DOI: 10.15514/ISPRAS-2026-38(3)-39

# Hybrid Augmented Model for Detecting Pedestrian Attention

*Z. Wang, ORCID: 0009-0004-3683-1233 <wangzhan@itmo.ru>*
*D.D. Zhdanov, ORCID: 0000-0001-7346-8155 <ddzhdanov@mail.ru>*
*A.D. Zhdanov, ORCID: 0000-0002-2569-1982 <adzhdanov@itmo.ru>*

*ITMO University,*
*Kronverkskiy Avenue, 49, Saint Petersburg, 197101, Russia.*

**Abstract.** Pedestrian safety on the road largely depends on how accurately a pedestrian perceives the traffic situation and how reliably a vehicle can assess the pedestrian's level of situational awareness. Estimating pedestrian gaze and attention in real-world street environments is therefore an important but challenging problem due to small face resolution, motion blur, occlusions, and extreme illumination variations. In previous studies, face-based gaze estimation and pupil tracking were shown to provide fine-grained attention cues under favorable visual conditions. However, these approaches remain fragile in unconstrained street scenes, where eye visibility and facial landmark detection frequently fail, leading to unreliable or missing outputs in safety-critical situations. In this paper, we present a robust hybrid framework that extends our earlier work by introducing a pipeline of pedestrian attention estimation methods with explicit error control at each stage to enable continuous gaze direction estimation under complex urban conditions. In addition to a learning-based gaze estimator, we incorporate geometric methods and a lightweight head pose estimation network to compensate for the breakdown of individual methods under noisy input data. The proposed framework integrates face detection, gaze estimation, geometric head pose computation, and learning-based head pose prediction, including the prediction of head position and orientation at the next time step. Unlike prior gaze-centric systems, neural networks in this design are not only used to improve accuracy but are elevated to ensure continuous video processing under real-world visual degradation associated with complex traffic scenarios. Extensive experiments on open datasets collected in real street environments demonstrate that the proposed method substantially enhances the success rate and accuracy of long-distance gaze recognition, thereby improving the feasibility of end-to-end assessment of pedestrian situational awareness while maintaining accuracy comparable to gaze estimation under favorable conditions.

**Keywords:** pedestrian attention; gaze estimation; head pose; autonomous driving.

**For citation:** Wang Z., Zhdanov D.D., Zhdanov A.D. Hybrid Augmented Model for Detecting Pedestrian Attention. Trudy ISP RAN/Proc. ISP RAS, vol. 38, issue 3, part 3, 2026, pp. 115-124. DOI: 10.15514/ISPRAS-2026-38(3)-39.





# Гибридная дополненная модель для обнаружения внимания пешехода

*Ч. Ван, ORCID: 0009-0004-3683-1233 <wangzhan@itmo.ru>*
*Д.Д. Жданов, ORCID: 0000-0001-7346-8155 <ddzhdanov@mail.ru>*
*А.Д. Жданов, ORCID: 0000-0002-2569-1982 <adzhdanov@itmo.ru>*

*Университет ИТМО,*
*Россия, 197101, Санкт-Петербург, Кронверкский проспект, д.49, литер А.*

**Аннотация.** Безопасность пешеходов на дороге во много зависит от того насколько пешеход правильно оценивает дорожную ситуацию и насколько транспортное средство в состоянии оценить уровень его контроля дорожной ситуации. Одним из средств оценки внимания пешехода может служить анализ направления его взгляда. Однако низкое разрешение изображения лица пешехода, размытие в движении, сложные условия освещения и частичное перекрытие лиц затрудняют эту оценку. В результате предыдущих исследований было показано, что в благоприятной обстановке оценка направления взгляда по ориентации лица и положению зрачков обеспечивают детальные данные о направлении внимания пешехода. Однако эти решения ненадежны в условиях реальной уличной обстановки. В данной статье представлено надежное расширенное гибридное решение, которое использует конвейер методов оценки внимания пешеходов с контролем ошибок каждого из них для непрерывного определения направления взгляда пешеходов в сложных уличных условиях. В дополнение к предложенному ранее нейросетевому методу оценки направления взгляда добавляются геометрические и легковесные нейросетевые методы оценки ориентации головы, компенсирующие потенциальные отказы методов в случаях зашумленных входных данных. Предложенное решение интегрирует детектирование лиц, оценку направления взгляда, расчет положения и ориентации головы, включая прогнозирование ее положение и ориентацию в следующий момент времени. В отличие от предыдущих решений, ориентированных исключительно на оценку направления взгляда, построенные в данной работе нейронные сети позволяют не только повысить точность, но также обеспечить непрерывную обработку видеоданных в условиях реальных визуальных помех, связанных со сложной дорожной ситуацией. Многочисленные эксперименты на открытых наборах данных в условиях реальной уличной обстановки демонстрируют, значительное превосходство предложенного решения на дальних дистанциях, что повышает доступность сквозной оценки уровня контроля дорожной ситуации при сохранении точности оценки на уровне, полученном по определению направления взгляда в благоприятных условиях.

