

Person re-identification in images based on a conditional hypermodel

Roman Kvyetnyy^a, Oleksandr Kyrylenko^a, Volodymyr Garmash^a, Ilona Bogach^a,
Tetiana Piliavoz^a, Iryna Shakhina^b, Daniel Sawicki^{*c}, Gulmira Bazil^d

^aVinnitsia National Technical University, 95 Khmelnytsky highway, Vinnitsa, 21021, Ukraine

^bVinnitsia Mykhailo Kotsiubynskyi State Pedagogical University, 32 Ostrozhskogo str.,
21001, Vinnitsia, Ukraine,

^cLublin University of Technology, Nadbystrzycka 38D str., 20-618, Lublin, Poland

^dKazakh National Technical University, 22 Satpayev str., Almaty, 050013, Kazakhstan

ABSTRACT

This paper provides a systematic review of modern adaptive hypermodel methods for the task of person re-identification. The analysis is based on comparing the effectiveness of models tested in both controlled laboratory conditions and real-world scenarios. Considerable attention is paid to key performance indicators—recognition accuracy (mAP, CMC) and data processing speed—which enables a comprehensive assessment of the methods under study.

The focus of the research is the study of the impact of dynamic parameter changes on system performance, as well as the analysis of incremental learning strategies aimed at minimizing the risk of catastrophic forgetting during adaptation to new conditions without the need for full retraining. This approach ensures a rapid response to variations in shooting conditions—changes in lighting, angles, and other characteristics—which is critically important for video surveillance.

Based on the obtained results, promising directions for further research are identified, including optimization of adaptive learning algorithms, design of new architectural schemes, and scaling of systems while maintaining performance. The implementation of these ideas will contribute to the creation of more reliable and efficient solutions for re-identification in modern information platforms. A hybrid architecture is proposed that combines a conditional hypermodel with a base deep neural network. The main advantages of this approach lie in high adaptability and training stability, achieved through dynamic adjustment of parameters via the hypermodel. The combination of loss functions—cross-entropy and triplet—contributes to forming compact and well-separated features for different identities, enhancing the model's ability to correctly identify subjects under high variability of input data. Experimental results confirm the feasibility of integrating adaptive mechanisms into re-identification systems, providing enhanced robustness to environmental changes and the high performance necessary for successful implementation in information technologies.

Keywords: person re-identification, hypermodels, dynamic adaptation, incremental learning, CNN, OSNet, Market-1501, DukeMTMC-ReID, MSMT17.

INTRODUCTION

With the increasing importance of security systems and the concept of “smart” cities, the task of person re-identification becomes especially relevant. It finds application not only in video surveillance but also in analyzing visitor behavior in commercial and public spaces. The main challenge lies in enabling algorithms to definitively determine whether a frame or video sequence depicts the same person as in other images or database entries, despite significant differences in viewing angle, lighting, or changes in appearance due to accessories and clothing¹.

Significant progress in this area has been made through the implementation of deep neural networks—primarily convolutional architectures (CNNs) and transformers—trained on widely recognized datasets such as Market-1501, DukeMTMC-ReID, and MSMT17⁷. These models show high mAP and CMC Rank-1 scores in laboratory experiments. However, real-world systems face additional challenges: video streams come from new cameras with different angles and characteristics, shooting conditions change dynamically (seasonal variations, weather conditions, interior decorations), and full retraining is not feasible due to limited computational resources and time costs.

Moreover, classical incremental learning techniques often suffer from catastrophic forgetting, where adaptation to new data leads to a decline in performance on previously learned samples. A promising solution is the use of conditional hypermodels, which generate or adjust the parameters of the main network based on environmental changes while retaining previous configurations. This approach allows for flexible adaptation without full retraining, which is particularly important for scalable real-time video surveillance systems².

