

ИНТЕЛЛЕКТУАЛЬНЫЕ ТРАНСПОРТНЫЕ СИСТЕМЫ/INTELLIGENT TRANSPORT SYSTEMS

DOI: https://doi.org/10.60797/IRJ.2026.170.48 EDN: DDOUAZ

МОДИФИЦИРОВАННЫЙ MAMBA-КОНТРОЛЛЕР С СЕМАНТИЧЕСКОЙ АДАПТАЦИЕЙ ДЛЯ ДИФФЕРЕНЦИРУЕМОЙ РЕГИСТРАЦИИ ЛИЦ В ИНТЕЛЛЕКТУАЛЬНЫХ ТРАНСПОРТНЫХ СИСТЕМАХ

Научная статья

Куликов А.А.[1, *]

[1] ORCID : 0000-0002-8443-3684;

[1] Научно-исследовательский и проектно-конструкторский институт информатизации, автоматизации и связи на железнодорожном транспорте, Москва, Российская Федерация

[1] Финансовый университет, Москва, Российская Федерация

* Корреспондирующий автор (tibult41[at]gmail.com)

Предложена: 13.06.2026; Принята: 22.07.2026; Опубликована: 17.08.2026

Аннотация

В работе рассматривается проблема накопления ошибок в классических двухэтапных конвейерах распознавания лиц. Проблема обусловлена разрывом графа вычислений между этапами геометрической регистрации и идентификации. Ошибка локализации ключевых точек необратимо снижает качество эмбеддинга, поскольку градиент функции потерь распознавания не передается модулю выравнивания. Этот эффект особенно критичен для видеоподсистем интеллектуальных транспортных систем (ИТС), функционирующих в условиях вибрации, переменного освещения и агрессивного сжатия видеопотока. Предложен модифицированный контроллер дифференцируемой регистрации на основе модели пространства состояний (State Space Model, SSM) архитектуры Mamba.

Ключевой вклад — механизм семантической адаптации — обогащение последовательности 68 ключевых точек обучаемыми поточечными эмбеддингами их анатомической роли (зоны глаз, носа, рта, контура), type-conditioned селективность и иерархическая инициализация скоростей затухания скрытого состояния, согласованная со стабильностью анатомических зон. Контроллер интегрирован в единый дифференцируемый конвейер с CNN-энкодером MobileNetV3-Small и обучается end-to-end с регуляризованной функцией потерь, сочетающей ArcFace и геометрический штраф. На бенчмарке LFW модель показывает сопоставимую с современными решениями точность 99.89% при 1.6 млн параметров, 0.26 GFLOPs и времени инференса 4.3 мс; по предварительным данным (один прогон, seed=42) на авторском деградированном наборе (JPEG q=30, шум, аффинные искажения) она сохраняет 98.4% точности против 71.8% у канонической связки ArcFace+Dlib, что соответствует разрыву 26.6 процентного пункта. Сложность контроллера линейна по числу ключевых точек, что существенно для плотных схем разметки; при используемых 68 точках главный выигрыш составляют параметрическая эффективность и учет анатомической структуры.

Ключевые слова: Mamba, State Space Models, дифференцируемая регистрация, распознавание лиц, семантическая адаптация, анатомические эмбеддинги, type-conditioned селективность, интеллектуальные транспортные системы, робастность к JPEG-сжатию, edge-устройства.

MODIFIED MAMBA CONTROLLER WITH SEMANTIC ADAPTATION FOR DIFFERENTIABLE FACE REGISTRATION IN INTELLIGENT TRANSPORT SYSTEMS

Research article

Kulikov A.A.[1, *]

[1] ORCID : 0000-0002-8443-3684;

[1] Scientific Research and Design Institute of Informatization, Automation and Communication in Railway Transport, Moscow, Russian Federation

[1] Financial University, , Russian Federation

* Corresponding author (tibult41[at]gmail.com)

Suggested: 13.06.2026; Accepted: 22.07.2026; Published: 17.08.2026

Abstract

The paper addresses the problem of error accumulation in classical two-stage face recognition pipelines caused by the computational-graph break between the geometric registration and the identification stages: a keypoint localisation error irreversibly degrades the embedding because the gradient of the recognition loss is not propagated to the alignment module. This effect is especially critical for the video subsystems of intelligent transport systems (ITS) operating under vibration, variable illumination and aggressive video-stream compression. A modified differentiable-registration controller based on a State Space Model (SSM) of the Mamba architecture is proposed.

The key contribution is a semantic-adaptation mechanism: enrichment of the sequence of 68 keypoints with learnable per-point embeddings encoding their anatomical roles (eye, nose, mouth and contour zones), type-conditioned selectivity, and a hierarchical initialisation of hidden-state decay rates consistent with the stability of the anatomical zones. The controller is integrated into a single differentiable pipeline together with a MobileNetV3-Small CNN encoder and is trained end-to-end with

a regularised loss combining ArcFace and a geometric penalty. On the LFW benchmark the model achieves an accuracy of 99.89% comparable to state-of-the-art solutions, with 1.6M parameters, 0.26 GFLOPs and a 4.3 ms inference time; according to preliminary data (a single run, seed=42), on a custom degraded set (JPEG q=30, noise, affine distortions) it preserves 98.4% accuracy versus 71.8% for the canonical ArcFace+Dlib pipeline, which corresponds to a 26.6 percentage-point gap. The controller's complexity is linear in the number of keypoints, which matters for dense markup schemes; at the 68 points used, the main gains are parameter efficiency and the use of anatomical structure.

Keywords: Mamba, State Space Models, differentiable registration, face recognition, semantic adaptation, anatomical embeddings, type-conditioned selectivity, intelligent transport systems, JPEG-compression robustness, edge devices.

Введение

Обеспечение безопасности на объектах транспортной инфраструктуры все чаще связывают с переходом от реактивных мер к превентивному непрерывному мониторингу. Видеоподсистемы интеллектуальных транспортных систем (ИТС) — бортовые и салонные камеры мониторинга состояния водителя, терминалы биометрического контроля доступа, камеры пропускных и досмотровых пунктов — функционируют в существенно более тяжелых условиях, чем офисные сценарии. Например, низкая и переменная освещенность, вибрация, смаз от движения, загрязнение оптики и агрессивное сжатие видеопотока в каналах ограниченной пропускной способности [1], [2]. При этом ошибки идентификации в транспортном контуре напрямую затрагивают безопасность участников движения, что выводит задачу устойчивого к деградации изображения распознавания лиц в число прикладных задач интеллектуализации управления транспортными процессами и средствами. Эволюция систем биометрической идентификации в обеспечении безопасности железнодорожного транспорта — от классических конвейеров к дифференцируемым гибридным архитектурам — рассмотрена в [3]. Вопросы автоматизации технологических процессов и сквозного контроля участников перевозочного процесса рассматривались, в частности, в работах [4], [5]. Предлагаемая архитектура ориентирована прежде всего на видеоподсистемы ИТС, сохраняя при этом применимость и в других областях биометрической идентификации.