**Ключевые слова:** внимание пешеходов; оценка направления взгляда; положение головы; автономное вождение.

**Для цитирования:** Ван Ч., Жданов Д.Д., Жданов А.Д. Гибридная дополненная модель для обнаружения взгляда пешехода. Труды ИСП РАН, том 38, вып. 3, часть 3, 2026 г., стр. 115–124 (на английском языке). DOI: 10.15514/ISPRAS–2026–38(3)–39.

## *1. Introduction*

Understanding pedestrian attention is a key requirement for autonomous driving systems operating in complex urban environments. Compared with highways, urban roads involve dense pedestrian traffic, frequent interactions, and highly unpredictable behaviors. In such scenarios, recognizing whether a pedestrian is aware of an approaching vehicle–and whether an imminent crossing action is likely – can substantially improve planning safety and reaction timing.

Traditional pedestrian perception pipelines primarily rely on full-body detection and motion-based cues. While effective for coarse localization, these methods often generate large regions of interest (ROI), increasing computational overhead and diluting intention-related information. Moreover, full-body pose estimation provides only indirect evidence of attention and suffers from large angular errors, which limits its effectiveness for fine-grained intention prediction, especially at long distances.

In our previous work, we investigated face-based and eye-based perception as a more direct indicator of pedestrian attention. In our early research, we explored gaze-based pedestrian analysis as a more direct indicator of attention. We proposed a facial perception-based processing pipeline that uses a YOLOv8s-face detector trained on WIDER_FACE to narrow down the ROI [1], enabling gaze estimation for pedestrians at a distance. Subsequently, we developed an appearance-based pupil and gaze tracking model built on EfficientNet-B0 [2], which achieved low angular error and high real-time performance under controlled and near-field conditions.

Despite these advances, real-world deployment revealed a critical limitation shared by most gaze-centric systems: gaze estimation is inherently brittle in unconstrained street scenes. Eye regions are frequently degraded by motion blur, occlusions, sunglasses, extreme illumination, and small face size. In such cases, both eye-based gaze estimation and facial landmark–based geometric head pose estimation may fail entirely, resulting in missing outputs. From a system perspective, such silent failures are unacceptable for safety-critical autonomous driving applications.

Motivated by this limitation, this paper presents a paradigm shift from gaze-centric pipelines toward a failure-aware, learning-centered perception framework. In addition to gaze estimation, we introduce a learning-based head pose estimation network as a complementary neural component. Unlike geometric head pose methods that depend on reliable landmark detection, the proposed head pose network directly operates on face crops and is trained with weak supervision. By integrating gaze estimation, geometric head pose computation, and neural head pose prediction within a reliability-gated pipeline, the proposed framework guarantees continuous directional output for every detected face. This design represents a natural extension of our earlier work toward a robust, deployable pedestrian attention perception system.

## *2. Related Work*

Pedestrian intention analysis in autonomous driving spans multiple research areas, including object detection, gaze estimation, head pose estimation, and hybrid perception systems [3].

Early pedestrian perception methods focused on full-body detection and motion analysis, leveraging handcrafted features or deep convolutional models. With the advent of single-stage detectors such as YOLO, real-time pedestrian detection has become feasible even in dense urban scenes. Recent studies have shown that narrowing the perception focus to facial regions can significantly reduce ROI size and computational overhead, particularly for intention-related analysis. Training face detectors on large-scale datasets such as WIDER_FACE enables robust detection under extreme variations in scale, pose, and occlusion.

Gaze estimation has evolved rapidly with the availability of large-scale datasets such as MPIIGaze and Gaze360 [4]-[5]. Appearance-based methods that directly regress gaze direction from eye images have become dominant due to their flexibility under varying lighting and head poses. In our prior pupil tracking work, we demonstrated that lightweight attention mechanisms and distance-simulation modules can substantially improve robustness for low-resolution eye regions. Nevertheless, most gaze estimation approaches implicitly assume reliable eye visibility and degrade silently when this assumption is violated.

