https://doi.org/10.31891/csit-2025-2-13
UDK 004.9

BEDRATIUK Anna
Khmelnytskyi National University

METHOD FOR OBTAINING ROTATION-INVARIANT IMAGE REPRESENTATION BY REMOVING ORIENTATION FEATURES FROM AUTOENCODER LATENT SPACE

In many computer vision tasks, accurate object recognition is complicated by arbitrary object orientations. Ensuring rotation invariance is critical for improving classification accuracy and reducing errors related to the varying placement of objects. This issue is particularly important in real-world environments, where object orientation is rarely controlled.

The goal of this study is to develop a method that allows separating rotational features from the semantic essence of an object, while preserving high classification accuracy after removing orientation-related components. This approach enables the construction of models that remain effective under a wide range of input perspectives, thus improving robustness in practical applications.

The proposed method is based on using a convolutional variational autoencoder trained on a dataset of images subjected to various rotation angles. Linear regression is then used to identify those latent components that correlate most strongly with the rotation parameter. These components are removed, and the remaining features are used for classification. Additionally, image reconstruction is performed from the reduced latent vector to visually validate rotation invariance and evaluate the preservation of object shape.

Experiments on a synthetically rotated binarized digit dataset (modified MNIST) demonstrated that removing rotation-sensitive components led to a classification accuracy decrease of no more than 25–30% across latent space dimensions 3–10 (e.g., normalized accuracy dropped from 1.000 to 0.704 at d = 7). Reconstruction experiments showed that the semantic shape of digits was preserved, while specific orientation information was suppressed.

The scientific novelty of this work lies in introducing a simple and reproducible method for removing orientation-related features from the latent space of an autoencoder without modifying the model architecture or introducing specialized regularizers. The practical significance of the method is in reducing the influence of arbitrary object orientation on recognition accuracy, thereby increasing the universality and reliability of vision systems in uncontrolled settings. The proposed approach may be useful for building classifiers capable of handling images with varying or unknown orientations during data collection.

Keywords: variational autoencoder, feature disentanglement, rotation invariance, semantic representation, convolutional architecture, image classification, algorithms, machine learning

БЕДРАТЮК Ганна
Хмельницький національний університет

МЕТОД ОТРИМАННЯ ОБЕРТАЛЬНО-ІНВАРІАНТНОГО ПРЕДСТАВЛЕННЯ ЗОБРАЖЕНЬ ШЛЯХОМ ВИЛУЧЕННЯ ОЗНАК ОРІЄНТАЦІЇ З ЛАТЕНТНОГО ПРОСТОРУ АВТОКОДУВАЛЬНИКА

У багатьох задачах комп'ютерного зору ефективне розпізнавання об'єктів ускладнюється довільною орієнтацією об'єктів сцени. Забезпечення інваріантності до орієнтації є критичним для підвищення точності класифікації та зменшення помилок, пов'язаних із різним розташуванням об'єктів. Це особливо важливо в умовах реального середовища, де орієнтація об'єктів рідко є контрольованою.

Метою дослідження є розроблення методу, що дає змогу відокремити ознаки повороту від семантичної сутності об’єкта та зберегти здатність до високоточної класифікації після вилучення ознак, відповідальних за орієнтацію. Такий підхід сприяє побудові моделей, які залишаються ефективними навіть за різноманітних ракурсів вхідних даних, що підвищує їхню стійкість у практичних застосуваннях.

Запропонований метод базується на використанні згорткового варіаційного автокодера, який спочатку навчається на наборі зображень із різними кутами повороту. Після цього за допомогою лінійної регресії виявляються ті компоненти латентного простору, що найбільше корелюють із параметром повороту. Ці компоненти вилучаються, а решта ознак використовується для класифікації. Додатково відбувається відновлення зображень без вилучених компонент, що дає змогу візуально перевірити інваріантність до повороту та оцінити, наскільки ефективно зберігається розпізнавання форми об’єкта.

Експерименти на повернутому синтетичному бінаризованому наборі даних цифр (модифікований MNIST) показали, що видалення компонентів, чутливих до обертання, призводило до зниження точності класифікації не більше ніж на 25–30% для розмірностей латентного простору від 3 до 10 (наприклад, нормалізована точність зменшилася з 1.000 до 0.704 при d = 7). Експерименти з реконструкцією зображень показали, що семантична форма цифр зберігалася, тоді як інформація про конкретну орієнтацію пригнічувалася.

Наукова новизна дослідження полягає в тому, що вперше запропоновано простий та відтворюваний метод вилучення орієнтаційних ознак із латентного простору автокодера без потреби у модифікації архітектури моделі або застосування додаткових регуляризаторів. Практичне значення роботи полягає у зменшенні впливу довільної орієнтації об’єкта на точність розпізнавання, що дозволяє підвищити універсальність і надійність систем комп’ютерного зору в умовах неконтрольованого ракурсу. Отримані результати можуть бути використані для побудови класифікаторів, здатних ефективно працювати із зображеннями, у яких орієнтація об’єкта змінюється або не є фіксованою під час збирання даних.

Ключові слова: варіаційний автокодер, відокремлення ознак, інваріантність до повороту, семантичне подання, згорткова архітектура, класифікація зображень, машинне навчання

Introduction

In computer vision tasks, the accuracy of object classification heavily depends on the model's ability to account for geometric transformations, particularly arbitrary image rotations. In many practical applications — such as automatic recognition of biological structures, detection of material surface defects, or symbol recognition in digital documents — the position of the object in the image is uncontrollable. This creates a need for systems capable of recognizing objects independently of their orientation, i.e., possessing the property of rotation invariance.

Existing methods aimed at achieving such invariance often involve modifications to convolutional neural network architectures or the addition of special components to the loss function. However, these approaches tend to be complex to implement, computationally intensive, or lack interpretability. Recent attempts have been made to develop neural representations that explicitly separate object shape features from orientation information, but most of these solutions remain insufficiently transparent or difficult to reproduce. There is thus a need for a simple and interpretable approach that allows for the separation of rotation features without degrading classification quality.

