[20] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic, "In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images," in *Proc. CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 12 607–12 616.

[21] W. Song, S. Workman, A. Hadzic, X. Zhang, E. Green, M. Chen, R. R. Souleyrette, and N. Jacobs, "FARSA: fully automated roadway safety assessment," *CoRR*, vol. abs/1901.06013, 2019. [Online]. Available: http://arxiv.org/abs/1901.06013

[22] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," *CoRR*, vol. abs/1901.11504, 2019.

[23] A. R. Zamir, A. Sax, W. B. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proc. IJCAI*, S. Kraus, Ed. ijcai.org, 2019, pp. 6241–6245. [Online]. Available: https://doi.org/10.24963/ijcai.2019/871

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, Y. Bengio and Y. LeCun, Eds., 2015.

[25] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *CoRR*, vol. abs/1502.03044, 2015.

[26] M. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *CoRR*, vol. abs/1508.04025, 2015.

[27] I. Kreso, D. Causevic, J. Krapac, and S. Segvic, "Convolutional scale invariance for semantic segmentation," in *Proc. GCPR*, B. Rosenhahn and B. Andres, Eds., 2016.

[28] Y. Cui, M. Jia, T. Lin, Y. Song, and S. J. Belongie, "Class-balanced loss based on effective number of samples," *CoRR*, vol. abs/1901.05555, 2019.

[29] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. ICLRW*, 2014.