CrossFeat: Semantic Cross-modal Attention for Pedestrian Behavior Forecasting

CrossFeat: Semantic Cross-modal Attention for Pedestrian Behavior Forecasting

Francesco Marchetti, Taylor Mordan, Federico Becattini, Lorenzo Seidenari, Alberto Del Bimbo, Alexandre Alahi, *Member, IEEE*

***Abstract*—Forecasting pedestrian behaviors is essential for autonomous vehicles to ensure safety in urban scenarios. Previous works addressed this problem based on motion alone, omitting several additional behavioral cues helping understanding pedestrians' true intentions. We address the problem of forecasting pedestrian actions through joint reasoning about pedestrians' past behaviors and their surrounding environments. For this, we propose a Transformer-based feature fusion approach, where multi-modal inputs about pedestrians and environments are all mapped into a common space, then jointly processed through self and cross-attention mechanisms to take context into account. We also use a semantic segmentation map of the current input frame, rather than the full temporal visual stream, to further focus on semantic reasoning. We experimentally validate and analyze our approach on two benchmarks about pedestrian crossing and Stop&Go motion changes, which rely on several standard self-driving datasets centered around interactions with pedestrians (JAAD, PIE, TITAN), and show that our semantic joint reasoning yields state-of-the-art results.**

***Index Terms*—Autonomous Driving, Deep Learning, Behavior Prediction**

I. Introduction

In autonomous driving, planning beforehand is fundamental to avoid collisions: the autonomous agent must be equipped with forecasting modules to predict in advance a set of possible future scenarios taking into account the trajectories and behaviors of others. In practice, a large crop of literature already exists regarding trajectory forecasting [1]–[8] and it is foreseeable that interactions with other vehicles will be eased by intra-vehicle cooperation [9]–[11] as automotive pervasiveness increases. On the other hand, safely interacting with pedestrians is not straightforward. Their motion can be erratic and they may abruptly take decisions that are not inferrable from past trajectories alone. Autonomous vehicles, just like human drivers, must understand behaviors and forecast intents to avoid danger. Works in the literature have addressed understanding human behaviors [12], [13] and predicting their intents [14], [15], often focusing on forecasting pedestrian crossings [16], [17] or, more recently, state changes in their motion [18]. To obtain such a fine comprehension of human agents, cues from the environment can be used, *e.g.*, it is more likely that a pedestrian might cross in front of the vehicle in the presence of zebra crossings or when traffic lights allow it.

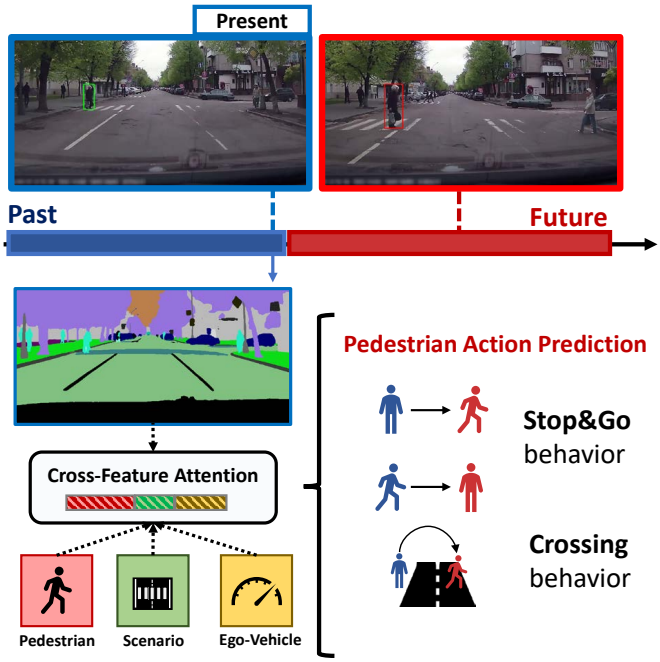


Fig. 1: Pedestrian behavior forecasting. Given the scene information at the current instant and the past temporal data referring to the pedestrian and the ego-vehicle, CrossFeat predicts Stop&Go and Crossing behaviors that may occur. A multi-modal cross-attention mechanism relates the features and determines which is most relevant to the final output.

In our work, we address the problem of forecasting pedestrian behaviors by relating their motion patterns to environmental features, as illustrated in Fig. 1. We propose an architecture that exploits cross-modality attention by leveraging at the same time information about the ego-vehicle, the pedestrian and the surrounding context. Ego-vehicle data can be easily provided by the car itself, whereas pedestrian position and motion can be reliably inferred via detection or tracking [19], [20]. Useful attributes describing the pedestrian can also be derived with vision models [12]. We focus on the reasoning and interaction part of the problem rather than on the feature extraction part, so we work under the assumption that such quantities are given or can be easily inferred if needed as done in prior works [17], [18], [21]. As for the surrounding context, we choose to adopt

Francesco Marchetti, Lorenzo Seidenari and Alberto Del Bimbo are with the University of Florence, 50121 Firenze, Italy (e-mail: francesco.marchetti@unifi.it; lorenzo.seidenari@unifi.it; alberto.delbimbo@unifi.it).

Federico Becattini is with the University of Siena, 53100 Siena, Italy (e-mail: federico.becattini@unisi.it).

Taylor Mordan and Alexandre Alahi are with the École polytechnique fédérale de Lausanne, Switzerland (email: alexandre.alahi@epfl.ch, taylor.mordan@epfl.ch)

only semantic segmentation maps instead of relying on raw RGB inputs as done by prior work. This choice is motivated by several aspects. First of all, no additional sensor is required to obtain semantic maps, which can be obtained onboard with pre-trained segmentation models, such as [22]. Segmentations offer a higher comprehension of the semantics of the scene while being less affected by noise (*e.g.*, shadows) and lighting conditions. Furthermore, learning from segmentations is easier since a shallow convolutional network can be exploited rather than training from scratch or fine-tuning large networks to model the world.

Overall, our proposed approach focuses on spatio-temporal relations between pedestrians and the environment. This is achieved by explicitly feeding as input the semantics of the scene, allowing the model to establish correlations between the presence (or absence) of elements in the surrounding area with information gathered from both the pedestrians and the ego-vehicle. This distinguishes our approach from prior works, which mainly focus on temporal relations alone [1], [16], [17], [23], [24]. Notably, the usage of multi-modal attention mechanisms makes the whole model interpretable by design. We show that meaningful visual explanations can be derived, that underline the relative importance between elements in the scene and the motion of pedestrians.

The main contributions of our paper are the following:

1. We present CrossFeat, an encoder-decoder transformer-based architecture leveraging multi-modal input features and establishing relative importance between semantic image regions and pedestrian features to forecast road crossings and changes of state in motion patterns.

2. The decisions of the model are interpretable thanks to visual explanations obtained from multi-modal cross-attention. We show that relevant scene elements are taken into account by the model to make predictions.

3. We experiment on two different benchmarks concerning pedestrian behavior prediction. We report state-of-the-art results on the recently introduced Stop&Go benchmark and on-par or superior results on pedestrian crossing benchmarks.