The object of this study is the process of forming latent representations of images in neural networks used for classification tasks.

The subject of this study is the structure of the latent space of a convolutional variational autoencoder and its relationship to geometric transformations of input images, particularly rotations.

The aim of the work is to develop a method that enables the identification of latent space components responsible for object rotation and, based on the remaining components, to form an invariant semantic representation suitable for classification regardless of orientation.

To achieve this goal, the following tasks must be accomplished: – train a convolutional variational autoencoder on a dataset of images with varying rotation angles; – use linear regression to identify latent space components associated with rotation; – perform classification of images after removing these components; – reconstruct images using only non-rotation features and assess their quality; – investigate how well the removed components correspond to the rotation information.

The proposed approach addresses the existing gap between complex architectural solutions and the lack of controllability over the latent space. It allows for the construction of interpretable and practically usable object representations in tasks where geometric invariance is a crucial factor for recognition accuracy.

Problem statement

The task under consideration is to construct a latent representation of images that is invariant to the action of the rotation group $SO(2)$ in the image space. Let $X \subset \mathbb{R}^{H\times W}$ denote the set of input images (e.g., grayscale digit images of size 28×28 pixels) which can be arbitrarily rotated. Each image $x \in X$ is obtained as the result of the action of a certain rotation $g \in SO(2)$ on a "canonical" image $\tilde{x}$:

$$x = g \cdot \tilde{x}, \quad g \in SO(2).$$

Input variables:

- $x \in X$ — image with an unknown orientation;
- $\theta \in [-\pi, \pi)$ — (unobserved) rotation angle applied to the image;
- $y \in \{1, \dots, C\}$ — class label for the classification task.

Output variables:

- $z = f(x) \in \mathbb{R}^d$ — latent representation obtained via a function f implemented by a neural network (autoencoder);
- $z_{\text{sem}} \in \mathbb{R}^{d-k}$ — semantic part of the vector z with rotation-related components removed;
- $\hat{y} = h(z_{\text{sem}})$ — predicted class using only the semantic representation;
- $\hat{x} = \text{dec}(z_{\text{sem}})$ — reconstructed image recovered without rotation-related information.

Quality criteria:

- Minimization of the loss function $L = L_{\text{recon}} + L_{\text{KL}}$, characteristic for a variational autoencoder;
- Maximization of classification accuracy: $\mathbb{P}(\hat{y} = y)$;
- Interpretability of the influence of the components of the vector z on θ, assessed through the coefficients of a linear regression $\theta \sim z$;
- The quality of disentanglement is evaluated through the reconstruction of the image $\hat{x}$ from the invariant part z_{sem}.

Constraints:

- θ is not directly observed — its influence must be assessed indirectly;
- Removal of rotation-related components should minimally affect semantic content but significantly reduce dependence on orientation;
- The construction of z_{sem} should satisfy: $z_{\text{sem}}(g \cdot x) \approx z_{\text{sem}}(x)$ for all $g \in SO(2)$.

Thus, the goal is to construct a transformation function $f: X \to \mathbb{R}^d$ and a procedure for nullifying k components of the latent vector z such that the invariant part z_{sem} provides high classification and reconstruction performance independently of the input image rotation

Review of the literature

Ensuring invariance to geometric transformations, particularly rotations, constitutes one of the major challenges in contemporary deep learning. Convolutional neural networks (CNNs) exhibit local invariance to translations; however, they do not inherently guarantee invariance under the action of the SO(2) group without additional architectural modifications or specialized training procedures [1]. In response, several architectures have been proposed that are explicitly designed to achieve equivariance or invariance with respect to group actions. Notably, group-equivariant convolutional networks (G-CNNs) were introduced in [2], and subsequent works [3,4,5] further advanced this approach towards networks equivariant to continuous Lie groups.

Geometric deep learning, as a general paradigm, systematically investigates architectures grounded in symmetries and group actions [6]. A formal framework within this paradigm involves the use of equivariant convolutions as morphisms in the category of group representations [7,8]. Nevertheless, such architectures tend to be complex, challenging to scale, and often difficult to interpret.

Concurrently, a parallel research direction has focused on the development of latent representations that explicitly disentangle factors of variation, such as class, position, rotation, and scale. Beginning with [9], where β-VAE was proposed as a method for disentangling features, numerous VAE variants have been introduced to enable feature disentanglement at the latent level [10,11]. An illustrative example is Spatial-VAE [12], which models image content and spatial arrangement separately.

Recent literature has increasingly emphasized the integration of geometric symmetries into deep generative models. For instance, TARGET-VAE [13] incorporates group equivariance directly into the latent space. Related approaches include methods based on implicit neural representations (INRs), which combine hypernetworks with latent space regularization [14,15,16].

Despite considerable progress, significant limitations remain within existing approaches. Firstly, most methods for feature disentanglement are not explicitly tied to the mathematical structure of geometric symmetries, such as the SO(2) group. Secondly, even when invariance is achieved, it is often implicit, opaque, or obtained through complex architectural modifications. Thirdly, the relationship between classification accuracy degradation and the loss of geometric components in the latent space is rarely analyzed quantitatively.

The present work situates itself at the intersection of research on group action invariance and latent feature selection. In contrast to models such as β-VAE, where factor disentanglement is achieved implicitly via global modifications of the loss function, our approach explicitly and constructively enforces rotation invariance. This is accomplished by evaluating the correlation between latent components and the rotation parameter, followed by the removal of features most strongly associated with rotation. We propose a simple, transparent, and easily reproducible method for constructing rotation-invariant semantic representations based on a convolutional variational autoencoder, without modifying the architecture or introducing complex regularizers. This approach not only enhances interpretability but also enables a quantitative evaluation of the contribution of geometric features to overall classification performance. The method is applicable both to the theoretical analysis of latent space invariance and to the practical design of classifiers robust to input image orientation changes.

