# Backward-Compatible Aligned Representations via an Orthogonal Transformation Layer

Simone Ricci ⬤, Niccolo Biondi ⬤, Federico Pernici ⬤, and Alberto Del Bimbo ⬤

DINFO (Department of Information Engineering), University of Florence, Italy,
MICC (Media Integration and Communication Center),
`{name}.{surname}@unifi.it`

**Abstract.** Visual retrieval systems face significant challenges when updating models with improved representations due to misalignment between the old and new representations. The costly and resource-intensive backfilling process involves recalculating feature vectors for images in the gallery set whenever a new model is introduced. To address this, prior research has explored backward-compatible training methods that enable direct comparisons between new and old representations without backfilling. Despite these advancements, achieving a balance between backward compatibility and the performance of independently trained models remains an open problem. In this paper, we address it by expanding the representation space with additional dimensions and learning an orthogonal transformation to achieve compatibility with old models and, at the same time, integrate new information. This transformation preserves the original feature space's geometry, ensuring that our model aligns with previous versions while also learning new data. Our Orthogonal Compatible Aligned (OCA) approach eliminates the need for re-indexing during model updates and ensures that features can be compared directly across different model updates without additional mapping functions. Experimental results on CIFAR-100 and ImageNet-1k demonstrate that our method not only maintains compatibility with previous models but also achieves state-of-the-art accuracy, outperforming several existing methods. Code at: GitHub repository.

**Keywords:** Deep Learning · Representation Learning · Compatible Representation Learning · Orthogonal Transformation
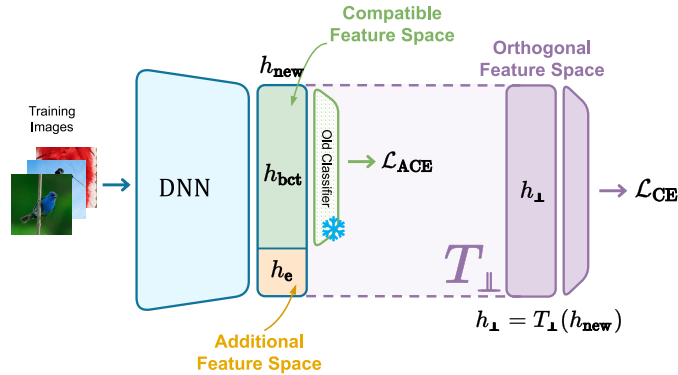
## 1 Introduction

Visual retrieval systems operate by matching images from a stored dataset (the gallery set) to input images (the query set). This process involves using a trained representation model to encode all gallery images into feature representations. When queries are available, the system retrieves the most similar gallery representations.

With advancements in the expressive power of representation models, updating the gallery with newer models is necessary to obtain improved performance [18]. This is particularly challenging when the new model is trained independently of the old one, or they have different network architectures, resulting in completely different and incompatible representations. Consequently, recomputing the feature vectors for all images in the gallery set, a process known as backfilling or re-indexing, becomes

essential. However, this can be prohibitively expensive or even infeasible for real-world galleries containing vast numbers of images.

Recent research addresses the challenge of avoiding the backfilling of the gallery set by learning model representations that can be directly compared without reprocessing the gallery data with an improved model. These representations are referred to as compatible [1, 3, 12, 14, 18, 23, 25]. The seminal work by [18] proposed learning compatible representations using an influence loss into the training objective of the new model, which aligns the new representation with the previous one. However, training the new model using this loss reduces its performance compared to training the same model independently [14]. To overcome this issue, subsequent studies [12, 23, 24] have proposed different loss functions, but these efforts have met with limited success. In another line of research, [14, 20] have explored learning a lightweight transformation between old and new model representations, aiming to fully leverage the improvements provided by the independently trained version of the new model. However, learning these transformations still demands additional training time and a resource-intensive process of mapping all the data in the gallery with these transformation functions. More recently, the conflict between backward compatibility and new model performance has been addressed by [25]. This approach expands the representation space with additional dimensions, allowing the old portion of the feature space to align with the old model while incorporating new knowledge using the independently trained version of the new model in the remaining feature dimensions. In their method, the backbone generates a representation optimized to align with a newly trained independent model's representation through a matching and classification loss. A subset of this representation undergoes a learnable basis transformation, preserving information from the new representation. The new representation is then projected into a compact embedding space and merged with part of the transformed new representation. A second basis transformation is applied to this merged space to match the old model's representation. The part of the transformed new representation that is not merged captures additional information that may be incompatible with the old model but is useful for improving representation quality.

In this paper, we address the dual challenges of learning backward-compatible representations while maintaining performance comparable to the independently trained version of the new model. To achieve this, we expand the feature space of the new model by adding extra dimensions relative to the old model. New information can be integrated into this representation space without necessitating a new independently trained model by learning an orthogonal transformation function (Fig. 1). This orthogonal transformation preserves the geometry of the compatible learned feature space, ensuring that the new model aligns with the old one. Meanwhile, in the additional dimensions not affected by the compatible training, the model incorporates new information. At inference time, the orthogonal transformation is discarded, and visual search is performed with features extracted from the original feature space prior to the transformation. This approach ensures that features generated with our method can be directly compared across multiple model updates without the need for composing mapping functions when comparing non-sequential representations.

