



OPEN A novel competitive learning prototype for image ordinal classification

Chao Zhang^{1,2}, Chao Feng³, Jianmei Cheng^{1,2}✉ & Keren Fu⁴

Image ordinal classification (IOC) assigns a discrete yet ordinal scalar label to an input image, such as age estimation. Learning similarity metric, as a fine separation tool, is a prevailing solution to this problem. However, it is non-trivial to effectively incorporate the ordinal information of inputs into the classification task, and it is overlooked recently. In this work, we adopt the concept of Mixing techniques, including Mosaic, CutMix and Mixup, which are powerful data augmentation strategies that blend two or more samples and their corresponding labels in a proportional manner during the training process, to bolster the task of ordinal classification. In particular, we delve into the mixing policy from the perspective of metric learning instead of data augmentation, which naturally and simply implements classification learning and difference regression learning, and propose a novel Competitive Learning (CL) prototype for IOC. Firstly, we embed the ordinal measurements of two input images into the training procedure by a virtual combination of image and label, to enhance the discriminative power of local features. We also conduct an in-depth analysis on the impact and advantages associated with implementing mixing policy. Secondly, the difference between two ordinal values is utilized to help measure the subtle difference in IOC, which is integrated with mixing policy properly. Finally, for further improvement, dual augmentation and random augmentation are proposed to enlarge and diversify the disparity of two objects, so as to exploit more robust feature. Comprehensive experiments on two ordinal classification tasks, i.e., age estimation and car date estimation, demonstrate the effectiveness and robustness of our novel approach, which outperforms previous methods with a relatively large margin.

Image ordinal classification (IOC) involves predicting a discrete target value that exhibits an ordering correlation among image categories. A quintessential example of this is age estimation, where one individual's age can be unambiguously determined to be younger or older than another's. The fundamental distinction between IOC and conventional recognition tasks lies in whether the input data encompasses ordinal information across all samples. When IOC is approached using a standard classification framework, the critical ordinal relationships among samples or categories are overlooked, leading to suboptimal performance. As illustrated in Fig. 1a, the distinction between the three classes - dog, cat and bear - is unrelated to their unique identity labels. Conversely, in Fig. 1b, the ordinal relationship among labels becomes a pivotal factor in both feature learning and classifier training. Despite the remarkable success of deep CNNs and Transformers^{1–4} in general recognition tasks over the past decade, the unique characteristics of ordinal information in IOC problems have largely been neglected.

How to fully use ordinal relationship in IOC is of utmost importance, and a model with a more refined separation capability is much necessary. Metric learning/CL with general pairing is exactly a promising solution to this problem, see Fig. 1c. However, we have observed that conventional metric learning methods are often too fragile due to their inflexible structure, see Fig. 1c, which heavily relies on an ideal sample-pairing strategy within the input module. In other words, only an effective sampling policy can lead to the desired performance in metric learning. In essence, metric learning is to learn distance measurement in which some similar inputs always belong to the same class while examples from different classes are separated by a large margin. Conventional methods mainly focus on the embedding space, and train a Siamese or Triplet network that still suffers limited pairing and soft estimating operation. The sampling and soft estimating strategy is crucial for metric learning.

A natural question arises: Does a model that can emulate efficient and effective metric learning function so as to learning a ideal distance metric in the feature space such that similar samples are close to each other

¹Department of Traffic Engineering, Sichuan Police College, Luzhou 646000, China. ²Intelligent Policing Key Laboratory of Sichuan Province, Luzhou 646000, China. ³Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor 48109, USA. ⁴School of Computer Science, Sichuan University, Chengdu 610065, China. ✉email: freecjm2003@scpolicec.edu.cn

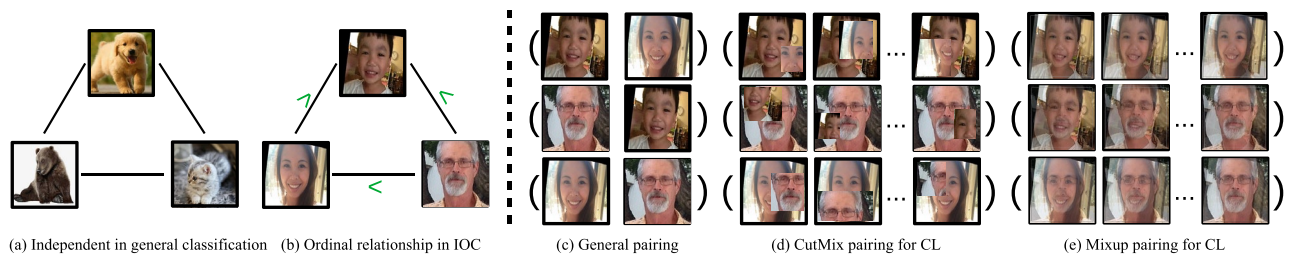


Figure 1. Motivation. (a) In the general classification task, targets are categorized into distinct classes with no inherent ordinal relationship among them. (b) The image ordinal classification task, however, requires classifying targets based on their ranked ordering. A key challenge lies in effectively leveraging this ordinal information to improve model performance, as conventional methods often struggle to capture the subtle differences between ordered categories. (c) Traditional pairing strategies for metric learning, such as contrastive or triplet loss, are designed to minimize intra-class variance and maximize inter-class margins. However, generating diverse image-pairs or patch combinations for training is computationally demanding and impractical, especially in scenarios requiring fine-grained discrimination. (d,e) Unlike general pairing, CutMix/Mixup can also serve as scalable and efficient pairing mechanism by dynamically blending image pairs or patches with stochastic mixing ratios and pairing possibilities, thereby generating diverse hybrid training samples. This diversity facilitates effective CL in IOC. (All the face images in this paper are from literature ⁵).

