Privacy-Aware Traffic Re-Identification with Interpretable Sparse Autoencoders

Joris Heemskerk[1], Martijn Folmer[3], Michael Dubbeldam[3], Rianne van Os[1], and Henry Maathuis[1,2]

[1] Utrecht University of Applied Science,
Research Group Artificial Intelligence, Utrecht, The Netherlands
[2] Jheronimus Academy of Data Science, 's-Hertogenbosch, The Netherlands
[3] Technolution, Gouda, The Netherlands
joris_heemskerk@outlook.com, {martijn.folmer, michael.dubbeldam}@technolution.nl, {riannne.vanos, henry.maathuis}@hu.nl

Abstract. Artificial Intelligence (AI) systems have become increasingly powerful tools used by businesses and governments to increase productivity, earnings, efficiency, and more. With this significant power comes the responsibility of balancing innovation with ethical considerations. We aim to increase transparency and trust in Computer Vision embedding models by suppressing unwanted identifiable features from a model trained on re-identification of traffic participants. This is achieved by using Sparse Autoencoders as a dictionary learning technique to extract highly interpretable features from our model. Unwanted identifiable features are suppressed and we analyse the effects on performance. Using this technique, we demonstrate that it is possible to create a transparent and highly interpretable model with a limited reduction in performance (a decrease from 0.98 mAP@0.6 to 0.90 mAP@0.6).

Keywords: Mechanistic Interpretability · Sparse Autoencoder · Open-World Re-Identification

1 Introduction

The use of Artificial Intelligence (AI) systems has expanded significantly in recent years, with applications emerging across sectors such as finance [4], healthcare [2], and government [30]. While these systems provide major benefits, they also bring new challenges, one of the most pressing being *privacy concerns*. Addressing these concerns is crucial both for ensuring compliance with regulations such as GDPR [7] and the EU AI Act [6], and for building public trust in AI technologies. Central to this challenge is the need for *explainability and transparency*. If the internal workings of an AI model can be examined, biases and unwanted correlations can be identified and potentially removed. This is the central aim of the field of *mechanistic interpretability*, which seeks to uncover the internal reasoning behind AI decisions. One domain where privacy and interpretability intersect strongly

is re-identification (re-ID). Re-ID refers broadly to the task of determining whether two observations belong to the same entity, even if they are recorded under different circumstances [14]. It has applications ranging from security and surveillance to biometric authentication and multimedia retrieval [12]. In computer vision, re-ID is commonly formulated as an image-based task, in which the objective is to match entities across variations in viewpoint, illumination, occlusion, and temporal conditions.

An important application of re-ID is traffic participants, where systems recognize and track pedestrians or vehicles across cameras to support urban mobility, safety, and intelligent transportation [29]. However, while such systems have achieved remarkable accuracy, they raise two pressing concerns. First, these models often act as "black boxes", offering little insight into which visual features drive their decisions. Second, re-ID by nature touches on *privacy-sensitive information*, since it relies on identifying and matching individuals. Together, these issues make transparency and accountability essential in deploying re-ID systems responsibly.

To address this, research on explainability and transparency in AI offers a promising perspective. If we can understand which visual features an embedding model relies on when matching two images, then biases, identifiable attributes, or privacy-sensitive patterns become more susceptible to detection and, if necessary, suppression. This could support compliance with regulations such as GDPR [7] and the EU AI Act [6], and more broadly inform societal debates on the responsible use of AI-based surveillance technologies.

Building on these insights, our research applies the concept of interpretable sparse features to *computer vision embedding models for re-ID of traffic participants*. Our aim is to create a concise set of visual features that explain how the model matches two images, and then to investigate whether identifiable or undesired features can be selectively suppressed without severely impacting re-ID performance. More concretely, we ask how an interpretable set of features can be constructed to increase transparency in re-ID while minimising the cost to performance, and how identifiable or undesired features can be suppressed once such transparency is achieved. Finally, we present both quantitative and qualitative results before discussing the broader implications for trustworthy re-ID problems.

2 Related Work

Privacy concerns in AI systems can be understood along three key stages: *data gathering*, *data processing*, and *data outcomes* [8][17][1]. At the stage of data gathering, issues arise around the collection of personal information such as facial images, addresses, or financial records. For example, facial recognition in mobile devices or police surveillance systems raises questions about storing and comparing sensitive data [20]. At the level of data outcomes, concerns emerge when sensitive information is revealed or can be inferred from a model's predictions, as in cooperative UAV systems whose real-time object identification

and re-identification outputs can expose individuals' identities and trajectories if not properly safeguarded [24]. Our research, however, focuses on the stage of data processing: how models use internal representations to make predictions, and how potentially undesirable features can be identified and suppressed.

Within computer vision, privacy concerns are particularly acute in re-ID, where models aim to match individuals across different cameras or time points. Re-ID has evolved from hand-crafted descriptors to deep embedding models optimised via metric learning and, more recently, transformer-based architectures. For a comprehensive overview of modern techniques, datasets, and evaluation protocols in pedestrian re-ID, we refer to Sun et al.'s recent survey [25]. While this literature establishes strong performance baselines, it devotes comparatively limited attention to the transparency of the learned embeddings. As a result, the features driving re-ID decisions remain largely opaque, highlighting the need for methods that *explain* and *control* internal representations. This is precisely where the field of *mechanistic interpretability* becomes relevant.