II. Related Works

Pedestrian Crossing Prediction. Most works forecasting pedestrian crossings focus on effective methods to fuse and relate multi-modal inputs gathered from the scene and the pedestrian. Several approaches have been proposed, such as stacked recurrent neural networks that gradually fuse data from different modalities [25] and asymmetrical Bi-RNNs that correlate the temporal sequence of skeletal points with the speed of the ego-vehicle. As in [25], sequences of human poses are often used, either alone [26] or along with the local context [27], [28] or bounding boxes [29]. The environment plays a crucial role in forecasting human behaviors. In [30], the model uses the features of the different composition of the road context like zebra crossing and road lane. PCPA [31] uses a 3D convolutional neural network to embed the sequence of visual information and dedicated RNNs for high-level inputs describing the pedestrian and the ego-vehicle. In [32], the model uses a monocular depth estimation map to better understand the distances and relationships between the target pedestrian and the other interacting road users. Attention modules have also been used in [31] to determine the temporal importance of features. Attention is also used to merge features of various high-level and visual inputs to generate the final output. Extending on this idea, Song *et al.* [23] added a graph model to understand pedestrian interactions and an Interframe and Intraframe Gated Recurrent Unit (II-GRU) to deal with long-term temporal dependencies. In [33], a graph convolutional model is developed to understand spatio-temporal relationships in the scene across video frames. More recently, hybrid fusion models have been developed where the non-visual and visual features are first analyzed separately, generating their own features, and then merged to generate the output [24], [34]–[36]. In [37], [38] a multitask learning approach is explored: in addition to predicting the action, the position and trajectory of the pedestrian are also predicted. In [39], starting from a single frame, both pose and intention to cross of the pedestrian are estimated at the same time. In [40], the model is composed of two stages where predicting future video frames also helps to predict the pedestrian's future action.

A common way to model dependencies between multiple temporal features is the usage of transformers. In crossing prediction tasks, features representing the pedestrian and the ego-vehicle have been concatenated and given as input to a transformer encoder where an attention process takes place between the features at different timesteps [16], [41], [42]. Meanwhile, a video encoder analyzes the sequence of frames captured by the ego-camera. In [17], a Temporal Adaptive Mask Transformer is introduced to weigh past and present features differently. Instead, [21] focuses on the importance of pedestrian's bounding boxes as input to a transformer. In general, such works exploit the transformer's attention to learn the importance of different tokens of a time sequence. We instead leverage attention to establish the relative importance of multi-modal tokens. We show that mapping different modalities to a common representation space in which to perform attention provides effective predictions as well as a simple way to inspect relative importance across heterogeneous inputs.

Stop-and-Go Pedestrian Behavior. As a more fine-grained extension of the pedestrian crossing prediction task, the Stop-and-Go forecasting problem was studied by Guo *et al.* [18], introducing the Stop&Go benchmark. Inspired by [34], the authors separately process visual features and high-level descriptors of pedestrians to then fuse them with a fusion hybrid model. Stop and go movements are harder to spot compared to crossings due to their abruptness and less clear correlation with the previous motion of the pedestrian: it is when analyzed in the context of the surrounding environment that such motion patterns acquire meaning and become easier to foresee.

The most similar work to ours is [18], which, like our proposed method, addressed the problem of forecasting state changes in pedestrian motion. Nonetheless, the approach proposed in [18] leverages RGB streams, whereas we use a single semantic segmentation image to relate semantic entities to higher-level features, such as pedestrian motion or presence/absence of scene elements, through attention. Another notable difference is the nature of the proposed model. In

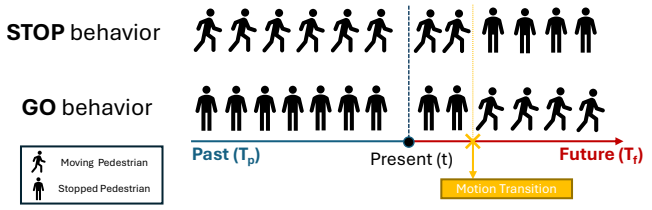


Fig. 2: Stop&Go forecasting task. In a timestep between t and T_f, the motion state of the pedestrian changes compared to that between T_p and t.

fact, CrossFeat is trained end-to-end, without requiring a two-phase optimization like [18], where a ResNet18 and recurrent modules modeling high-level features are trained separately.

In addition, differently from prior work, CrossFeat does not processes past images sequentially, but only the semantic segmentation at the current instant is needed, making the model simpler yet effective. This also helps to interpret the model's decisions since attention weighs important image regions rather than locating relevant timesteps. Furthermore, to understand the dependencies and spatial information within the semantic frame, it is sufficient to use a shallow CNN instead of a deep CNN pre-trained on other datasets.

III. Problem Formulation and Benchmarks

We address the problem of pedestrian behavior forecasting, *i.e.*, given a set of T_p past, possibly multimodal, observations of a scene until time t, predict if a given behavior will happen in the future temporal interval $[t, t+T_f]$. Specifically, we address three behavior prediction problems: crossing, Stop and Go. The crossing behavior has been first analyzed in [31]. Guo et al. [18] extended [31] by forecasting Stop and Go behaviors (Fig. 2). Forecasting a Stop behavior means to detect if a pedestrian moving until timestep t will *Stop* in the interval $[t, t+T_f]$, while for the Go behavior we must detect the opposite transition: from a still state to a future moving state. For the crossing task instead, one must predict whether a pedestrian will start to traverse the road. The tasks are defined at the pedestrian level and we assume its location to be known in all frames up to t. We address all three tasks separately as binary classification problems. Given past observations in $[t-T_p, t]$ and a single pedestrian, we predict if a given behavior will happen in the future time interval $[t, t+T_f]$.

Data Setting. We can formulate the three distinct problems in a single way, having identical inputs and the same binary output, yet with different meanings. For each example, multi-modal inputs represent the urban scenario, the pedestrian and ego-vehicle information. First of all, we have static information at the present time t describing the composition of the urban context and where the observed agent is located within it: (i) an RGB frame of size $[H, W, 3]$ captured by the camera mounted in the ego-vehicle; (ii) six high-level attributes describing the context, namely the number of traffic lanes, the presence or absence of an intersection, a zebra crossing, or a traffic signal and the lateral or longitudinal direction that pedestrians can follow in the scene; (iii) the bounding box of the pedestrian within the RGB frame.

TABLE I: Attributes processed by CrossFeat. Both static information as well as dynamic cues that are observable from the vehicle's point of view are fed to the decoder module.

Attribute	Description
Bounding box (BB)	Pedestrian position
Motion (M)	Sequence of past pedestrian bounding boxes
Behavior (B)	motion; gaze; nodding; hand gesture
Scene (S)	#lanes; intersection; crossing; traffic sign; direction
Velocity (V)	Ego-vehicle past speed/acceleration

Furthermore, we have temporal information relative to the past timesteps $[t-T_p, t]$ describing the dynamic aspects of the pedestrian and the ego-vehicle: (i) the motion of the pedestrian within the previous frames, represented as the sequence of bounding box coordinates plus their speed obtained as the difference between adjacent box centers; (ii) a sequence of binary attributes describing non-verbal behaviors of the pedestrian including whether it is moving, looking, nodding, or making hand gestures in every frame; (iii) the sequence of motion states of the ego-vehicle, describing speed or acceleration, depending on data available in each dataset.

All the attributes leveraged by CrossFeat are summarized in Tab. I. Given this temporal and static multi-modal information, our goal is to predict whether in the next $[t, t+T_f]$ seconds the pedestrian will start crossing the road or whether a motion transition (Stop $\rightarrow$ Go or Go $\rightarrow$ Stop) will occur. We can train the same architecture for both tasks on different benchmarks since they share the same output structure, *i.e.*, a probability that a behavior will happen.