Современные системы распознавания лиц достигли высокой точности на стандартных бенчмарках, однако в подавляющем большинстве строятся по двухэтапной схеме: детекция и геометрическое выравнивание лица по ключевым точкам, а затем — извлечение эмбеддинга и его сравнение с эталонами. Такая декомпозиция упрощает разработку, но порождает разрыв вычислительного графа между этапами: первый выступает «черным ящиком» для второго, и ошибки локализации ключевых точек на 2–5 пикселей, неизбежные при плохом освещении, закрытии лица или сильном ракурсе, приводят к некорректному аффинному преобразованию и каскадно переносятся на распознавание. Поскольку градиент от функции потерь идентификации не распространяется к модулю регистрации, эти искажения не компенсируются, а накапливаются [6]. При агрессивном JPEG-сжатии отказ детектора ключевых точек ведет к резкому падению точности распознавания.

Радикальным решением является переход к дифференцируемым (end–to–end) конвейерам, где параметры геометрической трансформации оптимизируются совместно с задачей распознавания. Базовая идея восходит к Spatial Transformer Networks (STN) [7], впервые предложившим дифференцируемый модуль аффинного преобразования, сохраняющий поток градиентов сквозь операцию выравнивания. Вместе с тем классический STN опирается на полносвязный локализатор, плохо моделирует нелинейные зависимости между удаленными точками лица и ограничен в типах преобразований [8]. Близкое современное решение задачи робастного дифференцируемого выравнивания — ARoFace [9] — не использует механизм моделей пространства состояний. Механизмы внимания трансформеров имеют квадратичную по длине последовательности сложность $O(N^2)$, что затрудняет их применение на периферийных устройствах.

В последние годы модели пространства состояний (State Space Models, SSM), прежде всего архитектура Mamba [10], продемонстрировали способность моделировать длинные зависимости при линейной сложности $O(N)$, конкурируя с трансформерами при меньших затратах ресурсов. Тем не менее прямое применение стандартного Mamba к задаче геометрической регистрации не учитывает анатомическую семантику данных: для модели все ключевые точки являются равнозначными векторами координат, хотя глаза, нос, рот и контур челюсти обладают разной степенью вариативности и информативности. Возможность применения SSM для адаптивного управления регистрацией по антропометрическим точкам остается практически неисследованной.

Целью настоящей работы является разработка и исследование модифицированного Mamba–контроллера с механизмом семантической адаптации, позволяющего дифференцированно обрабатывать анатомические зоны лица для повышения устойчивости регистрации в условиях деградации видеопотока ИТС. Научная новизна состоит в способе интеграции SSM в дифференцируемый конвейер регистрации лица: насколько известно автору, впервые Mamba используется как контроллер конвейера распознавания лиц, учитывающего выравнивание (alignment–aware), последовательность ключевых точек обогащается обучаемыми эмбеддингами анатомической роли точек, а скорости затухания скрытого состояния инициализируются иерархически с учетом стабильности зон лица. Практическая значимость определяется вычислительной эффективностью полученной модели (1,6 млн параметров, 0.26 GFLOPs, 4,3 мс) и количественной характеристикой ее робастности к контролируемой деградации входных изображений.

Обзор современных методов

Эволюция методов распознавания лиц прошла путь от ручных геометрических и статистических признаков (Eigenfaces, Fisherfaces) и эластичных графов к глубоким сверточным сетям (DeepFace, FaceNet). Качественный скачок обеспечили функции потерь с угловым маржином (ArcFace [11], CosFace, AdaFace, MagFace), ставшие стандартом де-факто; параллельно развивались легкие архитектуры (MobileFaceNet, MobileOne, EfficientFormer) для устройств с ограниченными ресурсами. Однако даже передовые модели сохраняют двухэтапную организацию, при которой

регистрация лица по ключевым точкам (Dlib, MTCNN, HRNet) и распознавание являются независимыми подзадачами, что воспроизводит описанный во введении разрыв.

Парадигма дифференцируемой регистрации устраняет этот разрыв, превращая геометрическое выравнивание из внешнего автономного алгоритма во встроенный обучаемый модуль. Первым и наиболее влиятельным шагом стала работа над STN [7], показавшая возможность сквозного распространения градиента через операцию выборки пикселей. Последующие расширения (TPS-STN, FiLM-модуляция, GridFace) и наиболее современное из них ARoFace [9] решают задачу робастного к выравниванию распознавания, однако ограничены классами преобразований либо действуют лишь на уровне функции потерь и не задействуют механизм SSM. Анализ показывает [8], что STN эффективен для аффинных преобразований, но плохо справляется с нелинейными деформациями мимики, а полносвязный локализатор снижает интерпретируемость и робастность предсказания параметров трансформации.

Модели пространства состояний предлагают иную алгоритмическую основу: последовательность описывается рекуррентным скрытым состоянием, кодирующим долгосрочные зависимости с линейной сложностью *O(N)*. Модель S4 [12] ввела структурированную HiPPO-параметризацию матрицы состояния, обеспечившую удержание информации на тысячах шагов и интерпретацию рекуррентной обработки как глобальной свертки. Mamba [10] устранила статичность классических SSM, сделав параметры *B, C* и шаг дискретизации *Δ* функциями входного сигнала (input-driven селективность): модель динамически выбирает, какие части истории запоминать. Семейство S5 [13] упрощает модель, используя диагональную поканальную матрицу состояния (как диагональную аппроксимацию HiPPO-инициализации); этот подход концептуально близок к применяемой в настоящей работе type-conditioned селективности — селективности, обусловленной типом ключевой точки.

В компьютерном зрении SSM адаптированы как самостоятельный визуальный backbone (опорную сеть): Vision Mamba [14] вводит двунаправленный SSM-блок, а VMamba — cross-scan механизм покрытия пространственных зависимостей; обе конкурентоспособны с трансформерами при меньших затратах. Формируется и тенденция гибридизации CNN и SSM, сочетающей локальную чувствительность сверток с глобальным контекстом SSM (U-Mamba, ConvMambaSR). Критический контраргумент дает MambaOut [15]: в классификации изображений SSM-блоки не приносят существенного выигрыша против тщательно настроенных сверточных моделей; эта критика, однако, относится к чистым SSM-классификаторам и не отменяет преимуществ SSM в задачах с явной последовательной структурой, какой является последовательность ключевых точек лица.