Mechanistic interpretability aims to uncover the internal workings of AI models, with the goal of making their decisions more transparent and controllable. By identifying which patterns are extracted and how they relate to outcomes, it becomes possible to directly target unwanted correlations or privacy-sensitive features, such as those that could reveal an individual's identity. While work on addressing data gathering concerns often removes biases from training datasets [28], this approach is limited: hidden correlations are hard to eliminate fully, and unwanted patterns can still appear in inference data. Analysing the processing stage through interpretability therefore offers a more direct way to expose and suppress problematic features, which is central to our approach.

Research in interpretability can be broadly grouped into three directions: interpretability by design, post-hoc explanations, and mechanistic analysis of generalisation [23][15]. Interpretability-by-design approaches create models whose internal reasoning is inherently transparent. These approaches, while powerful in terms of transparency, often involve trade-offs in performance or flexibility and are not easily applicable to embedding models like those used in re-ID.

Post-hoc methods, in contrast, seek to explain decisions of already-trained models. Techniques such as SHAP [16] and LIME [21] approximate feature contributions to individual predictions. In embedding spaces specifically, concept-based interpretability via Testing with Concept Activation Vectors (TCAV) probes model sensitivity to human-defined concepts [11], with recent guidance refining best practices and limitations for CAV-based analyses [18]. These methods are valuable for exposing major contributing concepts, but they generally probe sensitivity without offering a sparse, disentangled basis that can be *directly* modified to suppress undesired features.

Finally, mechanistic approaches focus on understanding how models internally represent and process data. Of particular relevance, Bricken et al. [3] and Huben et al. [10] demonstrated that sparse autoencoders can extract semantically interpretable features from large language models, even in the presence of superposition [5]. Crucially, activating or suppressing such features was shown to

consistently alter model behavior. Inspired by these advances, and complementary to concept-probing approaches like TCAV, we apply sparse autoencoders to computer vision embedding models, aiming to recover a sparse, interpretable feature basis that both *explains* re-ID decisions and enables *targeted suppression* of privacy-sensitive or otherwise undesired features.

3 Data

In this study, we employ a proprietary dataset supplied by Technolution[4], developed as part of a collaborative effort to investigate transparency in re-ID systems. The dataset was specifically constructed for training an embedding model on the task of re-ID of traffic participants. In total, it comprises 262,400 images, distributed across 12,396 unique tracks. On average, each track (a set of sequential images corresponding to a single individual or object) contains 21 images, each recorded from a single camera angle. The dataset covers three traffic modalities: 17% vehicles, 77% cyclists, and 6% pedestrians. Data were collected across multiple urban locations, ensuring diversity in terms of backgrounds, viewpoints, and environmental conditions.

Figure 1 illustrates the dataset. Each sample consists of an already cropped object of interest, resized to 224×224 pixels to match the model's input requirements. In other words, no detection within full-frame images is performed since the dataset provides only the isolated object crops. While this resizing introduces minor geometric distortions, it also minimises background context, encouraging the model to attend to salient object features rather than environmental cues [13,26,19].

In addition to the training set, Technolution supplied a dedicated test set consisting of 6,248 images across 772 tracks. The class distribution is comparable, with 13% vehicles, 79% cyclists, and 8% pedestrians. Although collected under similar conditions, the test set contains no overlap with the training set and includes additional out-of-distribution data involving novel locations, tracks with limited samples (as few as two images), and cases of occlusion or obstruction.

4 Methodology

In this section, we outline the methodology of our study. We begin with the backbone embedding model, which produces compressed intermediate representations for re-ID. These embeddings serve as input to the sparse autoencoder, which derives a sparse and interpretable feature basis. We then describe the evaluation setup, followed by the procedure for identifying and suppressing undesired or privacy-sensitive features.

4.1 Embedding Model

The backbone of our approach is a fine-tuned version of MobileNetV3 [9], trained on the dataset introduced in Section 3. Training was performed using the Triplet

[4] https://www.technolution.com/

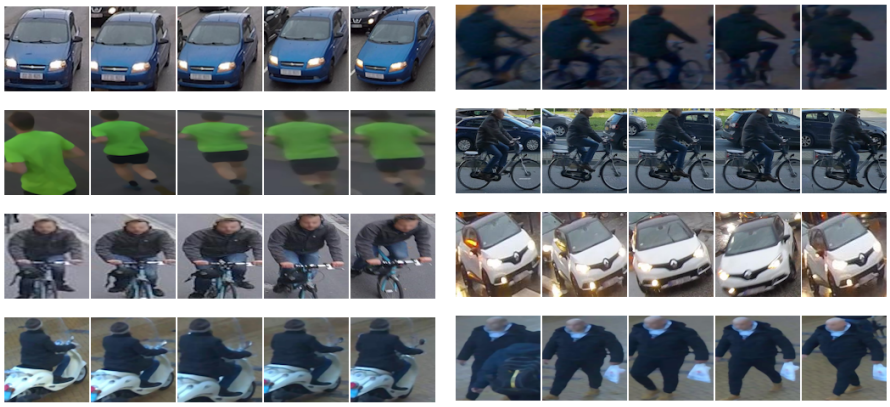

Fig. 1: Sample images from the dataset, illustrating different traffic modalities, viewpoints, lighting conditions, and backgrounds.