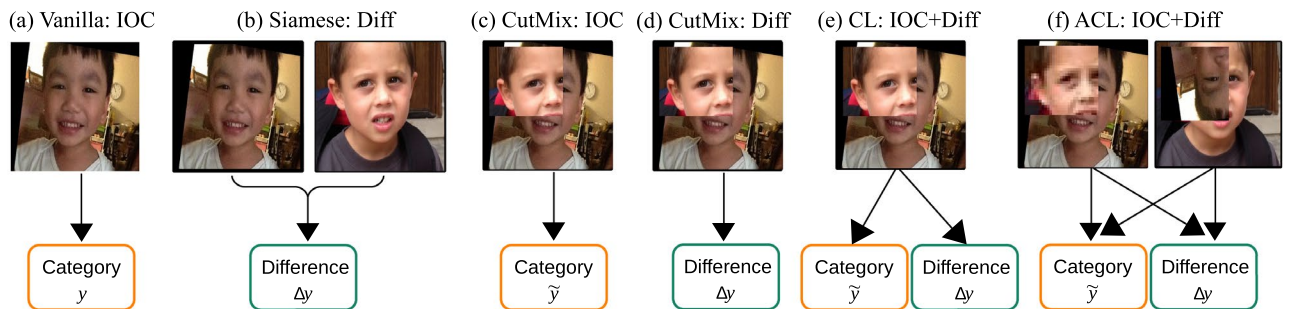


Figure 2. The overall architecture comparisons of various visual representation learning techniques for IOC. (a) The baseline typically employs vanilla CNN to solve IOC task. (b) Classical metric learning often adopts a Siamese architecture that takes two images as input and outputs ordinal difference Δy . (c) CutMix⁶ combines input images by overlaying one onto the other and merges their labels based on a mixing ratio to form a virtual label $\tilde{y}$, and then it only needs to process a single composite image. (d) CutMix can also be employed for metric learning to assess the dissimilarity or similarity between two images. This approach serves the same function as the Siamese architecture but achieves more efficient and effective pairing and learning. (e) Apart from category task $\tilde{y}$, we further supplement the task of label difference Δy to help better ordinal classification using CL. (f) ACL adopts extra augmentation method including dual augmentation and random augmentation, aiming to amplify the distinctions between the two image patches.

while dissimilar samples are far apart? Unlike coarse-grained object classification, effective pairing strategies must be employed in IOC to discern subtle differences between closely related ordinal categories. In fact, some traditional pairing strategies for metric learning is inflexible and inefficient (see Figs. 1 and 2). A kind of data augmentation method, CutMix⁶/Mixup⁷/Mosaic⁸, meets this requirement exactly. It's worth noting that both CutMix and Mixup can serve a similar purpose. As shown in Figs. 1d and 2e, CutMix randomly occludes part of the object and then combines two images, while Mixup overlays two images together in a certain proportion. In the subsequent section, we will solely focus on introducing the application of CutMix. Specifically, we interpret CutMix from a different perspective. It is a regularization method that implements soft estimation, but also be seen as an effective pairing strategy for metric learning. CutMix introduces a stochastic mixing mechanism that enables the input module to generate a diverse range of cropping and pairing combinations.

In this work, we first delve into the interpretation of CutMix from different perspectives. On the one hand, it can be viewed as a regularization technique that randomly dropouts patch with a specific ratio and paste with patch from another image. On the other hand, it can also be viewed as a competitive learning technique in IOC since it facilitates learning more discriminative feature from convex virtual image and label. This unique combination fosters a competitive learning environment between two objects, and compelling the model to meticulously learn nearly all spatial information, rather than merely focusing on the most discriminative regions.

In addition, CutMix uses combination ratio λ to mix two image patches, which introduces the stochasticity to help input module produce various patch combinations and pairing combinations.

Besides reinterpreting and implementing Mixing technique from a different view, we further propose two modified approaches for IOC (i.e., Competitive Learning (CL), and Augmented CL (ACL), which fully leverage the order information among images and labels. When compared to general classification task, as seen in Figs. 1 and 2, the label y in IOC possesses two attributes: category information and order information. The former comprises several rough clusters, with the label y merely assigned independent marks. Unlike category label being independent in classification, ordinal labels delineate neighboring ordinal relationships, offering more granular insights. According to the literature⁹, the coarse category information provides a foundation for the ordinal information, while simultaneously benefiting from the refined ordinal details.

Based on the discussion above, the ordinal information is not being fully leveraged. The label y of image x is not merely as a category mark, but also carries ordinal information. So the ordinal difference $\Delta y = |y - y'|$ between two samples (x, x') possesses an intuitive and meaningful interpretation. It will enhance feature learning by exploiting ordinal information, and also facilitate classification task by learning more discriminative feature. We propose to utilize ordinal difference information as a pivotal pretext to bolster the classification task, as shown in Fig. 2d. Although virtual label $\tilde{y}$ is introduced in Mixing technique, it is Not actually calculated in the training process. In fact, there are two classification mappings in the implementation of Mixing technique: $\tilde{x} \rightarrow y$ and $\tilde{x} \rightarrow y'$. That is true for both CutMix and Mixup. Two mappings share a common CNN model controlled by two different loss functions, as depicted in Fig. 2e, where the two objects essentially compete with each other. This scenario gives rise to CL prototype.

To further advance metric learning, we introduce an augmented CL technique that integrates pairs of dual image patches. This approach fully harnesses the potential of both individual image patches and their dual counterparts, while also exploiting the functionality of mixing ratios. During each training iteration, the model is guided by two pairs of images, each associated with distinct mixing ratios λ and λ' , to learn more discriminative and robust embeddings.