IV. Method

CrossFeat is a transformer-based model that, given a set of multi-modal inputs regarding the surrounding scene, a pedestrian and the ego-vehicle, generates the probability of crossing or changing state for the pedestrian, depending on the task for which it is trained. The set of inputs is detailed in Sec. III. The main idea is to take all inputs, which have different modalities, structures and meanings, and to map them into a common multi-modal space $\mathcal{S}$, enabling the usage of attention mechanisms across modalities (see Tab. I). We formalize $\mathcal{S}$ as

$$\mathcal{S} = \{Feature\,(i)\}, i = \texttt{RGB}, \texttt{M}, \texttt{B}, \texttt{S}, \texttt{V} \quad (1)$$

The overview of the architecture is shown in Fig. 3. First, the semantic segmentation of the current frame is fed to a Convolutional Neural Network (CNN) to generate feature maps that capture spatial information. Such feature maps are then enriched through self-attention and then related via cross-attention to additional multimodal inputs encoding scene attributes and pedestrian behaviors. A final attention-based module generates the output probability.

Image Segmentation. First of all, the RGB frame of size $[H, W, 3]$ at timestep t is processed by a model that generates the segmentation (size $[H, W, 1]$) of the urban context by assigning a semantic label to each pixel. Semantic labels are represented as one-hot encodings. We discard RGB information, relying only on semantic segmentation, since it captures higher-level content and it is easier to interpret for a deep learning

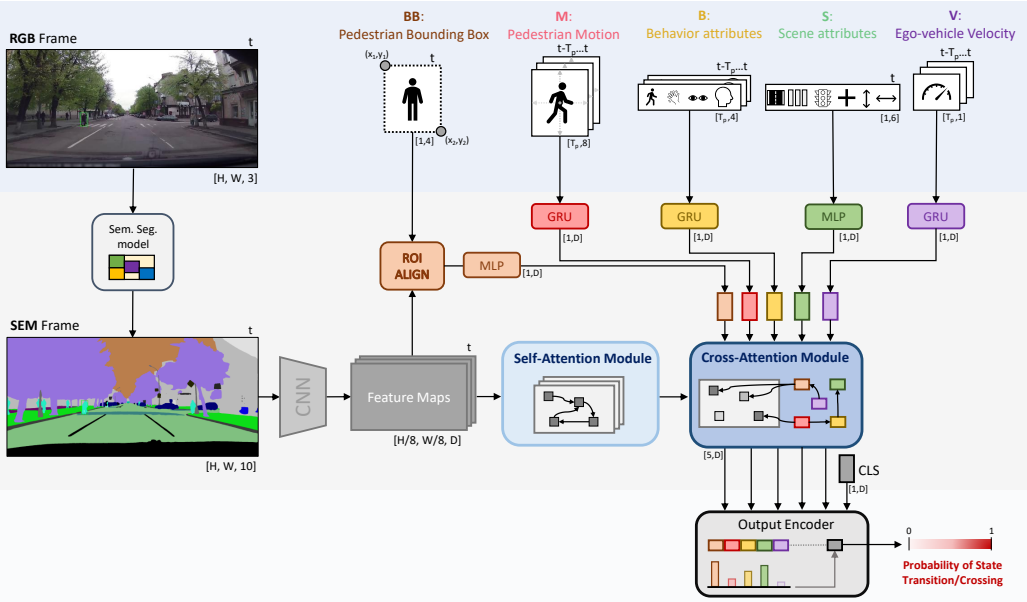


Fig. 3: The frame is converted into a semantic segmentation and is then fed to a CNN. A Self-Attention module (*i.e.*, a transformer encoder) is used to correlate spatial features among them. Multi-modal inputs are processed in parallel and used as queries for a Cross-Attention module (*i.e.*, a transformer decoder) to weigh relative importance with each other and the frame. Finally, a classification token (CLS) is fed to a transformer encoder classifier. D is the size of each feature. GRU is a Gated Recurrent Unit module. MLP is a MultiLayer Perceptron.

model, being less noisy and not affected by illumination changes. It has to be noted that such issues are actually still present in the overall system, since the whole pipeline starts from the RGB frame. However, we are shifting the responsibility of semantically characterizing the scene from the pedestrian behavior prediction model to the segmentation model. In this way, we decouple the effectiveness of the semantic segmentation from the analysis of the pedestrian. We will show in our experiments (see Sec. V-A) that: (i) any segmentation model can be adopted and that the quality of the segmentations reflects on the quality of the predictions; (ii) relying on a high-quality segmentation model makes the model more effective than directly processing RGB data.

Self-Attention Module. In our model, the semantic frame is first processed using a CNN to obtain a compact feature map. We use the pixels of the resulting feature map as separate inputs for the Self-Attention module. To this end, a positional encoding is added to the convolutional feature map, to retain the spatial information of each location. As positional encoding, we used the Sinusoidal Positional Embedding. In fact, as Self-Attention module we employ a transformer encoder that processes the inputs in parallel, disregarding the ordering of the data. Similar approaches have been used previously in vision transformers [43], [44]. The need for a Self-Attention module is motivated by the need to model spatial dependencies in the input, *i.e.*, obtaining a meaningful representation that is spatially aware of the surrounding elements in the scene. Then, each output of the module is independently projected into the space $\mathcal{S}$ in order to be jointly processed with the additional multi-modal inputs through cross-attention.

Data Encoding. The multi-modal inputs describing the pedestrian, the scene and the ego-vehicle are processed by separate learnable modules. In the following, we refer to each feature with symbols referring to their structure (Tab. I): `BB` for the bounding box of the pedestrian, `M` for pedestrian motion, `B` for behavioral attributes, `S` for scene attributes and `V` for the ego-vehicle velocity. The bounding box (`BB`) of the pedestrian is used in combination with ROI-Align [19] to extract a partial feature map from the convolutional one derived from the frame, focusing on the location of the pedestrian. A Multilayer Perceptron (MLP) then maps it into the multi-modal feature space $\mathcal{S}$. The pedestrian motion (`M`), the behavior attributes (`B`) and the ego-vehicle velocity (`V`) are given as input to dedicated recurrent neural networks (we use Gated Recurrent Units - GRU [45]) to generate a descriptor that summarizes the temporal information up to the current timestep t for each modality. The scene attributes (`S`) instead, being referred only to the current instant t, are processed by a Multilayer Perceptron (MLP). In this way, we obtain embeddings describing the temporal evolution of the pedestrian motion with respect to the camera, of the non-verbal body behaviors of the pedestrian, of the ego-vehicle speed and the embedding describing the key elements of the scene. All of these embeddings are projected from their own latent space into the common latent space $\mathcal{S}$. Each feature is provided as a query to the Cross-Attention module.

Cross-Attention Module. To relate the multi-modal features to each other and to the input frame, we rely on a Cross-Attention module. The goal is to enrich each multi-modal feature by discovering correlations among them, while also attending relevant elements (both semantically and spatially) in the frame. To perform cross-attention we use a transformer decoder, feeding the multi-modal features as input tokens (or

queries) and the frame tokens processed by the Self-Attention module as context. Note that a transformer encoder could also have been used instead of a transformer decoder by using self-attention rather than cross-attention, processing all the tokens in parallel. In our experiments, however, we show that a decoder structure outperforms an encoder in practice (Tab. VI). Moreover, the decoder structure has two relevant aspects: *(i)* it performs an initial self-attention step across queries and it keeps the number of outputs limited; *(ii)* all spatial information from the input frame is discarded, yet by relating the queries with spatial semantic information, we incorporate it into the output tokens.