Таким образом, ни один из существующих подходов не обеспечивает одновременно высокой точности, вычислительной эффективности и робастности к геометрическим искажениям в условиях неконтролируемой съемки, а гибридные CNN-SSM архитектуры, явно спроектированные для совместной дифференцируемой регистрации и распознавания лиц, в литературе не описаны: существующие гибриды заменяют внимание в задачах классификации и сегментации общего назначения, не интегрируя геометрическое выравнивание. Этот научный пробел определяет постановку настоящего исследования.

Математическая модель семантически адаптированного SSM

3.1. Дифференцируемая регистрация

Для устранения указанного пробела сначала формализуем дифференцируемую регистрацию, а затем введем семантически адаптированную SSM, управляющую ее параметрами. Пусть входное изображение лица задается тензором $I \in \mathbb{R}^{H\times W\times C}$. Цель — получить нормализованное представление, инвариантное к аффинным искажениям, но сохраняющее семантические особенности лица; в отличие от классического подхода, где выравнивание выполняется внешним алгоритмом, весь процесс полностью дифференцируем. Параметры аффинного преобразования предсказываются обучаемым локализатором $\theta = f_{loc}(I;\varphi)$, где $\theta \in \mathbb{R}^6$ формирует матрицу

$$T(\theta) = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \in \mathbb{R}^{2\times 3}.$$

Координаты выходной сетки отображаются в исходное изображение по обратному (backward) правилу $[x,y]^T = T(\theta)\left[x',y',1\right]^T$, после чего значение пикселя вычисляется билинейной интерполяцией

$$I'\left(x',y'\right) = \sum_{i=\lfloor x\rfloor}^{\lfloor x\rfloor+1}\sum_{j=\lfloor y\rfloor}^{\lfloor y\rfloor+1} \omega(i,j \mid x,y) I(i,j), \quad \omega(i,j \mid x,y) = (1-|i-x|)(1-|j-y|).$$

Билинейная интерполяция дифференцируема по *(x,y)*, поэтому градиент функции потерь распространяется к параметрам трансформации и далее к параметрам локализатора по цепному правилу:

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial I'}\cdot\frac{\partial I'}{\partial G}\cdot\frac{\partial G}{\partial \theta}, \quad \frac{\partial L}{\partial \varphi} = \frac{\partial L}{\partial \theta}\cdot\frac{\partial \theta}{\partial \varphi}.$$

В качестве основной компоненты функции потерь принят ArcFace [11], вводящий угловой маржин m между классами в нормализованном пространстве признаков:

$$L_{\text{ArcFace}} = -\log\frac{e^{s\cos\left(\theta_{\text{arc},y}+m\right)}}{e^{s\cos\left(\theta_{\text{arc},y}+m\right)}+\sum_{j\neq y}e^{s\cos\theta_{\text{arc},j}}},$$

где s — масштаб, $\theta_{arc,j}$ — угол между текущим эмбеддингом и центроидом класса j (обозначение θ_{arc} отделено от вектора параметров трансформации θ). В отличие от полносвязного локализатора STN, функция f_{loc} реализуется как гибридный CNN–SSM-блок, кодирующий долгосрочные зависимости между всеми ключевыми точками, что повышает устойчивость предсказания θ при частичном закрытии лица и сильном ракурсе.

3.2. Семантическое обогащение последовательности ключевых точек

В предлагаемой архитектуре SSM получает на вход не последовательность патчей изображения, а упорядоченную последовательность координат лицевых ключевых точек $\mathcal{L} = \left\{p_i\right\}_{i=1}^{n}, \quad p_i = \left(x_i,y_i\right)^T \in \mathbb{R}^2, \quad n = 68,$ и переосмысливается как инструмент моделирования пространственной геометрии объекта. Поскольку координаты сами

по себе несут мало семантической информации — модель не «знает», является ли точка центром глаза или уголком рта, — каждой из 68 точек схемы разметки сопоставляется собственный обучаемый эмбеддинг e_{tk}, кодирующий ее анатомическую роль (обучается 68 поточечных векторов); анатомические же группы (глаза, нос с бровями, рот, контур) используются далее только для иерархической инициализации спектра Λ и интерпретации. Входной вектор шага k формируется конкатенацией спроецированных координат и эмбеддинга:

$$x_k = \begin{bmatrix} W_p & p_k \\ W_e & e_{t_k} \end{bmatrix} \in \mathbb{R}^{2d_p},$$

где $W_p \in \mathbb{R}^{d_p \times 2}$ — проекция координат, $W_e \in \mathbb{R}^{d_p \times d_e}$ – проекция эмбеддинга типа. В реализации $d_p = d_e = 32$, так что размерность токена равна $2d_p = 64$, после чего он линейно проецируется в скрытую размерность $D_{model} = 512$. Введение поточечных эмбеддингов дает сети информацию об анатомической роли точки, отсутствующую в стандартном Mamba.

3.3. Type-conditioned селективность

Дискретная SSM описывается рекуррентным уравнением $h_k = \bar{A}_k h_{k-1} + \bar{B}_k x_k, y_k = Ch_k$. В каноническом Mamba селективны параметры. В каноническом Mamba селективны параметры B, C и шаг дискретизации Δ, задаваемые линейными проекциями текущего входа x_k. В предлагаемой модификации селективность сохранена для B и Δ, которые обусловливаются обучаемым эмбеддингом анатомического типа точки e_{tk}, а не значением входа; роль выходного отображения C выполняет обучаемая readout-проекция g_ψ конечного состояния h_n (см. раздел 4. Реализация):

$$\bar{B}_k = \sigma\left(W_r e_{t_k}\right) \odot B, \quad \Delta_k = \delta \cdot \mathrm{Softplus}\left(W_\Delta e_{t_k}\right)$$

где $\sigma(\cdot)$ — сигмоида,

W_r, W_Δ — обучаемые проекции,

$\delta > 0$ — масштабирующая константа.

Базовая матрица входа B при этом не статична: в реализации она порождается из того же эмбеддинга типа двухслойным MLP $f_{\mathrm{MLP}}\left(e_{t_k}\right) = W^{(2)}\,\mathrm{GELU}\left(W^{(1)} e_{t_k} + b^{(1)}\right) + b^{(2)}$, после чего гейтируется множителем $\sigma(W_r e_{tk})$ и дискретизируется с шагом. Такой подход позволяет контроллеру дифференцировать зоны: для стабильных модель интегрирует информацию на длинных интервалах, для подвижных – быстрее обновляет состояние, реагируя на мимику. Вычислительная сложность снижается относительно канонического Mamba за счет того, что проекции вычисляются от 32-мерного эмбеддинга типа, а не от 512-мерного входного токена; интерпретируемость при этом повышается. Термин «Mamba-контроллер» сохранен, поскольку сохранен ключевой принцип селективности Mamba — зависимость B и Δ от контекста, — однако условие перенесено со значения входа на дискретный анатомический тип; формально такая конструкция ближе к классу S5 [13].