Loss function [22], which enforces that embeddings of the same identity are pulled closer together while embeddings of different identities are pushed apart. As described in Section 3, the input images were provided at a resolution of 224×224 pixels, in line with MobileNetV3's architecture. The final model was trained for 250 epochs, showing stable convergence. The Adam optimiser was used with a learning rate of 0.0003.

The trained network outputs embeddings as 128-dimensional vectors, each normalised to the unit norm. After training, the model achieved a mean Average Precision (mAP)@0.6 score of 0.9760 on the training set and 0.9846 on the test set (see Section 4.3 for the formal definition of mAP). These embeddings form the input to the sparse autoencoder described below. In Section 6, we evaluate the effect of the autoencoder on both interpretability and re-ID performance.

4.2 Sparse Autoencoder

Our sparse autoencoder is inspired by the autoencoder introduced in Bricken et al. [3], tailored to our use case[5]. We apply the autoencoder to the final layer of our embedding model, which has a dimensionality of 128 (n). An overview of the model architecture can be found in Figure 2. It is worth noting that the output of the global average pooling layer of MobileNetV3, our embedding, also serves as the input layer for the sparse autoencoder. The dimensionality of the encoder layer in the sparse autoencoder is defined as m, where $m \geq n$. Additionally, we define W_e as the encoder weights ($W_e \in \mathbb{R}^{m \times n}$), W_d as the decoder weights ($W_d \in \mathbb{R}^{n \times m}$), b_e as the encoder bias ($b_e \in \mathbb{R}^m$), and b_d as the decoder bias ($b_d \in \mathbb{R}^n$).

[5] Example code can be found at https://github.com/JorisHeemskerk/sparse_autoencoder_example.

Firstly, for each input embedding x in dataset X, b_d is subtracted using

$$\bar{x} = x - b_d,$$

which will later be added back in the decoding step. The output of the encoder (f) is defined as

$$f = \mathrm{ReLU}\big(W_e\,\bar{x} + b_e\big).$$

This is where the sparse features live. The encoder output gets translated back towards the model input with

$$\hat{x} = W_d\,f + b_d.$$

The model gets trained using an L_2 loss to enforce input reconstruction as well as an L_1 penalty to encourage sparsity. These are defined as

$$L_2 = \frac{1}{|X|}\|x - \hat{x}\|_2^2, \text{ and}$$

$$L_1 = \lambda\|f\|_1.$$

where the L_1 penalty can be scaled using λ. The total loss is the sum of the individual losses.

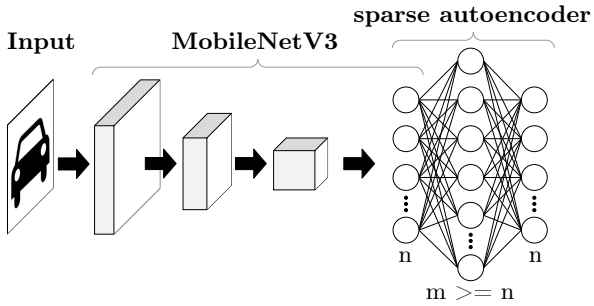

Fig. 2: A general overview of our model configuration, showing a simplified version of the MobileNetV3 CNN backbone architecture, along with the new sparse autoencoder head.

Dead neuron resampling During training, we performed neuron resampling in accordance with Bricken et al., doing so every 25,000 global steps. At each of these checkpoints, we identified all neurons $i \in \{1, \ldots, m\}$ for which the activation $f_i(x) = 0$ for all x over the past 12,500 training steps. These neurons were labelled as *dead* and all such neurons were resampled. Here, $f(x)$ denotes the encoder output for input x, and $f_i(x)$ is its ith component.

To resample a dead neuron i, we first evaluated the model's reconstruction loss over a randomly sampled subset $\mathcal{B} = \{x_1, x_2, \ldots, x_n\}$, consisting of 10% of the training data. For each input sample $x_j \in \mathcal{B}$, we computed its reconstruction

$\hat{x}_j = f(x_j)$ and assigned it a sampling probability p_j proportional to the square of its total loss:

$$p_j = \frac{\|x_j - \hat{x}_j\|_2^2}{\sum_{x_k \in \mathcal{B}} \|x_k - \hat{x}_k\|_2^2},$$

where x_j and x_k denote individual input samples from the batch $\mathcal{B}$, and $\hat{x}_j$, $\hat{x}_k$ are their corresponding reconstructions.