In summary, we conclude the contributions as follows.

- By adopting Mixing technique in IOC, we find that it is not merely an efficient augmentation trick but also a natural metric learning approach. Moreover, this approach enables the generation of diverse patch combinations and pairing configurations at very low cost, thereby enhancing the robustness of model.
- We propose a CL prototype for IOC, which learns the ordinal difference of two objects in CutMix by an extra regression task. It helps to pull similar-class samples closer and push dissimilar samples apart.
- We empirically find that CL implements classification task and difference regression task both on the context feature learning and the flat feature learning, which is exactly superior to Siamese network on recognition performance, flexibility and memory requirement.
- To strength the disparity among samples, augmented competitive learning (ACL) is further proposed to feed dual patch pairs with random augmentation.

Related work

Image ordinal classification

Ordinal classification assigns input data with scalar value label that is generally discrete but carry the ordering information among categories. Many applications, such as age estimation^{10,11}, image quality estimation¹², aesthetic estimation¹³, facial action unit intensity estimation¹⁴, belong to the task of ordinal classification. In traditional methods^{15–19}, image ordinal estimation usually uses hand-crafted features and individual classifier or regressor. With the increasing popularity of deep learning, all recent works^{5,10,11} apply CNNs¹ to deal with IOC, and naturally achieve better performance. There are three directions towards this issue. First, collecting sufficient training data with ordinal label is hard compared with generic classification task, some previous works¹³ use patch-based method to improve recognition accuracy and reduce overfitting. Second, some works^{20–23} focus on model designing, i.e., utilizing modified base model and complementing some multi-scale and multi-layer network architecture. Third, some works design different loss functions^{24,25} for metric learning, such as triplet loss²⁶, weighted loss²⁷ and quadruplet loss²⁸.

Data augmentation

Data augmentation is essential to the success of many modern neural network approaches across numerous applications. For visual recognition task, data augmentation approaches can be divided into three categories: geometric transformation, dropout operation and linear combination. Typical geometric transformation methods operate on a single image and preserve its initial label for the transformed image, eg., image shifting/rotation/flipping and other affine transformation. Dropout includes two aspects: neuron dropout and patch dropout. The former is to regularize the networks by randomly suppressing some neurons or connection during training, eg. Dropout²⁹, DropConnect³⁰ and FastDrop³¹, etc. The latter is a regional dropout strategies such as Cutout³², Random Erasing³³. The linear combination methods^{8,34–36}, such as Mixup⁷, CutMix⁶ and Mosaic⁸, are to generate a virtual combination that consists of several training images, and its label is also a convex combination of several labels³⁷. This strategy randomly combines training images, which is a simple yet surprisingly effective improving method. Recently, a large variety of regional dropout strategies have been proposed by partially occluding the discriminative parts of objects. Motivated by Mixup⁷, CutMix⁶ is an augmentation strategy incorporating region-level replacement, in which patches from one image are randomly cut and pasted onto the other image along with the ground truth labels being mixed together proportionally to the area of patches. In this work, we take this method as an example to implement competitive learning.

Metric learning

Metric learning aims at learning distance measures in which similar inputs are mapped to nearby points on a manifold and dissimilar ones are mapped apart from each other. In metric learning, some similar inputs belong to the same class while examples from different classes are separated by a large margin. Many of the metric learning losses such as contrastive loss³⁸, triplet loss³⁹, Lifted structured loss⁴⁰, N-pair loss⁴¹ and so on, take pairs of samples to optimize the loss with the desired properties⁴². Bromley et al.⁴³ firstly propose Siamese network for metric learning on signature verification, and then they utilize this method in face verification⁴⁴ for its strong discrimination capability. Large margin nearest neighbour (LMNN)⁴⁵ is a statistical machine learning algorithm for metric learning, which learns a pseudometric designed for k-nearest neighbour classification. It is based on semidefinite programming, a sub-class of convex optimization. Song et al.⁴⁰ describe an algorithm for taking full advantage of the training batches in the neural network training by lifting the vector of pair-wise distances within the batch to the matrix of pairwise distances. The work⁴⁶ use the conception of triplet loss²⁶ to learn ranking function. Triplet loss defines a loss function based on triplets of images in which two from the same class and one from a different class, and penalizes the network if it does not map the same class inputs to be closer than the different classes⁴⁷. Xuan et al.⁴⁷ propose a novel, generalizable and fast method to define a family of embedding functions that can be used as an ensemble to give improved results. Each embedding function is learned by randomly bagging the training labels into small subsets. No Fuss Embeddings⁴⁸ use the output layer of classification networks to provide very useful embedding functions for one-shot learning and image retrieval tasks.

Adopting mixing on IOC

Before describing proposed approach on IOC, we intend to explain why Mixing strategy fits well on metric learning, not just be seen as a data augmentation technique. Then we will analyse that IOC and Mixing policy exhibit a high degree of compatibility and rationality. For ordinal recognition problem, mixing policy like CutMix/Mixup is expected to become a standard learning prototype due to its effectiveness and efficiency on exploiting ordinal relationship. In fact, this approach is also poised to emerge as a prototypical implementation for all metric learning problems. In the subsequent sections, without loss of generality, we will use CutMix as an example of Mixing policies.

Why does mixing fit IOC well?

IOC task tends to assign a discrete yet ordinal label to each image, in which the classes to be predicted are discrete and ordered in some fashion. For IOC, the feature should have the property of both class identity and ordinal ranking. Hence the key to IOC is to obtain both a better category feature space by classification learning and a similarity embedding space by ordinal regression learning, as shown in Fig. 2a,b. In general classification, the classes are not ordered and independent. The similarity embedding learning is necessary to maximize the inter-class variations and minimizes the intra-class variation. As shown in Fig. 2b,d, Siamese network and CutMix can achieve the same function on similarity embedding learning.