**Fig. 1: Overview of our method.** The DNN backbone generates representations in a feature space $h_{\text{new}}$. This feature space is divided into two different parts: $h_{\text{btc}}$ is the learned compatible representation space according to $\mathcal{L}_{\text{ACE}}$, while $h_{\text{e}}$ is an extra feature space used to learn new information from new data without negatively affecting the old feature space configuration. $h_{\text{new}} = [h_{\text{btc}}|h_{\text{e}}]$ is then transformed with $T_{\perp}$ into $h_{\perp}$ and then used for classification using $\mathcal{L}_{\text{CE}}$.

Our experimental results on CIFAR-100 and ImageNet-1k show that, using the proposed approach, the new representation is compatible with the old one while achieving the state-of-the-art accuracy against several compared methods.

In summary, the contributions of this paper are:

– We introduce a method that expands the feature space of models to allow the integration of new information without losing backward compatibility or degrading performance.
– We employ an orthogonal transformation that preserves the geometry of the original feature space, ensuring that new models align with older versions for consistent results.
– We demonstrate the effectiveness of our approach through experiments on CIFAR-100 and ImageNet-1k, achieving state-of-the-art accuracy and proving compatibility across model updates.

## 2 Related Works

Compatible training aims to learn representations that can be used interchangeably when updating a model. The objective is to establish a unified representation space where it is possible to directly compare representations from various models. Methods in this area can be categorized into *direct* comparison between old and new representations [18, 23, 24] or mapping-based approaches [5, 12, 15, 20]. Direct compatibility can be obtained typically involving the usage of auxiliary loss functions. This is introduced to align representation with the old classifier prototypes [18] or to refine the prototype neighbor structure with a fully-connected graph [24]. Mapping-based methods differ in how they learn the transformation function used to compare updated and old representations. In

particular, [12] also allows for direct comparison by simply imposing the transformation module to be the identity.

Nonetheless, certain notable drawbacks associated with these methodologies persist. Relying on an auxiliary loss hinders the new model's ability to achieve comparable performance with the independently trained version of the new model, while the mapping-based approaches require additional training to learn the mapping function after the training of the new models and then a lightweight backfilling process [14] to extract compute the mapping of the existing gallery features. In addition to this, when the model is subjected to multiple updates, to compare non-sequential representations the solution is to concatenate multiple mapping functions, which again increases the re-processing cost. The research in [25] shows that a trade-off exists within the concept of backward-compatibility, as delineated in [18], and the performance that the new model has. To address this issue, they expand the feature space of the new model to retain new information in an extra space while aligning it to the old one. This work has connections with [1, 3]. This work theoretically and empirically showed that the stationarity of the representations (i.e., remaining aligned across several model updates) is crucial to achieving compatibility. To achieve stationarity, the feature space of the model is pre-allocated since the beginning of training according to a simplex configuration [13] to accommodate current data while keeping some free space for future classes.

Several other works have delineated new definitions of compatibility [4,8,17,21,23] or studied compatibility under a continual learning scenario [2,6,9,19]. In this paper, we follow the definition given by [18] and we update the model retraining from scratch the model every time new data is available using also the whole old data, i.e., therefore avoiding the catastrophic forgetting issue [11, 16].

## 3   Methodology

### 3.1   Backward-Compatible Training

Let $\phi_{\text{old}}$ be the initial representation model trained using cross-entropy loss on an initial labeled training set $\mathcal{D}_{\text{old}} = \{\mathbf{x}_i, y_i\}_{i=1}^{C_{\text{old}}}$, where $\mathbf{x}_i$ is a generic image with label $y_i$. After training, $\phi_{\text{old}}$ is used to extract features from a gallery $\mathcal{G} = \{\mathbf{x}_i, y_i\}_{i=1}^{C}$ and a query set $\mathcal{Q} = \{\mathbf{x}_i, y_i\}_{i=1}^{C}$, with $C$ being an arbitrary number of classes. In the following, we refer to the set of features from the gallery and the query set obtained with $\phi_{\text{old}}$ as $\phi_{\text{old}}^{\mathcal{G}}$ and $\phi_{\text{old}}^{\mathcal{Q}}$, respectively. When a new set of images $\mathcal{X}$ becomes available, $\phi_{\text{new}}$ is trained using $\mathcal{D}_{\text{new}} = \mathcal{D}_{\text{old}} \cup \mathcal{X}$. The newly added data $\mathcal{X}$ can have a similar distribution to the previous data $\mathcal{D}_{\text{old}}$ or a completely different one. In this paper, we assume the worst-case scenario where $\mathcal{X}$ belongs to a different and non-overlapping set of classes than $\mathcal{D}_{\text{old}}$, such that $\mathcal{D}_{\text{new}} = \{\mathbf{x}_i, y_i\}_{i=1}^{C_{\text{new}}}$. Backward-compatible training aims to learn $\phi_{\text{new}}$ in a way that allows direct comparison of the features of the query set extracted with the new model $\phi_{\text{new}}^{\mathcal{Q}}$ with the features of the gallery set obtained with the old model $\phi_{\text{old}}^{\mathcal{G}}$, thus avoiding the need to reprocess the gallery set with the new model $\phi_{\text{new}}^{\mathcal{G}}$.