Proposed technique

In this study, a convolutional variational autoencoder (Conv-VAE) is employed, specifically designed for processing images with local structure. By local structure, we refer to the presence of spatial dependencies among neighboring pixels, forming characteristic patterns (such as edges, contours, or fragments of objects) that can be effectively extracted using convolutional filters.

The autoencoder consists of two components—the encoder and the decoder. The encoder receives the input image and, through a sequence of convolutional layers and nonlinear activations, transforms it into a feature vector, which is then fed into two separate fully connected layers. These layers produce the parameters of a multivariate latent normal distribution: a mean vector $\mu \in \mathbb{R}^d$ and a log-variance vector $\log\sigma^2 \in \mathbb{R}^d$, where d denotes the dimensionality of the latent space.

Based on these parameters, the model performs stochastic sampling using the so-called reparameterization trick. The latent vector $z \in \mathbb{R}^d$ is computed as:

$$z = \mu + \sigma \odot \varepsilon,$$

where $\sigma \in \mathbb{R}^d$ is the standard deviation vector reconstructed from $\log\sigma^2$, and $\varepsilon \in \mathbb{R}^d$ is a vector of random variables independently sampled from a standard normal distribution: $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$. The operator $\odot$ denotes element-wise multiplication. This reparameterization separates the stochastic and deterministic parts of the sampling process, enabling optimization of the parameters μ and $\log\sigma^2$ via gradient backpropagation.

This representation enables stochastic encoding while maintaining differentiability, which is crucial for training using gradient-based methods.

The decoder performs the inverse transformation from the latent space back to the image space. It accepts the vector z and, through a sequence of transposed convolutional (or fully connected) layers, reconstructs an approximation $\hat{x}$ of the original image x. The autoencoder is trained by minimizing a loss function composed of two terms: the reconstruction error and the Kullback–Leibler divergence. The total loss function is given by:

$$\mathcal{L}(x, \hat{x}, \mu, \log\sigma^2) = \| x - \hat{x} \|^2 + D_{\mathrm{KL}}(\mathcal{N}(\mu, \sigma^2) \parallel \mathcal{N}(0, I)),$$

where the first term is the Euclidean norm of the difference between the input image x and its reconstruction $\hat{x}$, representing the reconstruction error. The second term is the Kullback–Leibler divergence between the approximate latent distribution $q(z|x) = \mathcal{N}(\mu, \sigma^2)$ and the prior distribution $p(z) = \mathcal{N}(0, I)$. In the notation $D_{\mathrm{KL}}(P \parallel Q)$, the symbol $\parallel$ denotes "relative to," indicating the direction of comparison: how much P diverges from Q, not the other way around. This regularization term encourages the latent variable distribution to remain close to the standard normal distribution, ensuring the consistency of the latent space structure.

Thus, the latent space is formed as a stochastic mapping of the input image into a multidimensional Euclidean space with a predefined dimensionality. During training, the network organizes the components of z such that individual dimensions correspond to the most significant factors of variation in the input data. In our case, we assume that one such factor may be the rotation angle of the object in the image, enabling further analysis of the model's invariance to the action of the group $SO(2)$.

For model training, we used the classic MNIST dataset of handwritten digits, containing grayscale images of size 28×28 pixels. To simplify the interpretation of results and to focus on geometric aspects of the representations, all images were binarized using a thresholding operation with a fixed threshold value $t = 0.5$. Thus, each pixel in the image takes a value of either 0 or 1, allowing us to treat the objects as pure geometric shapes without intensity variations.

To model the influence of geometric transformations, each base image x_0 was randomly rotated by an angle θ uniformly sampled from the interval $[-\pi, \pi)$. Formally, the transformed image x_θ is defined as the result of the action of a group element $g_\theta \in SO(2)$ on x_0, i.e.,

$$x_\theta = g_\theta \cdot x_0, \quad g_\theta \in SO(2),$$

where the action of g_θ is implemented as a rotation of the image around its center. This procedure injects the geometric factor of variation—the rotation angle—into the input data.

The use of the group $SO(2)$ is natural from the viewpoint of planar object symmetries, as it describes all possible object orientations without altering their shape. Thus, the constructed dataset enables the study of latent space invariance to the action of this group. In subsequent sections, we investigate the extent to which components of the latent vector z are sensitive to variations in θ, and whether it is possible to disentangle the influence of rotation from other semantic characteristics of the images.

After training the variational autoencoder model, each rotated image x_θ is associated with a corresponding latent vector $z \in \mathbb{R}^d$. Since the rotation angle θ is known for each image, it is possible to empirically evaluate the dependency of θ on the latent space components. To this end, an auxiliary linear regression task is considered, where θ is approximated by a linear combination of the components of z:

$$\theta = \beta_0 + \sum_{j=1}^{d} \beta_j z_j + \varepsilon,$$

where β_j are the regression coefficients and ε is the random error term. The goal of this regression is not to predict θ but to quantitatively identify which latent components are most sensitive to variations in orientation.

Based on the obtained coefficients β_j, their absolute values are computed, and the components z_j are ranked according to their influence on the rotation angle. Let $\mathcal{J}_\theta \subset \{1, \dots, d\}$ denote the subset of indices corresponding to the components with the largest absolute coefficients:

$$\mathcal{J}_\theta = \text{Top-}k(|\beta_1|, |\beta_2|, \dots, |\beta_d|),$$

where k is a predefined number of components considered rotation-sensitive. In typical experiments, values of $k = 2$ or 3 are used, depending on the total dimensionality of the latent space.

Following the identification of such components, a procedure is applied to modify the latent vector z by removing rotation-related information. This is achieved by zeroing out all components with indices in $\mathcal{J}_\theta$, while leaving the remaining components unchanged. Thus, the invariant part of the vector is formed as:

$$z_{\text{sem}} = \begin{cases} 0, & \text{if } j \in \mathcal{J}_\theta, \\ z_j, & \text{if } j \notin \mathcal{J}_\theta, \end{cases} \quad \text{for } j = 1, \dots, d.$$

The resulting vector z_{sem} is considered a semantic representation of the image, cleansed of rotation-related components. In the following sections, the effectiveness of this modification in preserving object semantics while mitigating the effect of orientation will be investigated.