3.4. Иерархическая инициализация спектра Λ

Матрица состояния использует структурированную диагональ-плюс-низкоранговую параметризацию HiPPO/S4, $A = \Lambda - pq^T$, $\Lambda = diag(\lambda_1,\ldots,\lambda_N)$, $\lambda_i < 0$. Элементы λ_i интерпретируются как скорости затухания информации о различных геометрических модах. Авторская иерархическая инициализация связывает значения λ_i с анатомическими группами точек:

$$\lambda_{\mathrm{eye}} = -0,5, \quad \lambda_{\mathrm{nose}} = -0,3, \quad \lambda_{\mathrm{mouth}} = -1,0, \quad \lambda_{\mathrm{contour}} = -0,8$$

Меньший модуль соответствует стабильной области (нос, глаза) и длительному удержанию состояния; больший — вариативной зоне (рот, контур), для которой задается быстрое затухание шума артикуляции. Динамическая устойчивость требует условия $\left|\lambda_i\left(\bar{A}_k\right)\right| < 1$, которое достигается параметризацией $\bar{A}_k = \left(I + \Delta_k A\right) \odot \sigma(S)$: множитель $\sigma(S) \in (0, 1)^N$ удерживает собственные значения внутри единичного круга, а инициализация (приведенная выше) задает согласованные начальные скорости затухания. Безусловно (при любом $\Delta_k > 0$) оно выполняется лишь для точной ZOH-дискретизации; для используемой в реализации Euler-аппроксимации при $\mathrm{Re}(\lambda_i(A)) < 0$ устойчивость гарантируется при ограниченном сверху шаге $0 < \Delta_k \le \Delta_{max}$ (из требования $\left|1 + \Delta_k \lambda_i\right| < 1$); при рабочих $\Delta_k \in \left[10^{-3}, 10^{-1}\right]$ и $\mathrm{Re}\left(\lambda_i\right) \le -0,1$ численная устойчивость подтверждена эмпирически, что обеспечивает ограниченность $\|h_k\|$ при ограниченном входе.

3.5. Регуляризатор и гибридный блок

Предложенный контроллер интегрируется в конвейер из пяти ступеней:

(1) CNN-энкодер MobileNetV3-Small извлекает признаковую карту F;

(2) keypoint-голова (голова предсказания ключевых точек) регрессирует координаты 68 точек через heatmap-представление (тепловую карту);

(3) координаты объединяются с эмбеддингами типов;

(4) Mamba-блок обрабатывает обогащенную последовательность и предсказывает $\theta \in R^6$;

(5) дифференцируемый сэмплер применяет θ к карте F, формируя выровненную карту F', на которую действует голова распознавания. Чтобы предотвратить вырожденные преобразования (коллапс изображения в линию или точку), в функцию потерь введен регуляризатор

$$L_{\mathrm{reg}} = \left\|\theta - \theta_{\mathrm{id}}\right\|_F^2 + \alpha \cdot \frac{1}{\varepsilon + \det(M)^2}, \quad \alpha = 0,1, \varepsilon = 10^{-6}$$

где M — линейная часть аффинной матрицы;

$\theta_{id} = [1, 0, 0, 0, 1, 0]^T$ — параметры тождественного преобразования.

Первый член (норма Фробениуса) штрафует отклонение от тождественного преобразования и обнуляется при identity-старте; второй препятствует обращению $|det(M)|$ в нуль, сохраняя обратимость и ориентацию; использование

$det(M)^2$ вместо $|det(M)|$ устраняет излом производной в нуле. Полная потеря имеет аддитивный вид $L_{total} = L_{ArcFace} + \lambda \cdot L_{reg}$, $\lambda = 10^{-4}$; начальное значение λ устанавливается эвристической балансировкой норм градиентов по аналогии с GradNorm [16], после чего фиксируется.

Реализация

Модель реализована на PyTorch 2.1.0 с CUDA 12.1 и библиотекой mamba_ssm 1.1.0. Входное изображение $I \in \mathbb{R}^{H \times W \times 3}$ преобразуется CNN-энкодером MobileNetV3-Small [17] в признаковую карту с $D = 576$ каналами; выбор обусловлен пригодностью для edge-устройств: число операций более чем втрое ниже, чем у ResNet-50, при сопоставимом качестве признаков. Heatmap-голова извлекает из карты координаты 68 точек (каждый канал отвечает за одну точку, координаты — взвешенное среднее активаций). Последовательность передается Mamba-контроллеру с описанной выше type-conditioned селективностью и иерархической λ-инициализацией. Readout-проекция конечного состояния g_ψ реализована двухслойным MLP с «бутылочным горлышком», выдающим вектор $\theta \in \mathbb{R}^6$.

Важным элементом реализации является identity-инициализация выходного слоя контроллера: параметры последнего линейного слоя g_ψ инициализируются так, чтобы при начале обучения θ_0 соответствовало тождественному аффинному преобразованию ($M_0 = I$, нулевые сдвиги). Это дает сети «безопасную» точку отсчета $F' \approx F$ и предотвращает разрушение предобученных сверточных признаков на ранних эпохах. Группы анатомических индексов в финальной конфигурации (68-точечная разметка Multi-PIE) распределены следующим образом: контур лица — точки 0–16 с $\lambda = -0{,}8$; брови и нос -– точки 17–35 с $\lambda = -0{,}3$; глаза — точки 36–47 с $\lambda = -0{,}5$; рот — точки 48–67 с $\lambda = -1{,}0$. Объединение бровей с группой носа отражает то, что диапазон 17–35 в основном покрывает переносицу и спинку носа, а инициализация задает лишь стартовые значения, корректируемые обучением.

Обучение проводится в два этапа. На стадии warm-up (разогрев, 5 эпох) CNN-энкодер заморожен, обучаются keypoint-голова, контроллер и голова распознавания; identity-инициализация θ стабилизирует начальную фазу. На этой стадии к функции потерь добавляется компонента $L_{kp} = \left\| p - p_{\text{gt}} \right\|_2$ — L2-расстояние до ground-truth ключевых точек из WFLW-аннотаций; внешняя разметка используется только для инициализации keypoint-головы; в режиме инференса внешний детектор не применяется, что сохраняет тезис об end-to-end-обучении. На стадии полного обучения (full training, 50 эпох) все слои размораживаются и дообучаются совместно оптимизатором AdamW (начальная скорость обучения, learning rate, $3 \cdot 10^{-4}$, коэффициент затухания весов (weight decay) 10^{-4}, scheduler ReduceLROnPlateau, размер батча 128) на NVIDIA A100. Без регуляризатора нормы градиентов по θ на порядок превышают нормы по весам CNN, и задача регистрации доминирует; введение L_{reg} с $\lambda = 10^{-4}$ выравнивает масштабы. Воспроизводимость обеспечивается фиксированным seed = 42, детерминированным режимом cuDNN и закрепленными версиями библиотек. Triton-ядра не гарантируют побитовую детерминированность: повторные запуски с тем же seed = 42 дают разброс точности на LFW в пределах ±0,05%; эта оценка характеризует недетерминизм вычислений и не заменяет валидацию по нескольким независимым значениям seed. Обученная модель занимает около 6.5 МБ (FP32, 1.6 млн параметров) и около 1.7 МБ при квантизации INT8.

