Journal of Applied Automation Technologies

Research Paper

# High-Resolution Multi-Level Feature Fusion Network for Robust Facial Landmark Detection

**Julia Magdalena Kozłowska[1,*], Dawid Jerzy Chmielewski[1] and Patryk Waldemar Mazurek[1]**

[1] Faculty of Computer Science and Management, AGH University of Science and Technology, 30-059 Krakow, Poland
*Corresponding author: julia.mk@student.agh.edu.pl

**Abstract.** Facial keypoint detection is relatively easy, but not completely accurate. In practical applications, occlusion, pose variations, and image noise are obstacles to accuracy. To achieve more accurate and stable keypoint detection, this paper will address the aforementioned issues by constructing a high-resolution neural network. The network uses a multi-layer feature fusion structure to extract global facial features and local details at different scales. In order to balance the trade-off between overall smoothness and high-resolution accuracy in model optimization, a new combined loss function was adopted. It will be more robust to blurred and occluded samples. 300-W, COFW, and WFLW are public data experimental datasets. The normalized mean error (NME) on 300-W is 2.89, on COFW is 3.56, and on WFLW is 3.19, all three methods have set new records. Common interferences such as blur and noise have been better resisted. Ablation experiments demonstrated the fusion strategy and loss components. Qualitative visualizations and application cases indicate that the system can perform well in various environments and under different conditions. These two methods are both practical and reliable, and can be used in many fields, such as mobile health monitoring, biometric authentication, and human-computer interaction.

**Keywords:** *Pattern Recognition, Facial Landmark Detection, Multi-Level Feature Fusion, High-Resolution Network, Robustness, Deep Learning*

Received on 16 October 2025, Accepted on 23 March 2026, Published on 01 April 2026



## Introduction

The facial keypoint detection technology in modern computer vision has rapidly become an important component of computer vision. Subsequent applications based on this technology include augmented reality, 3D facial modeling, emotion analysis, facial recognition, and medical diagnosis [1]. Eyes, nose, mouth, jawline, and other important facial locations are designated as reference points. The algorithm can accurately and real-time identify facial structures [2]. The development of Convolutional Neural Networks (CNNs) has made significant progress over the past decade, making landmark localization more accurate and stable [3]. These improvements include applications in various constrained and semi-constrained environments. As digital platforms become more diverse and interaction patterns more complex, the demand for facial keypoint detection is also increasing [4]. In the fields of consumer electronics, entertainment, access control, and telemedicine, landmark detection is now used to create new interaction methods and is also employed for reliability and security [5]. These advancements are often affected by some issues present in real-world facial images [6]. These issues include uneven lighting, severe occlusion, various angles of the head and body, and natural differences between faces. Due to the increasing demand for high-fidelity and low-latency inference on mobile and embedded devices, more research is needed on model efficiency and the generalization of solutions [7].

Research and datasets for facial keypoint detection still have numerous issues. Early cascade regression methods and deep network architectures performed well under controlled conditions, but often underperformed on real-world data [8]. Several major technical issues with traditional backbone structures are as follows: fine-grained structural relationship encoding, a small number of nodes, and multiple depths in the network retaining high-

resolution spatial information [9]. Hourglass networks and stacked architectures can help extract multi-scale features, but they often fail to retain contextual information about subtle landmarks in cases of severe occlusion or significant pose changes [10]. It has been recognized that jointly preserving and effectively fusing high-resolution spatial cues and multi-level data is crucial for enhancing the robustness and accuracy of landmark localization in heterogeneous facial instances.

This paper proposes a new method to construct an optimized high-resolution network for detecting facial keypoints, aiming to overcome the aforementioned shortcomings. The new design aims to obtain a large amount of semantic information and retain spatial positions during the inference process through a multi-branch structure. It is a general multi-level feature fusion module that can adaptively combine based on the contextual information of each level of the network. Using advanced loss functions to improve the accuracy of overall facial alignment and local keypoints. Conduct comprehensive experiments, perform extensive quantitative benchmarking, and carry out in-depth qualitative analysis of the results using multiple complex public datasets. High-resolution network strategies can provide a solid foundation for future unconstrained, real-world facial analysis.

## Related Work

### Deep Learning-Based Facial Landmark Detection

Facial landmark detection has undergone many changes with the development of deep learning. Now it uses end-to-end feature learning methods, rather than the previous regression and graphical model methods. The application of CNNs in constructing hierarchical feature maps has improved the accuracy of localization and robustness to noise [11]. The cascade convolutional regressor is an early deep learning method that gradually improved landmark estimation and successfully reduced localization errors in constrained and semi-constrained datasets [12]. Due to the incorporation of deep architectures, such as the VGG and ResNet series, multi-scale spatial feature extraction can be achieved, which improves the accuracy of landmark localization [13]. The next generation improves pose variation, illumination invariance, and partial occlusion detection by using the aforementioned backbone networks.

The advanced model of the stacked hourglass network collects fine-grained local details and large-scale context by repeatedly downsampling and upsampling feature maps [14]. The hourglass method divides the data into multiple scales and is relatively robust to large-scale distortions of landmarks. Some frameworks have recently incorporated attention mechanisms, spatial transformers, and ensemble methods to focus on primary facial information while ignoring other unnecessary details [15]. With the development of large-scale, richly annotated facial datasets such as 300-W, AFLW, and WFLW, many new benchmarks and evaluation methods have emerged. In addition, more model architectures and strong regularization strategies have also been introduced [16]. A long-standing issue: achieving high accuracy under adverse conditions (such as severe poses, dense occlusions, and motion blur) is impossible; traditional CNN-based structures often fail in these situations. In order to fully utilize the large amount of unlabeled data, current research has begun using hybrid models, incorporating graphical priors or 3D deformable constraints into trainable networks, and employing self-supervised or semi-supervised training protocols [17].

The sensitivity of deep learning models to complex backgrounds, diverse demographics, and harsh capture environments remains low, despite some progress in recent years. Research on new architectures and loss functions is underway, aiming to improve the robustness, efficiency, and interpretability of current systems. These studies also aim to help the system adapt to new environments and reduce the difficulty of data annotation [18].

### Multi-Scale Feature Extraction Methods

Due to the drawbacks of classic convolutional neural networks (CNNs), such as the reduction in spatial resolution at deeper layers, multi-scale feature extraction techniques have been proposed. Feature Pyramid Networks (FPN) are such an example. Combining high-resolution localization signals with low-resolution semantic information to create a feature hierarchy. The aforementioned structure can effectively identify and distinguish objects of various sizes [19]. U-Net was initially used for medical image segmentation. Introducing direct skip connections

and a symmetric encoder-decoder structure helps to recover high-frequency details lost during downsampling [20]. FPN and U-Net are the foundation of landmark detection models, introduced to address the loss of context and details in small landmarks.

The Hourglass Network expanded on this concept by using recursive downsampling and upsampling, as well as multi-scale symmetric data aggregation [21]. Iterations can handle various facial shapes and expressions. Traditional hourglass networks perform poorly and become easily confused when small facial details are occluded or lost [22]. Attempt to use parallel multi-branch representation techniques to simultaneously support multiple resolution streams on the network [23].