Output Encoder. Finally, the multi-modal features generated by the Cross-Attention module are fed to another transformer-based encoder. Along the input tokens, we feed a trainable classification token (CLS) initialized randomly and concatenated to the features, which weighs the importance of the features and generates an output. A fully-connected layer, followed by a sigmoid function, generates the binary classification output. A similar classification mechanism based on transformer encoder layers can be found in BERT [46].

V. Experiments

Datasets and Metrics. The experiments were performed using the data split and metrics proposed in the papers [18] and [31]. To evaluate the performance of the model, Average Precision (AP) is used in Stop&Go benchmark [18], while Accuracy (ACC), Area under the ROC Curve (AUC) and F1 metrics are used in the pedestrian crossing prediction benchmark [31]. The two benchmarks are based on three real public datasets dedicated to autonomous driving in urban scenarios, each with its own peculiarities and characteristics. Joint Attention for Autonomous Driving (JAAD) [47] is composed of short-duration HD videos captured at 30 FPS (5-10 seconds) of pedestrians near areas where it is possible to cross. Pedestrian Intention Estimation (PIE) [15] has a similar structure to JAAD but with videos of greater length (about 10 minutes). Trajectory Inference using Targeted Action priors Network (TITAN) [48] focuses on all moving agents, both pedestrians and vehicles, in a highly populated urban environment. In particular, Stop&Go is based on JAAD, PIE and TITAN, whereas the crossing prediction benchmark is based only on JAAD and PIE. While JAAD and PIE are dedicated strictly to pedestrian crossings on the street, TITAN is a more generic dataset covering the movements of people who do not necessarily interact with traffic. In the crossing benchmark, the experiments are performed on two different splits of the JAAD dataset. JAAD_{all} includes all pedestrians identified by the camera while JAAD_{beh} only considers those pedestrians who are close to the road and have the intention to cross. For each dataset, the RGB frames captured by the camera mounted in the ego-vehicle and the 2D bounding boxes of pedestrians are provided. PIE and JAAD also provide information about pedestrian behavior and scene characteristics.

Implementation and Training Details. In the crossing task, we observe 0.5s in the past (15 timesteps at 30 fps) to predict if the crossing occurs between the next 1s and 2s. This leaves a gap of at least 1s prior to the event, so to react safely. In Stop&Go, sequential data is downsampled compared to the crossing dataset: we observe 1s in the past (5 timesteps at 5fps) to predict state transitions in the next 2 seconds. We used Mask2Former [22] pre-trained on Mapillary Vistas [49] to obtain 10-channel[1] segmentations out of 1281x481 RGB frames. The CNN used to process segmentations has 3 layers, each with Batch Normalization, 0.5 dropout and ReLUs. Each layer has kernel 5×5 and in the first layer a 2D max-pooling is used. The channel dimensions for the 3-layer CNN are 16, 32, 64 respectively. The embeddings of each input generated by the dedicated GRU recurrent networks (`M`, `B`, `V`), the MLP modules (`BB`, `S`) and the CNN have dimension 64. In the Self-Attention and Cross-Attention modules we use a single transformer layer with 4 attention heads. In the Output Encoder module, we use a single layer with a single attention head. Training is end-to-end with Adam using batch size 8 and learning rate 1e-4. We used a weighted Binary Cross-Entropy loss since the dataset is unbalanced towards negative examples (*i.e.*, where crossing or change of state does not occur). The weights are chosen as the inverse of the number of examples for each class.

TABLE II: Inference time breakdown for CrossFeat and all the additional modules that provide the inputs.

Module	Inference time
CrossFeat	15ms
Mask2Former	200ms
Attribute classifier [47]	25ms
Pedestrian tracker [50]	52ms
Pedestrian behavior classifier [12]	76ms
Overall	368ms

Parameters and Inference Time. The CrossFeat model was developed to have the advantage of using the attention mechanisms of transformers but also the smallest possible number of parameters. In fact, for the Self- and Cross-Attention modules, we use a single attention layer with 4 heads. In total, the number of trainable parameters of CrossFeat is 366,610. The inference time of the model is 15ms based on having the semantic image and the attributes of the scene, bounding boxes and the behaviors of the pedestrian available. So the model can process inputs at around 66fps. The inference time to generate the semantic segmentation by Mask2Former is 0.20s. So, the images can be generated at around 5fps. In a real scenario, where all the inputs for CrossFeat are not available (see Tab. I), we can leverage additional modules to infer them. Scene attributes can be obtained with the AlexNet-based architectures proposed in [47], which runs on our hardware at 39.5fps (0.0253s per image). Similarly, a pedestrian multi-object tracking module [50] could be used to detect and track pedestrians and pedestrian behaviors can be obtained using [12], which is based on OpenPifPaf [51]. The models run respectively at 19fps and 13fps. In Tab. II we show an inference time breakdown of CrossFeat and all the additional modules, with an Intel(R) Core(TM) i7-2600K CPU@3.40GHz, and a GeForce RTX 2080 Ti GPU, 12GB of RAM.

[1]10 labels: *Road*, *Buildings*, *Vegetation*, *Ego*, *Car*, *Pedestrian*, *Sidewalk*, *Crosswalk*, *Traffic Line*, and *Other*

TABLE III: Quantitative results on Stop&Go benchmark (test set) in Average Precision (%).

	Go			Stop		
Model	JAAD	PIE	TITAN	JAAD	PIE	TITAN
Static [18]	73.3	61.2	60.9	58.7	62.5	59.1
CrossFeat Static	**74.6**	60.4	**63.2**	**69.2**	**67.2**	**63.7**
Video [18]	76.4	64.7	62.9	62.9	64.2	61.7
Hybrid [18]	85.9	**70.2**	65.1	67.8	65.4	63.6
TED [21]	62.4	59.9	65.0	60.8	57.8	59.1
MTL [38]	62.0	63.3	64.5	67.6	59.6	56.7
CrossFeat (Ours)	**88.9**	68.1	**70.1**	**75.4**	**71.0**	**67.3**

TABLE IV: Quantitative results: crossing action prediction benchmark (test set).

	PIE			**JAAD**$_{all}$			**JAAD**$_{beh}$		
	Acc	AUC	F1	Acc	AUC	F1	Acc	AUC	F1
PCPA [31]	0.86	0.86	0.77	0.85	0.86	0.68	0.58	0.5	0.68
ATGC [47]	0.59	0.55	0.39	0.64	0.60	0.53	0.48	0.41	0.62
SF-GRU [25]	0.86	0.83	0.75	0.83	0.77	0.58	0.58	0.56	0.65
I3D [52]	0.79	0.75	0.64	0.82	0.75	0.55	0.62	0.51	0.75
TroupSPI-Net [53]	0.88	0.88	0.80	0.85	0.73	0.56	0.64	0.56	0.76
TAMformer [17]	0.88	0.86	0.79	0.88	0.83	0.68	0.73	0.69	0.80
IntFormer [41]	0.89	0.92	0.81	0.86	0.78	0.62	0.59	0.54	0.69
BiPed [37]	0.91	0.90	0.85	0.83	0.79	0.60	0.68	0.60	0.78
PedGraph+ [54]	0.89	0.90	0.81	0.86	0.88	0.65	0.70	0.70	0.76
MTL [38]	0.91	0.93	0.82	**0.90**	**0.95**	**0.76**	0.63	0.65	0.77
SGM [23]	**0.92**	0.91	**0.86**	0.87	0.81	0.65	0.68	0.60	0.78
TED [21]	0.91	0.91	0.83	-	-	-	-	-	-
CrossFeat (Ours)	0.90	**0.94**	0.81	**0.90**	0.94	0.74	**0.76**	**0.75**	**0.83**

TABLE V: Ablation study. We train different models by removing one feature at a time to assess its importance.