*d.sawicki@pollub.pl

MATERIALS AND RESEARCH METHODS

Existing approaches to building hypermodels differ significantly in structure and in how they generate the parameters of the target network. Some of the earliest solutions proposed using a separate compact network to generate full weight matrices for the main model². This architecture made it possible to significantly reduce the number of trainable parameters and quickly adapt to new conditions. However, it showed instability when working with large datasets or high diversity of input images, as the complexity of simultaneously aligning all spatial-channel filter relationships exceeded the hypermodel's capabilities.

To achieve finer tuning for specific frames, the approach of dynamic filter networks was proposed, where convolutional kernels are created directly based on the input data³. This flexibility allows the model to account for the angle, lighting, and other individual characteristics of each image, but it also introduces considerable computational overhead and requires careful regularization to avoid overfitting and excessive variability in the generated filters.

In response to computational complexity challenges, the concept of conditional modular hypermodels emerged, which adjust only selected components of the target network or operate as an additional module introducing corrections to intermediate features extracted by the base architecture⁷. This approach combines the high stability of the base model with sufficient adaptability, as only the parameters most critical to image processing outcomes are updated—particularly important under resource constraints and real-time requirements.

Alongside the development of conditional hypermodels, methods for factorizing and low-dimensional approximations of weight tensors (e.g., based on CP or Tucker decompositions) have been developed. These significantly reduce memory load during parameter generation and are often integrated with conditional schemes to balance adaptability and performance⁸.

Comparative analysis shows that full-scale hypermodels provide maximum flexibility but are limited by computational cost and potential instability. Dynamic filter networks enhance the model's ability to react to shooting conditions but require additional resources to generate filters in real time. Factorization approaches optimize memory usage and speed up inference but don't always ensure sufficient adaptability on their own. In contrast, conditional modular hypermodels offer an optimal compromise, enabling on-the-fly adjustment of key parameters without full retraining, reducing the risk of catastrophic forgetting, and maintaining high system throughput⁹.

In the context of person re-identification, where it is necessary to quickly process streams from heterogeneous cameras with varying angles and lighting conditions, the conditional modular hypermodel appears to be the most suitable solution.

The combination of an adaptive module that generates corrective weights for key intermediate layers of the main network and a stable base architecture ensures high feature extraction accuracy while maintaining computational efficiency required for real-time video surveillance platforms¹⁰.

Purpose of work. The aim of this study is to improve the accuracy of person re-identification through the use of adaptive hypermodels and deep neural networks, and to evaluate the effectiveness of these models for further application in information technologies.

MODEL EXPERIMENT

In the proposed method, the conditional hypermodel acts as an overlay to the base OSNet architecture, providing dynamic correction only of key network parameters depending on the shooting conditions.

Let $x \in R^{H \times W \times 3}$ be the input image, a $c \in R^{d_c}$ be a context vector that may include information about the camera type, exposure, timestamp, etc. Initially, c is transformed into a compact representation $z \in R^{d_z}$ using a multilayer perceptron E with parameters Φ :

$$z = E(c; \Phi) = \sigma(W_E^{(2)} \sigma(W_E^{(1)} c + b_E^{(1)}) + b_E^{(2)}), \quad (1)$$

де $W_E^{(1)} \in R^{h \times d_c}$, $W_E^{(2)} \in R^{d_z \times h}$, $b_E^{(1)} \in R^h$, $b_E^{(2)} \in R^{d_z}$, and σ — is a nonlinearity (ReLU)².

The resulting z is passed to the conditional hypermodel H with parameters Θ , which generates adjustment ΔW only for the weight tensors of those Omni-Scale blocks in OSNet that are most sensitive to context changes.

Let the indices of these layers be denoted by the set S , and the total dimensionality of their weights by $\sum_{i \in S} |W_{0,i}|$. Then:

$$\Delta W = H(z; \Theta) = W_H^{(L)} \phi(W_H^{(L-1)} \phi(\dots \phi(W_H^{(1)} z + b_H^{(1)}) \dots) + b_H^{(L-1)}) + b_H^{(L)}, \quad (2)$$

where each $W_H^{(l)}$ and $b_H^{(l)}$ define hypermodel layers with $l = 1, \dots, L$, a ϕ — ReLU activation and Dropout + BatchNorm to prevent overfitting¹⁵.