Siamese network seems proper to be designed for IOC tasks. However, there are three major issues: 1) computational memory overhead; 2) sampling strategy inefficiency; 3) model flexibility limitations. In the training procedure, multiple pairs of images are considered as input, and share same architecture and weights. The network is trained to output low similarity scores for similar pairs and high scores for dissimilar ones. In fact, utilizing Siamese network or triplet loss usually yields suboptimal result because mining hard triplet is always difficult and time-consuming.

Coincidentally, CutMix/Mixup satisfies these requirements properly, as shown in Fig. 2c–f, which perfectly fits the IOC task. Instead of removing randomly selected patches, it replaces the deleted region with a patch from another image. CutMix exploits much richer feature information among multiple samples, and is a natural pairing-cropping strategy, see Fig. 2c. In addition, from the perspective of learning, it also provides a smoother estimate of uncertainty among different classes/levels rather than hard separations.

IOC: CutMix classification

Following CutMix⁶, given two images $x_n, x_{n'}$ and their ordering ground-truth labels $y_n, y_{n'}$, the combined virtual training image and its virtual label can be denoted as

$$\tilde{x}_n = \mathbf{M} \odot x_n + (\mathbf{1} - \mathbf{M}) \odot x_{n'}, \quad \tilde{y}_n = \lambda y_n + (1 - \lambda) y_{n'}, \quad (1)$$

where $\mathbf{M}$ is a binary mask, $\tilde{x}_n$ and $\tilde{y}_n$ are the combined image and soft label. Soft label is critical for IOC to capture ordinal dependencies but is redundant in general recognition tasks, which treat classes as discrete and unrelated. Two images as well as their category labels are combined with a specific proportion λ , see Fig. 2c. From the perspective of data augmentation, CutMix strategy prevents CNN from focusing too much on a small number of intermediate neurons or on a small region of the input image. From the perspective of machine learning, it is a soft mapping technique.

For classification task, two mappings: $\tilde{x} \rightarrow y$ and $\tilde{x} \rightarrow y'$ are implemented. Two mappings share with the same CNN backbone model but the different loss functions. We represent the classification loss of CutMix as

$$L_{IOC} = \lambda L_{cla} + (1 - \lambda) L_{cla'} = \lambda L(\tilde{x}_n, y_n) + (1 - \lambda) L(\tilde{x}_n, y_{n'}) \quad (2)$$

where λ is combination ratio as well as the area ratio. The label of the training image in input module is soft, and does not have exact hard category. In fact, the virtual label $\tilde{y}_n$ defined in CutMix is Not computed in the

training process. From the perspective of classifier learning, it provides a smoother estimate of uncertainty among different classes rather than hard separations.

Analysis on mixing pairing and cropping

Compared with Vanilla CNN, CutMix's numerous pairing combinations play an important role in dataset diversity, which enables learning robust feature. In input procedure, every two samples are randomly selected from batch-size N_t images, to be mixed by the combination ration λ . There are $P_{N_t}^2 = \binom{N_t}{2}$ possibilities to combine these images in each iteration. For the cropping, there are also $(H \times W)^2$ possibilities on the cropping area, where H and W are the height and width of image, respectively. In total, there are $P_{N_t}^2 \times (H \times W)^2$ possibilities in each iteration, which largely diversifies the dataset and get more generalizable model.

Comparing with general metric learning model like Siamese network, the advantages of mixing technique are reflected in three aspects. Firstly, it has diversified a large number of competitive combinations, which can further enhance the generalization of model. Secondly, it considerably simplifies the implementation of metric learning, especially in the design of the inputting procedure. Thirdly, CutMix enables extensive neuron connections on both spatial context and flat feature (see Fig. 5).

Competitive learning prototype

In this section, we introduce CL Prototype in details, which is a double task learning architecture including category classification (IOC) and difference regression (Diff), see Fig. 2e. To help IOC, Diff with auxiliary ordinal difference is proposed to measure two images' ordinal distinction by CutMix strategy. In fact, CL consists of two parts: IOC and Diff. On one hand, besides the individual classification, two patches are competed with each other in CutMix. On the other hand, difference regression, when framed as a similarity learning task, inherently involves competitive learning mechanism.

Competitive learning: supplementing difference regression

CL has the same input module and classification function as CutMix. In the training, the input $\tilde{x}_n$ is a combined image, and the groundtruth $\Delta y_n = |y_n - y_{n'}|$ is ordinal difference. We tend to learn the ordinal difference of various image pairs. This can be seen as a competitive metric learning approach. As shown in Fig. 2, ordinal difference Δy_n between two images serves as an important affiliated cue that can help learning better feature and classifier. CutMix and CL have the same function for classification task, while CL additionally possesses the property of difference learning that measures two images' dissimilarity.

The contrastive ordinal loss L_{Diff} measures the difference of two images, and is implemented by L2 distance (MSE loss),

$$L_{Diff}(\Delta y_n, \Delta \hat{y}_n) = \|\Delta y_n - \Delta \hat{y}_n\|_2^2, \quad (3)$$

where Δy_n and $\Delta \hat{y}_n$ are the difference label and difference prediction, respectively.

Following the discussion above, CutMix and CL both implement IOC, while CL additionally implements difference regression (Diff) with MSE loss. The affiliated task Diff forces main task IOC learn more refined ordinal information, not just on the rough category. The total loss L_{CL} consisting of two parts: L_{IOC} and L_{Diff} , omitting parameters for simplicity of notation, is given by

$$\begin{aligned} L_{CL} &= L_{IOC} + \gamma L_{Diff} = L_{IOC}(\tilde{x}_n, \tilde{y}_n) + \gamma L_{Diff}(\Delta y_n, \Delta \hat{y}_n) \\ &= \lambda L(\tilde{x}_n, y_n) + (1 - \lambda) L(\tilde{x}_n, y_{n'}) + \gamma L_{Diff}(\Delta y_n, \Delta \hat{y}_n) \end{aligned} \quad (4)$$

where L and γ are the generic cross-entropy loss and the balance factor, respectively.