According to [18, 20], the following definition of compatibility between representation models holds:

**Definition 1 (Backward Compatibility [18])** *Two representation models $\phi_{\text{old}}$ and $\phi_{\text{new}}$ are compatible if $\forall\, \mathbf{x}_i, \mathbf{x}_j$ samples from the distribution of interest, with $i \neq j$, holds that:*

$$\mathrm{d}\big(\phi_{\text{old}}(\mathbf{x}_i), \phi_{\text{new}}(\mathbf{x}_j)\big) \leq \mathrm{d}\big(\phi_{\text{old}}(\mathbf{x}_i), \phi_{\text{old}}(\mathbf{x}_j)\big) \tag{1a}$$
$$\text{with } y_i = y_j$$
$$\mathrm{d}\big(\phi_{\text{old}}(\mathbf{x}_i), \phi_{\text{new}}(\mathbf{x}_j)\big) \geq \mathrm{d}\big(\phi_{\text{old}}(\mathbf{x}_i), \phi_{\text{old}}(\mathbf{x}_j)\big) \tag{1b}$$
$$\text{with } y_i \neq y_j$$

*where $y_i, y_j$ are the corresponding labels of $\mathbf{x}_i, \mathbf{x}_j$ and $\mathrm{d}(\cdot, \cdot)$ is a distance function.*

An intuitive interpretation of Def. 1 is that the new model should perform as well as, or better than, the old model in grouping data of the same class. This implies that the distances between new and old feature points of the same class should be less than or equal to those between the old feature vectors (Eq. 1a). Simultaneously, the new model should be better or at least equal to the old model in discriminating data belonging to different classes. Consequently, distances between new and old feature points of different classes should be greater than or equal to those between the old feature vectors (Eq. 1b).

However, it is important to note that Def. 1 is not practical for real-world applications as it requires evaluating the pairwise distances between all data points. This requirement makes the criterion computationally intensive and challenging to implement at scale. Therefore, as suggested by [18], Def. 1 is relaxed into the following Empirical Compatibility Criterion:

$$M\big(\phi_{\text{new}}^{\mathcal{Q}}, \phi_{\text{old}}^{\mathcal{G}}\big) > M\big(\phi_{\text{old}}^{\mathcal{Q}}, \phi_{\text{old}}^{\mathcal{G}}\big) \tag{2}$$

where $M$ is a performance metric.

## 3.2 Backward-Compatibility via Representations Alignment and Orthogonal Transformation

In this section, we present how we train the new model $\phi_{\text{new}}$ to learn a representation that is backward-compatible with $\phi_{\text{old}}$ according to Def. 1, while improving the discrimination capability of the new model using the new incoming data. Fig. 1 shows an overview of our approach that is motivated in the following.

Theoretical [3] and empirical investigations [1, 2, 12, 18, 20] have shown that to achieve compatibility between $\phi_{\text{new}}$ and $\phi_{\text{old}}$, the new representation space should be aligned as closely as possible to the old one. To this end, we align the representation space of $\phi_{\text{new}}$ with the old one by minimizing the distance between the features and the class prototypes of the old classifier, which is kept fixed during the learning of the new model. An old class prototype $W_{\text{old}}^{(y)}$ is obtained by averaging all features extracted from the old network for each image of the class $y$. In particular, we optimize the influence loss via cross-entropy loss $\mathcal{L}_{\text{CE}}(W_{\text{old}}, \phi_{\text{new}}(\mathbf{x}))$ between features extracted with the new model $\phi_{\text{new}}$ and the old classifier prototypes $W_{\text{old}}$ as in [18]. Motivated by [22], which shows that cross-entropy optimization does not achieve optimal alignment between learnable features and fixed class prototypes, we enforce feature stationarity by

optimizing the cosine distance ($\mathcal{L}_\angle$) between the newly learned representations and their corresponding fixed old classifier prototypes, thereby directly enhancing the alignment property of the cross-entropy loss. To achieve backward-compatibility, we optimize the following Aligning Compatible Embedding (ACE) loss:

$$\mathcal{L}_{\text{ACE}} = \lambda_1 \cdot \mathcal{L}_{\text{CE}}(W_{\text{old}}, \phi_{\text{new}}(\mathbf{x})) + \lambda_2 \cdot \mathcal{L}_\angle \tag{3}$$

where $\lambda_1$ and $\lambda_2$ are two weighting factors, $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss defined as

$$\mathcal{L}_{\text{CE}} = - \sum_{(\mathbf{x},y)\in B} \sum_{i=1}^{C_{\text{new}}} y_i \log \left( \frac{\exp(\phi_{\text{new}}(\mathbf{x}) \cdot W_{\text{old}}^{(i)})}{\sum_{j=1}^{C_{\text{new}}} \exp(\phi_{\text{new}}(\mathbf{x}) \cdot W_{\text{old}}^{(j)})} \right) \tag{4}$$
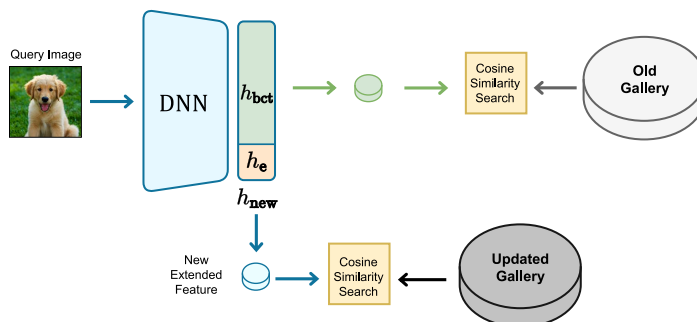
and $\mathcal{L}_\angle$ is defined as