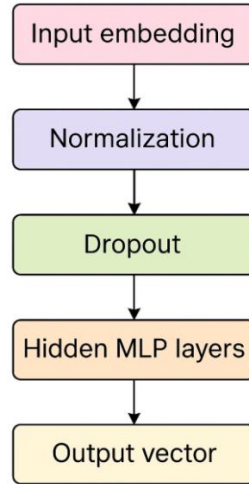


Figure 1. Structure of the conditional hypermodel

After computing $\Delta W \in R^{\sum_{i \in S} |W_{0,i}|}$ only the selected weight tensors are updated:

$$W_i(c) = W_{0,i} + \Delta W_i, \quad \forall i \in S, \quad (3)$$

where ΔW_i — is the subvector of ΔW , corresponding to layer i . Other OSNet parameters remain fixed. The new full network weights are denoted $W(c)^{11}$.

Next, with these adaptive weights, OSNet outputs a multi-scale feature vector¹⁴:

$$f(x, c) = f_{OS}(x; W(c)), \quad (4)$$

where each Omni-Scale block implements parallel convolution operations of size 1×1 , 3×3 and 5×5 , with the results concatenated and normalized.

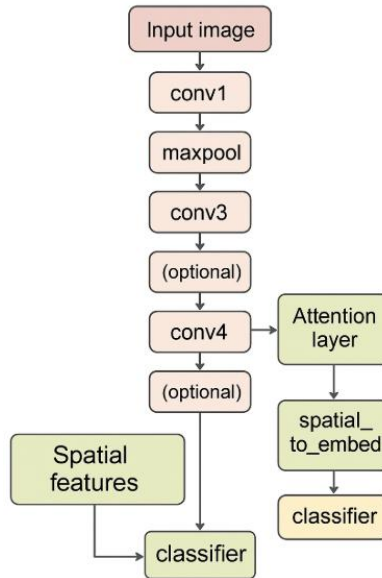


Figure 2. Scheme of the base OSNet network

The training objective is defined by a combined loss function:

$$\mathcal{L}(\Phi, \Theta, W_0) = - \sum_{k=1}^K 1\{y = k\} \log \frac{\exp(w_k^\top f(x, c))}{\sum_{j=1}^K \exp(w_j^\top f(x, c))} + \lambda [\|f(a, c) - f(p, c)\|_2^2 - \|f(a, c) - f(n, c)\|_2^2 + m]_+, \quad (5)$$

where w_k — are the weights of the final linear classification layer, m — is the margin, λ — is the balancing coefficient. Gradients of the function $\mathcal{L}$ simultaneously update parameters $\{W_0, \Phi, \Theta\}$ using the Adam optimizer¹⁸:

$$\begin{aligned}
W_0 &\leftarrow W_0 - \eta \frac{\partial \mathcal{L}}{\partial W_0}, \\
\Phi &\leftarrow \Phi - \eta \frac{\partial \mathcal{L}}{\partial \Phi}, \\
\Theta &\leftarrow \Theta - \eta \frac{\partial \mathcal{L}}{\partial \Theta}.
\end{aligned} \tag{6}$$

which ensures consistent optimization of both the adaptive (hypermodel) and base architecture layers.

This step-by-step division of responsibilities guarantees that the conditional hypermodel learns to generate minimal but highly relevant corrections to OSNet's weights, allowing the network to remain stable while still adapting to changes in angles, lighting, or other external conditions..