In the training process, the hyperparameter γ should not be set too large as this may undermine the performance of the main task of IOC. The second term γL_{Diff} is used to be an accessory, which helps learn much more discriminative feature. If the parameter γ is too large, the result of classification will be decreased obviously. We desire that the latter can help the former learn more discriminative feature. In order to jointly learn classification task and regression task, the dynamic balance factor of two losses is much essential. In later experiment, we set it as a dynamic value that is decreased gradually.

Implementation of competitive learning

Here we describe how to implement competitive learning in IOC. Though label of the virtual-combined image is represented as Eq. 4, the implementation on $\tilde{y}_n$ in the training is much different. This representation initially comes from the combination loss consisting of two separated ones with parameter λ in Eq. 2. In the implementation process, the label $\tilde{y}_n$ has Not been involved in the computation of classification loss L_{IOC} . Actually, there are two separated classification losses $L_{cla}, L_{cla'}$, based on only one feature. However, it corresponds to two ground-truth labels $y_n, y_{n'}$. After forwarding the whole network, the input image $\tilde{x}_n$ has two predictive outputs $\hat{y}_n, \hat{y}_{n'}$ which correspond to two mappings $L_{cla}(\tilde{x}_n, y_n), L_{cla'}(\tilde{x}_n, y_{n'})$. For two ground-truth labels, they share exactly the same network with the same fully connected feature. Competition between two labels $y_n, y_{n'}$ is clearly observed. When the ordinal difference $\Delta y_n = |y_n - y_{n'}|$ is utilized in CL and ACL, their competitive relationship will be further exploited. That is to say, both classification learning task and difference regression learning task are competitive. That is why we call our proposed method CL, especially on the adjective 'competitive'.

Comparison with Siamese networks

We compare vanilla CNN, Siamese network, CL from the point of recognition performance, flexibility and memory requirement. Here we provide some details on three Siamese networks. For the sake of fairness, we set two paths in the last fully connected layer for all three versions, as shown in Fig. 3a–c. They all include two loss terms: a CutMix classification term and an extra difference/verification term. The subtle difference among them lies in their different loss on the second term (see Fig. 3a–c). Comparing with Siamese network, CL is much easier and flexible while requiring lower memory overhead. Beyond its simplicity, it will also produce diversified pairing combinations. In experiment, fair comparisons are provided as realistic as possible.

Siamese-binary (Fig. 3a): Its second term is a binary classification task that judges whether two images belong to the same class by cross-entropy loss, instead of a difference regression task. The binary label $\tilde{y}$ is 1 or 0 according to whether two images belong to the same class or not. $\tilde{y} = 0$ if two images are deemed same label, $\tilde{y} = 1$ if they are deemed different label.

Siamese-difference (Fig. 3b): Its second term is same as difference loss $L_{Diff}(\Delta y_n, \Delta \hat{y}_n)$ based on difference label $\Delta y = |y - y'|$, as shown in Eq. 3 and Eq. 4. In fact, its total loss function is same as that of CL, with the sole distinction lying in the input procedure. In CL (Fig. 3d), the mixing policy, such as CutMix and Mixup, replaces the conventional twin-input approach in Siamese networks. From ablation study in Table 4, CL is exactly superior to Siamese network in learning the disparity between two objects.

Siamese-contrastive (Fig. 3c): Its second term is the contrastive loss function proposed in literature³⁸. Its definition in related with binary label $\tilde{y}$ can be denoted as follows,

$$L_{Diff} = (1 - \tilde{y}) \frac{1}{2} (D_W)^2 + \tilde{y} \frac{1}{2} \max\{0, m - D_W\}^2 \quad (5)$$

where D_W is the euclidean distance between two fully connected features, and $m > 0$ is a margin to define a radius around the features.

In Mixing policy, combination on two images constitutes the competitive learning between two original objects, and forces the model to learn almost all spatial information in detail instead of the most discriminative area. No pixel will be ignored during the training process because partial patch area will be dropped out with a certain percentage. Because there are two objects in a single image, an enormous number of neurons among different areas on the spatial are connected to implement competitive learning in CutMix. While for Siamese network, only limited neurons without spatial information are connected in the last fully connected layers. That is to say, CutMix implements competitive learning both on the spatial context (i.e., conv feature) and the concatenated neurons (i.e., fully connected feature). In addition, CutMix produces much more patch combinations and pairing combinations than that of Siamese network. It obviously obtains better recognition performance.

Augmented competitive learning

To further strengthen CL, we propose an extended version ACL that diversifies patch-pairing in CL by dual cropping-pasting and random augmentation, as shown in Fig. 4. ACL fully uses dual information without any intersection, and enlarges the visual difference between two objects.