For each dead neuron i, we then sampled an input $x^{(i)}$ from $\mathcal{B}$ according to this distribution $\{p_j\}$, and computed its normalised form as

$$\tilde{x}^{(i)} = \frac{x^{(i)}}{\|x^{(i)}\|_2}.$$

The decoder weight vector $W_d^{(:,i)}$ was then set to $\tilde{x}^{(i)}$, effectively initialising the neuron to reconstruct $x^{(i)}$. For the encoder weights, we scaled the direction of $\tilde{x}^{(i)}$ by a small factor relative to the average norm of the active encoder weights:

$$W_e^{(i,:)} = \alpha \cdot \bar{w} \cdot \tilde{x}^{(i)}, \text{ where } \bar{w} = \frac{1}{|A|} \sum_{j \in A} \|W_e^{(j,:)}\|_2 \text{ and } \alpha = 0.2,$$

and $A = \{j : f_j(x) > 0 \text{ for some } x \in X\}$ denotes the set of active neurons. The factor $\alpha = 0.2$ serves as a scaling constant (following [3]) to ensure that the newly initialised encoder weight has a relatively low norm, encouraging the neuron to activate weakly at first. The encoder bias term was reset to $b_e^{(i)} = 0$. Finally, we reinitialised the optimiser state (e.g., Adam moment estimates) for all modified weights and biases associated with neuron i, resulting in faster convergence.

Hyperparameters We found a large variation between different hyperparameter configurations in terms of interpretability and re-ID performance. Because of this, we decided to analyse a multitude of different values for some of the hyperparameters to establish an optimal configuration for our model.

Firstly, for our encoder dimensions m we decided on the following four values: 128, 256, 512, and 1024. According to Bricken et al., features found in encoders with a lower dimensionality split into more specific features when more encoder neurons are added. For example, in our situation this could mean that lower dimensional models learn more general features like a feature looking at all grey vehicles, where a higher dimensional model might split this feature into grey cars, grey vans and grey trucks. By analysing different encoder sizes we aim to find a model with the right level of detail regarding features, such that they are easily interpretable.

Secondly, finetuning the *L1* penalty is crucial for balancing model performance and interpretability. A λ that is too low causes the model to not be sparse, leading to no or less interpretable features. A λ that is too high results in the model becoming 100% sparse by not using the encoder at all, relying instead on the decoder weights to always predict the average input embedding. For our λ we used values ranging from $5e^{-3}$ to $1e^{-7}$, alternatingly dividing by 5 and 2 for

each step. This allowed us to quickly find a range striking a balance between re-ID performance and interpretability, which we then zoomed in on, taking more granular steps of $1e^{-a}$. This approach enabled us to decide on a specific model to manually interpret.

All other hyperparameters were held constant, as preliminary experiments indicated no meaningful effect on interpretability or re-ID performance. A batch size of 200 was selected as it maximised GPU utilisation without exceeding memory constraints. Each model was trained for 100 epochs, at which point both reconstruction and sparsity losses had stably converged. The learning rate was fixed at 1×10^{-5}, chosen to ensure gradual convergence and stability during training. All autoencoders were optimised using Adam, with an 80/20 split between training and validation data. Specifically, 210,628 images (80% of the available tracks) were used for training and 51,772 images (20%) for validation, with the split kept identical across all configurations to ensure comparability of results.

4.3 Evaluation Set-Up

To evaluate our trained sparse autoencoders, we used two quantitative metrics to establish model performance and manually interpreted and compared the features of the highest performing models to validate interpretability. We describe these metrics in the next section. In addition, we describe our method for identifying and suppressing features.

Quantitative Evaluation We evaluated the sparse autoencoders on two quantitative metrics. Firstly, we validated the re-ID performance using the mean Average Precision (mAP). We did so by comparing a set of query images Q to the full dataset X, specifically at a certain Euclidean distance threshold. We define this set as S_q for each query image $q \in Q$, and arrange it in ascending order. For example, mAP@0.5 means a threshold of 0.5 was used. For each item in S_q, the relevance is defined as

$$r_q(k) \begin{cases} 1 & \text{if the identities of } (S_q)_k \text{ and } q \text{ are the same,} \\ 0 & \text{otherwise.} \end{cases}$$

The precision at rank k for query q is defined as

$$P_q(k) = \frac{\sum_{i=1}^{k} r_q(i)}{k},$$

which measures the fraction of retrieved items in top k that are relevant. The Average Precision (AP) is the average of the precisions, computed at every rank k where a relevant item is found. The number of relevant items found is defined as

$$R_q = \sum_{k=1}^{|S_q|} r_q(k).$$

If at least one relevant item is retrieved (i.e. $R_q > 0$), then

$$AP_q = \frac{1}{R_q} \sum_{k=1}^{|S_q|} (r_q(k) \cdot P_q(k)),$$

otherwise $AP_q = 0$. Lastly, the mAP is defined as

$$mAP = \frac{1}{|Q|} \sum_{q \in Q} AP_q.$$

The highest mAP score here would be 1.0, as it is on the unit norm, which would only be achieved if a model was able to have S_q to consist of exclusively the identity of q, for every $q \in Q$.

For our results we primarily use a threshold of 0.6, as this is the previously used standard by Technolution, providing them the best real-world performance. They came to this threshold by consulting the number of False Positives and False Negatives given the available recorded annotated tracks. For completeness however, we also compute results with thresholds of 0.1, 0.3, 0.5, 0.7, and 0.9.