Методика эксперимента

Экспериментальное исследование проводилось на унифицированном вычислительном стенде (Intel Xeon Gold 6248R, NVIDIA A100 40 ГБ, 192 ГБ ОЗУ) с фиксированными версиями программного обеспечения и seed = 42. Для предобучения CNN-энкодера и головы распознавания использовалась очищенная исследовательская версия набора MS-Celeb-1M (в литературе известная как MS1MV2) [19]; для обучения и оценки keypoint-головы и контроллера – WFLW (10 000 изображений in-the-wild, то есть снятых в неконтролируемых условиях; первые 68 точек разметки Multi-PIE); для финальной верификации — LFW по протоколу «unrestricted with labeled outside data» (6000 пар; точность вычисляется при пороге, соответствующем точке равенства FAR и FRR). Качество локализации оценивалось метрикой NME, нормализованной по межзрачковому расстоянию, и метрикой PCKh; вычислительная эффективность — временем инференса полного цикла от загрузки изображения до получения эмбеддинга (усреднение по 1000 прогонов), числом параметров и FLOPs. Замеры выполнялись в типовых для каждого конвейера режимах исполнения: GPU-этапы — на A100, CPU-этап Dlib — на процессоре Intel Xeon; сравнение времени относится к полным конвейерам при гетерогенном исполнении.

Для оценки робастности в условиях, приближенных к реальной эксплуатации, сформирован авторский деградированный набор на основе LFW по методологии Hendrycks & Dietterich [18]: к каждому изображению одновременно применялись JPEG-сжатие с качеством 30 (имитация узких каналов передачи), гауссов шум со стандартным отклонением $\sigma = 0{,}01$ и случайные аффинные искажения (поворот ± 30˚, масштабирование [0.8,1.2], сдвиг до 10% размера изображения). Набор не использовался при обучении; случайные аффинные преобразования генерируются детерминированно при фиксированном seed = 42, а скрипт генерации набора и конфигурация деградаций входят в экспериментальный протокол работы и доступны у автора по обоснованному запросу. В качестве базовых моделей выбраны представители ключевых парадигм: канонический двухэтапный конвейер ArcFace+Dlib (детекция и локализация 68 точек Dlib, выравнивание методом наименьших квадратов, распознавание ArcFace на ResNet-100), легкая одностадийная MobileFaceNet, дифференцируемая STN-ArcFace (полносвязный локализатор STN после MobileNetV3-Small) и SSM-архитектура Vision Mamba (Vim-Tiny, привлекается для оценки применимости класса архитектур, а не как SOTA-baseline). Все базовые модели переобучены и протестированы в идентичных условиях единого стенда, поэтому их показатели могут отличаться от значений, опубликованных в оригинальных

работах в иных конфигурациях обучения; сравнение корректно внутри единого протокола. Сводные характеристики приведены в таблице 1.

Таблица 1 - Сравнительная характеристика базовых моделей и предложенного решения

DOI: https://doi.org/10.60797/IRJ.2026.170.48.1

Модель	Тип подхода	Энкодер	Метод регистрации	Параметры, млн	FLOPs, G	Inference, мс
ArcFace + Dlib	двухэтапный	ResNet-100	внешний (Dlib)	55,7	11,3	42,1
MobileFaceNet	одностадийный	MobileNet-like	нет	0,98	0,22	3,8
STN-ArcFace	дифференцируемый	MobileNetV3-Small	STN (MLP)	1,5	0,25	4,1
Vision Mamba (Vim)	SSM-based	Vim-Tiny	нет	5,8	1,7	18,5
Предложенная модель	гибридный (CNN+SSM)	MobileNetV3-Small	Mamba-контроллер	1,6	0,26	4,3

Как видно из таблицы 1, по вычислительной сложности предложенная модель сопоставима с легкими решениями (MobileFaceNet, STN-ArcFace), но вводит принципиально иной механизм управления регистрацией. Обработка биометрических данных регулируется в Российской Федерации Федеральным законом № 152-ФЗ, в международной практике — GDPR (ст. 9); все эксперименты выполнены на публично доступных академических наборах в исследовательских целях без идентификации частных лиц. Исходный набор MS-Celeb-1M был отозван правообладателем (Microsoft) в 2019 г. по соображениям защиты персональных данных, поэтому его очищенная исследовательская версия (MS1MV2) использована только для предобучения, без распространения исходных изображений.

Результаты

На стандартном бенчмарке LFW предложенная модель показала точность 99,89%, сопоставимую с лучшими из сравниваемых архитектур (таблица 2). Разница с ArcFace+Dlib (+0,06 п.п.) находится в пределах наблюдаемой дисперсии запусков; результат интерпретируется как сопоставимая точность при существенно более легком энкодере (MobileNetV3-Small против ResNet-100).

Таблица 2 - Результаты верификации на наборе LFW

DOI: https://doi.org/10.60797/IRJ.2026.170.48.2

Модель	Accuracy, %	AUC
Предложенная модель	99,89	0,9995
ArcFace + Dlib	99,83	0,9993
Vision Mamba (Vim)	99,30	0,9982
STN-ArcFace	98,95	0,9963
MobileFaceNet	98,72	0,9951

Ключевые различия проявляются на авторском деградированном наборе (таблица 3). По проведенным экспериментам (seed=42) предложенная модель сохранила точность 98,4%, тогда как у ArcFace+Dlib она снизилась до 71,8%; межмодельный разрыв составляет 26,6 процентного пункта. Время инференса предложенной модели (4,3 мс) при этом существенно меньше, чем у двухэтапной связки (42,1 мс); в последнюю входят CPU-этап Dlib (детекция и локализация точек) и GPU-этап распознавания, поэтому разрыв отражает в том числе гетерогенность исполнения, а не только глубину сетей. Результат отражает преодоление ограничения жесткого разделения этапов: обученный end-to-end Mamba-контроллер адаптивно корректирует выравнивание, компенсируя искажения, тогда как внешний детектор Dlib и жесткий STN менее устойчивы.