$$\mathcal{L}_\angle = - \sum_{(\mathbf{x},y)\in B} \left( 1 - \frac{\phi_{\text{new}}(\mathbf{x}) \cdot W_{\text{old}}^{(y)}}{\|\phi_{\text{new}}(\mathbf{x})\|\|W_{\text{old}}^{(y)}\|} \right) \tag{5}$$

on the mini-batch $B$. However, there is an inherent trade-off [25] between training backward-compatible representations, which avoid the need for backfilling, and achieving the performance of a model trained directly on $\mathcal{D}_{\text{new}}$. This is because the new backward-compatible feature space $h_{\text{bct}}$ is constrained to align with the old one, which prevents the model from learning a new feature space structure that could accommodate all the new information from the latest data. Due to this, there is a drop in performance as the new model cannot properly assimilate the new knowledge from the incoming data. To address this issue while preserving the compatibility of the representations, we extend the feature space with extra dimensions $h_{\text{e}}$ alongside the aligned compatible embeddings $h_{\text{bct}}$. These additional dimensions $h_{\text{e}}$ are not optimized by the ACE loss since they are used to accommodate new data.

As demonstrated by [25], directly training the new model on the expanded feature space $h_{\text{new}} = [h_{\text{bct}}|h_{\text{e}}]$ can lead to incompatible representations. This is because training in a higher-dimensional space tends to alter the geometric structure of the previous space, even if $h_{\text{bct}}$ is constrained to align with the old one. Consequently, we apply an orthogonal transformation to $h_{\text{new}}$, resulting in $h_\perp$, which lies in a representation space with the same dimensionality and geometric configuration as $h_{\text{new}}$. This transformation is achieved through a learnable orthogonal transformation layer $T_\perp$, which is used to obtain the new orthogonal feature space $h_\perp = T_\perp(h_{\text{new}})$. We define the weight of the linear transformation $T_\perp$ as $Q$, making $T_\perp$ a learnable orthogonal transformation that ensures $T_\perp^\top T_\perp = I$, where $I$ is as the identity matrix. To constraint a fully-connected layer to learn such a transformation, we use a matrix $A$, where $A$ is any skew-symmetric matrix (so that $A^\top = -A$) with learnable parameters randomly initialized. The orthogonal matrix $Q$ is then obtained using the exponential map $Q = e^A$. Applying the orthogonal transformation $T_\perp$ to any representation preserves all geometrical information and maintains the quality of the representations, as it holds that:

$$h_i^\top h_j = T_\perp(h_i)^\top T_\perp(h_j)$$

where $h_i$ and $h_j$ are two generic learned representations.

**Fig. 2: Overview of our method at inference time.** The DNN backbone model produces representations within a feature space $h_{\text{new}}$. This space is divided into two parts: $h_{\text{btc}}$ is the compatible representation space. Its representations are used to perform visual search directly with the old gallery features without using the orthogonal transformation function that we discard after training. Representations $h_{\text{new}} = [h_{\text{bct}}|h_{\text{e}}]$ are instead used to match with the updated gallery to be as close as possible to the performance of the independently trained version of the new model.

The transformed embeddings $h_{\perp}$ are finally optimized through the learned classifier $W$ to learn incoming information from the new data. The orthogonal transformation, due to its orthogonal column constraint imposed by $T_{\perp}^{\top} T_{\perp} = I$, ensures that the angles and norms of the input space are preserved, thereby maintaining the previous geometry in the learned feature space $h_{\text{new}}$, especially for the compatible learned subspace $h_{\text{bct}}$. This constrains the model to learn new information in the additional space $h_{\text{e}}$ without disrupting the geometric structure of $h_{\text{bct}}$. The orthogonal transformation layer ensures that the geometry of the compatible learned feature space $h_{\text{bct}}$ remains unchanged. The cross-entropy loss, computed using the new classifier $W$, refines the new knowledge in the extra embedding dimensions $h_{\text{e}}$.

The overall loss used in our approach is thus:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(W, \phi_{\text{new}}(\mathbf{x})) + \mathcal{L}_{\text{ACE}} \tag{6}$$

It is worth noticing that, after learning, we use only $h_{\text{new}}$ at inference time to perform image search/retrieval as shown in Fig. 2, while $T_{\perp}$, $h_{\perp}$, and the new classifier $W$ are discarded as they are used only to learn additional knowledge from the new data $\mathcal{D}_{\text{new}}$.