In addition to the mAP, we computed a statistic that helped in our qualitative analysis. Namely the median number of features used to encode a given input. If the model uses too many to encode an input image, this might result in an overload of explanations. Ideally, we are looking for a small set of features, combined with a high mAP.

Feature Interpretation In order to validate a feature on interpretability, we relied on manual qualitative analysis. For this, we considered the entire dataset and fed it through the different models. For each encoder neuron, we visualised the top 0.1% of all images, by neuron activation, and grouped them by track (saving only the highest activating sample per track). This results in a set of images from unique identities, supposedly sharing similar attributes. For each encoder neuron we compared these images (target set) with an equal set of random images with an activation value of zero (control set). An example can be found in Figure 4. This way each neuron could be manually analysed and we could validate consistency and exclusivity for each neuron's feature interpretability. For example, for a feature seemingly describing vehicles, we analyse how consistently different attributes are present in the target set, that simultaneously are not present in the control set (e.g. colour, position, or shape).

Identifiable Features The notion of (re-)identifiability presents a conceptual challenge, as virtually all features can be construed as identifying in some form. For example, a feature that detects yellow cars is inherently identifying those vehicles, yet such a definition lacks analytical precision. To refine this concept, we define a feature as identifiable if it either (1) encodes information that can directly reveal the identity of a specific individual (e.g., license plate recognition), or (2) captures a highly distinctive attribute that is shared only by a very limited

subset of individuals (e.g., unique facial structures or company-specific branding). Within this framework, features can be systematically analysed and filtered. Practically, this can be achieved by applying a binary mask at the output of the encoder layer, thereby suppressing identifiable features through zero-assignment. While this suppression may reduce the accuracy of reconstructed embeddings for instances previously reliant on such features, it effectively mitigates the risk of the model exploiting personally identifiable attributes.

4.4 Experimental Set-Up

We systematically evaluated combinations of the hyperparameters for encoder size (m) and the sparsity penalty λ. As stated in Section 4.2, for m we decided on values 128, 256, 512, and 1024. For the sparsity penalty we used 20 values ranging from $5e^{-3}$ to $1e^{-7}$, resulting in a total of 80 experiments. Using the metrics described in Section 4.3, we report three comparative analyses. Firstly, we compare the L_1 coefficient (λ) and mAP, to analyse the trade-off between sparsity and re-ID performance. Secondly, we compare the L_1 coefficient and median number of features used to encode a given input, thereby quantifying dataset coverage per neuron under different sparsity levels. Finally, we compare the mAP and the median number of features used, with the aim of identifying the optimal balance between re-ID performance and dataset coverage.

5 Results

The quantitative outcomes, based on the test dataset described in Section 3, are presented in Figure 3 across three comparative analyses as described in Section 4.4.[6]

The first graph shows the relationship between the L_1 coefficient and re-ID performance, measured by mAP@0.6. Reducing the sparsity constraint improves performance, whereas overly stringent constraints render the models unable to complete the task. Re-ID performance begins to emerge at an L_1 value of approximately $1e^{-5}$. At $1e^{-6}$, the models reach near-maximal mAP@0.6, approaching the performance of the encoder input (i.e., the output of the MobileNetV3-based embedding model), represented by the dashed black reference line. Encoder size has only a limited effect on mAP@0.6.

The second graph reports the relationship between the L_1 coefficient and the median number of features required to represent an image. Relaxing the sparsity constraint increases the number of features utilised per image, with each encoder neuron contributing to multiple representations. In contrast, strong sparsity constraints lead to near-complete suppression of features, as the model minimises loss primarily through the sparsity objective. At very low sparsity constraints, the influence of encoder size becomes more pronounced, with larger encoders employing proportionally more features per image.

[6] The full interactive figure can be found at https://jorisheemskerk.github.io/identifiable_feature_suppression_in_traffic_re-identification/.

The third and last graph examines the trade-off between interpretability and performance by comparing the median number of features per image with mAP@0.6. Configurations using approximately 25 median features achieve an effective balance, as further increases in features provide only marginal gains in mAP@0.6. Across these configurations, models attain an mAP@0.6 of approximately 0.9, compared to 0.985 for the encoder input. This difference reflects a reduction of roughly 0.085 mAP@0.6, attributable to the increased interpretability of the representations. Encoder size has no substantial effect on this trade-off.

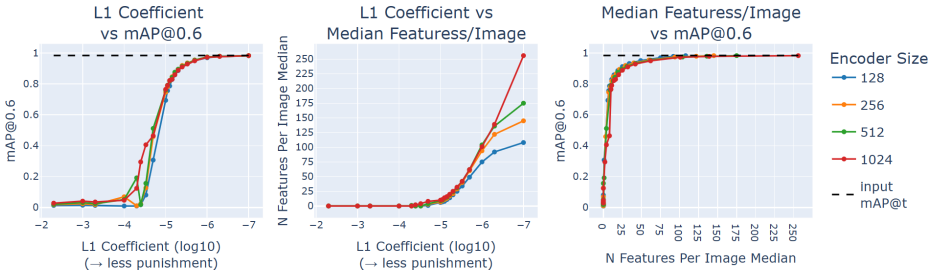

Fig. 3: Quantitative results of variations in encoder dimension, L_1 coefficient and the median number of features used (non-zero activation value) per image of test dataset.[6]

Table 1 reports the mean Average Precision (mAP) scores across multiple thresholds for both the suppressed and non-suppressed configurations. Overall, the results indicate minimal differences between the two configurations, suggesting that suppressing identifiable features has little impact on model performance. A more detailed discussion and contextualization of these findings can be found in Section 6.2.