Head pose estimation is commonly addressed either through geometric formulations based on facial landmarks and Perspective-n-Point optimization or through learning-based regression models [6]. While head orientation does not always coincide with eye gaze, it provides a stable proxy when eye information is unreliable. Several works have explored combining gaze and head pose cues; however, most treat head pose as an auxiliary signal rather than a functional fallback.

Hybrid perception frameworks are increasingly adopted in autonomous driving, particularly in the context of multi-sensor fusion [7]. In contrast, fewer studies focus on semantic hybridization between gaze and pose estimation. The present work differs from existing approaches by explicitly elevating neural networks to a robustness-driven role, using learning-based estimators to compensate





for geometric and appearance-based failures. This failure-aware design directly addresses limitations observed in our previous gaze-centric systems.

## *3. Method*

### 3.1 Problem Formulation

Given an input image $I$, the objective of the proposed framework is to detect all visible faces and output a reliable direction vector for each detected face. The direction represents either the estimated gaze direction or a head-orientation surrogate when gaze evidence is unreliable. The system outputs

$$\mathcal{O} = \{(b_i, \boldsymbol{d}_i)\}_{i=1}^{N}$$

where $N$ denotes the number of detected faces. Due to severe challenges in real-world street imagery – including small face size, motion blur, occlusions, and extreme illumination–direct gaze estimation is often unreliable. Therefore, we explicitly design the output direction $\boldsymbol{d}_i$ to be adaptively selected between gaze and head pose estimates according to their reliability, as shown in Fig. 1.

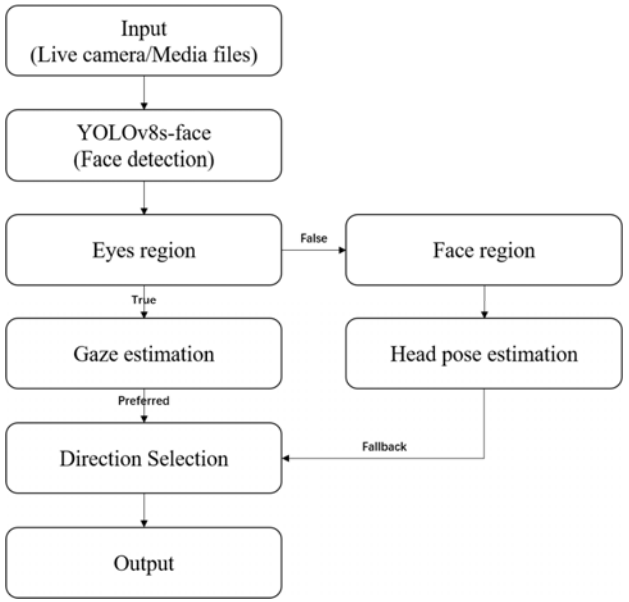


*Рис. 1. Блок-схема гибридной модели.*
*Fig. 1. Hybrid Model Workflow.*

### 3.2 Overview of the Hybrid Pipeline

Our approach follows a hybrid analysis paradigm that integrates detection, geometric reasoning, and learning-based estimation in a unified pipeline. First, a one-class face detector $F_\theta$ based on YOLO is applied to the input image to produce a set of face bounding boxes:

$$\{b_i, s_i\} = F_\theta(I)$$

where $s_i$ denotes the detection confidence. Each detected face region is then processed independently by two complementary branches: a gaze estimation branch and a head pose estimation branch. A reliability gate determines which branch provides the final output direction.

This hybrid design is motivated by the observation that gaze estimation is more semantically meaningful but brittle, whereas head pose estimation is geometrically stable but less precise in reflecting actual eye attention. By combining both, the proposed system achieves robust end-to-end direction estimation in the wild.

### 3.3 Gaze Estimation Branch

For each detected face, the system attempts gaze estimation as the primary branch. When facial landmarks are available, eye regions are localized, normalized, and fed into a learning-based gaze estimator $G_\phi$ built upon an EfficientNet-B0 backbone. The network regresses a three-dimensional gaze vector:

$$\hat{\boldsymbol{g}}_i = \frac{G_\phi\big(eye(I, b_i)\big)}{\|G_\phi(\cdot)\|}$$

The gaze estimator is trained on the MPIIGaze dataset following the appearance-based paradigm established in our previous work, with additional robustness mechanisms for low-resolution eye regions.