## 4    Experimental Results

In this section, we present our experimental results that (1) evaluate our approach alongside established backward compatible representation learning techniques according to the criteria outlined in Eq. 2, (2) examine our approach's capacity to address incremental scenarios in data, or what [18] identified as open classes (for example, where the previous model was trained on 500 ImageNet classes while the updated model uses 1000 ImageNet classes)

### 4.1   Datasets

In this paper, we make use of the following datasets:

– **CIFAR-100** [10]: It is a small image classification dataset of 100 classes divided into 50000 images of the training set and 10000 of the test set. We will refer to CIFAR-50 as the partition consisting of all the samples from the first 50 classes.
– **ImageNet-1k** [7]: It is a large-scale image recognition dataset proposed for the ILSVRC 2012 challenge. It has 1000 image classes with about 1k images per class. We follow the same partitioning as in [14]. We consider the images from the first 500 classes as ImageNet-500.

### 4.2   Evaluation Metrics

**Mean Average Precision (mAP):** The mAP is a standard metric in compatibility, where precision and recall evaluations are summarized by calculating the area under the precision-recall curve. The average precision within the recall interval [0.0, 1.0] is computed.

**Cumulative Matching Characteristics (CMC):** CMC refers to the top-k accuracy, where gallery representations are ranked based on their similarity to the query representation. A match is correct if a representation of the same class appears within the first k gallery entries. We report CMC-1 (top-1 accuracy) for all models.

We construct a distance matrix between all the representations obtained from the query and gallery sets under consideration. We use the cosine similarity between two feature vectors as the measure of distance. On top of this matrix, we evaluate the mAP and CMC-1 metrics. During the retrieval process, for a method that utilizes additional dimensions, it is relatively simple to differentiate between the gallery samples featuring the old representations and those displaying the concatenated representations, thanks to the difference in size. We zero-pad the old representations stored in the gallery during comparison with the new representations.

### 4.3   Compared Methods

We compare our method OCA against the following approaches:

– **Independent**: $\phi_{\mathrm{new}}^{\mathrm{I}}$ is trained from scratch with new data, without taking into consideration any backward compatibility method.
– **BCT** [18]: It is a widely adopted baseline in recent studies [14]. BCT employs a classification loss regularized by an "influence loss" during the training process to ensure backward compatibility. In BCT, denoted $W$ as the trainable classifier of $\phi_{\mathrm{new}}^{\mathrm{BTC}}$, and $W_{\mathrm{old}}$ as the fixed old classifier obtained by the old representation $\phi_{\mathrm{old}}$, the loss function comprises two terms:

$$\mathcal{L}_{\mathrm{BCT}}(\phi, W, \mathbf{x}) = \mathcal{L}_{\mathrm{CE}}(W, \phi|\mathbf{x}) + \lambda\mathcal{L}_{\mathrm{CE}}(W_{\mathrm{old}}, \phi|\mathbf{x}),$$

where $\lambda$ is a hyperparameter that weights the influence loss.

**Table 1:** Results of the mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) metrics from trials carried out on CIFAR-50 and CIFAR-100 datasets. These experiments utilized the Resnet50-128 architecture in both old and new models. The / symbol distinguishes the model that processes the query (left of /) from the model that processes the gallery set (right of /).

| Method | Case | mAP@1.0 | CMC-1 |
|---|---|---|---|
| Initial Model | $\phi_{old}/\phi_{old}$ | 23.32 | 31.32 |
| Independent | $\phi_{new}^{I}/\phi_{old}$ | 01.29 | 01.02 |
|  | $\phi_{new}^{I}/\phi_{new}^{I}$ | 45.35 | 56.75 |
| BCT [18] | $\phi_{new}^{BCT}/\phi_{old}$ | 25.14 | 36.73 |
|  | $\phi_{new}^{BCT}/\phi_{new}^{BCT}$ | 43.89 | 54.76 |
| BT$^2$ [25] | $\phi_{new}^{BT^2}/\phi_{old}$ | 26.05 | 38.64 |
|  | $\phi_{new}^{BT^2}/\phi_{new}^{BT^2}$ | 50.36 | 61.77 |
| **OCA** | $\phi_{new}^{OCA}/\phi_{old}$ | **26.35** | **41.37** |
|  | $\phi_{new}^{OCA}/\phi_{new}^{OCA}$ | **52.06** | **62.02** |

– **BT$^2$** [25]: It employs an embedding dimension expansion of 32, utilizing a combination of cosine similarity loss and BCT influence loss for matching $\phi_{old}$. Additionally, it enhances the learned representation by matching $\phi_{new}^{I}$ with an additional cosine similarity loss. The supplementary feature space is learned through two trainable changes of basis to prevent the introduction of information that may disrupt the compatible learned information.

### 4.4 Implementations Details

All baselines and our method employ ResNet50 models with an initial embedding size of 128 as backbone, trained using the Adam optimizer, with a learning rate of 0.001 and a batch size of 128 over 100 epochs. For our method, we expand the embedding dimension by 32 and set $\lambda_1 = 10$ and $\lambda_2 = 5$.

### 4.5 Experimental Results

**CIFAR-50 to CIFAR-100.** For this experiment, the $\phi_{old}$ model is trained using the CIFAR-50 dataset, while the $\phi_{new}$ model is trained on the complete CIFAR-100 dataset. We conduct retrieval tasks for metric evaluation on the CIFAR-100 validation set, which serves as both the gallery and query sets.