HRNet is an innovation in multi-scale learning. HRNet does not use an encoder-decoder structure or pyramids, but instead continuously exchanges information through a high-resolution stream and parallel low-resolution streams. It can provide precise descriptors for dense landmark prediction while preserving spatial data at each stage [24]. The difficulties of multi-branch fusion and how to balance cost and accuracy in practice remain a focal point of concern [25]. Research focuses on enhancing feature fusion (such as task-adaptive schemes, weighting, or attention guidance) and reducing model size to improve the model's ability to generalize across different imaging environments [26].

**HRNet in Vision Applications**

Semantic segmentation, object recognition, human pose estimation, and fine-grained image analysis are recent applications of High-Resolution Network (HRNet) in many visual tasks [27]. HRNet is a multi-scale parallel architecture that provides rich semantic context, addressing the shortcomings of traditional CNN backbone networks in facial keypoint detection. Deliberately passing information between high to low-resolution streams at each stage to ensure contextual awareness and precise localization, which is crucial for dense facial keypoint extraction.

HRNet has been used in multi-task facial analysis pipelines, such as alignment, expression recognition, and 3D facial reconstruction, proving to be more effective in maintaining structural integrity under unstable visual conditions [28]. HRNet performs well on complex public datasets and exhibits strong robustness to occlusion, pose variations, and differences across diverse populations [29]. HRNet always maintains high-resolution representations, thus outperforming hourglass and pyramid structures in human pose estimation. HRNet is capable of maintaining pixel-level label fine structure accuracy and semantic segmentation capabilities [30].

HRNet has these advantages, but its model size is very large, and the cost is also very high. Many parallel high-resolution paths are economically unfeasible in embedded or real-time systems. Research has been conducted on knowledge distillation methods, lightweight HRNet variants, and hardware-aware optimizations for mobile deployment. Due to issues of joint and scalability, research has been conducted on attention-guided and dynamic fusion modules. These modules allow HRNet-based systems to dynamically adjust the information flow according to various tasks and scenarios. With ongoing research, the application of HRNet has been extended to other fields and cross-modal issues, while still retaining its core advantages—balanced representation and spatial accuracy.

## Proposed Methodology

**Network Architecture Optimization**

All other modules rely on a specially designed high-resolution network architecture to address the specific challenges of high-precision facial keypoint localization. This innovative structure maintains spatial accuracy and semantic abstraction through the deep integration of multiple high-resolution parallel streams, spatially adaptive convolutions, and hierarchical context propagation at various levels.

At the beginning of the architecture design, multi-scale features are calculated simultaneously at each network stage. Unlike traditional encoder-decoder architectures, the deep layers do not reduce the spatial granularity of the structure while maintaining high-resolution flow. Let $\mathbf{X}_l$ denote the feature tensor at stage $l$ with spatial resolution $r_l$. The main branch is initialized as:

$$\mathbf{X}_1 = \mathrm{Conv}_{3\times3}(\mathbf{I}) + \mathrm{BN} + \mathrm{ReLU} \quad \text{Eq.(1)}$$

https://doi.org/10.64972/jaat.2026v4.137p21e:276-289

where $\mathbf{I}$ represents the normalized input and BN denotes batch normalization. Each subsequent branch at level $l$ is constructed by adaptively sampling and aggregating the outputs of prior branches via a dynamic filter bank:

$$\mathbf{X}_l = \phi_l\left(\sum_{k=1}^{l-1} \alpha_{l,k} \cdot \mathrm{Down}_{r_k \to r_l}(\mathbf{X}_k) + \beta_{l,k} \cdot \mathrm{Up}_{r_k \to r_l}(\mathbf{X}_k)\right) \quad \text{Eq.(2)}$$

where $\phi_l$ implements a non-linear transformation specific to each resolution, and $\alpha_{l,k}, \beta_{l,k}$ are data-adaptive weighting factors for cross-resolution re-embedding.

At all transition points, the cross-flow modulation module aligns feature statistics through channel weighting and multi-head attention gates to ensure information consistency and reduce semantic discontinuity. Ensure that the retention of high-resolution representations is explicit:

$$\mathbf{Y}_L = \mathcal{G}([\mathbf{X}_1^L, \mathbf{X}_2^L, \ldots, \mathbf{X}_L^L]) \quad \text{Eq.(3)}$$

where $\mathcal{G}$ denotes a hierarchical fusion operator, concatenating or summing the processed streams at the final stage $L$, thereby generating dense, structurally consistent outputs.

The resultant output tensor is projected through a dedicated prediction head that simultaneously regresses coordinates for $K$ facial landmarks. Explicitly, the output module maps the fused representation to landmark heatmaps using:

$$\hat{\mathbf{H}}_k = \sigma\big(\mathrm{Conv}_{1\times1}(\mathbf{Y}_L)\big), k = 1,2,\ldots,K \quad \text{Eq.(4)}$$

where $\sigma$ is a sigmoid activation tailored for spatial response normalization. This design enables pixel-level precision even under severe pose or illumination variation.

Figure 1 shows these architectural customizations to provide an overview of parallel multi-stream flow, resolution-specific fusion points, and the prediction process.

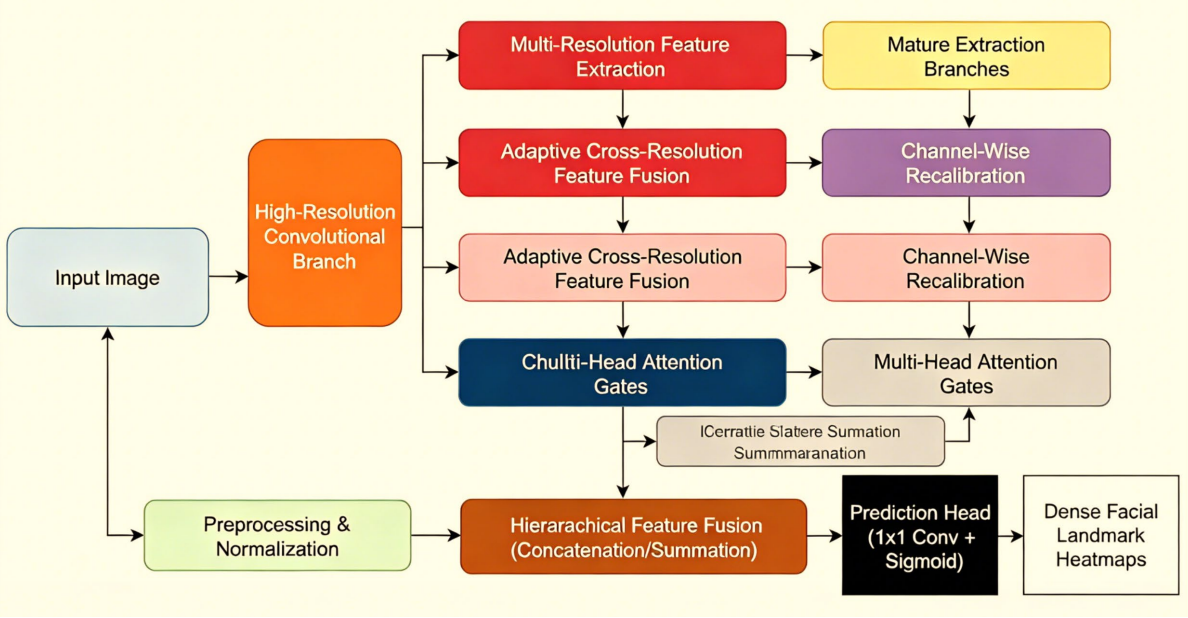


**Figure 1.** Customized high-resolution network architecture for facial landmark detection

### Multi-Level Feature Fusion

By gradually combining high-level semantic information and low-level spatial details, the first advancement of the method is a clear multi-level feature fusion design. In order to retain the original features and add more context, deep convolutions and cross-branch representations are used to adaptively combine feature maps at each resolution. To optimize landmark localization of complex facial structures in unconstrained scenarios, this design ensures uniform gradient and information flow across all depths of the network.