Despite its semantic relevance, this branch is sensitive to landmark fitting quality and image degradation. In practice, failure cases frequently arise when eye regions are heavily blurred, occluded, or extremely small, which motivates the introduction of a complementary head pose estimation mechanism.

### 3.4 Head Pose Estimation and Pose-Based Fallback

When gaze estimation becomes unreliable or unavailable, the system falls back to head pose estimation. Two complementary head pose estimators are employed.

If facial landmarks are successfully detected, a geometric head pose is computed using a Perspective-n-Point formulation. The resulting rotation is projected onto the image plane to produce a two-dimensional head orientation vector $\hat{\boldsymbol{h}}_i^{geo}$. This method provides accurate pose estimates when landmark detection is reliable but degrades sharply under occlusion or severe blur.

To address landmark failure cases, we introduce a lightweight learning-based head pose estimator $H_\psi$. This network directly operates on the face crop and predicts discretized yaw and pitch angles, which are subsequently mapped to a direction vector $\hat{\boldsymbol{h}}_i^{net}$. The head pose network is trained using weak supervision: pseudo labels are generated by applying the geometric head pose estimator to reliably detected faces in the WIDER_FACE dataset. This strategy enables the network to learn head orientation cues without requiring additional manual annotations.

Importantly, the learning-based head pose estimator is not designed as an auxiliary refinement module. Instead, it serves as a functional replacement when geometric methods fail, ensuring that the perception pipeline does not collapse due to missing landmarks.

### 3.5 Reliability-Gated Direction Selection

The final output direction is determined by a reliability-gated mechanism:

$$\boldsymbol{d}_i = \begin{cases} \hat{\boldsymbol{g}}_i, & if\ r(\hat{\boldsymbol{g}}_i) \geq \tau, \\ \hat{\boldsymbol{h}}_i^{geo}, & if\ landmarks\ available, \\ \hat{\boldsymbol{h}}_i^{net}, & otherwise, \end{cases}$$

where $r(\cdot)$ denotes a confidence measure derived from eye-region validity and landmark completeness. This hierarchical selection strategy ensures that gaze estimation is used whenever reliable, geometric head pose is preferred when landmarks are available, and learning-based head pose estimation guarantees a valid output under extreme degradation.





Unlike conventional systems where neural networks primarily serve accuracy optimization, the proposed framework assigns neural estimators a system-level responsibility: maintaining operational continuity under failure.

### 3.6 Training Strategy

The proposed system is trained in a multi-stage manner. The face detector is trained with strong supervision on the WIDER_FACE dataset, focusing on high recall under small-scale and occluded faces [8]. The gaze estimator is trained using MPIIGaze with supervised gaze annotations and auxiliary pupil localization when available. Finally, the learning-based head pose estimator is trained on WIDER_FACE using pseudo labels obtained from geometric pose estimation, enabling robust head orientation prediction even in the absence of facial landmarks.

## *4. Experiments*

Given that this study intends to tackle the challenges of autonomous driving under complex road conditions, distance limitations have been imposed on data collection in accordance with the speed limits of actual urban roads. For a typical family car equipped with an Anti-lock Braking System (ABS) on dry asphalt or concrete roads, the typical deceleration achievable during full braking is approximately 7 to 9 m/s². We adopt the mid-point value of 8 m/s² for our computations. Taking the speed limit of 60 km/h in most towns as an instance, the distance required to decelerate from 60 km/h to 0 km/h is:

$$s = \frac{v_0^2}{2|a|},$$

$$s = \frac{(\frac{60}{3.6})^2}{2 \times 8} \approx 17.36\ meters$$

Considering the practical application of this part, we used a long focal length camera to collect street view materials of 15 meters or more for verification.

Furthermore, we conduct a comparison and evaluation between the original hybrid script and an enhanced version that integrates head pose and face determination on 101 real-world street view and crowd gathering photos. When comparing the Hybird system without the incorporation of head pose constraints with the Kingstand system after the integration of the HeadPose module using the same street scene data, a notable performance disparity can be discerned, as shown in Table. 1.