Таблица 3 - Результаты на авторском деградированном наборе (JPEG q = 30 + шум + аффинные искажения)

DOI: https://doi.org/10.60797/IRJ.2026.170.48.3

Модель	Accuracy, %	NME, %	PCKh, %	Inference, мс
Предложенная	98,4	3,82	94,3	4,30

Модель	Accuracy, %	NME, %	PCKh, %	Inference, мс
модель				
ArcFace + Dlib	71,8	5,67	82,1	42,10
Vision Mamba (Vim)	91,1	4,25	89,7	18,50
STN-ArcFace	85,3	4,51	87,2	4,10
MobileFaceNet	82,0	-	-	3,80

Примечание: прочерки у MobileFaceNet означают, что метрики локализации неприменимы – модель не использует ключевые точки

Для локализации источника робастности проведена серия экспериментов с разными типами деградации (таблица 4). Необходимо оговорить протокол: значения строки «JPEG (качество = 30)» таблицы 4 поэлементно совпадают с таблицей 3 (98,4 / 71,8 / 91,1) — это то же измерение на авторском наборе, обозначаемое по доминирующему фактору деградации; изолированный вклад сжатия без шума и аффинных искажений отдельно не оценивался, и данную строку следует читать как оценку на комбинированном наборе, тогда как строки с шумом, поворотом и масштабом характеризуют отдельные факторы. Вследствие различия протоколов строки таблицы 4 корректно сравнивать между моделями, но не между собой. На комбинированном наборе точность всех моделей оказывается выше, чем при изолированном повороте на ±30°. Преимущество предложенной модели сохраняется во всех сценариях, но наиболее выражено при геометрических искажениях (поворот, масштаб), что напрямую подтверждает эффективность адаптивного управления геометрией через Mamba-контроллер.

Таблица 4 - Точность моделей при различных типах деградации (на LFW)

DOI: https://doi.org/10.60797/IRJ.2026.170.48.4

Тип деградации	Предложенная модель, %	ArcFace + Dlib, %	Vision Mamba, %
Исходный LFW	99,89	99,83	99,30
JPEG (качество = 30)	98,40	71,80	91,10
Гауссов шум (σ = 0,01)	98,95	85,20	93,50
Поворот (± 30˚)	97,80	68,40	89,20
Масштаб (0,8–1,2)	98,10	75,60	90,80

Примечание: строка «JPEG (качество = 30)» соответствует авторскому комбинированному деградированному набору (JPEG-сжатие + шум + аффинные искажения; те же значения, что в таблице 3), а не изолированному JPEG-сжатию; остальные строки характеризуют каждый фактор деградации по отдельности, поэтому строки таблицы сопоставимы между моделями, но не между собой

Вклад компонентов оценен абляционным исследованием (таблица 5) с поэтапным наращиванием архитектуры от MLP-контроллера к полной модели. Добавление блока Mamba дает прирост точности на деградированных данных почти на 9 п.п. относительно MLP, однако ключевой вклад вносит семантическая адаптация: эмбеддинги типов (Type Embeddings) добавляют еще ~2,9 п.п., подтверждая, что знание анатомической роли точки помогает фильтровать шум, а иерархическая инициализация Λ стабилизирует обучение и дает дополнительный прирост робастности при сильных геометрических искажениях. Эти выводы согласуются с дополнительной серией абляций, выполненной по схеме исключения одного компонента (leave-one-out) из полной модели. В отличие от таблицы 5, где архитектура наращивается от MLP-базлайна и каждый прирост измеряется относительно предыдущей конфигурации, в этой серии вклад компонента оценивается как падение точности при его удалении из полной модели; различием схем измерения и объясняется несовпадение числовых значений двух рядов. Так, замена эмбеддингов типов на простую конкатенацию снижает точность на 4,2 п.п., замена Mamba-контроллера на полносвязный слой — на 6,8 п.п., а отключение регуляризатора L_{reg} вызывает нестабильность обучения и вырождение трансформаций в 12% случаев (приросты и потери точности здесь и выше указаны в процентных пунктах).

Таблица 5 - Результаты абляционного исследования компонентов Mamba-контроллера

DOI: https://doi.org/10.60797/IRJ.2026.170.48.5

Модель	Accuracy LFW (clean), %	Accuracy Degraded, %	NME WFLW, %	Параметры, млн
Baseline (MLP-контроллер)	98,95	85,30	4,51	1,5
Standard Mamba	99,70	94,20	3,95	1,6

Модель	Accuracy LFW (clean), %	Accuracy Degraded, %	NME WFLW, %	Параметры, млн
+ Type Embeddings	99,82	97,10	3,88	1,6
+ Hierarchical Λ Init	99,85	97,80	3,85	1,6
Full Proposed Model	99,89	98,40	3,82	1,6

Анализ гиперпараметров показал, что оптимально 68 ключевых точек: меньшее количество (20) дает недостаток информации о геометрии, большее (98) — шум от менее устойчивых точек (например, на контуре волос) и снижение точности на проверочной выборке; недостаточный размер скрытого состояния не позволяет кодировать зависимости между точками, чрезмерный — ведет к переобучению. Конфигурация $D_{model} = 512$, $d_{state} = 16$, $n = 68$ принята как компромисс.

Обсуждение и ограничения

Предложенный механизм семантической адаптации отвечает на ключевую проблему применения SSM к геометрическим данным: в отличие от однородных токенов текста, каждая точка лица несет уникальную смысловую нагрузку. Type-conditioned селективность реализует аналог «структурного внимания» с линейной сложностью *O(N)*. Оговорим пределы этого аргумента: при *n = 68* выигрыш асимптотики невелик (матрица сходства 68 × 68 вычислительно дешева), и практическое преимущество составляют параметрическая эффективность и индуктивное смещение анатомической структуры; линейная сложность значима для плотных схем разметки (98 и более точек, 3D-представления) и видеопоследовательностей. При этом собственный анализ показал, что 98 точек снижают точность, поэтому масштабирование по числу точек – направление дальнейшей работы, а не доказанное преимущество. Профиль эффективности (1,6 млн параметров, 0.26 GFLOPs, 4,3 мс, 1,7 МБ в INT8) по порядку величины соответствует возможностям встраиваемых платформ класса NVIDIA Jetson и NPU бортовых видеорегистраторов, однако фактические замеры на периферийном оборудовании остаются необходимым условием практического внедрения.

Полученная робастность имеет прямое прикладное значение для видеоподсистем ИТС: терминалы биометрического контроля на пропускных пунктах и в депо, турникетные комплексы вокзалов и салонные камеры мониторинга водителя работают именно при агрессивном сжатии видеопотока и геометрических искажениях (поворот ± 30˚ — прокси вибрационных смещений камеры), где преимущество модели выражено сильнее всего. Количественная оценка сложных сценариев очерчивает границы применимости: частота ошибок предложенной модели составляет 8,2% при частичном закрытии лица маской, 12,1% при контровом свете, 18,9% при экстремальном профиле (> 60˚) и 15,4% при сильном размытии движения — против 22,5%, 35,7%, 48,2% и 41,0% соответственно у ArcFace+Dlib. Подчеркнем также, что деградированный набор производен от LFW (та же популяция лиц), поэтому сохранение 98,4% точности не переносится автоматически на иные домены съемки.