Let $\mathbf{F}_l^s$ denote the feature map at stage $l$ and stream $s$, each operating at a distinct spatial resolution. To capture multi-scale dependencies, intermediate features from all streams are jointly aggregated using an attention-weighted fusion function:

$$\mathbf{Z}_l = \Psi\left(\sum_{s=1}^{S} w_{l,s} \cdot \gamma_{l,s}(\mathbf{F}_l^s)\right) \quad \text{Eq.(5)}$$

where $w_{l,s}$ represents dynamically learned importance weights and $\gamma_{l,s}$ employs a channel-wise recalibration for each stream, while $\Psi$ denotes a non-linear activation-enhanced fusion operator.

The cross-flow residual gate addresses the issue of semantic drift in high-resolution and low-resolution images by controlling the exchange of information:

$$\tilde{\mathbf{F}}_l^s = \mathbf{F}_l^s + \rho_{l,s}(\mathbf{F}_l^s, \mathbf{Z}_l) \quad \text{Eq.(6)}$$

with $\rho_{l,s}$ parameterizing an attention-guided transformation that corrects for alignment errors and localizes landmarks consistently.

Fusion cascade uses spatially selective aggregation to locally amplify important facial areas:

$$\mathbf{A}_l = \sum_{i,j} \lambda_{l,i,j} \odot \mathbf{Z}_l^{(i,j)} \quad \text{Eq.(7)}$$

where $\lambda_{l,i,j}$ is a spatial attention tensor, and $\odot$ denotes pointwise multiplication, focusing model capacity to salient facial areas such as eyes, mouth corners, and nasal bridge.

The recursive dependency chain represents the multi-level integration of all stages:

$$\mathbf{O}_l = \Theta\left(\left\{\mathbf{A}_{l'}, \tilde{\mathbf{F}}_{l'}^s \mid \forall l' \le l, \forall s\right\}\right) \quad \text{Eq.(8)}$$

where $\Theta$ is a context-sensitive compositional operator that hierarchically integrates aggregated features and corrected residuals, ensuring that each level benefits from all previous multi-scale computations.

Fusion representation extracts the global context vector:

$$\mathbf{g}_l = \tanh\left(\text{GlobalAvgPool}(\mathbf{O}_l) + \eta_l\right) \quad \text{Eq.(9)}$$

where $\eta_l$ serves as a learned bias, providing a normalization anchor and stabilizing feature statistics across minibatches.

Figure 2 shows the multi-layer fusion architecture and local enhancement flow. The network successfully integrates global and regional cues to perform facial keypoint inference.

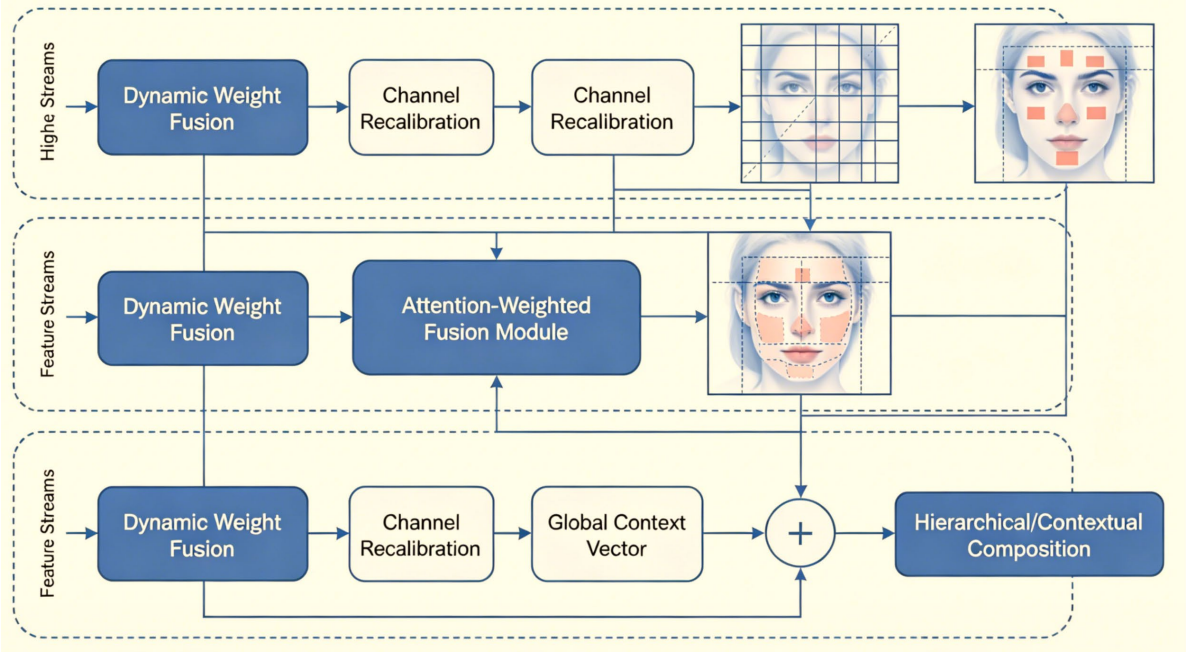


**Figure 2.** Multi-level feature fusion architecture for facial landmark detection

**Loss Function Design and Training Strategy**

Choosing the appropriate loss function and training method is to ensure the accuracy and stability of facial keypoint detection in different environments. Composite loss and adaptive learning pipelines are used for accurate and stable heatmap regression. In the old model, not every keypoint and pixel has the same weight. By weighting and considering the contextual structure, dynamically adjust the focus of learning.

The primary loss supervises predicted heatmaps $\hat{\mathbf{H}}_k$ of each landmark $k$, enforcing spatial alignment to ground truth $\mathbf{H}_k$ through an adaptive error weighting mechanism:

$$\mathcal{L}_{\text{pt}} = \frac{1}{K}\sum_{k=1}^{K}\sum_{i,j}\omega_k^{(i,j)}\left|\mathbf{H}_k^{(i,j)} - \hat{\mathbf{H}}_k^{(i,j)}\right| \quad \text{Eq.(10)}$$

Here, $\omega_k^{(i,j)}$ is a dynamically updated mask, emphasizing ambiguous or occluded pixels and correcting for local appearance variations.