Table 1: mAP before the use of an autoencoder, and mAP of model 512 before and after suppressing identifiable features. We observe a negligible drop in performance after suppressing identifiable features.

Model information	Dataset	mean Average Precisions (mAPs) @ thresholds					
Non-Interpretable		@0.1	@0.3	@0.5	**@0.6**	@0.7	@0.9
Without autoencoder	Train	1.0	0.996	0.982	**0.976**	0.968	0.961
	Test	1.0	0.995	0.988	**0.985**	0.978	0.963
Interpretable		@0.1	@0.3	@0.5	**@0.6**	@0.7	@0.9
Not Suppressed	Train	0.999	0.950	0.898	**0.895**	0.895	0.895
	Test	0.999	0.977	0.905	**0.896**	0.892	0.892
Suppressed	Train	0.999	0.950	0.898	**0.895**	0.894	0.894
	Test	0.999	0.977	0.904	**0.895**	0.892	0.891

6 Discussion

This study set out to apply sparse autoencoders to learn a concise set of visual features that explain how two images are matched within *computer vision embedding models for re-ID of traffic participants*. And then to investigate whether identifiable or undesired features can be selectively suppressed without severely impacting re-ID performance. Our results only show quantitative analysis of model sparsity and performance. In this section, we discuss our qualitative findings by manually analysing the models with the strongest quantitative performance, through a qualitative assessment of their extracted features. Additionally, we demonstrate the use of a binary mask to suppress unwanted features, without measurable harm to overall re-ID accuracy. Lastly, we describe the general observed patterns and discuss their meaning.

6.1 Interpretability–Specificity Trade-Off

To assess interpretability across architectures, we analysed one representative model per encoder size, each selected for achieving a median of approximately 25 features per image and an mAP@0.6 close to 0.9. For each model, the twenty most prominent features were manually examined, with representative examples presented in Figure 4.

Models with smaller encoders exhibited features that were comparatively global and abstract, thereby limiting interpretability. As illustrated in Figure 4a, the extracted feature corresponded primarily to black vans, but also included unrelated vehicle types (e.g., a white car), suggesting insufficient specificity.

By contrast, models with larger encoders yielded features of high specificity, which were also difficult to interpret. Figure 4c exemplifies this pattern, combining the presence of a centrally positioned face, a dark jacket, and a lighter-colored central object. Such highly specific features are less accessible to human interpretation, as they require extensive sampling of the feature space for accurate characterisation.

Overall, the encoder of size 512 achieved the most favorable balance between interpretability and representational specificity. Trained with an L_1 coefficient of $5e^{-6}$, this model attained an mAP@0.6 of 0.896 with a median of 24 features per image. As demonstrated in Figure 4b, the extracted features retained sufficient global structure to remain interpretable, while also exhibiting a degree of specificity absent in smaller encoders. Accordingly, the encoder size 512 configuration was selected for subsequent analyses.

6.2 Suppressing Identifiable Features

As established in Section 6.1, the encoder of size 512 yielded the most interpretable feature representations. For the remainder of this analysis, we therefore refer to this configuration as *Model 512*. To systematically investigate the presence of identifiable features, we manually analysed all extracted features. This procedure revealed a small subset that can be considered potentially identifiable.

Fig. 4: Samples from target (T) and control (C) image sets from the models with encoder sizes 128, 512 and 1024.

An illustrative case is feature 77, as shown in Figure 5, which corresponds to grey taxis from multiple companies characterised by green logos and roof-mounted signs. The degree of specificity exhibited by this feature, combined with its likely rarity in real-world traffic distributions, qualifies it as identifiable under our definition. Accordingly, we conclude that this feature should be suppressed.

A second case is feature 410, also shown in Figure 5. This feature appears to capture either (1) a specific brand logo displayed on vehicles, which coincidentally also activates in instances of bicycles carrying red or pink bags, or (2) red and pink bags, which coincidentally co-occur with the aforementioned brand logo. Owing to this semantic ambiguity, we adopt a precautionary stance and classify this feature as identifiable, thereby warranting its suppression.

In *Model 512*, features 77 and 410 were suppressed by applying a binary mask to the encoder layer, thereby setting the corresponding neuron activations to zero and preventing their participation in the construction of the final embedding. The outcomes, summarised in Table 1, demonstrate that suppression does not incur any statistically meaningful degradation in performance, with observed differences limited to fluctuations at the fourth decimal place.

A detailed analysis indicates that feature 77 is expressed in 7,598 training images (2.90%), while feature 410 is expressed in 12,811 training images (4.88%). Given that the median number of features utilised per image in this configuration is 25, we infer that the relative contribution of these features is insufficient to produce a measurable impact on aggregate performance metrics.