*Table 1. Experimental comparison results.*

| Model | Fps | vr | VFR | TTFD | lat mean | p50 |
|---|---|---|---|---|---|---|
| Hybird | 2.29 | 0.950 | 2.17 | 450.0 | 294.8 | 342.2 |
| Kingstand | 10.39 | 0.990 | 10.29 | 264.5 | 118.5 | 126.5 |

The average processing frame rate of the Hybird system on this dataset is 2.29 frames per second (fps), and the Valid Frame Rate (VFR) is 2.17. When the valid ratio is 0.950, the overall processing delay is substantial, with an average delay of 294.8 milliseconds (ms). Moreover, there remains a notable tail delay in the high quantiles (407.7 ms at the 90th percentile, 438.0 ms at the 99th percentile). The Time to First Detection (TTFD) is 450.0 ms, suggesting that the system requires a considerable initialization and calculation time before reaching a stable operation state in a multi-target street scene.

In comparison, the Kingstand system exhibits a substantial improvement in overall performance on the same dataset. The average frame rate has increased to 10.39 fps, with the corresponding VFR reaching 10.29 – approximately 4.5 times higher than that of the original script. Concurrently, the valid detection ratio has improved to 0.990, indicating that detection accuracy has been preserved despite the performance acceleration. System latency characteristics have also been significantly

enhanced: the average delay has decreased from 294.8 ms to 118.5 ms, the p90 latency has reduced from 407.7 ms to 159.6 ms, and the p99 latency has declined from 438.0 ms to 255.4 ms, reflecting a more consistent and reliable real-time response in complex urban street environments. Notably, the TTFD has been reduced from 450.0 ms to 264.5 ms, suggesting that the integration of head pose estimation and frontal view determination mechanisms not only enhances steady-state efficiency but also enables faster achievement of effective detection. The experimental results are shown in Fig. 2 and 3.


*Fig. 2 and 3. Visualization of results.*

It should be noted that we have added a compensation mechanism to the algorithm, which determines the direction of gaze by comparing the black pixels of the pupil with the white pixels of the sclera when the human eye region can be completely recognized. As shown in the figure, because the human sclera is not completely symmetrical, this compensation mechanism may sometimes lead





to a small probability of misjudgment. In general, we have achieved the function of vision detection at a medium and long distance and greatly improved the distance of vision detection.

Furthermore, approximately 75.2% of the faces in the dataset were classified as nearly frontal (frontal% = 0.752). In such cases, the system outputs a neutral label rather than generating an enforced directional estimate, thereby avoiding unreliable predictions when gaze cues are insufficient. This finding demonstrates that by incorporating a head-pose-based frontal detection mechanism, the system can substantially reduce unnecessary computational overhead in large-scale street view analysis, allowing resources to be allocated more efficiently to samples with clearer pose or gaze signals. As a result, an improved balance among processing speed, system stability, and output reliability is achieved.

## 5. Discussion

The introduction of a learning-based head pose estimator fundamentally changes the role of neural networks in the perception pipeline. In our earlier studies, neural networks were primarily employed to improve gaze estimation accuracy under favorable conditions. In contrast, the present framework leverages multiple neural networks to explicitly absorb uncertainty and compensate for structural failure at different perceptual stages. The model replaces geometric reasoning when landmark detection collapses, a scenario frequently encountered in real-world street imagery. By training the model with weak supervision derived from geometric estimates, the system bridges the gap between data-driven learning and classical geometry. This design enables graceful degradation: as visual evidence deteriorates, the system transitions from gaze estimation to geometric head pose, and finally to neural head pose prediction, without producing missing outputs.

Importantly, the introduction of an explicit frontal state further refines this degradation process by allowing the system to abstain from unreliable directional inference when head orientation remains near frontal. Rather than forcing uncertain predictions, the system outputs a neutral representation, which both improves perceptual reliability and significantly reduces unnecessary computational overhead. This behavior is reflected in the substantial gains in processing speed and latency stability observed in large-scale street experiments.

From a system design perspective, this work demonstrates that robustness in autonomous driving perception cannot be achieved by accuracy-driven optimization alone. Instead, neural networks must be integrated as functional components responsible for sustaining system availability, regulating decision confidence, and preserving temporal stability under adverse and information-degraded conditions.

## 6. Conclusion

This paper presents a robust hybrid framework for pedestrian direction analysis in autonomous driving scenarios. Building upon our prior work on face-based gaze detection and pupil tracking, we introduce a failure-aware perception pipeline that integrates gaze estimation, geometric head pose computation, and a weakly supervised learning-based head pose estimator. By elevating neural networks to system-level components responsible for both accuracy and robustness, the proposed approach significantly improves end-to-end direction availability while preserving gaze precision whenever reliable eye evidence exists.