	Go			Stop		
CrossFeat	JAAD	PIE	TITAN	JAAD	PIE	TITAN
w/o ROI-Align	84.2	67.1	59.7	67.8	64.0	60.8
w/o Motion	81.1	66.4	68.2	73.2	67.5	63.7
w/o Behavior	79.5	67.3	-	71.1	67.9	-
w/o Scene	73.0	67.5	-	**75.4**	70.3	-
w/o Velocity	83.9	62.7	65.8	74.1	70.0	64.5
All features	**88.9**	**68.1**	**70.1**	**75.4**	**71.0**	**67.3**

A. Results

The results obtained by the proposed model on both benchmarks are reported in Tab. III and Tab. IV. In Tab. III, we can see the performance of CrossFeat trained to forecast the Stop and Go behaviors. The models were evaluated on the 3 datasets present in the benchmark: JAAD, PIE and TITAN. The comparison is made with the results obtained from the model proposed in [18] and two other state-of-the-art models from the literature, TED [21] and MTL [38], that we trained from scratch for this task by adapting them from the crossing prediction task to Stop&Go forecasting. CrossFeat is able to better classify Stop and Go behaviors in most of the datasets in the benchmark. Averaging across configurations, we have an increase of about 4.6% in Average Precision. Furthermore, a comparison between our model and the one proposed in [18] was carried out using a static setting. The static setting consists of using only the frame and the bounding box of the pedestrian at the present timestep. From the comparison, we can deduce that even our static model performs better in almost all the datasets in the benchmark, demonstrating the effectiveness of using semantic segmentation and an attention-based model for the present timestep. CrossFeat performs well also in crossing prediction (Tab. IV) achieving on-par or better results than state-of-the-art models, especially in the JAAD$_{beh}$ dataset.

It is worth to notice the results for [21] and [38] in both benchmarks. Although they achieve high results in the crossing benchmark, they have low performances in the Stop&Go benchmark. Instead, our model manages to obtain high results in both. The main difference is the lack of scene information in [21] and [38] (neither RGB nor semantic labels). This highlights the importance of using road scenario information when a pedestrian exhibits different motion patterns between past and future.

Ablation Study. Ablation studies related to the Stop&Go benchmark were carried out to analyze the effectiveness of each feature input to the model (Tab. V), of the need for each different module in the model (Tab. VI), of semantic segmentation in comparison with RGB (Tab. VII) and of choice of the segmentation module (Tab. IX). The feature analysis in Tab. V was performed by removing one feature at a time and training and testing the model again. We can observe that each feature is relevant to perform in the best possible way and that features can have different importance in different configurations. For example, in Go PIE the use of ego-vehicle speed is important while in Stop JAAD and Stop PIE the performance degradation is small when removing the velocity. On the other hand, in Stop JAAD and Stop PIE more degradation occurs if the model does not utilize the Behavior or Motion features.

Tab. VI shows the results using baselines to process the multi-modal features. In the *Single Query* case, each feature is flattened, concatenated, processed with an MLP, and then input as a single query to the Cross-Attention module. In *Concatenation*, instead of using Self-Attention and Cross-Attention modules, we concatenate all the features generated by single input data modules. We feed this concatenation into an MLP for binary classification. In *Self-attention decoding*, we feed all the features together to a transformer encoder – those generated by the Self-Attention module starting from the semantic segmentation image with those elaborated by the modules that observe the attributes of the pedestrian, scene and ego-vehicle. The resulting features are then fed to an Output Encoder with a CLS token for classification. From the overall worsening of the results, the importance of using different pipelines for each feature is highlighted, sending them as multiple queries to the cross-attention module. Moreover, the importance of using attention mechanisms between multi-modal features instead of simple concatenation has been demonstrated. Finally, to make queries interact with image features it is better to use a transformer decoder.

In Tab. VII, the experiments were performed using different types of frame inputs (Semantic Segmentation or RGB) and different backbones (Imagenet pre-trained ResNet18 and CNN from scratch described in Sec. V). In all configurations, the performance of the model using semantic segmentation and CNN is better than those using the RGB frame with both a

TABLE VI: Ablation study. We use different baseline modules to treat the multi-modal input.

	Go			Stop		
CrossFeat	JAAD	PIE	TITAN	JAAD	PIE	TITAN
Single query	66.6	66.8	62.3	55.6	60.1	60.5
Concatenation	83.6	66.8	68.3	69.8	67.8	65.7
Self-attention decoding	**88.9**	63.1	63.5	73.6	59.9	60.5
Complete	**88.9**	**68.1**	**70.1**	**75.4**	**71.0**	**67.3**

TABLE VII: Ablation study. We train models on RGB and semantic segmentations. Segmentations can effectively be leveraged by a light-weight CNN, differently from RGB.

		Go			Stop		
Image	Backbone	JAAD	PIE	TITAN	JAAD	PIE	TITAN
RGB	ResNet18	79.7	65.5	61.6	70.7	63.1	64.1
RGB	CNN	86.3	66.4	63.9	69.5	64.3	62.8
Sem. Seg.	CNN	**88.9**	**68.1**	**70.1**	**75.4**	**71.0**	**67.3**

finetuned ResNet18 and a from-scratch CNN. This means that giving the model a semantic segmentation permits to have better information on how the scene is composed.

Further experiments were carried out taking another segmentation module, DeepLabv3+ MobileNet [55]. Semantic images were extracted using a model pre-trained on Cityscapes. To compare the two segmentation models pre-trained on the same dataset, we used the segmentations of Mask2Former also pre-trained on Cityscapes. As we can see from Tab. IX, the segmentations generated by DeepLabv3+ lead to worse performance than using a better segmentation module such as Mask2Former, suggesting that the quality of the segmentation is crucial, as we can also see from the examples in Fig. 6.

Finally, we carried out an experiment giving as input only the present timestep for all the features. In fact, despite our main model assumes to use only the frame at the current timestep to understand the structure of the scene, the features of the pedestrian's movement and behavior and the speed of the ego-vehicle are always sequential data regarding the past up to the present timestep. From Tab. X, we can observe that in some datasets we have only a slight decrease in performance (in one we even have a slight increase). This shows that even with non-sequential data about the past it is possible to forecast pedestrian behaviors effectively.

Noisy Inputs Our evaluation was initially carried out following the protocol defined by the benchmarks, in order to establish a fair comparison with prior work. We carry out a further evaluation by applying noise to the multi-modal input to see how the model behaves when the different inputs are unreliable. The bounding boxes have been perturbed by moving them of a random amount between $[-r * W, r * W]$ along the x-axis, and between $[-r * H, r * H]$ along the y-axis, where W and H are width and height of the bounding box. We varied the perturbation parameter r between 0.1 and 0.5.

In addition to the perturbing bounding boxes, we carried out experiments by perturbing the attributes that describe the scene (Scene attributes S) and those that indicate the non-verbal behaviors of the pedestrian (behavior attributes B). The parameters of the scene and that of the pedestrians' behavior were varied randomly with a probability of 0.5 by flipping the value for binary quantities. The only non-binary attribute is the one describing the number of lanes, which was varied randomly from the minimum number, 0, to the maximum number, 5.

TABLE VIII: Comparison of results between ground-truth and noisy data given as input to CrossFeat.