6.3 Central Observations

In the sections above, we highlight the trade-off between sparsity and performance. Reducing the sparsity constraint improves re-ID accuracy, whereas overly stringent constraints prevent meaningful feature utilisation. Re-ID performance begins to stabilise at an L_1 value of approximately $1e^{-5}$, with near-maximal mAP@0.6 reached at $1e^{-6}$. Across most configurations, encoder size has only a limited influence on performance. However, under sufficiently weak sparsity constraints, larger encoders employ proportionally more features to describe each image. This observation aligns with the expectation that increased model capacity enables

Fig. 5: Samples of the target (T) and control (C) image sets for features 77 and 410 of model 512.

the subdivision of features into more specific components. In such cases, the subdivided features may still be used jointly when representing images, resulting in a comparable amount of information, albeit in a more specialised and nuanced form.

Manual inspection of encoder features further supports this interpretation. Smaller encoders tend to learn global and abstract patterns, while larger encoders develop features of higher specificity that are, at times, more difficult to interpret. This difference likely stems from the limited representational capacity of smaller encoders, which necessitates broader feature coverage to sustain performance. By contrast, larger encoders, when using a similar median number of features per image, face fewer representational constraints and can allocate features to narrowly defined attributes without performance degradation. The emergence of such narrowly scoped attributes may, however, reflect a tendency toward overfitting to the training data distribution.

Taken together, our findings underscore the viability of sparse autoencoders as a principled means of embedding transparency into *computer vision embedding models for re-ID of traffic participants*. They address one of the key limitations of contemporary re-ID systems, their interpretability. In addition, the ability to suppress unwanted features may be relevant for regulatory contexts such as GDPR and the EU AI Act, and could inform broader discussions on trustworthy AI. Beyond traffic participant re-ID, the approach outlined here may also serve as a reference for introducing interpretability into other embedding-based applications.

7 Conclusion and Future Work

We have demonstrated that the sparse autoencoder technique introduced by Bricken et al. [3] provides an effective means of uncovering interpretable feature bases in re-ID embeddings, enabling targeted suppression of privacy-sensitive attributes while incurring only a minimal reduction in re-ID performance. Specifically, our best performing model produced highly interpretable features and achieved an mAP@0.6 of 0.90, compared to 0.98 for the non-interpretable baseline embedding model. Furthermore, we established that identifiable features can be effectively suppressed via masking at the encoder layer, obviating the need

for retraining and without inducing significant performance degradation. This was demonstrated through the identification and suppression of two features associated with company-specific logos, whose removal had no measurable impact on mAP performance suggesting that sparse autoencoder-based approaches provide a principled mechanism for constructing transparent and interpretable embedding models.

Several directions for future research emerge from this study. First, model selection was guided by configurations with a median of approximately 25 explanatory features per image, as adding further features yielded diminishing returns in performance. Although this threshold provided a practical trade-off between accuracy and interpretability, we were unable to establish a more principled justification for this number. Future research should examine how explanation size aligns with stakeholder needs, and to what extent semantic interpretability is preserved when models rely on varying numbers of features.

Second, the sparse autoencoder approach may allow semantic attributes to be distributed across feature activations, leading to poly-semantic representations. While our findings are broadly consistent with prior work suggesting that higher activations correspond to stronger feature presence, the persistence of low-level activations complicates interpretability. To encourage more mono-semantic features, future work could explore alternative sparsity constraints (e.g., penalising the number of datapoints activated rather than total activation magnitude) or investigate top-k activation mechanisms, such as those employed in VQ-VAE models [27].

Additionally, this research relied on the use of a proprietary dataset, as it contained potentially privacy-sensitive features. We encourage future research to develop a benchmark dataset for problems similar to ours, such that alternative methods can be developed and compared. We also suggest future research to apply the proposed technique to existing benchmark datasets, to both compare feature interpretability with other methods and to analyse the scalability of the proposed method.

Finally, the process of annotating and interpreting learned features was conducted manually, which introduces subjectivity and potential bias. To strengthen reliability, future work should incorporate controlled perturbations to input images. For example, removing license plates from vehicles or applying systematic augmentations to pedestrian faces to assess how these modifications affect feature activations. Such experiments would help to validate whether the learned features capture the intended semantic content or encode unintended identifiable attributes.

Acknowledgments. This work was supported by Technolution, who provided both funding and data for the research.

Disclosure of Interests. The authors declare no competing interests relevant to the content of this article.