Furthermore, by explicitly modeling near-frontal configurations as a distinct perceptual state, the system avoids unreliable direction inference in the absence of sufficient visual cues, leading to substantial improvements in processing efficiency and temporal stability in large-scale street environments. The results demonstrate that explicitly modeling failure and incorporating learning-based head pose estimation are critical steps toward deployable pedestrian attention analysis in complex urban environments.

## References

[1]. Wang Z., Zhdanov D.D., Zhdanov A.D. Research on pupil tracking techniques for autonomous driving systems. Proceedings of GraphiCon 2025, 2025, pp. 215-226 DOI: 10.25686/978-5-8158-2474-4-2025-215-226.

[2]. Wang Z., Zhdanov D., Zhdanov A. Research on pedestrian gaze detection in autonomous driving systems. Optoelectronic Imaging and Multimedia Technology XII. SPIE, 2025, vol. 13718, pp. 289-296. DOI: 10.1117/12.3074848.

[3]. Zhou H., Laval J., Zhou A., Wang Y., Wu W., Qing Z., Peeta S. Review of learning-based longitudinal motion planning for autonomous vehicles: research gaps between self-driving and traffic congestion. Transportation research record, 2022, 2676(1), pp. 324-341. DOI: 10.1177/03611981211035764.

[4]. Kellnhofer P., Recasens A., Stent S., Matusik W., Torralba A. Gaze360: Physically unconstrained gaze estimation in the wild. Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6912-6921. DOI: 10.1109/ICCV.2019.00701.

[5]. Zhang X., Sugano Y., Fritz M., Bulling A. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE transactions on pattern analysis and machine intelligence, 2017, 41(1), pp. 162-175. DOI: 10.1109/TPAMI.2017.2778103.

[6]. Rocca F, Mancas M, Gosselin B. Head pose estimation by perspective-n-point solution based on 2d markerless face tracking. International Conference on Intelligent Technologies for Interactive Entertainment. Cham: Springer International Publishing, 2014, pp. 67-76. DOI: 10.1007/978-3-319-08189-2_8.

[7]. Carrillo D., Nutt M., Meijer M., Khan J., Fu S., Yang Q. Bringing Different Views Together: A Hybrid Cooperative Perception Framework for Connected Autonomous Vehicles. IEEE Network, 2025. DOI: 10.1109/MNET.2025.3546821.

[8]. Yang S., Luo P., Loy C. C., Tang X. Wider face: A face detection benchmark. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5525-5533. DOI: 10.1109/CVPR.2016.596

### *Информация об авторах / Information about authors*

Чжань ВАН – аспирант факультета программной инженерии и компьютерной техники университета ИТМО. Сфера научных интересов: разработка компьютерных систем обеспечения безопасности дорожного движения в условиях ограниченной видимости и оценки поведения пешеходов.

Zhan WANG – postgraduate student at the Faculty of Software Engineering and Computer Systems of ITMO University. Research interests: development of computer systems for ensuring road safety in conditions of limited visibility and assessment of pedestrian behavior.

Дмитрий Дмитриевич ЖДАНОВ – кандидат физико-математических наук, доцент факультета программной инженерии и компьютерной техники университета ИТМО. Сфера научных интересов: Компьютерная графика, компьютерное зрение, вычислительная оптика, виртуальная и дополненная реальности, параллельные и распределенные вычисления.

Dmitrii Dmitrievich ZHDAHOV – Cand. Sci. (Phys.-Math.), Associate Professor at the Faculty of Software Engineering and Computer Systems of ITMO University. Research interests: Computer graphics, computer vision, computational optics, virtual and augmented reality, parallel and distributed computing.

Андрей Дмитриевич ЖДАНОВ – кандидат технических наук, доцент факультета программной инженерии и компьютерной техники университета ИТМО. Сфера научных интересов: компьютерная графика, компьютерное зрение, нейросетевые технологии, параллельные и распределенные вычисления.

Andrei Dmitrievich ZHDAHOV – Cand. Sci. (Tech.), Associate Professor at the Faculty of Software Engineering and Computer Systems of ITMO University. Research interests: computer graphics, computer vision, neural network technologies, parallel and distributed computing.