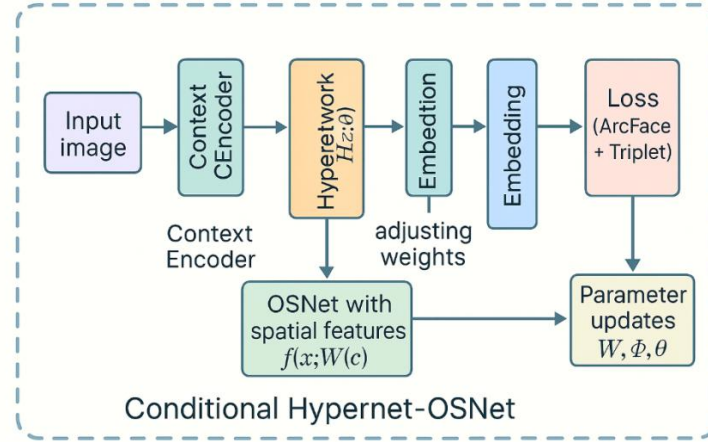


Figure 3. General scheme of the proposed method

Experimental evaluation was conducted using two widely accepted person Re-ID datasets — Market-1501 and DukeMTMC-ReID.

The Market-1501 dataset was created using video from six cameras positioned in front of a supermarket on a campus, including five 1280 × 1080 HD cameras and one 720 × 576 SD camera. Some cameras have overlapping fields of view. The dataset includes 32,668 images of 1,501 individuals. Each person is captured by up to six cameras. Every annotated person is seen by at least two cameras to enable cross-camera search⁴.

The DukeMTMC-reID dataset was collected at Duke University and contains over 14 hours of video sequences from 8 cameras. A random selection of 702 individuals forms the training set, and the remaining 702 form the test set. One query image is chosen per person per camera in the test set, while the rest of the images are used for the gallery. In total, there are 16,522 training images from 702 individuals, 2,228 query images from the remaining 702 individuals, and 17,661 gallery images^{5,19,20}.

Both datasets were preprocessed: all images were resized to 256×128 pixels, normalized using ImageNet statistics, and labeled with person and camera IDs¹⁶.

After each training epoch, mean Average Precision (mAP) and Cumulative Matching Characteristic Rank-1 metrics were calculated. Cumulative Matching Characteristics (CMC) are widely used for evaluating results in person re-identification tasks. These curves show the probability of correct identification based on the number of top-ranked gallery samples considered in the query analysis. During CMC computation, the model ranks gallery samples by similarity to the query image. Accordingly, the top-k CMC accuracy is defined as follows^{12,21}:

$$Acc_k = \begin{cases} 1 & \text{if the highest - rated gallery samples contain the query identifier} \\ 0, & \text{otherwise} \end{cases}$$

if at least one of the top-ranked gallery samples matches the target individual, the accuracy is set to one.

Mean Average Precision (mAP) is widely used to evaluate model performance in object detection and retrieval tasks, particularly where accurate person identification is required. It is based on precision and recall metrics. Precision is the

ratio of true positives (TP) to the sum of true positives and false positives (TP + FP). Recall is the ratio of true positives to the sum of true positives and false negatives (TP + FN).

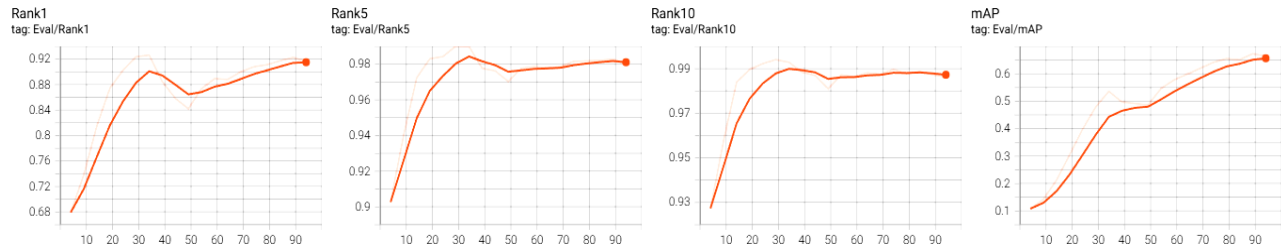


Figure 4. Changes in the model's quality metrics during training.