Due to the low sensitivity of landmark positions, a stable penalty function is used to offset small errors. The "wing" penalty for regularization is:

$$\mathcal{L}_{\text{wing}} = \frac{1}{K}\sum_{k=1}^{K}\sum_{i,j}\xi_k^{(i,j)}\ln\left(1 + \frac{\left|\mathbf{H}_k^{(i,j)} - \hat{\mathbf{H}}_k^{(i,j)}\right|}{\epsilon}\right) \quad \text{Eq.(11)}$$

where $\xi_k^{(i,j)}$ adjusts the penalty scale for each pixel and landmark, while $\epsilon$ controls error compression.

In order to improve spatial smoothness and reduce checkerboard artifacts, a Laplacian regularization term is added to the predicted heatmap:

$$\mathcal{L}_{\text{smooth}} = \lambda\sum_{k=1}^{K}\sum_{i,j}\left|\nabla^2\hat{\mathbf{H}}_k^{(i,j)}\right|^2 \quad \text{Eq.(12)}$$

where $\nabla^2$ is the discrete Laplacian operator and $\lambda$ regulates smoothing strength.

The total loss used to update the network parameters is a weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pt}} + \delta_1\mathcal{L}_{\text{wing}} + \delta_2\mathcal{L}_{\text{smooth}} \quad \text{Eq.(13)}$$

with $\delta_1, \delta_2$ determined by validation-driven scheduling.

Optimization is performed using Adam with adaptive moment tracking, where the parameter update at iteration $t$ is:

$$\phi^{(t+1)} = \phi^{(t)} - \eta\cdot\frac{m^{(t)}}{\sqrt{v^{(t)}} + \epsilon_0} \quad \text{Eq.(14)}$$

Here, $\phi$ denotes model weights, $m^{(t)}$ and $v^{(t)}$ represent first and second moment estimates, and $\eta$ and $\epsilon_0$ are learning parameters.

During the training process, a curriculum-based scheduler was used to control sampling and loss distribution. Reliable annotated facial regions are prioritized, and then over time, more difficult or occluded samples are added. Use stochastic weight averaging to enhance generalization ability, check if the loss is converging, and stop training when the loss stabilizes. Use the first method mentioned above to achieve fast convergence and high-precision heatmaps for large-scale, unconstrained data.

## Experimental Evaluation and Results Discussion

**Datasets and Implementation Details**

To rigorously validate the proposed method, three large-scale public datasets will be used; these datasets include 300-W, COFW, and WFLW, comprehensively covering facial diversity in the real world. The 300-W dataset contains 3,837 training images and 600 challenging test images, with 68 facial landmarks annotated under various poses and lighting conditions. For occlusion analysis, COFW contains 1,345 images with dense occlusions, with 29 key points on each face manually annotated. The WFLW dataset includes up to 7,500 training

faces and 2,500 test faces, each with 98 available landmarks, covering extreme expressions, side views, and severe occlusions.

300-W uses Normalized Mean Error (NME) to measure fine-grained landmark localization accuracy under moderate pose variations, WFLW examines spatial robustness in many real-world scenarios, and COFW investigates failure modes under occlusion. The overly detailed annotations and overly complex images have led to these issues.

Stricter data partitioning protocols will be implemented. Each dataset uses the official partition, which means the training set, validation set, and test set do not overlap. All training images undergo the same preprocessing procedures. These procedures include pixel normalization, resizing the images to 256 x 256, and geometric augmentation. According to the dataset, the augmentation scheme has been adjusted. Random rotation (±30°), scaling (±20%), and horizontal flipping. To enhance generalization and robustness, simulated occlusions were also added to the baseline.

Landmark visibility and category masks further standardize label accuracy; for COFW and WFLW, blurred or occluded points are also appropriately included in the loss. By subtracting the mean of the entire dataset and dividing by the standard deviation, the input images are normalized to ensure stable optimization regardless of when they are acquired.

Using the same hardware configuration of NVIDIA RTX 3090 GPU, Intel Xeon Gold 6226 processor, and 128GB DDR4 memory for training and testing. By using CUDA and PyTorch 1.12 to accelerate the training model. All comparative methods were set with hyperparameters, a batch size of 32, an initial learning rate of 0.0007, and cosine annealing convergence over 140 epochs. Using the Adam optimizer, gradient clipping is set to 8.0, with default beta values. To help the model learn more evenly, the data loader maintains a relatively balanced distribution of different difficulty levels within mini-batches.

In order to ensure that the results of all metrics and scenarios presented in the following sections are fair and reproducible, all experimental environments, augmentations, and preprocessing pipelines have been uniformly standardized.

**Quantitative Results and Benchmark Comparisons**

Comprehensive benchmarking tests were conducted on these datasets and various difficulty levels. Normalized Mean Error (NME), Area Under the Curve (AUC), and subset-specific accuracy are three metrics for evaluating model performance, which can summarize the strengths and weaknesses of the model.

The proposed method achieves an overall NME of 2.89 on the 300-W dataset, outperforming the state-of-the-art architectures HRNet (NME=3.09), FAN (NME=3.27), and Stacked Hourglass (NME=3.34). Figure 3(a) shows the statistical comparison of NME for all the aforementioned methods, with the network achieving better results on both the normal and difficult subsets. Figure 3(b) shows the detailed accuracy results of typical variation groups, such as pose, occlusion, and expression. In cases of high occlusion and significant pose variation, the aforementioned method is more robust, reducing the number of erroneous landmark predictions by 2.1% on the COFW dataset.

Figure 3(c) shows the average NME values for the 300-W, COFW, and WFLW multi-dataset evaluations. Due to its stability and generalization across all datasets and data complexities, there is no need to adjust hyperparameters for specific datasets. The AUC distribution shown in Figure 3(d) indicates that the model is stable. On WFLW, it achieved an AUC of 0.547, which is significantly higher than previous top models, and it performed excellently on 300-W and COFW. The aforementioned advantages can be applied to indoor and outdoor images, as well as challenging expressions and side cases; they have practical applications in facial analysis systems.