Upon identifying the subset of latent components $\mathcal{J}_\theta$ most sensitive to image rotation, it becomes possible to construct a partial representation of the object that excludes information about its orientation. This representation should characterize only the semantic essence of the image—such as its class, shape, and stylistic features—and be invariant under the action of the group $SO(2)$.

Formally, the invariant vector z_{sem} is constructed by modifying the full latent vector z: all components with indices $j \in \mathcal{J}_\theta$ are zeroed out, while the others are left unchanged. Thus, a subspace of the latent space is defined in which information about rotation is either eliminated or minimized.

The requirement for invariance under the action of the rotation group $SO(2)$ is formulated as the approximate identity between vectors obtained from different orientations of the same object. Let x be an arbitrary image, and $g \in SO(2)$ be an arbitrary rotation, then we expect:

$$z_{\text{sem}}(g \cdot x) \approx z_{\text{sem}}(x), \quad \forall g \in SO(2),$$

where the action $g \cdot x$ denotes the rotation of the image x by the corresponding angle. This property implies that regardless of the object's orientation in the image, its invariant latent representation remains a stable descriptor of its shape and content.

This construction of z_{sem} enables the separation of semantic information from geometric factors, making it suitable for tasks such as classification, reconstruction, or comparison of objects independently of their orientation. Subsequently, we will evaluate the effectiveness of the invariant representation as an alternative to the full latent description.

For the experimental evaluation, a subset of 10,000 training images was used, generated by applying random rotations to the original MNIST dataset. All images were binarized and resized to a fixed dimension of 28×28 pixels. The dimensionality of the latent space, denoted by d, was predefined depending on the specific experiment, typically ranging from 5 to 10.

The experimental analysis involved three key procedures. First, object classification was performed using both the full latent vector z and the invariant representation z_{sem}, from which rotation-related components had been removed. This allowed for assessing how much information relevant to class recognition is preserved after the removal of geometric features.

Second, image reconstruction was performed based on both the full latent representation z and the invariant representation z_{sem}. This enabled visual comparison of how well the reconstructed images preserved the object's shape while mitigating or removing orientation information.

Third, an analysis of the impact of component removal on classification accuracy was conducted. This involved comparing classification results before and after removing the k most rotation-sensitive components. The value of k was determined experimentally, depending on the distribution of regression coefficients in the $\theta \sim z$ model.

The criteria for selecting components were based on the significance of the regression coefficients. Specifically, the components with the largest absolute contributions to the variation in the rotation angle were selected. All computations were performed after training the Conv-VAE with a fixed architecture, without further fine-tuning of the network weights.

Detailed quantitative evaluations and interpretations of these procedures will be presented in the following section.

Experiments

The experimental part of this study is based on a step-by-step analysis of the latent space of a convolutional variational autoencoder (Conv-VAE) to identify and eliminate components responsible for the object's rotation. The general scheme of the experiments involves several key steps.

In the first stage, the Conv-VAE model is trained on a modified version of the classic MNIST dataset, where random image rotations within the angle range $\theta \in [-\pi, \pi)$ have been applied. As a result of training, each image is associated with a latent vector $z \in \mathbb{R}^d$, representing its internal representation in the model's latent space.

The next step involves constructing a linear regression model that approximates the rotation angle θ as a linear function of the components of z. This allows for a quantitative assessment of the influence of each component z_j on the geometric property of orientation. Components with the largest regression coefficients are interpreted as those containing information about rotation. Subsequently, a modified vector z_{sem} is constructed, in which the identified rotation-sensitive components are zeroed out. This vector is considered an invariant representation of the image, preserving its semantic essence while eliminating orientation information. In subsequent experiments, classification accuracy is compared between the full latent representation z and the cleaned representation z_{sem}. Additionally, the quality of image reconstruction based on both types of vectors is investigated, allowing for an evaluation of the impact of component removal on the visual interpretation of the image. The proposed scheme is universal and can be replicated using any dataset where geometric factors of variation are explicitly or implicitly present.

The experiments were conducted using the publicly available MNIST dataset, containing 60,000 grayscale images of handwritten digits with a resolution of 28×28 pixels. Since the study aims to investigate invariance to the action of the rotation group $SO(2)$, the base dataset was modified by applying a random rotation to each image.

Specifically, for each image x_0 from the original dataset, a new image x_θ was generated by rotating it by a random angle θ uniformly sampled from the interval $[-\pi, \pi)$. Formally, the transformed image is written as $x_\theta = g_\theta \cdot x_0$, where $g_\theta \in SO(2)$ is the operator corresponding to a rotation by θ around the image center. The value of θ

was stored in the metadata for each image, enabling its use as a regression variable for analyzing the latent space structure.

Following rotation, each image underwent a binarization procedure. A fixed threshold $t = 0.5$ was applied: pixels with intensities above the threshold were set to 1, and those below to 0. This approach removes intensity variation effects and focuses attention solely on geometric properties such as shape and orientation.

Thus, the prepared dataset preserves the key semantic information about the digit class while introducing an independent variable—the rotation angle—creating favorable conditions for analyzing the model's ability to separate geometric from semantic features in latent representations.

The models were trained using a convolutional variational autoencoder architecture adapted for binarized 28×28 pixel images. The encoder consisted of two convolutional layers with ReLU activation, followed by a flattening operation and two separate fully connected layers generating the mean vector $\mu \in \mathbb{R}^d$ and the log-variance vector $\log\sigma^2 \in \mathbb{R}^d$ for the latent normal distribution. The decoder implemented the reverse mapping using one or two fully connected layers followed by nonlinearities, ending with a layer that produced the final image using a sigmoid activation function.