	Go			Stop		
Data Setting	JAAD	PIE	TITAN	JAAD	PIE	TITAN
Ground-truth data	**88.9**	**68.1**	**70.1**	**75.4**	**71.0**	**67.3**
BB (10% noise)	87.8	67.7	69.2	75.2	70.1	66.7
BB (20% noise)	86.2	67.4	65.3	74.1	68.3	65.8
BB (30% noise)	85.5	67.4	62.7	73.2	66.7	64.6
BB (40% noise)	82.0	67.0	60.7	73.0	65.0	63.5
BB (50% noise)	80.1	66.9	58.8	72.4	64.3	63.2
Noisy Behavior	86.0	65.5	-	70.6	63.7	-
Noisy Scene	70.4	64.0	-	74.3	70.1	-

TABLE IX: Quantitative results using two different semantic segmentation modules (DeepLabv3+ and Mask2Former) trained on two different datasets (Cityscapes and Mapillary).

	Go			Stop		
Image Setting	JAAD	PIE	TITAN	JAAD	PIE	TITAN
DeepLabv3+ (Cityscapes)	86.6	66.3	70.0	69.5	68.6	61.9
Mask2Former (Cityscapes)	**88.9**	67.9	**70.7**	72.3	70.1	63.3
Mask2Former (Mapillary)	**88.9**	**68.1**	70.1	**75.4**	**71.0**	**67.3**

From Tab. VIII, we can observe that by adding noise the model gracefully degrades, still maintaining an accuracy similar to the original model tested on clean data.

Attention Analysis. Through the attention mechanisms of the model, it is possible to have a good degree of explainability of the generated outputs. In particular, we can observe the attention between the classification token of the Output Encoder with the multi-modal features to quantify the importance of each feature when generating the output (Fig. 4). Interestingly, the attention distribution is distributed differently on the splits of Stop&Go benchmarks. For example, if in the TITAN dataset the attention is distributed equally among all features, this is not the case in the other datasets. In JAAD, scene attributes have a high impact, while in Go PIE it is the ego-vehicle speed that contributes the most.

Qualitative Examples. In Fig. 5, we report qualitative samples. We show the frame at the current timestep, the frame in the future (after 2s), and the output of the model. Moreover, we show the attention between query features and tokens of the semantic segmentation and between the query features themselves in the Cross-Attention module. Finally, we show feature importance by observing the attention that is computed inside the Output Encoder. In Fig. 5 (top), the model must decide whether the pedestrian, who is stationary along the sidewalk, will cross the road in future timesteps (Go scenario). CrossFeat correctly predicts the crossing intention of the pedestrian. Furthermore, it is interesting to observe that the model focuses on the pedestrian crossing area and on the car that is moving away. The most important feature of this example is the one concerning the scene. In Fig. 5 (bottom),

TABLE X: Quantitative results varying the input timeframe for bounding boxes, pedestrian behaviors and ego-vehicle's speed.

	Go			Stop		
BB, Behavior, Speed	JAAD	PIE	TITAN	JAAD	PIE	TITAN
Temporal Past Sequence	**88.9**	**68.1**	**70.1**	75.4	**71.0**	**67.3**
Present Timestep	88.5	**68.1**	65.2	**77.3**	66.3	65.2

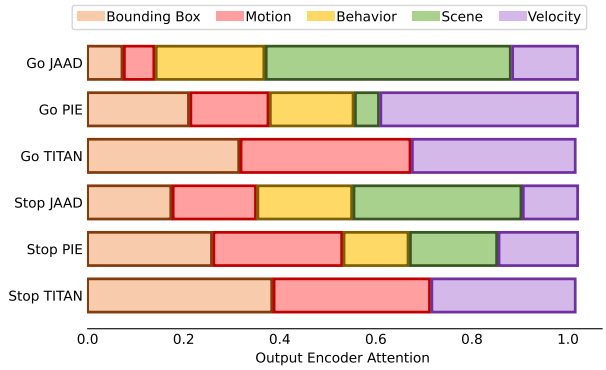


Fig. 4: Attention values of Output Encoder for each split of the Stop&Go benchmark. Each bar represents the attention between each multi-modal feature and classification token (the sum is 1). Values are averaged for all the test samples. The attention focus varies depending on the split the model is observing.

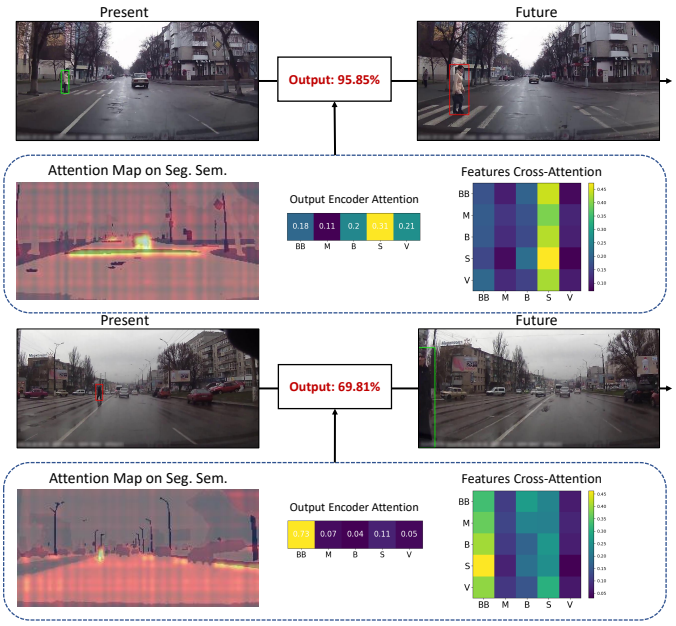


Fig. 5: Go (top) and Stop (bottom) behaviors. Green boxes indicate stationary pedestrians, red ones that are moving. Attentions of query features and semantic segmentation within the Cross-Attention module and Output Encoder are shown.

instead we have a pedestrian moving on the side of the road and the model has to decide if it will stop in the near future. CrossFeat is able to predict the correct action (Stop) with good confidence (69.81%). In this example, we have that the focus on semantic segmentation peaks on the location of the pedestrian and is uniformly distributed over the entire road area. The Output Encoder focuses decisively on the pedestrian's bounding box feature. We then analyze an example where a bad semantic segmentation does not allow for predicting the correct behavior of the pedestrian (Stop example, Fig. 6). The model predicts incorrectly with a probability of 39.09% given the first semantic segmentation as input, while it predicts the correct label with a probability of 75.38% if we use the second segmentation. A clean segmentation with correctly segmented sidewalks and pedestrians allows for a correct prediction.

Failures Cases Now we analyze two more examples where our full model fails to predict the correct classification of pedestrian behavior. In this false positive example (Fig. 7 left), CrossFeat predicts with a 93.61% probability that the pedestrian will start moving in the future (GO situation). Instead in the future, the pedestrian remains stationary. It is probable that the error depends on the fact that the pedestrian in front of him is moving and the network gets confused by the pedestrian crossing. In this false positive example in STOP setting (Fig. 7 right), a child was moving in the past and the model predicts that the child will stop in the future with 82.34% probability. In the future, however, the child continues to move.