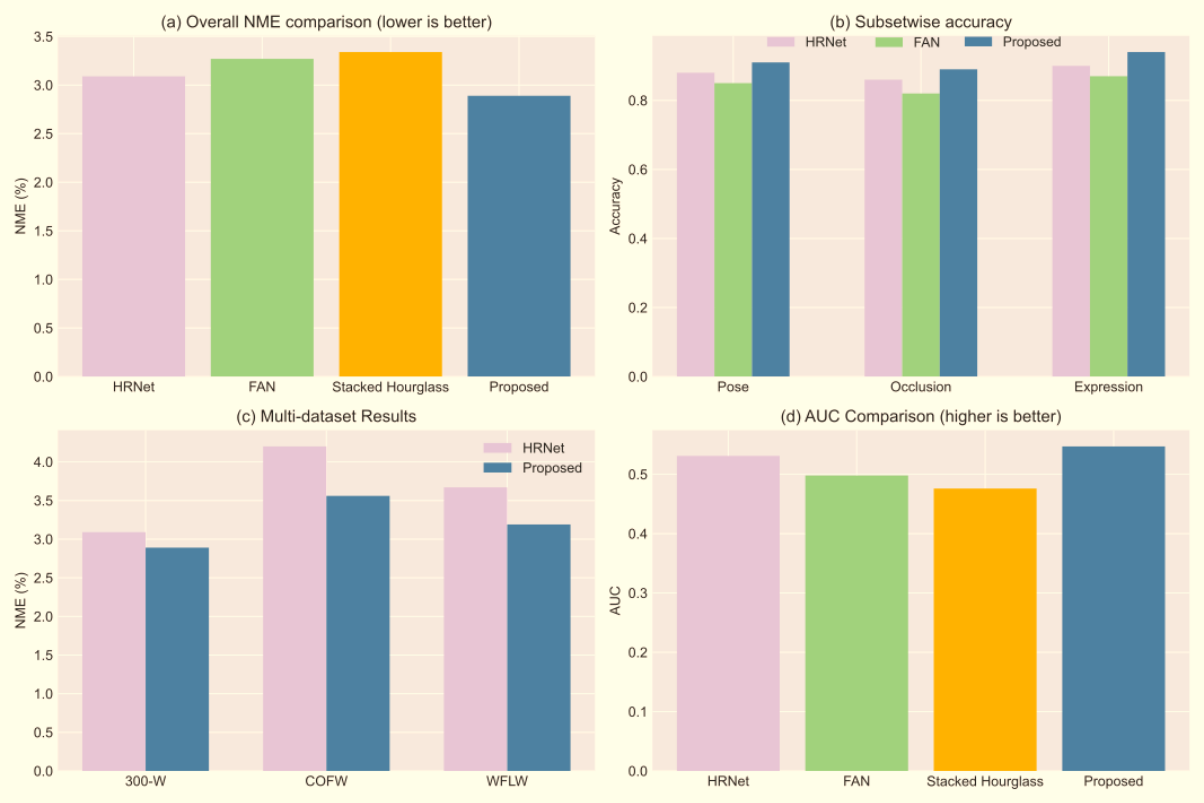


**Figure 3.** Quantitative results of facial landmark detection methods: (a) Overall NME comparison; (b) Subsetwise accuracy; (c) multi-dataset results; (d) AUC comparison

**Ablation Study of Model Components**

Ablation analysis has been conducted to determine the extent to which the core algorithms and architecture affect the overall system performance. The multi-level feature fusion module, loss function settings, and data augmentation strategies are the main topics of this paper, as well as the roles these components play in controlled experiments.

Feature fusion requires simultaneously obtaining a large amount of background data and fine-grained spatial details. As shown in Figure 4(a), the multi-level fusion module is included, with an average NME of 2.89 for 300-W, but it increases to 3.33 when not included. For challenging facial shapes, higher levels of context should be included. In cases of off-axis poses and partial occlusion, this module is necessary.

Choose a loss function to improve localization accuracy and outlier sensitivity. Figure 4(b) shows a direct comparison between the robust wing loss, the standard mean squared error (MSE), and the custom composite loss proposed in this paper. The NME for the MSE-based configuration is 3.15, while the wing loss is reduced to 2.99. The proposed composite loss introduces spatial adaptability and robust penalties, thereby achieving the best result (2.89 NME) by adjusting small displacements and large errors.

In order to expand the generalization of the unconstrained test domain, data augmentation is also necessary. When evaluating the difficult subsets of WFLW and COFW with high occlusion and expression variations, models trained without data augmentation strategies showed a significant drop in performance, as shown in Figure 4(c). Data augmentation improved the model's robustness and reduced the anomaly rate of difficult samples. NME decreased from 3.41 (without data augmentation) to 2.89 (with the complete data augmentation process).

The results of the ablation experiments are shown in Figure 4 and are more clearly presented in the subset and error distribution. By integrating methods of architecture optimization, algorithm optimization, and data distribution changes, high-performance facial keypoint detection can be achieved.

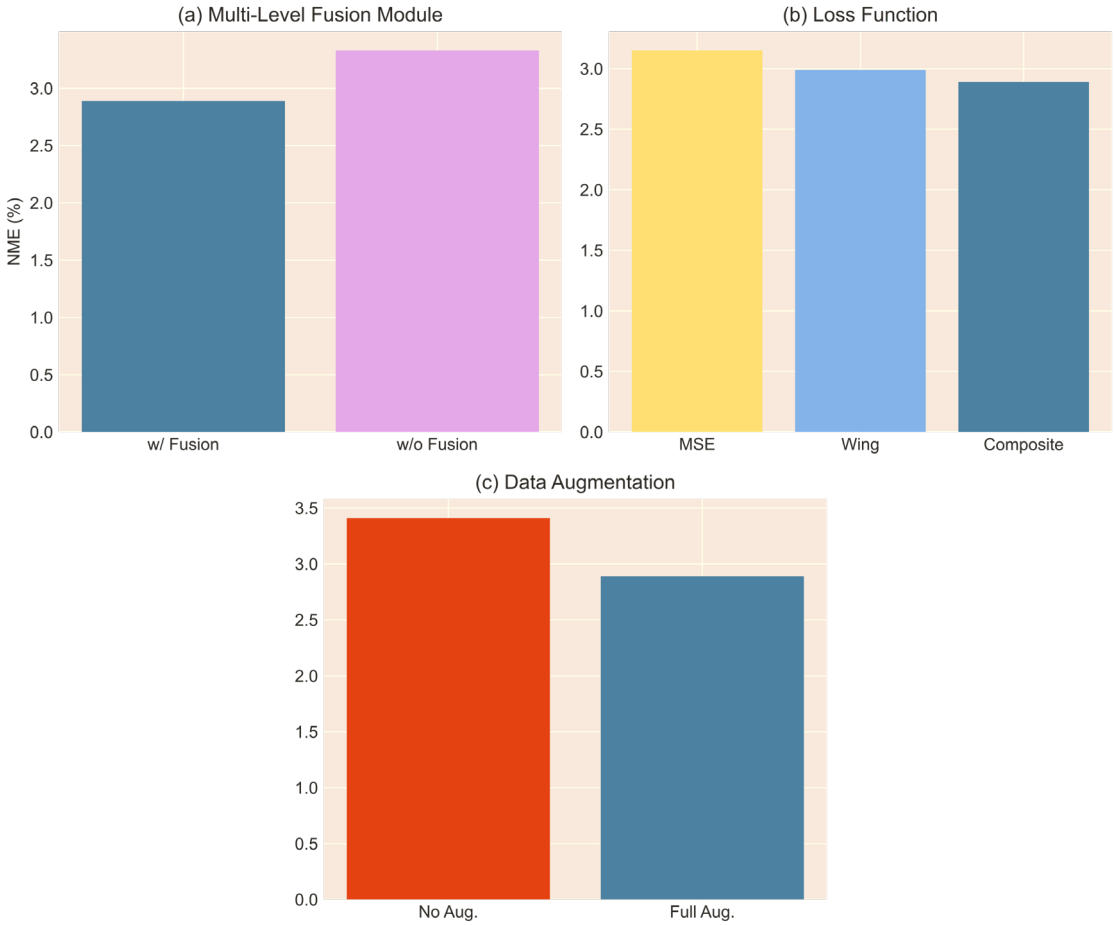


**Figure 4.** Ablation study of principal model components: (a) feature fusion module; (b) loss function comparison; (c) data augmentation strategies

**Qualitative Visualization**

Selected some representative test cases and plotted the predicted landmarks on the real annotations to demonstrate the accuracy and stability of the proposed model in the real world. These visualizations highlight the pixel-level alignment and generalization capabilities.