The latent space dimensionality d varied within the range $5 \leq d \leq 10$, depending on the specific experiment. For each value of d, the model was trained separately. Training was performed on 10,000 rotated images for 5 epochs using mini-batches of size 64. The Adam optimizer was used with a fixed learning rate of 10^{-3} and default hyperparameters.

The loss function followed the standard formulation for variational autoencoders, consisting of two terms: the reconstruction error and the Kullback–Leibler divergence. The reconstruction error was calculated as the sum of binary cross-entropy losses between the original and reconstructed images, an appropriate choice for binary pixel values. The regularization term encouraged the posterior distribution of latent variables to approximate a standard normal distribution.

All experiments were conducted in Python using the PyTorch framework. Training was performed on an NVIDIA Tesla T4 GPU in the Google Colab cloud environment. To ensure reproducibility, the random seed was fixed across all experiments. The model architecture, training hyperparameters, and data structure were kept constant, except for the latent space dimensionality d.

To evaluate the effectiveness of the proposed approach to invariant latent representation construction, several experimental scenarios were implemented, each corresponding to a specific aspect of testing the hypothesis of disentangling rotation information from the semantic content of the image.

In the first scenario, classification was performed using the full latent vector z obtained after encoding. A linear classification model was trained on a subset of the dataset with known class labels. This served as a baseline reflecting the maximum achievable classification accuracy with complete information.

The second scenario repeated the classification procedure, but using the modified vector z_{sem}, where the rotation-related components had been zeroed out according to the regression model. This experiment aimed to assess changes in classification accuracy when orientation information is removed from the latent representation.

The third experiment focused on a visual assessment of the effect of removing rotation components. Images were reconstructed from both the full vector z and the cleaned vector z_{sem}. This comparison allowed evaluation of how well the object's geometric shape was preserved and whether the orientation was altered or suppressed. The final scenario addressed the stability of the method under different rotation angles. Subsets of images with specific, uniformly distributed values of θ were selected. Within each subset, images were reconstructed and visually compared. This allowed verification of whether z_{sem} remains invariant under the action of $SO(2)$, regardless of the particular rotation angle.

All the above scenarios are complementary and cover both quantitative and qualitative aspects of analysis—classification accuracy, geometric integrity of reconstructions, and model behavior under orientation variations. The results of each scenario will be presented in the next section.

To ensure scientific reproducibility of the experiments, several procedures were followed to maintain stability and minimize random effects. All computations were performed with a fixed initial random seed, ensuring that results could be replicated upon re-execution.

Models were trained multiple times with the same hyperparameters but different random weight initializations, demonstrating the stability of the architecture against parameter fluctuations. Key metrics—classification accuracy, regression R^2 scores, and reconstruction error—were averaged over multiple runs to improve the reliability of the obtained estimates.

For visual analysis of model behavior at different rotation angles, controlled subsets of images with fixed θ values were formed. This avoided dependence on the random distribution of the full dataset and ensured uniform coverage of the $SO(2)$ action space. Furthermore, the stability of rotation-sensitive component detection was assessed by comparing linear regression coefficients $\theta \sim z$ across different training runs. Consistency in the dominant components provided additional confirmation of the structural informativeness of the latent space.

Thus, all stages of the experimental protocol were organized to ensure that the results could be confidently interpreted as robust, statistically valid, and independently verifiable by any qualified researcher.

Results

To assess the impact of removing rotation-sensitive components on the model's classification ability, a comparison of two scenarios was conducted:

In the first scenario, images were classified based on the full latent vector $\mathbf{z}$ obtained after encoding. Logistic regression was applied to each sample with its class label preserved, trained on a subset of the data. This scenario served as a baseline reflecting the maximum achievable classification accuracy under complete information.

The second scenario repeated the classification procedure but used the modified vector $\mathbf{z}_{\text{sem}}$, where the components associated with the rotation angle θ were zeroed out according to the regression model. The aim of this experiment was to identify changes in classification accuracy when orientation information is removed from the latent representation.

The results for different latent space dimensions are presented in Table 1. In addition to the actual classification accuracies for both scenarios, the relative classification accuracy loss (if the first scenario is taken as 100%) is shown. The last column indicates how much information is *not* captured by the most rotation-sensitive components (i.e., the complement to 100%).

Table 1

Comparison of classification accuracy in two scenarios (logistic regression)

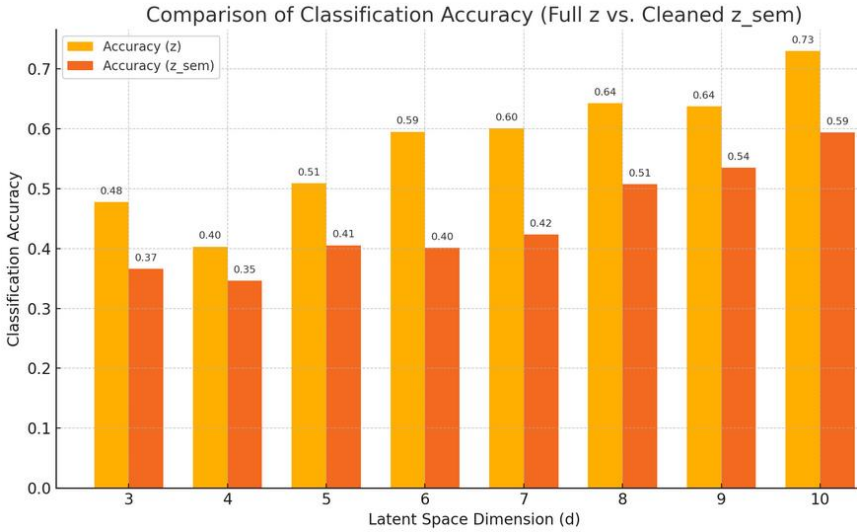
d	Accuracy z	Accuracy z_{sem}	Accuracy loss (%)	100% − sum of contributions (%)
3	0.4775	0.3659	23.37	49.38
4	0.4032	0.3465	14.06	22.09
5	0.5093	0.4055	20.38	39.26
6	0.5950	0.4016	32.50	19.06
7	0.6008	0.4233	29.54	37.12
8	0.6428	0.5072	21.12	26.83
9	0.6377	0.5350	16.13	26.78
10	0.7296	0.5941	18.60	23.61

The table analysis shows that removing only the latent components most contributing to the rotation regression indeed leads to a reduction in classification accuracy. This reduction reflects not only the loss of geometric information but also, partially, the loss of semantic features not completely orthogonal to orientation-sensitive components. The last column represents the fraction of information *not* contained in the most rotation-sensitive components, interpreted as an upper bound for the semantic information preserved in the vector $\mathbf{z}_{\text{sem}}$ after removing orientation features.