**ImageNet-500 to ImageNet-1k.** The initial model $\phi_{old}$ is trained with the ImageNet-500 dataset, followed by training of $\phi_{new}$ using the complete ImageNet-1k dataset. In our retrieval process evaluation, we use the ImageNet-1k validation set for both gallery and query purposes.

The results are shown in Table 1 and 2 for CIFAR-100 and ImageNet-1k, respectively. We observe that Independent training is the only method that fails to achieve compatibility with the gallery extracted by $\phi_{old}$ because it does not implement any compatible learning strategy. In contrast, BCT achieves compatibility of $\phi_{new}^{BCT}$ with

**Table 2:** Results of the mean Average Precision (mAP) and Cumulative Match Characteristic (CMC) metrics from trials carried out on ImageNet-500 and ImageNet-1k datasets. These experiments utilized the Resnet50-128 architecture in both old and new models. The / symbol distinguishes the model that processes the gallery (left of /) from the model that processes the query set (right of /).

| Method | Case | mAP@1.0 | CMC-1 |
|---|---|---|---|
| Initial Model | $\phi_{\text{old}}/\phi_{\text{old}}$ | 32.33 | 43.53 |
| Independent | $\phi_{\text{new}}^{\text{I}}/\phi_{\text{old}}$ | 00.14 | 00.12 |
| | $\phi_{\text{new}}^{\text{I}}/\phi_{\text{new}}^{\text{I}}$ | 55.53 | 69.11 |
| BCT [18] | $\phi_{\text{new}}^{\text{BCT}}/\phi_{\text{old}}$ | 35.81 | 48.56 |
| | $\phi_{\text{new}}^{\text{BCT}}/\phi_{\text{new}}^{\text{BCT}}$ | 54.44 | 67.57 |
| BT$^2$ [25] | $\phi_{\text{new}}^{\text{BT}^2}/\phi_{\text{old}}$ | 36.55 | 50.21 |
| | $\phi_{\text{new}}^{\text{BT}^2}/\phi_{\text{new}}^{\text{BT}^2}$ | 55.65 | 67.75 |
| **OCA** | $\phi_{\text{new}}^{\text{OCA}}/\phi_{\text{old}}$ | **36.71** | **50.73** |
| | $\phi_{\text{new}}^{\text{OCA}}/\phi_{\text{new}}^{\text{OCA}}$ | **56.82** | **69.73** |

the old gallery representations, but it shows a reduction in both performance metrics compared to Independent training when evaluating new query representations against new gallery representations, which aligns with the findings in [25]. BT$^2$ improves compatibility performance compared to BCT, thanks to an additional embedding dimension and the increased number of parameters provided by the two orthogonal matrices for the basis changes and a previously trained Independent model. Our method achieves the best results on both datasets, demonstrating its potential. Compared to BT$^2$, we add fewer parameters to the model because our Orthogonal Transformation layer and the auxiliary classifier are used only during training and are then completely removed. This allows us to leverage all the new information without hurting compatibility and increases model generalization, achieving also better performance than an Independent model when evaluating new query representations against new gallery representations.

### 4.6    Ablation Studies

In the following, we present ablation studies on the influence of the dimensionality of the extra space $h_{\text{e}}$ and each component of our method on compatible training on the CIFAR-100 dataset. Table 3 shows the results of our method compared to the BT$^2$ strategy in handling the additional representation space $h_{\text{e}}$. We observe that our method does not suffer from increased dimensionality of $h_{\text{e}}$, achieving consistent results across different sizes. Furthermore, we notice that increasing the dimensionality of the extra space allows the model to assimilate more new information, thereby achieving better results on both metrics when using new representations for both query and gallery sets. Instead, BT$^2$ struggles to learn compatible representations as the size of $h_{\text{e}}$ increases. This is related to their model's change of basis architecture and the cosine loss that tries to match the geometrical structure of an independently trained model. This demonstrates the stability of our method in managing the new extra space compared to BT$^2$.

**Table 3:** Ablation on the influence of the extra space $h_e$ dimentionality to the compatible training. Results of the mean Average Precision (mAP) and Cumulative Match Characteristic (CMC) metrics from trials carried out on CIFAR-50 and CIFAR-100 datasets. These experiments utilized the Resnet50-128 architecture in both old and new models. The / symbol distinguishes the model that processes the gallery (left of /) from the model that processes the query set (right of /).