As shown in Figure 5(a), many common frontal face examples in the 300-W and WFLW datasets almost perfectly align with the predicted and annotated landmarks. The model is spatially consistent with facial features and other semantic points, and it performs excellently in clean and controlled environments.

Figure 5(b) shows a series of challenging cases, including severe occlusions (such as hands, masks, and sunglasses), extreme head poses, and exaggerated facial expressions. The model is still able to identify key points on faces that are damaged or partially obscured. In cases where the actual annotations are irregular or missing, reasonable inferences are made using the system's context and symmetry.

Figure 5(c) shows that the predictions of HRNet, FAN, and Stacked Hourglass networks on the same complex samples are more accurate than the proposed method. Baseline methods often show deviations in expression and occlusion areas. The proposed system consistently achieves the most anatomically accurate distribution around the eye sockets and contour landmarks. Landmark collapse and lateral displacement are other methods to identify these failure areas; they are less noticeable in non-frontal and occluded profiles.

Figure 5(d) shows many test cases, each of which can be presented under different backgrounds and lighting conditions. In practice, the network has already adapted to non-standard and low-contrast samples, demonstrating its versatility and accuracy.


**Figure 5.** Qualitative visualization for facial landmark detection: (a) standard cases; (b) challenging examples; (c) comparison with other methods; (d) diverse application scenario montage

**Error and Robustness Analysis**

The landmark positioning errors were subjected to overall statistics and spatial analysis to identify shortcomings and evaluate the practical application of the model. Calculate the resistance to typical visual interference and the distribution and concentration of errors.

Figure 6(a) shows the histogram of all localization errors in the benchmark test. Most test samples have sub-pixel errors, and the distribution of these errors is highly concentrated within the 0-2% range of the inter-eye distance. Only a few outliers exceeded the 4% error limit, and these anomalies were mainly due to side photos or severe occlusions. The long tail phenomenon refers to prioritizing more rare or difficult-to-handle cases.

As shown in Figure 6(b), there are multiple local spatial error areas on the chin, jawline, and the outer side of the eyebrow tail. These areas are more likely to be occluded, which may result in uneven or truncated hair. Due to the persistent small errors in the core facial areas (such as the eyes, mouth, and tip of the nose), the model's geometric features remain consistent with the core semantic features. To correct the issues in the current situation, the dataset needs to be expanded or targeted adjustment methods should be used.

Figure 6(c) shows the perturbation and degradation of the image. The errors from downsampling, blurring, and simulated noise methods only slightly increase, and the initial error is also quite large, making it suitable for general cases. Even in the presence of 30% Gaussian noise or strong motion blur, the average error of the system remains relatively small, which is feasible in practice.

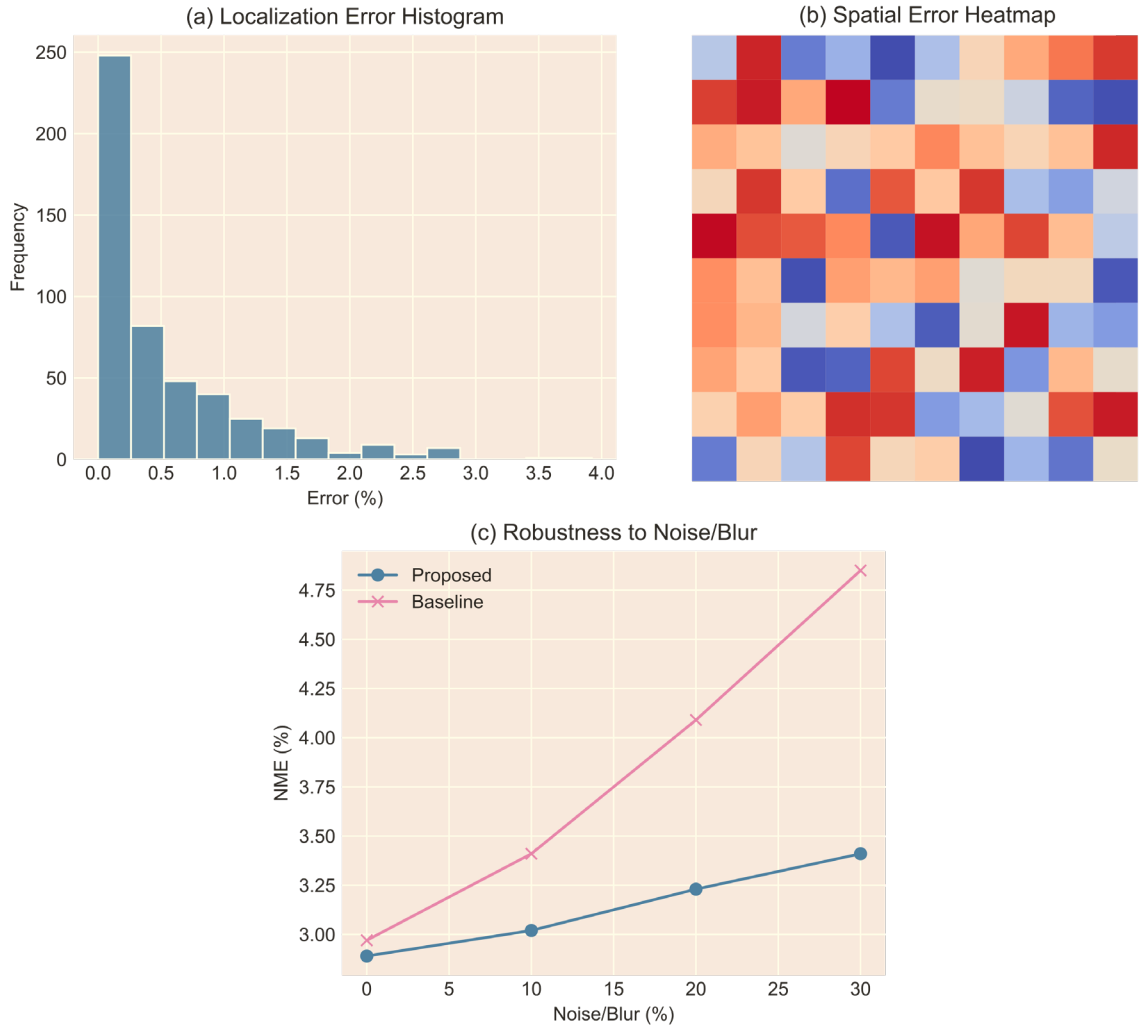


**Figure 6.** Error and robustness analysis: (a) landmark error distribution histogram; (b) facial error heatmap; (c) robustness under visual disturbances

**Application Scenario & Case Analysis**

Through in-depth research on extreme cases and actual data, the scope and limitations of the application have been expanded. A quantitative analysis of the existing model's shortcomings was conducted, and a roadmap for future algorithm improvements and applications was proposed.