On the Market-1501 dataset, models with the adaptive hypermodel demonstrate a significant performance improvement compared to the baseline OSNet model without a hypermodel¹⁴. Specifically, the mean average precision (mAP) rose from 73.5 % to the range of 75.4–78.3 %, and the Rank-1 accuracy increased from 90 % to approximately 96–97 %. This indicates effective extraction of relevant features and adaptive adjustment of representations, enabling better discrimination between identities even under high image variability.



Figure 5. An example of the model's operation on the Market-1501 dataset.

Average Precision (AP) for a specific class is calculated as the mean precision at various recall levels, expressed as:

$$AP = \frac{1}{R} \sum_{r \in R} \text{Precision}(r). \quad (7)$$

Mean Average Precision (mAP), in turn, is the average of APs across all classes and serves as the final aggregate performance indicator. This metric is actively used to compare and evaluate the performance of modern models in person re-identification and object detection tasks due to its ability to reflect feature distribution quality across all categories^{12, 17}.

$$\text{mAP} = \frac{1}{M} \sum_{m \in M} AP(m). \quad (8)$$

The model consisted of a base OSNet with initial weights W_0 , a context extractor E , implemented as a two-layer MLP with 256 and 128 neurons and ReLU activation, and a hypermodel H composed of three MLP layers sized $512 \rightarrow 512 \rightarrow |W_0|$ with Dropout regularization.

Training was conducted on two GPUs using the Adam optimizer¹⁸, with an initial learning rate of 3×10^{-4} that was reduced by a factor of 10 every 30 epochs. The loss-balance coefficient λ was set to 1.0, and the training ran for a total of 200 epochs.

A similar trend is observed on the DukeMTMC-reID dataset, where the baseline OSNet network achieves an mAP of 62.81 % and a Rank-1 accuracy of 81.7 %. Models augmented with the hypermodel reach an mAP in the range of 65.1–67.6 % and a Rank-1 accuracy from 83.5 % to 84.5 %, indicating an enhanced ability of the system to recognize individuals in more complex and diverse data, and confirming the effectiveness of the conditional hypermodel for adaptively tuning network parameters under varying scenarios.

CONCLUSION

This study conducted a systematic analysis of modern adaptive hypermodel methods for the task of person re-identification. The effectiveness of the models was compared based on their testing in both laboratory conditions and real-world scenarios. Particular attention was paid to key quality metrics—recognition accuracy (mAP, CMC) and data processing speed—which together provide a comprehensive evaluation of method performance.

The datasets considered included human attribute datasets: Market-1501 and DukeMTMC-reID, and the proposed method was tested on these datasets.

The experimental research was conducted using the Python programming language and the PyTorch deep learning framework. Experiments on the Market-1501 dataset demonstrated that integrating the adaptive hypermodel significantly improves performance compared to the base OSNet model without the hypermodel.

Mean Average Precision (mAP) increased from 73.5% to a range of 75.4–78.3%, while Rank-1 rose from 90% to approximately 96–97%, indicating more effective extraction of relevant features and adaptive adjustment of representations, which allows for more accurate identity differentiation even under conditions of high image variability.

A similar trend was observed on the DukeMTMC-reID dataset, where the base OSNet model achieved an mAP of 62.81% and Rank-1 of 81.7%, while models with the hypermodel reached mAP values in the range of 65.1–67.6% and Rank-1 between 83.5% and 84.5%.

These results confirm that the use of adaptive hypermodels enhances the system’s ability to recognize individuals in more complex and diverse data, paving the way for their further implementation in real-world video surveillance and security systems.