Результаты имеют ряд ограничений. Во-первых, все числовые показатели получены в одном прогоне обучения с seed = 42 и носят предварительный характер; разброс ±0,05% на LFW от недетерминизма triton-ядер характеризует лишь вычислительный шум, а строгая валидация по нескольким независимым значениям seed (3 и более) с построением доверительных интервалов запланирована, но не выполнена. Во-вторых, пригодность к edge-развертыванию опирается на профиль сложности, измеренный на серверном GPU, и не подтверждена замерами на периферийных устройствах. В-третьих, метод предполагает знание топологии лица (фиксированную схему разметки): в биометрии это требование обычно выполняется, но оно ограничивает перенос подхода на произвольные классы объектов. Наконец, приведенные выше частоты ошибок в экстремальных сценариях показывают, что точность модели снижается при сильном закрытии лица и экстремальных ракурсах, отражая зависимость от наличия достаточного количества видимых ключевых точек для корректного анализа геометрии. Применение к мониторингу состояния водителя (DMS) рассматривается как потенциальное направление и экспериментально в настоящей работе не проверялось.

Заключение

В работе предложен и исследован модифицированный Mamba-контроллер с механизмом семантической адаптации для дифференцируемой регистрации лиц в составе единого end-to-end конвейера с CNN-энкодером MobileNetV3-Small.

Экспериментальное исследование показало, что предложенные модификации обеспечивают сопоставимую с современными решениями точность на чистом бенчмарке LFW (99,89%) при существенно более легком энкодере и, по проведенным экспериментам (seed=42) имеет повышенную робастность к деградации входных изображений: на авторском деградированном наборе модель сохраняет 98.4% точности против 71,8% у канонической связки ArcFace+Dlib (разрыв 26,6 процентного пункта) при времени инференса 4,3 мс против 42,1 мс у двухэтапной связки. Необходимо подчеркнуть, что это сопоставление времени получено при гетерогенном исполнении конвейеров (CPU-этап локализации Dlib против полностью GPU-исполнения предложенной модели), поэтому наблюдаемый разрыв во времени отражает не только различие архитектур, но и условия исполнения. Вычислительная эффективность (1,6 млн параметров, 0.26 GFLOPs) делает архитектуру потенциально пригодной для систем реального времени, в том числе для видеоподсистем интеллектуальных транспортных систем; вместе с тем тезис о пригодности к edge-развертыванию опирается на профиль сложности, измеренный на серверном GPU, и требует подтверждения прямыми замерами на встраиваемом оборудовании. Дальнейшие направления работы включают статистическую валидацию по нескольким

независимым значениям seed с построением доверительных интервалов, замеры на периферийных устройствах и экспериментальную проверку применимости подхода к мониторингу состояния водителя.

Конфликт интересов

Не указан.

Рецензия

Нуриев М.Г., Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ, Казань Российская Федерация
DOI: https://doi.org/10.60797/IRJ.2026.170.48.6

Conflict of Interest

None declared.

Review

Nuriev M.G., Kazan National Research Technical University named after A.N. Tupolev – KAI, Kazan Russian Federation
DOI: https://doi.org/10.60797/IRJ.2026.170.48.6

Список литературы / References

1. Hu C.-H. Toward Driver Face Recognition in the Intelligent Traffic Monitoring Systems / C.-H. Hu, Y. Zhang, F. Wu [et al.] // IEEE Transactions on Intelligent Transportation Systems. — 2020. — Vol. 21, № 12. — P. 4958–4971. — DOI: 10.1109/TITS.2019.2945923.

2. Yang L. Video-Based Driver Drowsiness Detection With Optimised Utilization of Key Facial Features / L. Yang, H. Yang, H. Wei [et al.] // IEEE Transactions on Intelligent Transportation Systems. — 2024. — Vol. 25, № 7. — P. 6938–6950. — DOI: 10.1109/TITS.2023.3346054.

3. Куликов А.А. Эволюция систем биометрической идентификации в обеспечении безопасности железнодорожного транспорта: от классических конвейеров к дифференцируемым гибридным архитектурам / А.А. Куликов // Транспортное дело России. — 2026. — № 3. — С. 153–155. — EDN: DXLOUP.

4. Кулагин М.А. Интеллектуальная система анализа и прогнозирования нарушений при управлении подвижным составом : дис. … канд. техн. наук : 05.13.06 / М.А. Кулагин. — Москва : РУТ (МИИТ), 2022.

5. Замышляев А.М. Автоматизация процессов комплексного управления техническим содержанием инфраструктуры железнодорожного транспорта : дис. … д-ра техн. наук : 05.13.06 / А.М. Замышляев. — Москва : НИИАС, 2014.

6. Bulat A. How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230 000 3D Facial Landmarks) / A. Bulat, G. Tzimiropoulos // Proceedings of the IEEE International Conference on Computer Vision (ICCV). — 2017. — P. 1021–1030. — DOI: 10.1109/ICCV.2017.116.

7. Jaderberg M. Spatial Transformer Networks / M. Jaderberg, K. Simonyan, A. Zisserman [et al.] // Advances in Neural Information Processing Systems 28 (NeurIPS 2015). — 2015. — P. 2017–2025. — arXiv:1506.02025.

8. Finnveden L. Understanding when spatial transformer networks do not support invariance, and what to do about it / L. Finnveden, Y. Jansson, T. Lindeberg // 25th International Conference on Pattern Recognition (ICPR). — Milan, Italy, 2021. — P. 3427–3434. — DOI: 10.1109/ICPR48806.2021.9412997.

9. Saadabadi M.S.E. ARoFace: Alignment Robustness to Improve Low-Quality Face Recognition / M.S.E. Saadabadi, S.R. Malakshan, A. Dabouei, N.M. Nasrabadi // Computer Vision – ECCV 2024. — Cham : Springer, 2024. — P. 308–327. — DOI: 10.1007/978-3-031-73414-4_18.

10. Gu A. Mamba: Linear-Time Sequence Modeling with Selective State Spaces / A. Gu, T. Dao. — arXiv, 2023. — arXiv:2312.00752.

11. Deng J. ArcFace: Additive Angular Margin Loss for Deep Face Recognition / J. Deng, J. Guo, N. Xue [et al.] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — 2019. — P. 4690–4699. — DOI: 10.1109/CVPR.2019.00482.

12. Gu A. Efficiently Modeling Long Sequences with Structured State Spaces (S4) / A. Gu, K. Goel, C. Ré // International Conference on Learning Representations (ICLR). — 2022. — arXiv:2111.00396.

13. Smith J.T.H. Simplified State Space Layers for Sequence Modeling (S5) / J.T.H. Smith, A. Warrington, S.W. Linderman // International Conference on Learning Representations (ICLR). — 2023. — arXiv:2208.04933.