Dual augmentation

Based on CL, we propose an extended version ACL that considers dual image patches and random augmentation, as shown in Fig. 4. In CutMix/CL, each image is divided into two parts: $x_n = \mathbf{M} \odot x_n + (\mathbf{1} - \mathbf{M}) \odot x_{n'}$, $x_{n'} = \mathbf{M} \odot x_{n'} + (\mathbf{1} - \mathbf{M}) \odot x_n$. There is no intersection between two combined image $\tilde{x}_n, \tilde{x}_{n'}$, i.e., $\tilde{x}_n \cap \tilde{x}_{n'} = \emptyset$. The dual information can be fully utilized. In training procedure, two combined images $\tilde{x}_n, \tilde{x}_{n'}$, are processed by a unified framework, where they share an identical CNN model, along with the same classification loss function and regression loss function. Here we denote cutmix classification loss as $L(\tilde{x}_n, y_n, y_{n'})$, then its dual loss can be represented as $L(\tilde{x}_{n'}, y_n, y_{n'})$. There is no appearance intersection, but we tend to build a representation relationship between them on CutMix classification loss. The

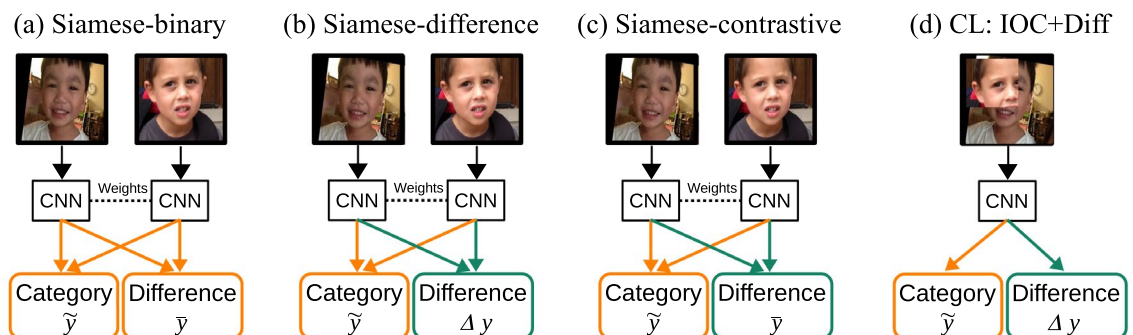


Figure 3. Three different Siamese networks and CL.

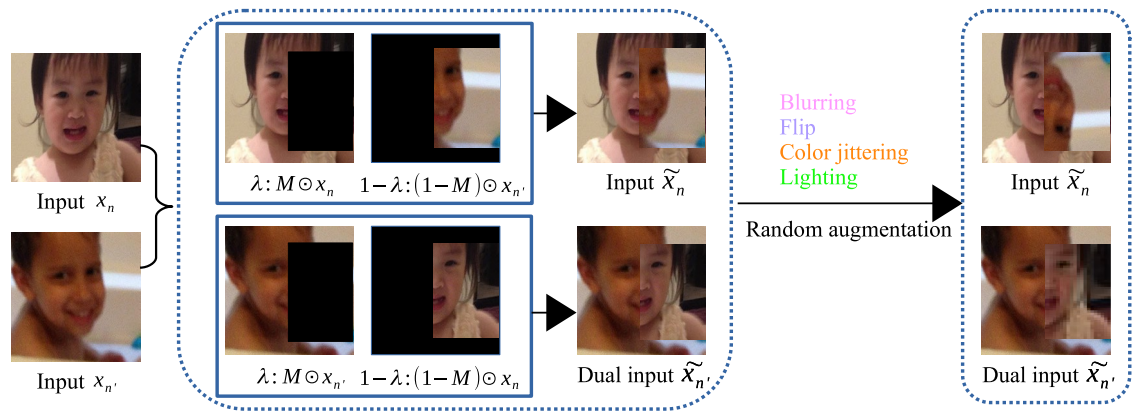


Figure 4. ACL pipeline. ACL consists of two steps: dual augmentation and random augmentation. Because no intersection is existed between two combined images, the dual information can be fully exploited. The disparity of two objects will be further diversified through random augmentation, thereby leading to a more intense competitive learning.

training on two mixed images could achieve better performance than that on the initial image or on a single mixed image.

The total classification loss L_{IOC_total} of this dual enhancement strategy can be formulated as follow:

$$L_{IOC_total} = L_{IOC} + L_{IOC_dual} = L(\tilde{x}_n, y_n, y_{n'}) + L(\tilde{x}_{n'}, y_n, y_{n'}) \quad (6)$$

According to Eq. 2 and Fig. 4, we might mistakenly believe that their summation is equivalent to the classification loss of two full images, because we have prior knowledge $\tilde{x}_n + \tilde{x}_{n'} = x_n + x_{n'}$, and the parts of two original images and combination images are exchanged.

$$\begin{aligned} \tilde{x}_n &= \mathbf{M} \odot x_n + (\mathbf{1} - \mathbf{M}) \odot x_{n'}, \\ \tilde{x}_{n'} &= \mathbf{M} \odot x_{n'} + (\mathbf{1} - \mathbf{M}) \odot x_n, \\ x_n &= \mathbf{M} \odot x_n + (\mathbf{1} - \mathbf{M}) \odot x_n, \\ x_{n'} &= \mathbf{M} \odot x_{n'} + (\mathbf{1} - \mathbf{M}) \odot x_{n'}, \\ \tilde{x}_n + \tilde{x}_{n'} &= x_n + x_{n'}. \end{aligned} \quad (7)$$

However, their corresponding loss can Not be transferred by the sum of different patches directly, as shown in Fig. 4. That is to say,

$$L(\tilde{x}_n, y_n, y_{n'}) + L(\tilde{x}_{n'}, y_n, y_{n'}) \neq L(x_n, y_n) + L(x_{n'}, y_{n'}) \quad (8)$$