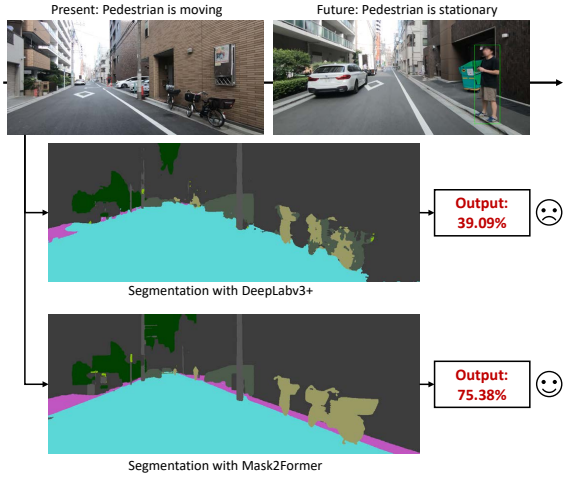


Fig. 6: Effect of semantic segmentation. The pedestrian is moving at the present timestep and will stop in the future. A state-of-the-art segmentation model (Mask2Former) leads to the correct prediction.

VI. Limitations and Conclusions

We proposed CrossFeat, a pedestrian behavior predictor based on multi-modal attention. The proposed model is simple yet effective and works with a single segmentation frame rather than a history of RGB frames, as typically done in the literature. The proposed model uses a pre-trained state-of-the-art model to extract segmentations and leverages environmental and behavioral labels available in the benchmarks. In a realistic scenario, a module would be required to obtain this information from the frame and the sensors on the ego-vehicle. Furthermore, experiments were performed on real datasets each with different models for crossing and Stop and Go prediction. As a future development, it would be interesting to train and test the models

Fig. 7: Failure cases. GO (left): The pedestrian (green) is stationary on the sidewalk in the present and remains stationary also in the future. STOP (right): The child (red) is moving in the present and moves also in the future.

on all datasets together to bridge the differences in features that can be observed in each dataset and task.

Acknowledgments This work was supported by the European Commission under European Horizon 2020 Programme, grant number 951911—AI4Media. This work was partially supported by the Piano per lo Sviluppo della Ricerca (PSR 2023) of the University of Siena - project FEATHER: Forecasting and Estimation of Actions and Trajectories for Human-robot intERactions.

REFERENCES

[1] Z. Yin, R. Liu, Z. Xiong, and Z. Yuan, "Multimodal transformer networks for pedestrian trajectory prediction," in *Proceedings of the Thirtieth International Joint Conf. on Artificial Intelligence, IJCAI-21*, Z.-H. Zhou, Ed. International Joint Conf.s on Artificial Intelligence Organization, 8 2021, pp. 1259–1265, main Track.

[2] Z. Su, G. Huang, S. Zhang, and W. Hua, "Crossmodal transformer based generative framework for pedestrian trajectory prediction," in *2022 International Conf. on Robotics and Automation*, 2022, pp. 2337–2343.

[3] L. Li, M. Pagnucco, and Y. Song, "Graph-based spatial transformer with memory replay for multi-future pedestrian trajectory prediction," in *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2022, pp. 2231–2241.

[4] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, "Transformer networks for trajectory forecasting," in *2020 25th international Conf. on pattern recognition (ICPR)*. IEEE, 2021, pp. 10 335–10 342.

[5] C. Xu, W. Mao, W. Zhang, and S. Chen, "Remember intentions: Retrospective-memory-based trajectory prediction," in *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2022, pp. 6488–6497.

[6] F. Marchetti, F. Becattini, L. Seidenari, and A. Del Bimbo, "Smemo: social memory for trajectory forecasting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[7] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE Conf. on computer vision and pattern recognition*, 2016, pp. 961–971.

[8] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Multi-agent generative trajectory forecasting with heterogeneous data for control," *arXiv preprint arXiv:2001.03093*, 2020.

[9] M. Gerla, E.-K. Lee, G. Pau, and U. Lee, "Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds," in *2014 IEEE world forum on internet of things (WF-IoT)*. IEEE, 2014, pp. 241–246.

[10] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2vnet: Vehicle-to-vehicle communication for joint perception and prediction," in *Computer Vision–ECCV 2020: 16th European Conf., Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 605–621.

[11] N. Wang, X. Wang, P. Palacharla, and T. Ikeuchi, "Cooperative autonomous driving for traffic congestion avoidance through vehicle-to-vehicle communications," in *2017 IEEE Vehicular Networking Conf. (VNC)*. IEEE, 2017, pp. 327–330.

[12] T. Mordan, M. Cord, P. Pérez, and A. Alahi, "Detecting 32 pedestrian attributes for autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems (T-ITS)*, 2021.

[13] G. Singh, S. Akrigg, M. Di Maio, V. Fontana, R. J. Alitappeh, S. Khan, S. Saha, K. Jeddisaravi, F. Yousefi *et al.*, "Road: The road event awareness dataset for autonomous driving," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 1036–1054, 2022.

[14] Z. Sui, Y. Zhou, X. Zhao, A. Chen, and Y. Ni, "Joint intention and trajectory prediction based on transformer," in *2021 IEEE/RSJ International Conf. on Intelligent Robots and Systems*, 2021, pp. 7082–7088.

[15] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *2019 IEEE/CVF International Conf. on Computer Vision*, 2019, pp. 6261–6270.

[16] J. Lorenzo, I. P. Alonso, R. Izquierdo, A. L. Ballardini, A. H. Saz, D. F. Llorca, and M. A. Sotelo, "Capformer: Pedestrian crossing action prediction using transformer," *Sensors*, vol. 21, no. 17, 2021.

[17] N. Osman, G. Camporese, and L. Ballan, "Tamformer: Multi-modal transformer with learned attention mask for early intent prediction," in *ICASSP 2023 - 2023 IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[18] D. Guo, T. Mordan, and A. Alahi, "Pedestrian stop and go forecasting with hybrid feature fusion," in *2022 International Conf. on Robotics and Automation*, 2022, pp. 940–947.

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international Conf. on computer vision*, 2017, pp. 2961–2969.

[20] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international Conf. on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[21] L. Achaji, J. Moreau, T. Fouqueray, F. Aioun, and F. Charpillet, "Is attention to bounding boxes all you need for pedestrian action prediction?" in *2022 IEEE Intelligent Vehicles Symposium*, 2022, pp. 895–902.

[22] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," 2022.

[23] X. Song, M. Kang, S. Zhou, J. Wang, Y. Mao, and N. Zheng, "Pedestrian intention prediction based on traffic-aware scene graph model," in *2022 IEEE/RSJ International Conf. on Intelligent Robots and Systems*, 2022, pp. 9851–9858.

[24] D. Yang, H. Zhang, E. Yurtsever, K. A. Redmill, and U. Ozguner, "Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 2, pp. 221–230, 2022.

[25] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Pedestrian action anticipation using contextual feature fusion in stacked rnns," 2020.

[26] Z. Wang and N. Papanikolopoulos, "Estimating pedestrian crossing states based on single 2d body pose," in *IEEE/RSJ International Conf. on Intelligent Robots and Systems*, 2020, pp. 2205–2210.

[27] J. Lorenzo, I. Parra, F. Wirth, C. Stiller, D. F. Llorca, and M. A. Sotelo, "Rnn-based pedestrian crossing prediction using activity and pose-related features," in *IEEE Intelligent Vehicles Symposium*, 2020, pp. 1801–1806.

[28] A. Singh and U. Suddamalla, "Multi-input fusion for practical pedestrian intention prediction," in *2021 IEEE/CVF International Conf. on Computer Vision Ws*, 2021, pp. 2304–2311.

[29] F. Piccoli, R. Balakrishnan, M. J. Perez, M. Sachdeo, C. Nuñez, M. Tang, K. Andreasson, K. Bjurek, R. Dass Raj, E. Davidsson, C. Eriksson, V. Hagman, J. Sjöberg, Y. Li, L. Srikar Muppirisetty, and S. Roychowdhury, "Fussi-net: Fusion of spatio-temporal skeletons for intention prediction network," in *2020 54th Asilomar Conf. on Signals, Systems, and Computers*, 2020, pp. 68–72.