14. Zhu L. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model / L. Zhu, B. Liao, Q. Zhang [et al.] // International Conference on Machine Learning (ICML). — 2024. — arXiv:2401.09417.

15. Yu W. MambaOut: Do We Really Need Mamba for Vision? / W. Yu, X. Wang. — arXiv, 2024. — arXiv:2405.07992. — URL: https://arxiv.org/abs/2405.07992 (accessed: 08.06.2026).

16. Chen Z. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks / Z. Chen, V. Badrinarayanan, C.-Y. Lee [et al.] // Proceedings of the 35th International Conference on Machine Learning (ICML). — PMLR, 2018. — Vol. 80. — P. 794–803.

17. Howard A. Searching for MobileNetV3 / A. Howard, M. Sandler, G. Chu [et al.] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). — 2019. — P. 1314–1324. — DOI: 10.1109/ICCV.2019.00140.

18. Hendrycks D. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations / D. Hendrycks, T. Dietterich // International Conference on Learning Representations (ICLR). — 2019. — arXiv:1903.12261.

19. MS1M-ArcFace Dataset // Kaggle. — URL: https://www.kaggle.com/datasets/yakhyokhuja/ms1m-arcface-dataset?resource=download (accessed: 08.06.2026).

Список литературы на английском языке / References in English

1. Hu C.-H. Toward Driver Face Recognition in the Intelligent Traffic Monitoring Systems / C.-H. Hu, Y. Zhang, F. Wu [et al.] // IEEE Transactions on Intelligent Transportation Systems. — 2020. — Vol. 21, № 12. — P. 4958–4971. — DOI: 10.1109/TITS.2019.2945923.

2. Yang L. Video-Based Driver Drowsiness Detection With Optimised Utilization of Key Facial Features / L. Yang, H. Yang, H. Wei [et al.] // IEEE Transactions on Intelligent Transportation Systems. — 2024. — Vol. 25, № 7. — P. 6938–6950. — DOI: 10.1109/TITS.2023.3346054.

3. Kulikov A.A. Evolyutsiya sistem biometricheskoy identifikatsii v obespechenii bezopasnosti zheleznodorozhnogo transporta: ot klassicheskikh konveyerov k differentsiruyemym gibridnym arkhitekturam [Evolution of biometric identification systems in ensuring railway transport security: from classical conveyors to differentiable hybrid architectures] / A.A. Kulikov // Transportnoye delo Rossii [Transport Business of Russia]. — 2026. — № 3. — P. 153–155. — EDN: DXLOUP. [in Russian]

4. Kulagin M.A. Intellektual'naya sistema analiza i prognozirovaniya narusheniy pri upravlenii podvizhnym sostavom [Intelligent system for analysis and prediction of violations in rolling stock control] : diss. ... Cand. Tech. Sciences : 05.13.06 / M.A. Kulagin. — Moscow : RUT (MIIT), 2022. [in Russian]

5. Zamyshlyaev A.M. Avtomatizatsiya protsessov kompleksnogo upravleniya tekhnicheskim soderzhaniem infrastruktury zheleznodorozhnogo transporta [Automation of processes for integrated management of technical maintenance of railway transport infrastructure] : diss. ... dr. of Tech. Sciences : 05.13.06 / A.M. Zamyshlyaev. — Moscow : NIIAS, 2014. [in Russian]

6. Bulat A. How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230 000 3D Facial Landmarks) / A. Bulat, G. Tzimiropoulos // Proceedings of the IEEE International Conference on Computer Vision (ICCV). — 2017. — P. 1021–1030. — DOI: 10.1109/ICCV.2017.116.

7. Jaderberg M. Spatial Transformer Networks / M. Jaderberg, K. Simonyan, A. Zisserman [et al.] // Advances in Neural Information Processing Systems 28 (NeurIPS 2015). — 2015. — P. 2017–2025. — arXiv:1506.02025.

8. Finnveden L. Understanding when spatial transformer networks do not support invariance, and what to do about it / L. Finnveden, Y. Jansson, T. Lindeberg // 25th International Conference on Pattern Recognition (ICPR). — Milan, Italy, 2021. — P. 3427–3434. — DOI: 10.1109/ICPR48806.2021.9412997.

9. Saadabadi M.S.E. ARoFace: Alignment Robustness to Improve Low-Quality Face Recognition / M.S.E. Saadabadi, S.R. Malakshan, A. Dabouei, N.M. Nasrabadi // Computer Vision – ECCV 2024. — Cham : Springer, 2024. — P. 308–327. — DOI: 10.1007/978-3-031-73414-4_18.

10. Gu A. Mamba: Linear-Time Sequence Modeling with Selective State Spaces / A. Gu, T. Dao. — arXiv, 2023. — arXiv:2312.00752.

11. Deng J. ArcFace: Additive Angular Margin Loss for Deep Face Recognition / J. Deng, J. Guo, N. Xue [et al.] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — 2019. — P. 4690–4699. — DOI: 10.1109/CVPR.2019.00482.

12. Gu A. Efficiently Modeling Long Sequences with Structured State Spaces (S4) / A. Gu, K. Goel, C. Ré // International Conference on Learning Representations (ICLR). — 2022. — arXiv:2111.00396.

13. Smith J.T.H. Simplified State Space Layers for Sequence Modeling (S5) / J.T.H. Smith, A. Warrington, S.W. Linderman // International Conference on Learning Representations (ICLR). — 2023. — arXiv:2208.04933.

14. Zhu L. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model / L. Zhu, B. Liao, Q. Zhang [et al.] // International Conference on Machine Learning (ICML). — 2024. — arXiv:2401.09417.

15. Yu W. MambaOut: Do We Really Need Mamba for Vision? / W. Yu, X. Wang. — arXiv, 2024. — arXiv:2405.07992. — URL: https://arxiv.org/abs/2405.07992 (accessed: 08.06.2026).

16. Chen Z. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks / Z. Chen, V. Badrinarayanan, C.-Y. Lee [et al.] // Proceedings of the 35th International Conference on Machine Learning (ICML). — PMLR, 2018. — Vol. 80. — P. 794–803.

17. Howard A. Searching for MobileNetV3 / A. Howard, M. Sandler, G. Chu [et al.] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). — 2019. — P. 1314–1324. — DOI: 10.1109/ICCV.2019.00140.

18. Hendrycks D. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations / D. Hendrycks, T. Dietterich // International Conference on Learning Representations (ICLR). — 2019. — arXiv:1903.12261.

19. MS1M-ArcFace Dataset // Kaggle. — URL: https://www.kaggle.com/datasets/yakhyokhuja/ms1m-arcface-dataset?resource=download (accessed: 08.06.2026).