Then the total classification loss of dual enhancement strategy can be represented as follows,

$$\begin{aligned} L_{IOC_total} &= L_{IOC} + L_{IOC_dual} = L(\tilde{x}_n, y_n, y_{n'}) + L(\tilde{x}_{n'}, y_n, y_{n'}) \\ &= \lambda L(\tilde{x}_n, y_n) + (1 - \lambda)L(\tilde{x}_n, y_{n'}) + \lambda L(\tilde{x}_{n'}, y_{n'}) + (1 - \lambda)L(\tilde{x}_{n'}, y_n) \\ &= \lambda L(\tilde{x}_n, y_n) + (1 - \lambda)L(\tilde{x}_{n'}, y_n) + \lambda L(\tilde{x}_{n'}, y_{n'}) + (1 - \lambda)L(\tilde{x}_n, y_{n'}) \end{aligned} \quad (9)$$

However, sum of classification loss on two full images can be denoted as follows,

$$L(x_n, y_n) + L(x_{n'}, y_{n'}) = \lambda L(x_n, y_n) + (1 - \lambda)L(x_n, y_n) + \lambda L(x_{n'}, y_{n'}) + (1 - \lambda)L(x_{n'}, y_{n'}) \quad (10)$$

Although $\tilde{x}_n + \tilde{x}_{n'} = x_n + x_{n'}$, but their corresponding losses are not equal, $L(\tilde{x}_n, y_n, y_{n'}) + L(\tilde{x}_{n'}, y_n, y_{n'}) \neq L(x_n, y_n) + L(x_{n'}, y_{n'})$. The training on virtual mixed image could achieve better and more robust performance than that on full image.

Random augmentation

In order to further enhance CNN model on CL, an image transformation set is built to diversify image patch pairs, which enables difference learning task to become much more harder. Widening the disparity between patch $\mathbf{M} \odot x_n$ and patch $(\mathbf{1} - \mathbf{M}) \odot x_{n'}$, as well as the disparity between $\mathbf{M} \odot x_{n'}$ and patch $(\mathbf{1} - \mathbf{M}) \odot x_n$, helps Cutmix classification and difference regression to extract robust features. Thus we construct an augmentation set $\mathcal{V}$ including four operations: blurring, horizontal/vertical flip, color jittering, lighting. We randomly select one for a specific patch, see horizontal/vertical flip and blurring in Fig. 4.

Experiments

This section demonstrates the effectiveness of CL and ACL under various settings on two applications: age estimation with Adience dataset (AdDb⁵) and date estimation on Car dataset (CarDb¹⁹). First, comparisons with SOTA approaches on two datasets are given. Then we implement CutMix on IOC and explain why CutMix fits IOC well from the experiment. To further validate CL and ACL, two ablation studies are conducted with various result analysis. Finally, we have analysed some examples and network visualization on feature map and CAM.

Datasets and criteria

Here two datasets (AdDb and CarDb) and two different evaluation metrics for IOC are introduced in detail.

AdDb⁵ includes roughly 26,000 images from 2,284 persons. All the images in Adience are unconstrained, which makes the experiment very challenging. It gives different levels y_n to each image, in which person ages are divided into eight levels $\{0, 1, \dots, 7\}$ (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60-), in terms of age value. Following the work⁵, training and testing on age estimation are performed using a standard five-fold cross validation: Cross0, Cross1, Cross2, Cross3, Cross4. We evaluate the capability of different models by recognition accuracy, some examples and visualization of CAM.

CarDb¹⁹ includes 13473 images in which cars are made from 1920 to 1999. Each image is labelled with an accurate production date that is a continuous value. In fact, the previous works except literature⁹ on this dataset carry out image estimation instead of image level classification. Following that work, all the dates in CarDb are divided into eight different decades, i.e., each group includes ten years $\{[i, i + 9] | 0, 10, \dots, 70\}$. Then the label in CarDb consists of 8 categories. For fair comparison, we adopt the commonly used splits in CarDb, in which 10343 images and 3340 images are used for training and testing.

In order to evaluate our proposed method, Top-1-Accuracy (Top-1-ACC) is the evaluation criteria for ordinal classification task. Moreover, we also report Top-3-ACC result since it computes the accuracy of correctly recognizing two highest probabilities in which top-1 prediction is next to the target label.

Implementation details

Our implementation is derived from the publicly available code CutMix⁶, but contains lots of significant modifications. Following this work, we train all the models for 300 epochs using standard SGD with the initial learning rate of 0.005 decayed by factor 0.5 at every 60 epochs. The batch size in two datasets is set to 32. To fairly evaluate all the approaches, we report and recode the best performance on CutMix and other models during training procedure. All experiments are implemented using PyTorch. In each section, all the experiments are fairly under the same condition, especially at the stage of ablation study.

The hyper-parameter γ that balances two losses in Eq. 4 is set to 0.25, and it is decayed by factor 0.5 at every 60 epochs, i.e., $\gamma = 0.25 * 0.5^{\# \text{epoch} / 60}$. The former loss is the main task, while the latter one is just affiliated as an accessory. In summary, γ is much important in the training. To ensure the stability of result, there are two fully connected layers following feature module in ResNet50, and the number of output channels are 512, 128 and 8, respectively. The margin value m in Siamese-contrastive network is set to 2 in Eq. 5. For each iteration, combination ratio λ is randomly selected from uniform distribution $\text{Unif}(0, 1)$.

Comparison against SOTAs

Comparison on AdDb. As shown in Table 1, CL and ACL get better result than other state-of-the-art methods, even without excessive training tricks and pretraining. In work⁴⁹, they used external face dataset MS-Celeb-1M for pre-training to improve recognition accuracy. However, our Top-1-ACC is still higher than other results by a large margin. That is to say, CL/ACL is a conceptually simple but effective method that can be used for both data augmentation and metric learning. In general, our methods outperform over other SOTAs by a large margin. We observe that CL and ACL achieves the state-of-the-art performance among the existing learning strategies on both VGGNet and ResNet50, with accuracy of 63.06% and 64.62%.