References

1. Agrawal, A., Gans, J., Goldfarb, A.: Prediction machines, updated and expanded: The simple economics of artificial intelligence. Harvard Business Press (2022)
2. Al Kuwaiti, A., Nazer, K., Al-Reedy, A., Al-Shehri, S., Al-Muhanna, A., Subbarayalu, A.V., Al Muhanna, D., Al-Muhanna, F.A.: A review of the role of artificial intelligence in healthcare. J Pers Med **13**(6) (Jun 2023)
3. Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J.E., Hume, T., Carter, S., Henighan, T., Olah, C.: Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread (2023), https://transformer-circuits.pub/2023/monosemantic-features/index.html
4. Cao, L.: Ai in finance: A review. Available at SSRN **3647625**, 1 (2020). https://doi.org/https://dx.doi.org/10.2139/ssrn.3647625
5. Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., Olah, C.: Toy models of superposition. Transformer Circuits Thread (2022), https://transformer-circuits.pub/2022/toy_model/index.html
6. European Parliament, Council of the European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (Artificial Intelligence Act) (2024-07-12), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689
7. European Parliament, Council of the European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (2016-05-04), https://data.europa.eu/eli/reg/2016/679/oj
8. Golda, A., Mekonen, K., Pandey, A., Singh, A., Hassija, V., Chamola, V., Sikdar, B.: Privacy and security concerns in generative ai: A comprehensive survey. IEEE Access **12**, 48126–48144 (2024). https://doi.org/10.1109/ACCESS.2024.3381611
9. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for mobilenetv3 (2019), https://arxiv.org/abs/1905.02244
10. Huben, R., Cunningham, H., Smith, L.R., Ewart, A., Sharkey, L.: Sparse autoencoders find highly interpretable features in language models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=F76bwRSLeK
11. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. pp. 2668–2677. PMLR (2018)
12. Leng, Q., Ye, M., Tian, Q.: A survey of open-world person re-identification. IEEE Transactions on Circuits and Systems for Video Technology **30**(4), 1092–1108 (2019)

13. Li, K., Wu, Z., Peng, K.C., Ernst, J., Fu, Y.: Tell me where to look: Guided attention inference network (2018), https://arxiv.org/abs/1802.10171
14. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2197–2206 (2015)
15. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable ai: A review of machine learning interpretability methods. Entropy **23**(1) (2021). https://doi.org/10.3390/e23010018, https://www.mdpi.com/1099-4300/23/1/18
16. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
17. Martin, K.D., Zimmermann, J.: Artificial intelligence and its implications for data privacy. Current Opinion in Psychology **58**, 101829 (2024). https://doi.org/https://doi.org/10.1016/j.copsyc.2024.101829, https://www.sciencedirect.com/science/article/pii/S2352250X24000423
18. Nicolson, A., Schut, L., Noble, J.A., Gal, Y.: Explaining explainability: Recommendations for effective use of concept activation vectors. arXiv preprint arXiv:2404.03713 (2024)
19. Ning, X., Gong, K., Li, W., Zhang, L., Bai, X., Tian, S.: Feature refinement and filter network for person re-identification. IEEE Transactions on Circuits and Systems for Video Technology **31**(9), 3391–3402 (2021). https://doi.org/10.1109/TCSVT.2020.3043026
20. Parliament, E., for Parliamentary Research Services, D.G., Madiega, T., Mildebrath, H.: Regulating facial recognition in the EU – In-depth analysis. European Parliament (2021). https://doi.org/doi/10.2861/140928
21. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier (2016), https://arxiv.org/abs/1602.04938
22. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). p. 815–823. IEEE (Jun 2015). https://doi.org/10.1109/cvpr.2015.7298682, http://dx.doi.org/10.1109/CVPR.2015.7298682
23. Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., Biderman, S., Garriga-Alonso, A., Conmy, A., Nanda, N., Rumbelow, J., Wattenberg, M., Schoots, N., Miller, J., Michaud, E.J., Casper, S., Tegmark, M., Saunders, W., Bau, D., Todd, E., Geiger, A., Geva, M., Hoogland, J., Murfet, D., McGrath, T.: Open problems in mechanistic interpretability (2025), https://arxiv.org/abs/2501.16496
24. Silva, S.H., Rad, P., Beebe, N., Choo, K.K.R., Umapathy, M.: Cooperative unmanned aerial vehicles with privacy preserving deep vision for real-time object identification and tracking. Journal of Parallel and Distributed Computing **131**, 147–160 (2019). https://doi.org/https://doi.org/10.1016/j.jpdc.2019.04.009, https://www.sciencedirect.com/science/article/pii/S0743731518308839
25. Sun, Z., Wang, X., Zhang, Y., Song, Y., Zhao, J., Xu, J., Yan, W., Lv, C.: A comprehensive review of pedestrian re-identification based on deep learning. Complex & Intelligent Systems **10**(2), 1733–1768 (2024)
26. Tian, M., Yi, S., Li, H., Li, S., Zhang, X., Shi, J., Yan, J., Wang, X.: Eliminating background-bias for robust person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5794–5803 (2018). https://doi.org/10.1109/CVPR.2018.00607

27. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
28. Verma, S., Ernst, M., Just, R.: Removing biased data to improve fairness and accuracy (2021), https://arxiv.org/abs/2102.03054
29. Zheng, W.S., Gong, S., Xiang, T.: Towards open-world person re-identification by one-shot group-based verification. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(3), 591–606 (2016). https://doi.org/10.1109/TPAMI.2015.2453984
30. Zuiderwijk, A., Chen, Y.C., Salem, F.: Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. Government Information Quarterly **38**(3), 101577 (2021). https://doi.org/https://doi.org/10.1016/j.giq.2021.101577, https://www.sciencedirect.com/science/article/pii/S0740624X21000137