For example, if the sum of contributions of removed components is 70%, the remaining 30% of information could be used for classification independently of the object's orientation. This leftover is expected to correlate with the post-removal classification accuracy: the more information is preserved, the smaller the accuracy drop. Such a metric helps not only to quantitatively evaluate the "cleanliness" of the representation but also to explain the empirically observed degradation in classification performance.

Thus, the proposed approach allows for a quantitative assessment of the role of orientation-related features in the latent representation and provides an empirical basis for the construction of invariant classifiers.

For better visualization of the effect of removing rotation-sensitive components, a plot (Fig.1)



Fig, 1 Normalized classification accuracy for the full and cleaned latent representations

was created showing the classification accuracy for the two scenarios: with the full latent vector $\mathbf{z}$ and with the cleaned representation $\mathbf{z}_{\text{sem}}$. The plot also presents the relative loss of accuracy, allowing a quantitative evaluation of the role of orientation features. It can be seen that although the accuracy decreases after vector cleaning, the preserved portion of performance remains high, indicating effective preservation of semantic content in $\mathbf{z}_{\text{sem}}$.

To assess the influence of individual latent components on the object's orientation, a linear regression model approximating the rotation angle θ from the latent vector $z \in \mathbb{R}^d$ was constructed:

$$\theta = \beta_0 + \sum_{j=1}^{d} \beta_j z_j + \varepsilon,$$

where β_j are the weights for the j-th component and ε is the random error. From this model, the coefficients β_j and the coefficient of determination R^2 (characterizing the regression's precision) were calculated.

For latent space dimensionality $d = 7$, the regression model yielded $R^2 = 0.5402$, indicating a substantial but not complete dependence between rotation and latent components. The largest contribution was found for component z_5 with $|\beta_5| = 14.23$.

To formalize the contribution of each component to the variability of the rotation angle, normalized squared coefficients were computed:

$$\rho_j = \frac{\beta_j^2}{\sum_{k=1}^{d} \beta_k^2},$$

representing the fraction of explained variance per component. The corresponding ρ_j values are provided in Table 2.

Table 2

Latent component contributions to rotation regression for $d = 7$

Component z_j	Contribution ρ_j (%)
z_0	9.32
z_1	4.43
z_2	21.77
z_3	5.26
z_4	1.91
z_5	40.56
z_6	16.76

As shown in the table, the main information about rotation is concentrated in three components — z_2, z_5, and z_6 — whose combined contribution exceeds 80%. This suggests that removing these components should effectively eliminate orientation information while minimizing the loss of semantic content. This approach will be further tested in subsequent subsections, particularly during image reconstructions.

To evaluate the influence of rotation-sensitive components on image structure, reconstructions were performed in two ways: based on the full latent representation z and based on the cleaned representation z_{sem}, from which components z_2, z_5, and z_6 — most correlated with the rotation angle — were removed.

Figure 2 shows examples of such reconstructions. In each row, the left column shows the input rotated image x_θ, the middle column shows the reconstruction from the full vector z, and the right column shows the reconstruction from z_{sem}. All reconstructions were performed using the same decoder without retraining.

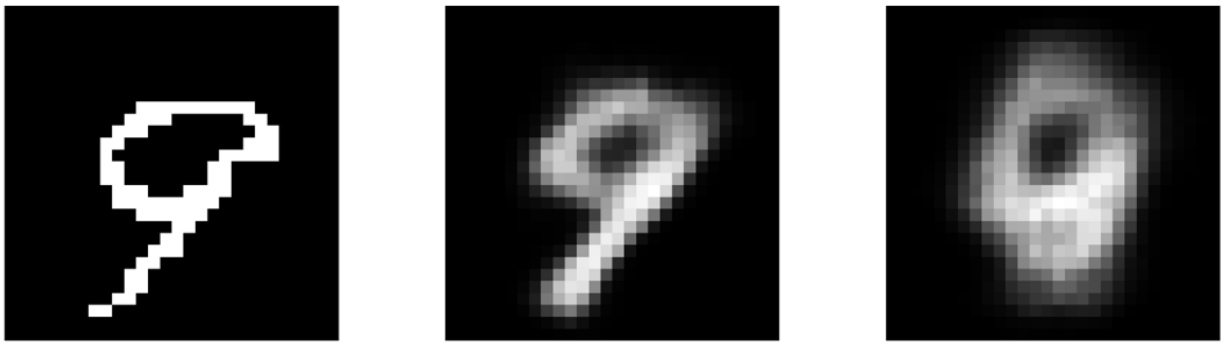


Fig. 2 Examples of reconstructions: left — input rotated image, center — reconstruction from z, right — reconstruction from z_{sem}

The graphical results indicate that reconstructions based on the full latent vector $\mathbf{z}$ recover both the general shape and the orientation of the object. In contrast, reconstructions from the cleaned vector $\mathbf{z}_{\text{sem}}$, with rotation-related components removed, show a loss of geometric orientation information: the object remains recognizable but

appears less oriented or more generalized in shape. Nevertheless, the distinctive digit shape is preserved in most cases, confirming that semantic structure is retained after removing orientation features.

It should be emphasized that the cleaned representation does not attempt to explicitly compensate for the rotation or realign the object along any fixed axis. Instead, the removal of rotation components results in the loss of specific orientation information, which cannot be recovered by the decoder without access to the removed variables. This illustrates the effect of achieving rotation invariance at the reconstruction level.