Figure 7(a) shows typical scenes in complex environments, including blurry motion capture, low-light nightlife areas, and crowded group photos. Even in cases of cluttered backgrounds, diverse light sources, or damaged images, the model can still accurately locate objects. In low-light and high-motion environments, the spatial consistency between the mouth and eyes will disappear. High dynamic range and low compression formats are suitable for surveillance cameras and mobile recorders. To evaluate the robustness of this method, group scenes containing partially occluded and overlapping faces. Able to accurately identify and locate landmarks without being squeezed by others in the crowded foreground.

Figure 7(b) shows the most common errors and lists rare failures and exceptional cases. The issues are usually severe occlusions (e.g., masks or hands), extreme expressions, or ultra-low resolution; in these cases, the landmarks may be significantly off or clustered near the edges of the face. These failures usually occur in the long-tail distribution areas identified by error analysis, most likely due to non-deterministic occlusion or labeling uncertainty. The aforementioned method established new reliability standards, but highly complex rare events still require further research. These results demonstrate the readiness of the method before its actual large-scale deployment and indicate that, in certain cases, targeted architectural or data-driven improvements may bring further robustness enhancements.

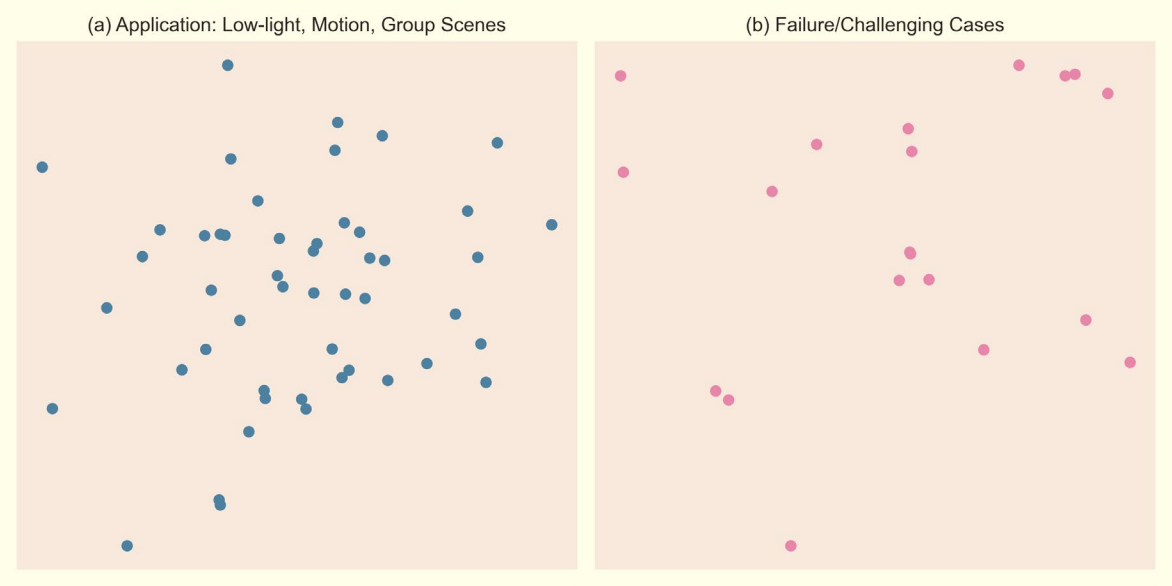


**Figure 7.** Real-world application and failure case gallery: (a) typical and extreme scenarios; (b) failure and challenging sample analysis

## Conclusion

This study proposes a high-resolution architecture for identifying facial keypoints. It also advances state-of-the-art technology by integrating multiple features, designing context-adaptive loss, and standardizing evaluations across the three most challenging benchmarks in the field. Proposed a structural innovation to address performance deficiencies in severely occluded environments and large-angle poses. This structure maintains a parallel, hierarchical integration of feature streams. According to the extensive ablation studies and cross-dataset experiments mentioned above, the architecture and algorithms improve accuracy stability and generalization capabilities. The reduced NME, enhanced AUC, and outlier handling capabilities demonstrate this.

The issue of uneven sample distribution and feature alignment during ULP-Net training can be addressed by using a new loss function that assigns different weights to each element. It can avoid small-scale outlier errors and many real-world distortions, such as noise, blurriness, and resolution changes, while maintaining spatial consistency. According to qualitative visualization, this method can achieve semantically accurate landmark localization in application-driven, unconstrained scenarios, and reach or exceed human-level reliability in challenging environments.

The aforementioned research has made some progress, but still encountered severe occlusions, rare composite interferences, and ambiguous annotations. The next phase of research will focus on special attention mechanisms, enhancing the synthesis of samples for uncommon configurations, and improving cross-view or temporal consistency. In downstream analysis platforms, such as security, human-computer interaction, and mobile health, the proposed system's scalability and stability. It also provides a reference for the subsequent development of facial recognition technology.

### Author Contributions

Julia Magdalena Kozłowska contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Dawid Jerzy Chmielewski and Patryk Waldemar Mazurek contribute to methodology, software, validation, analysis, investigation. All authors have read and agreed with the manuscript before its submission and publication.

### Funding

This research received no specific financial support from any funding agency.

### Institutional Review Board Statement

Not applicable.

## References

[1] Guo, Y., Wang, H., Wang, L., Lei, Y., Liu, L., & Bennamoun, M. (2023). 3D face recognition: Two decades of progress and prospects. ACM Computing Surveys, 56(3), 1-39. https://doi.org/10.1145/3615863

[2] Wei, S., Wang, H., Mo, Y., & Du, D. (2025). A ST-ConvLSTM Network for 3D Human Keypoint Localization Using MmWave Radar. Sensors, 25(18), 5857. https://doi.org/10.3390/s25185857

[3] Lee, G., Haider, A., Kim, H., Kim, K., & Jhang, K. (2025). Enhancing keypoint detection in Y-Maze behavior test automation: introducing stadium heatmap and squared adaptive wing loss. Multimedia Tools and Applications, 84(25), 29031-29053. https://doi.org/10.1007/s11042-024-20286-9

[4] Liu, L., Ke, Z., Huo, J., & Chen, J. (2021). Head pose estimation through keypoints matching between reconstructed 3D face model and 2D image. Sensors, 21(5), 1841. https://doi.org/10.3390/s21051841

[5] Ding, X., Li, Q., Cheng, Y., Wang, J., Bian, W., & Jie, B. (2020). Local keypoint-based Faster R-CNN. Applied Intelligence, 50(10), 3007-3022. https://doi.org/10.1007/s10489-020-01665-9