REFERENCES

- [1] Maslii, R, Krylenko, O., Marushchak, Y., “Analysis of methods of person re-identification in multi-camera environment,” Norwegian Journal of Development of International Science, no. 47, no. 1, 46-48 (2020). https://nor-ijournal.com/wp-content/uploads/2023/09/NJD_47_1.pdf
- [2] Ha, D., Dai, A., and Le, Q.V., “HyperNetworks,” arXiv preprint arXiv, 1609.09106 (2016). <https://arxiv.org/abs/1609.09106>
- [3] Jia, X., Wei, X., and Cao, X., “Dynamic Filter Networks,” arXiv preprint arXiv, 1605.09673 (2016). <https://arxiv.org/abs/1605.09673>
- [4] Market-1501 dataset.: <https://www.pkuvmc.com/dataset.html>

- [5] DukeMTMC-ReID dataset. https://github.com/layumi/DukeMTMC-reID_evaluation
- [6] MSMT17 dataset. <https://arxiv.org/abs/1711.08565>
- [7] Zhou, K., Yang, Y., Cavallaro, A., and Xiang, T., "Learning Generalisable Omni-Scale Representations for Person Re-Identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, 7593–7609 (Oct. 2021). <https://arxiv.org/abs/1910.06827>
- [8] Vaswani, A. et al., "Attention is All You Need," in *Adv. Neural Inf. Process. Syst.*, 5998–6008 (2017). <https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>
- [9] Goodfellow, I., Bengio, Y., and Courville, A., "Deep Learning. Cambridge", MA, USA: MIT Press (2016). <http://www.deeplearningbook.org/>
- [10] Hermans, A., Beyer, L., and Leibe, B., "In Defense of the Triplet Loss for Person Re-Identification", *arXiv preprint arXiv, 1703.07737* (2017). <https://arxiv.org/abs/1703.07737>
- [11] Ristani, E., Solera, F., Zou, R. S, Cucchiara, R., and Tomasi, C., "Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking," in *European Conf. Comput. Vis. Workshops*, (2016). <https://arxiv.org/abs/1609.01775>
- [12] Kyrylenko, O.M., "Development of a method for person re-identification," *Opto-Electron. Inf.-Energy Technol.*, vol. 41, no. 1, 25–32 (May 2022). <https://doi.org/10.31649/1681-7893-2021-41-1-25-32>
- [13] Zhou, K. and Xiang, T., "Torchreid: A Library for Deep Learning Person Re-Identification in PyTorch," *GitHub repository* (2019). <https://github.com/KaiyangZhou/deep-person-reid>
- [14] Zheng L., Yang Y., and Hauptmann A.G., "Person re-identification: Past, present and future," *arXiv preprint arXiv, 1610.02984* (2016). <https://arxiv.org/abs/1610.02984>
- [15] Kyrylenko, O., Kvyetnyy, R., and Maslii, R., "Research of human attributes for the problem of re-identification," *Information Technology and Computer Engineering*, vol. 49, no. 3, 4–13 (2020). <https://doi.org/10.31649/1999-9941-2020-49-3-4-13>
- [16] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 248–255 (2009). <https://doi.org/10.1109/CVPR.2009.5206848>
- [17] He, K., Zhang, X., Ren, S., and Sun, J., "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 770–778 (2016). <https://doi.org/10.1109/CVPR.2016.90>
- [18] Kingma, D. P., Ba, J., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2014). <https://arxiv.org/abs/1412.6980>
- [19] Romanyuk, O.N., Pavlov, S.V., Romanyuk, O.V., Smolarz, A., Bazarova, M., "A function-based approach to real-time visualization using graphics processing units", *Proc. SPIE 11581, 115810E* (2020).
- [20] Bisikalo, O., Kharchenko, V., Kovtun V., et al., "Parameterization of the Stochastic Model for Evaluating Variable Small Data in the Shannon Entropy Basis. *Entropy*. 25(2): 184. (2023). <https://doi.org/10.3390/e25020184>
- [21] Romanyuk, O.N, Pavlov, S.V, Melnyk, S., Romanyuk, S.O, Smolarz, A., Bazarova, M., "Method of anti-aliasing with the use of the new pixel model", *Proc. SPIE 9816, 981617* (2015).