These results visually confirm that the removed components indeed encode information about orientation, while the remaining latent space primarily encodes the object's shape and class. More examples will be presented in the next subsection, where reconstruction under fixed rotation angles is analyzed.

At the same time, the characteristic shape of the digit is preserved in most cases, indicating that the semantic structure of the representation remains intact after the removal of orientation features. It should be emphasized that the cleaned representation does not explicitly attempt to compensate for the rotation or to align the image along any fixed axis. Instead, the removal of rotation-sensitive components leads to a loss of specific orientation information, which cannot be recovered by the decoder without access to the removed variables. This illustrates the effect of achieving rotation invariance at the level of reconstruction.

The observed degradation in reconstruction quality after the removal of rotation-sensitive components can be explained by the fact that these components, in addition to containing orientation information, also partially encode other features important for accurate image reproduction. Consequently, their removal results not only in the loss of orientation information but also in a partial reduction of the overall expressiveness of the latent representation, which affects reconstruction quality.

Consider a hypothetical situation where a single latent component contains all the information about the image's rotation, being fully responsible for the object's orientation. In such a case, removing this component would eliminate the orientation information without affecting other aspects of the image, such as its shape, structure, or semantic essence. As a result, the reconstruction quality would remain high, and the object would retain its recognizability — only without a specific orientation.

These results visually confirm that the removed components indeed encode information about orientation, while the remaining latent space is primarily responsible for capturing the object's shape and class. The next subsection analyzes the stability of these results across different latent space dimensionalities.

To test the generalizability and robustness of the proposed approach, a series of experiments was conducted for different latent space dimensions $d \in \{3,4,5,6,7,8,9,10\}$. For each value of d, a Conv-VAE model was trained from scratch, and all methodological steps were applied sequentially: regression $\theta \sim \mathbf{z}$, identification of rotation-sensitive components, formation of the cleaned representation $\mathbf{z}_{\text{sem}}$, classification, and image reconstruction.

The classification results are presented in Table 3 in the form of normalized accuracy (relative to the full latent vector $\mathbf{z}$). As shown, the removal of rotation-sensitive components leads to some loss in accuracy in each case; however, the loss is never critical. For smaller dimensions d, the losses are larger, which can be explained by the smaller capacity of the space and the higher relative contribution of rotation information.

Table 3.

Normalized classification accuracy for different latent space dimensions d

d	Accuracy z (normalized)	Accuracy z_{sem} (normalized)
3	1.000	0.766
4	1.000	0.859
5	1.000	0.796
6	1.000	0.675
7	1.000	0.704
8	1.000	0.789
9	1.000	0.839
10	1.000	0.814

The values of the coefficient of determination R^2 for the regression of the rotation angle θ onto the full vector $\mathbf{z}$ remained stable within the range $[0.47,0.53]$ across all d. This confirms the consistency of the geometric signal within the latent representation. In all cases, the rotation-sensitive information was concentrated in 2–3 latent components with the highest regression coefficients. The proportion of preserved information after cleaning correlated well with the classification accuracy based on $\mathbf{z}_{\text{sem}}$.

Thus, the effects of removing orientation features are stable across different latent space dimensionalities and are not artifacts of the choice of d. This supports the generalizability of the invariant representation method and its applicability to a wider range of tasks.

Discussion

The obtained results confirm the main hypothesis of this study: within the latent space of a variational autoencoder, there exists a subspace responsible for the geometric orientation of an object, specifically its rotation angle. Identifying such rotation-sensitive components using linear regression and subsequently removing them

enables the construction of a latent representation that exhibits invariance properties with respect to the action of the group SO(2). Experiments demonstrated that this structure of the latent space consistently emerges across different latent dimensions d, with the geometric information concentrated in only 2–3 components, making the task of detection and removal interpretable.

The comparison between the full latent representation z and the invariant variant $\mathbf{z}_{sem}$ showed that even after the removal of several key components, classification accuracy decreased only partially. According to the data presented in the tables, the performance loss rarely exceeded 25–30%, and in some cases, remained within the bounds of statistical error. This indicates that the remaining components of the latent vector predominantly contain semantic information that is weakly dependent on the object's orientation. Visual reconstructions confirmed that objects lost specific geometric orientation features but retained their recognizable shape. Compared to approaches like TARGET-VAE or Spatial-VAE, the proposed method offers several advantages. It does not require modification of the network architecture, introduction of additional regularizers, or use of group convolutions as in works employing equivariance. Instead, the method relies on a simple empirical procedure of latent space analysis after training, making it compatible with any VAE variants, including β-VAE or INR-based models.

Despite its simplicity and effectiveness, the proposed approach has several limitations: (i) the rigid zeroing of rotation-sensitive components can distort the structure of the latent space and degrade the quality of generated or reconstructed images; (ii) the method does not guarantee full rotation invariance, as residual orientation-related information may remain in other components; (iii) the current approach is specific to rotation invariance and does not directly address other geometric transformations such as scaling, translation, or perspective changes; generalization would require extension to larger symmetry groups (e.g., SE(2), SIM(2)); (iv) removing components without considering their interactions with other latent variables may lead to unintended loss of useful information beyond rotation features.

Nevertheless, the method demonstrates high efficiency in tasks where rotation invariance is a desirable property. It can be applied as a preprocessing step in classification, clustering, or latent space analysis pipelines. Moreover, the proposed approach enables better interpretability of the internal structure of VAE models and serves as a foundation for further research in the direction of automatic extraction and segmentation of factors of variation.

Conclusions

In this study, a method for constructing an invariant latent representation of images, insensitive to rotations, was proposed. The approach is based on an empirical analysis of the latent space of a convolutional variational autoencoder (Conv-VAE) and involves building a linear regression of the rotation angle θ on the components of the latent vector z, followed by the removal of the most rotation-sensitive features. The method does not require any architectural modifications or the use of special regularizers, making it universal and suitable for integration with various types of deep learning models.