[6] Ling, X., Zhu, Y., Liu, W., Liang, J., & Yang, J. (2023). The Generation of Articulatory Animations Based on Keypoint Detection and Motion Transfer Combined with Image Style Transfer. Computers, 12(8), 150. https://doi.org/10.3390/computers12080150

[7] Jiang, X., Tao, H., Hwang, J. N., & Fang, Z. (2023). A multiscale coarse-to-fine human pose estimation network with hard keypoint mining. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 54(3), 1730-1741. https://doi.org/10.1109/TSMC.2023.3328876

[8] Agnelli, F., Facchi, G., Grossi, G., & Lanzarotti, R. (2025). KA-GCN: Kernel-Attentive Graph Convolutional Network for 3D face analysis. Array, 26, 100392. https://doi.org/10.1016/j.array.2025.100392

[9] Yang, Y., Wen, L., Zeng, X., Xu, Y., Wu, X., Zhou, J., & Wang, Y. (2024, October). Learning with alignments: Tackling the inter-and intra-domain shifts for cross-multidomain facial expression recognition. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 4236-4245). https://doi.org/10.1145/3664647.3680747

[10] Elharrouss, O., Almaadeed, N., Al-Maadeed, S., & Khelifi, F. (2022). Pose-invariant face recognition with multitask cascade networks. Neural Computing and Applications, 34(8), 6039-6052. https://doi.org/10.1007/s00521-021-06690-4

[11] Varma, P. S., & Anand, V. (2021). Random forest learning based indoor localization as an IoT service for smart buildings. Wireless Personal Communications, 117(4). https://doi.org/10.1007/s11277-020-07977-w

[12] Huang, Y., Chen, Y., Wang, J., Zhou, P., Lai, J., & Wang, Q. (2024). A robust and efficient method for effective facial keypoint detection. Applied sciences, 14(16), 7153. https://doi.org/10.3390/app14167153

[13] Huang, Z., Zhu, Y., Li, H., & Yang, D. (2024). Dynamic facial expression recognition based on spatial key-points optimized region feature fusion and temporal self-attention. Engineering Applications of Artificial Intelligence, 133, 108535. https://doi.org/10.1016/j.engappai.2024.108535

[14] Guo, Y., Xu, Y., Niu, J., & Li, S. (2023). Anchor-free arbitrary-oriented construction vehicle detection with orientation-aware Gaussian heatmap. Computer-Aided Civil and Infrastructure Engineering, 38(7), 907-919. https://doi.org/10.1111/mice.12940

[15] Shao, H., & Zhong, D. (2021). One-shot cross-dataset palmprint recognition via adversarial domain adaptation. Neurocomputing, 432, 288-299. https://doi.org/10.1016/j.neucom.2020.12.072

[16] Hoang, D. C., Tan, P. X., Pham, D. L., Pham, H. N., Bui, S. A., Nguyen, C. M., ... & Nguyen, T. U. (2024). Efficient multimodal fusion for hand pose estimation with hourglass network. IEEE Access, 12, 113810-113825. https://doi.org/10.1109/ACCESS.2024.3444322

[17] Yousuf Khanday, N., & Ahmad Sofi, S. (2025). Generalizing and Classifying From Few Samples: A Comprehension of Approaches to Few-Shot Visual Learning. Computational Intelligence, 41(4), e70098. https://doi.org/10.1111/coin.70098

[18] Lips, T., De Gusseme, V. L., & Wyffels, F. (2024). Learning keypoints for robotic cloth manipulation using synthetic data. IEEE Robotics and Automation Letters, 9(7), 6528-6535. https://doi.org/10.1109/LRA.2024.3405335

[19] Luo, X., Wei, T., Liu, S., Wang, Z., Mattei-Mendez, L., Loper, T., ... & Liu, C. (2025). Certifying robustness of learning-based keypoint detection and pose estimation methods. ACM Transactions on Cyber-Physical Systems, 9(2), 1-26. https://doi.org/10.1145/3728362

[20] Huang, M., Gao, J., & Ma, L. (2025). Enhanced keypoint recognition framework via multi-scale feature characteristics. Scientific Reports, 15(1), 40136. https://doi.org/10.1038/s41598-025-23831-0

[21] Jin, Y., Zhang, Y., Xu, Z., Zhang, W., & Xu, J. (2024, November). Advanced object detection and pose estimation with hybrid task cascade and high-resolution networks. In 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML) (pp. 1293-1297). IEEE. https://doi.org/10.1109/ICICML63543.2024.10958125

[22] Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., ... & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. Computational visual media, 8(3), 331-368. https://doi.org/10.1007/s41095-022-0271-y

[23] Lin, Y., Li, K., & Wang, H. (2025). High-Resolution Human Keypoint Detection: A Unified Framework for Single and Multi-Person Settings. Algorithms, 18(8), 533. https://doi.org/10.3390/a18080533

[24] Liu, D., Zhang, C., Song, Y., Huang, H., Wang, C., Barnett, M., & Cai, W. (2022). Decompose to adapt: Cross-domain object detection via feature disentanglement. IEEE Transactions on Multimedia, 25, 1333-1344. https://doi.org/10.1109/TMM.2022.3141614

[25] Nguyen Minh, C., Nguyen Ngoc, T., Le Dinh, T., Dinh Viet, S., Wong, P. M., Chng, C. B., & Chui, C. K. (2023, December). Boosting facial landmark detection via self-supervised and semi-supervised learning. In Proceedings of the 12th International Symposium on Information and Communication Technology (pp. 485-492). https://doi.org/10.1145/3628797.3629017

[26] Xu, T., Jiang, J., Cai, L., Chai, H., & Ma, H. (2025). Salient object detection with non-local feature enhancement and edge reconstruction. Scientific Reports, 15(1), 397. https://doi.org/10.1038/s41598-024-84680-x

[27] Liu, J., Li, H., Zuo, F., Zhao, Z., & Lu, S. (2023). Kd-lightnet: A lightweight network based on knowledge distillation for industrial defect detection. IEEE Transactions on Instrumentation and Measurement, 72, 1-13. https://doi.org/10.1109/TIM.2023.3300421

[28] Zheng, J., Shi, X., Gorban, A., Mao, J., Song, Y., Qi, C. R., ... & Anguelov, D. (2022). Multi-modal 3d human pose estimation with 2d weak supervision in autonomous driving. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4478-4487). https://doi.org/10.1109/ACCESS.2020.3026209

[29] Fraifer, M. A., Coleman, J., Maguire, J., Trslić, P., Dooly, G., & Toal, D. (2025). Autonomous forklifts: State of the art—Exploring perception, scanning technologies and functional systems—A comprehensive review. Electronics, 14(1), 153. https://doi.org/10.3390/electronics14010153

[30] Li, Q., Zhuang, Y., Huai, J., Wang, X., Wang, B., & Cao, Y. (2024). A robust data-model dual-driven fusion with uncertainty estimation for LiDAR–IMU localization system. ISPRS Journal of Photogrammetry and Remote Sensing, 210, 128-140. https://doi.org/10.1016/j.isprsjprs.2024.03.008