| Ext. Dim. | Method | Case | mAP@1.0 | CMC-1 |
|---|---|---|---|---|
|  | Initial Model | $\phi_{\text{old}}/\phi_{\text{old}}$ | 23.32 | 31.32 |
|  | Independent | $\phi_{\text{new}}^{\text{I}}/\phi_{\text{old}}$ | 01.29 | 01.02 |
|  |  | $\phi_{\text{new}}^{\text{I}}/\phi_{\text{new}}^{\text{I}}$ | 45.35 | 56.75 |
| +1 | BT$^2$ [25] | $\phi_{\text{new}}^{\text{BT}^2}/\phi_{\text{old}}$ | 25.58 | 34.79 |
|  |  | $\phi_{\text{new}}^{\text{BT}^2}/\phi_{\text{new}}^{\text{BT}^2}$ | 44.42 | 59.84 |
|  | OCA | $\phi_{\text{new}}^{\text{OCA}}/\phi_{\text{old}}$ | 27.09 | 43.95 |
|  |  | $\phi_{\text{new}}^{\text{OCA}}/\phi_{\text{new}}^{\text{OCA}}$ | 50.87 | 61.04 |
| +32 | BT$^2$ [25] | $\phi_{\text{new}}^{\text{BT}^2}/\phi_{\text{old}}$ | 26.05 | 38.64 |
|  |  | $\phi_{\text{new}}^{\text{BT}^2}/\phi_{\text{new}}^{\text{BT}^2}$ | 50.36 | 61.77 |
|  | OCA | $\phi_{\text{new}}^{\text{OCA}}/\phi_{\text{old}}$ | 26.35 | 41.37 |
|  |  | $\phi_{\text{new}}^{\text{OCA}}/\phi_{\text{new}}^{\text{OCA}}$ | 52.06 | 62.02 |
| +64 | BT$^2$ [25] | $\phi_{\text{new}}^{\text{BT}^2}/\phi_{\text{old}}$ | 22.60 | 24.17 |
|  |  | $\phi_{\text{new}}^{\text{BT}^2}/\phi_{\text{new}}^{\text{BT}^2}$ | 50.36 | 62.87 |
|  | OCA | $\phi_{\text{new}}^{\text{OCA}}/\phi_{\text{old}}$ | 26.76 | 42.26 |
|  |  | $\phi_{\text{new}}^{\text{OCA}}/\phi_{\text{new}}^{\text{OCA}}$ | 51.69 | 61.03 |
| +128 | BT$^2$ [25] | $\phi_{\text{new}}^{\text{BT}^2}/\phi_{\text{old}}$ | 12.43 | 08.98 |
|  |  | $\phi_{\text{new}}^{\text{BT}^2}/\phi_{\text{new}}^{\text{BT}^2}$ | 48.64 | 62.37 |
|  | OCA | $\phi_{\text{new}}^{\text{OCA}}/\phi_{\text{old}}$ | 26.19 | 40.65 |
|  |  | $\phi_{\text{new}}^{\text{OCA}}/\phi_{\text{new}}^{\text{OCA}}$ | 52.12 | 61.82 |

We present in Table 4 the results of our method with each component turned on and off. The results show that adding the cosine distance loss $\mathcal{L}_{\angle}$ to the BCT head improves performance by directly inducing alignment of the newly learned representations with the old class representation prototypes. The orthogonality in the transformation $T_{\perp}$ also helps the model avoid disruption of compatible representations and inject new knowledge into the extra space, compared to a linear layer without orthogonality. When used together, the orthogonality of $T_{\perp}$ and $\mathcal{L}_{\angle}$ lead to state-of-the-art performance on the CIFAR-100 dataset.

## 5    Conclusion

This paper introduced an approach to manage the challenges associated with updating models in visual retrieval systems, particularly the need for backward compatibility and the high costs of backfilling. By expanding the feature space and applying an orthogonal transformation, our method allows for the integration of new information while maintaining compatibility with older models.

**Table 4:** Ablation study on the effect of orthogonality in $T_\perp$ and the cosine loss $\mathcal{L}_\angle$ in our method within a compatible learning setting. Results of the mean Average Precision (mAP) and Cumulative Match Characteristic (CMC) metrics are obtained from trials conducted on the CIFAR-50 and CIFAR-100 datasets. These experiments utilized the Resnet50-128 architecture for both the old and new models.

| Method | Case | mAP@1.0 | CMC-1 |
|---|---|---|---|
| Initial Model | $\phi_{\mathrm{old}}/\phi_{\mathrm{old}}$ | 23.32 | 31.32 |
| Independent | $\phi_{\mathrm{new}}^{\mathrm{I}}/\phi_{\mathrm{old}}$ | 01.29 | 01.02 |
| | $\phi_{\mathrm{new}}^{\mathrm{I}}/\phi_{\mathrm{new}}^{\mathrm{I}}$ | 45.35 | 56.75 |
| OCA w/o $T_\perp$, w/o $\mathcal{L}_\angle$ | $\phi_{\mathrm{new}}^{\mathrm{OCA}}/\phi_{\mathrm{old}}$ | 22.76 | 40.11 |
| | $\phi_{\mathrm{new}}^{\mathrm{OCA}}/\phi_{\mathrm{new}}^{\mathrm{OCA}}$ | 48.54 | 59.78 |
| OCA w/o $T_\perp$ | $\phi_{\mathrm{new}}^{\mathrm{OCA}}/\phi_{\mathrm{old}}$ | 25.89 | 40.11 |
| | $\phi_{\mathrm{new}}^{\mathrm{OCA}}/\phi_{\mathrm{new}}^{\mathrm{OCA}}$ | 50.23 | 61.18 |
| OCA w/o $\mathcal{L}_\angle$ | $\phi_{\mathrm{new}}^{\mathrm{OCA}}/\phi_{\mathrm{old}}$ | 26.13 | 40.34 |
| | $\phi_{\mathrm{new}}^{\mathrm{OCA}}/\phi_{\mathrm{new}}^{\mathrm{OCA}}$ | 51.06 | 61.90 |
| OCA | $\phi_{\mathrm{new}}^{\mathrm{OCA}}/\phi_{\mathrm{old}}$ | **26.35** | **41.37** |
| | $\phi_{\mathrm{new}}^{\mathrm{OCA}}/\phi_{\mathrm{new}}^{\mathrm{OCA}}$ | **52.06** | **62.02** |

Our approach has demonstrated the potential to reduce operational complexities and costs traditionally involved with model updates, particularly in large-scale image datasets. The effectiveness of this method was assessed using the CIFAR-100 and ImageNet-1k datasets, where it was found to maintain compatibility and improve accuracy compared to existing methods.