[30] S. Bonnin, T. H. Weisswange, F. Kummert, and J. Schmuedderich, "Pedestrian crossing prediction using multiple context-based models," in *17th International IEEE Conf. on Intelligent Transportation Systems (ITSC)*, 2014, pp. 378–385.

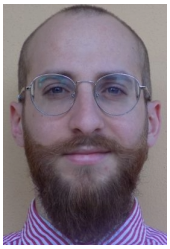
[31] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for evaluating pedestrian action prediction," in *2021 IEEE Winter Conf. on Applications of Computer Vision*, 2021, pp. 1257–1267.

[32] D. Zhang, F. Shi, Y. Meng, Y. Xu, X. Xiao, and W. Li, "Pedestrian intention prediction via depth augmented scene restoration," in *2021 5th CAA International Conf. on Vehicular Control and Intelligence (CVCI)*, 2021, pp. 1–6.

[33] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3485–3492, 2020.

[34] A. Rasouli, T. Yau, M. Rohani, and J. Luo, "Multi-modal hybrid architecture for pedestrian action prediction," in *2022 IEEE Intelligent Vehicles Symposium*, 2022, pp. 91–97.

[35] M. Dong, "Pedestrian cross forecasting with hybrid feature fusion," in *Asian Conf. on Machine Learning*. PMLR, 2024, pp. 327–342.
[36] J.-S. Ham, D. H. Kim, N. Jung, and J. Moon, "Cipf: Crossing intention prediction network based on feature fusion modules for improving pedestrian safety," in *2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Ws*, 2023, pp. 3666–3675.
[37] A. Rasouli, M. Rohani, and J. Luo, "Bifold and semantic reasoning for pedestrian behavior prediction," in *2021 IEEE/CVF International Conf. on Computer Vision*, 2021, pp. 15 580–15 590.
[38] D. Schorkhuber, M. Pröll, and M. Gelautz, "Feature selection and multi-task learning for pedestrian crossing prediction," in *2022 16th International Conf. on Signal-Image Technology and Internet-Based Systems (SITIS)*, 2022, pp. 439–444.
[39] H. Razali, T. Mordan, and A. Alahi, "Pedestrian intention prediction: A convolutional bottom-up multi-task approach," *Transportation Research Part C: Emerging Technologies*, vol. 130, p. 103259, 2021.
[40] M. Chaabane, A. Trabelsi, N. Blanchard, and R. Beveridge, "Looking ahead: Anticipating pedestrians crossing with future frames prediction," in *2020 IEEE Winter Conf. on Applications of Computer Vision*, 2020, pp. 2286–2295.
[41] J. Lorenzo Díaz and M.-A. Sotelo, "Intformer: Predicting pedestrian intention with the aid of the transformer architecture," 05 2021.
[42] A. Rasouli and I. Kotseruba, "Pedformer: Pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning," in *2023 IEEE International Conf. on Robotics and Automation*, 2023, pp. 9844–9851.
[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conf. on Learning Representations, May 3-7, 2021*.
[44] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Springer International Publishing, 2020, pp. 213–229.
[45] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conf. on Empirical Methods in Natural Language Processing*, Oct. 2014, pp. 1724–1734.
[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.
[47] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *2017 IEEE International Conf. on Computer Vision Ws*, 2017, pp. 206–213.
[48] S. Malla, B. Dariush, and C. Choi, "Titan: Future forecast using action priors," in *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 11 183–11 193.
[49] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *2017 IEEE International Conf. on Computer Vision*, 2017, pp. 5000–5009.
[50] C. Yan, C. Xu, R. Yuan, M. Li, X. Li, and H. Liu, "A pedestrian multi-object tracking algorithm based on centertrack for autonomous driving," in *2022 International Conf. on Virtual Reality, Human-Computer Interaction and Artificial Intelligence*. IEEE, 2022, pp. 189–194.
[51] S. Kreiss, L. Bertoni, and A. Alahi, "Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 13 498–13 511, 2021.
[52] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition*. Los Alamitos, CA, USA: IEEE Computer Society, jul 2017, pp. 4724–4733.
[53] J. Gesnouin, S. Pechberti, B. Stanciulcscu, and F. Moutarde, "Trouspi-net: Spatio-temporal attention on parallel atrous convolutions and u-grus for skeletal pedestrian crossing prediction," in *2021 16th IEEE International Conf. on Automatic Face and Gesture Recognition*. Los Alamitos, CA, USA: IEEE Computer Society, dec 2021, pp. 1–7.
[54] P. R. G. Cadena, Y. Qian, C. Wang, and M. Yang, "Pedestrian graph +: A fast pedestrian crossing prediction model based on graph convolutional networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 21 050–21 061, 2022.
[55] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conf. on computer vision*, 2018, pp. 801–818.



Francesco Marchetti received a PhD degree cum laude in 2024 in computer engineering from the University of Florence with the thesis "Trajectory and Behavior Forecasting in Autonomous Driving and Robotics". Currently he is a researcher at Media Integration and Communication Center (MICC) and the research work focuses on trajectories forecasting in the automotive field.

Taylor Mordan received the degree in engineering from ENSTA ParisTech, Paris, France, the M.S. degree in computer science from UPMC, Paris, in 2015, and the Ph.D. degree in computer science from Sorbonne University, Paris, in 2018. From 2015 to 2018, he was a Research Assistant with Thales LAS France. Since 2019, he has been a Post-Doctoral Researcher with VITA Laboratory, EPFL, Lausanne, Switzerland. His research interests include computer vision, multi-task learning, and perception in autonomous vehicles.

Federico Becattini is a Tenure-Track Assistant Professor at the University of Siena. His research focuses on computer vision and memory-based learning, with applications in automotive. He organized tutorials and workshops at ICPR2020, ICIAP2020, ACMMM2022, ICPR2022, ECCV2022, ECCV2024. He has co-authored more than 50 papers. He is Associate Editor of the International Journal of Multimedia Information Retrieval (IJMIR).

Lorenzo Seidenari is an Assistant Professor at the Department of Information Engineering of the University of Florence. He received his Ph.D. degree in computer engineering in 2012 from the University of Florence. His research focuses on deep learning for object and action recognition in video and images. He is an ELLIS scholar. He authored 16 journals and more than 40 peer-reviewed conference papers. He has an h-index of 25 with more than 2200 citations.

Alberto Del Bimbo is Full Professor of Computer Engineering interested in multimedia retrieval, pattern recognition, image analysis, and HCI. From 1996 to 2000, he was President of IAPR Italian Chapter and Member-at-Large of IEEE Publication Board from 1998 to 2000. He was General Co-Chair of ACMMM2010 and ECCV2012. He was nominated as ACM Distinguished Scientist in 2016. He received the SIGMM Technical Achievement Award for Outstanding Technical Contributions to Multimedia Computing, Communications and Applications. He is a IAPR Fellow, and an Associate Editor of several international journals.

Alexandre Alahi received the Ph.D. from EPFL. He was a Post-Doctoral Researcher and a Research Scientist at Stanford University for five years. He worked on the theoretical challenges and practical applications of socially-aware artificial intelligence. His research interests enable machines to perceive the world and make decisions in transportation and smart environments. He was awarded the Swiss NSF Early and Advanced Researcher Grants for his work on predicting human social behavior. He has co-founded multiple startups, such as Visiosafe and won startup competitions. He was one of the Top 20 Swiss Venture leaders in 2010.