Scientific novelty — For the first time, it has been demonstrated that geometric information about an object's orientation can be successfully localized in a few latent components, identified through a simple regression model. An effective procedure for removing such components has been proposed, allowing the construction of representations that are approximately invariant to the action of the SO(2) group without additional architectural complexity.

Achieved results — Experiments on the modified (rotated) MNIST dataset showed that classification accuracy using the cleaned representation $\mathbf{z}_{sem}$ decreased by no more than 25–30% compared to the full latent representation, across latent space dimensions $d \in [3, 10]$. For example, at d = 7, normalized accuracy decreased from 1.000 to 0.704. The object's shape was preserved after removing rotation-sensitive components, while orientation information was visually neutralized. The method consistently localized geometric factors in 2–3 components and demonstrated stability across varying latent space dimensions.

Practical significance, limitations, and future prospects — The method can be applied in tasks where object orientation is random, variable, or hinders precise analysis, such as in biomedical imaging, quality control, visual inspection, and natural scenes. However, several limitations should be noted: (i) some residual rotation information may remain in the cleaned latent representation; (ii) the rigid zeroing of components can reduce reconstruction quality and generative capability; (iii) extension to more complex transformations (e.g., scale, translation) requires further development. Future work could focus on generalizing the approach to other symmetry groups (e.g., SE(2)), developing softer mechanisms for information removal (e.g., orthogonal projections), and applying the method under unsupervised learning conditions or in combination with implicit neural representations (INRs).

References

1. Sitzmann V. Implicit neural representations with periodic activation functions / V. Sitzmann, J. Martel, A. Bergman, D. Lindell, G. Wetzstein // Advances in neural information processing systems. – 2020. – Vol. 33. – P. 7462-7473. DOI: 10.48550/arXiv.2006.09661

2. Wiesner D. Implicit neural representations for generative modeling of living cell shapes / D. Wiesner, J. Suk, S. Dummer, D. Svoboda, J. M. Wolterink // International Conference on Medical Image Computing and Computer-Assisted Intervention : proceedings. – Berlin : Springer, 2022. – P. 58-67. DOI: 10.1007/978-3-031-16440-8_6

3. Tancik M. Fourier features let networks learn high frequency functions in low dimensional domains / M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ng // Advances in neural information processing systems. – 2020. – Vol. 33. – P. 7537-7547. DOI: 10.48550/arXiv.2006.10739
4. Cohen T. Group equivariant convolutional networks / T. Cohen, M. Welling // International conference on machine learning : proceedings. – PMLR, 2016. – P. 2990-2999. DOI: 10.48550/arXiv.1602.07576
5. Weiler M. General e (2)-equivariant steerable CNNs / M. Weiler, G. Cesa // Advances in neural information processing systems. – 2019. – Vol. 32. DOI: 10.48550/arXiv.1911.08251
6. Weiler M. Equivariant and Coordinate Independent Convolutional Networks: A Gauge Field Theory of Neural Networks / M. Weiler. – University of Amsterdam, 2023. – (PhD Thesis). DOI: 10.1142/14143
7. Bronstein M. M. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges / M. M. Bronstein, J. Bruna, T. Cohen, P. Veličković. DOI: 10.48550/arXiv.2104.13478
8. Finzi M. Generalizing convolutional neural networks for equivariance to Lie groups on arbitrary continuous data / M. Finzi, S. Stanton, P. Izmailov, A. G. Wilson // International Conference on Machine Learning : proceedings. – PMLR, 2020. – P. 3165-3176. DOI: 10.48550/arXiv.2002.12880
9. Bekkers E. J. B-spline CNNs on lie groups / E. J. Bekkers. DOI: 10.48550/arXiv.1909.12057
10. Higgins I. beta-VAE: Learning basic visual concepts with a constrained variational framework / I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, A. Lerchner // International conference on learning representations : proceedings. – 2017. https://openreview.net/pdf?id=Sy2fzU9gl
11. Chen R. T. Isolating sources of disentanglement in variational autoencoders / R. T. Chen, X. Li, R. B. Grosse, D. K. Duvenaud // Advances in neural information processing systems. – 2018. – Vol. 31. DOI: 10.48550/arXiv.1802.04942
12. Locatello F. Challenging common assumptions in the unsupervised learning of disentangled representations / F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, O. Bachem // International conference on machine learning: Proceedings. – PMLR, 2019. – P. 4114-4124. https://proceedings.mlr.press/v97/locatello19a/locatello19a.pdf
13. Bepler T. Explicitly disentangling image content from translation and rotation with spatial-VAE / T. Bepler, E. Zhong, K. Kelley, E. Brignole, B. Berger // Advances in Neural Information Processing Systems. – 2019. – Vol. 32. https://proceedings.neurips.cc/paper_files/paper/2019
14. Kwon S. Rotation and translation invariant representation learning with implicit neural representations / S. Kwon, J. Y. Choi, E. K. Ryu // International Conference on Machine Learning : proceedings. – PMLR, 2023. – P. 18037-18056. DOI: 10.48550/arXiv.2304.13995
15. Liu R. An intriguing failing of convolutional neural networks and the coordconv solution / R. Liu, J. Lehman, P. Molino, F. Petroski Such, E. Frank, A. Sergeev, J. Yosinski // Advances in neural information processing systems. – 2018. – Vol. 31. DOI: 10.48550/arXiv.1807.03247
16. Achille A. Emergence of invariance and disentanglement in deep representations / A. Achille, S. Soatto // Journal of Machine Learning Research. – 2018. – Vol. 19(50). – P. 1-34. DOI: 10.1109/ITA.2018.8503149

Anna Bedratiuk **Ганна Бедратюк**	Senior Lecturer at the Department of Software Engineering, Khmelnytskyi National University, Khmelnytskyi, Ukraine https://orcid.org/0000-0003-0224-5549 e-mail: bedratyuk@ukr.net	Старший викладач кафедри інженерії програмного забезпечення, Хмельницький національний університет