# References

1. Biondi, N., Pernici, F., Bruni, M., Del Bimbo, A.: Cores: Compatible representations via stationarity. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 2, 4, 5
2. Biondi, N., Pernici, F., Bruni, M., Mugnai, D., Bimbo, A.D.: Cl2r: Compatible lifelong learning representations. ACM Transactions on Multimedia Computing, Communications and Applications **18**(2s), 1–22 (2023) 4, 5
3. Biondi, N., Pernici, F., Ricci, S., Del Bimbo, A.: Stationary representations: Optimally approximating compatibility and implications for improved model replacements. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 2, 4, 5
4. Budnik, M., Avrithis, Y.: Asymmetric metric learning for knowledge transfer. CVPR (2021) 4
5. Chen, K., Wu, Y., Qin, H., Liang, D., Liu, X., Yan, J.: R3 adversarial network for cross model face recognition. In: CVPR. pp. 9868–9876. Computer Vision Foundation / IEEE (2019) 3
6. Cui, Z., Zhou, J., Wang, X., Zhu, M., Peng, Y.: Learning continual compatible representation for re-indexing free lifelong person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16614–16623 (June 2024) 4

7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 8

8. Duggal, R., Zhou, H., Yang, S., Xiong, Y., Xia, W., Tu, Z., Soatto, S.: Compatibility-aware heterogeneous visual search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10723–10732 (2021) 4

9. Iscen, A., Zhang, J., Lazebnik, S., Schmid, C.: Memory-efficient incremental learning through feature adaptation. In: European Conference on Computer Vision. pp. 699–715. Springer (2020) 4

10. Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images. Technical report, Univ. Toronto (2009) 8

11. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation, vol. 24, pp. 109–165. Elsevier (1989) 4

12. Meng, Q., Zhang, C., Xu, X., Zhou, F.: Learning compatible embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9939–9948 (2021) 2, 3, 4, 5

13. Pernici, F., Bruni, M., Baecchi, C., Del Bimbo, A.: Regular polytope networks. IEEE Transactions on Neural Networks and Learning Systems (2021) 4

14. Ramanujan, V., Vasu, P.K.A., Farhadi, A., Tuzel, O., Pouransari, H.: Forward compatible training for large-scale embedding retrieval systems. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19386–19395 (2022) 2, 4, 8

15. Ramanujan, V., Vasu, P.K.A., Farhadi, A., Tuzel, O., Pouransari, H.: Forward compatible training for representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 3

16. Robins, A.: Catastrophic forgetting in neural networks: the role of rehearsal mechanisms. In: Proceedings 1993 The First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems. pp. 65–68. IEEE (1993) 4

17. Seo, S., Uzunbas, M.G., Han, B., Cao, S., Zhang, J., Tian, T., Lim, S.N.: Online backfilling with no regret for large-scale image retrieval. arXiv preprint arXiv:2301.03767 (2023) 4

18. Shen, Y., Xiong, Y., Xia, W., Soatto, S.: Towards backward-compatible representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6368–6377 (2020) 1, 2, 3, 4, 5, 7, 8, 9, 10

19. Wan, T.S.T., Chen, J.C., Wu, T.Y., Chen, C.S.: Continual learning for visual search with backward consistent feature embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16702–16711 (June 2022) 4

20. Wang, C., Chang, Y., Yang, S., Chen, D., Lai, S.: Unified representation learning for cross model compatibility. In: 31st British Machine Vision Conference 2020, BMVC 2020. BMVA Press (2020) 2, 3, 4, 5

21. Yan, S., Xiong, Y., Kundu, K., Yang, S., Deng, S., Wang, M., Xia, W., Soatto, S.: Positive-congruent training: Towards regression-free model updates. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14299–14308 (2021) 4

22. Yang, Y., Chen, S., Li, X., Xie, L., Lin, Z., Tao, D.: Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? Advances in Neural Information Processing Systems **35**, 37991–38002 (2022) 5

23. Zhang, B., Ge, Y., Shen, Y., Li, Y., Yuan, C., XU, X., Wang, Y., Shan, Y.: Hot-refresh model upgrades with regression-free compatible training in image retrieval. In: International Conference on Learning Representations (2021) 2, 3, 4

24. Zhang, B., Ge, Y., Shen, Y., Su, S., Yuan, C., Xu, X., Wang, Y., Shan, Y.: Towards universal backward-compatible representation learning. arXiv preprint arXiv:2203.01583 (2022) 2, 3

25. Zhou, Y., Li, Z., Shrivastava, A., Zhao, H., Torralba, A., Tian, T., Lim, S.N.: Btˆ 2: Backward-compatible training with basis transformation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11229–11238 (2023) 2, 4, 6, 9, 10, 11