Comparison on CarDb. As seen in Table 2, we have compared CL and ACL with some SOTAs on CarDb in the same way, and observe that ACL improves the performance with the same backbone model VGGNet. To establish a fair comparison with SOTAs, we train both the model with ResNet50 and VGGNet. For CarDb, some previous methods usually used this dataset as a regression task, few related works implemented classification. For IOC task, we aim at pursuing ideal result on classification accuracy, and use Acc/Top3-Acc as evaluation criteria. For both two evaluation criteria, ACL achieves the top results in most cases, with an accuracy close to 77%.

In this section, we compare it with R-CNN⁹ and DTCNN²⁵. We surpass these two works by a large margin, partly because we recode the best performance, which follow the work⁶. However, for all evaluation metrics, both CL and ACL exactly get superior results. To better evaluate the effectiveness of CL and ACL, some ablation studies are provided in the next section.

Ablation study: CL and ACL

To demonstrate the effectiveness of the proposed methods, we compare four groups of ablation studies on two datasets with ResNet50 under the same condition, including Baseline (Vanilla CNN), CutMix, CL and ACL. In contrast to the comparison with SOTA methods, conducting ablation studies under the same condition is more compelling in terms of demonstrating the effectiveness of CL and ACL. We show model's superiority from two aspects: classification accuracy and prediction examples.

Effectiveness of CutMix/CL/ACL. For top-1 accuracy, the result sequence of ablation study on all five cross validations is much more convincing in Table 3, i.e., Baseline $\prec$ CutMix $\prec$ CL $\prec$ ACL, corresponding to Fig. 2a,c,e,f. The best result is in bold type. CL and ACL exactly work well on deploying ordinal difference to implicitly promote IOC. ACL is much better than CL in many cases. For top-3 accuracy, their results can be

models	Mean-ACC (%)
LBP+FPLBP ²⁰	45.10
CNN (with GilNet) ⁵	50.70
Cumulative Attribute ⁵¹ + CNN (with VGGNet)	52.34
L-w/o-hyper (with VGGNet) ⁴⁹ (Pretrained)	49.46
L-w/o-KL (with VGGNet) ⁴⁹ (Pretrained)	54.52
Full model (with VGGNet) ⁴⁹ (Pretrained)	56.01
age-gender (with GilNet) ⁵²	51.40
age-gender (with VGGNet) ⁵²	53.60
R-CNN (with GilNet) ⁹	52.42
R-CNN (with VGGNet) ⁹	53.23
CNNPOR ⁵³	57.4
GP-DNNOR ⁵⁴	57.4
SORD ⁵⁵	59.6
Ranking ¹¹	56.7
Classification + POE ⁵⁶	60.5
CL (with VGGNet)(batch_size64)	62.59
CL (with ResNet50)(batch_size64)	62.72
ACL (with VGGNet)(batch_size32)	63.06
ACL (with ResNet50)(batch_size32)	64.62

Table 1. The result of AdDb. Compared with other SOTAs, CL and ACL achieve better performance on AdDb for both VGGNet and ResNet50 Significant values are in [bold].

Model	ACC(%)	Top-3-ACC(%)
CNN-cla ²⁵	52.02	88.18
DTCNN ²⁵	57.49	92.10
R-CNN ⁹	63.67	–
CL(VGGNet)	76.82	97.49
CL(ResNet50)	76.73	97.01
ACL(VGGNet)	77.00	97.23
ACL(ResNet50)	76.91	98.21

Table 2. The result of CarDb. Compared with other SOTAs, CL and ACL achieve better performance on CarDb Significant values are in [bold].

sequenced as CL, ACL, Baseline and CutMix. The anomalies of CutMix result on top-3 accuracy enable us to think about the reason behind its weakness. CutMix enables model to get powerful discriminative capability, while it is weak on learning the ordinal regression information. Because the virtual mixed image corresponds to two categories, its ordinal information is easily mistaken. In addition, its internal ordinal information has rarely been exploited. CL and ACL deploying ordinal difference happen to have this function to repair shortcoming in CutMix. About the result of ACL, some implementation details need to be further optimized. That is to say, Diff with ordinal information is exactly much helpful for IOC, and it exploits more richer ordinal information. In general, we have verified our proposed methods from multiple ablation studies on two datasets.

Examples of CL/ACL. In Fig. 6, we show six pairs of images and their mixed images using CL and ACL. We tend to study its feature map by implementing forward network. The difference prediction $\Delta\hat{y}$ is much close to the ground-truth Δy . This demonstrates that CL and ACL could separate two images with minor difference. CL/ACL exactly has learned ordinal difference information, which means adding the second term in Eq. 4 is necessary to supervise model. In addition, when ACL is implemented, the difference prediction is much more closer to the groundtruth in relation to CL. This demonstrate that ACL with augmented dual strategy is much better than CL on ordinal regression task.

Comparison with Siamese networks

In this section, we compare seven models on two datasets. To evaluate the applicability of various models, we will assess vanilla CNN, Siamese CNN, CutMix, CL, ACL, based on three characteristics - Performance, Flexibility, Memory requirement.

According to the results in Table 4, we observe that CutMix/CL/ACL, Siamese CNN and vanilla CNN obtain best, good and common performance, respectively. To be fair, here the results of three Siamese network versions (see Fig. 3) are obtained all under the same condition. We can see that Siamese network with binary classification as well as with difference, both get inferior result on AdDb-0 and CarDb. From the empirical study, their poor

Dataset	Model	ACC(%)	Top-3-ACC(%)
AdDb0	Baseline	68.10	94.37
	CutMix	69.28	94.79
	CL	69.86	95.95
	ACL	70.64	95.11
AdDb1	Baseline	58.72	92.16
	CutMix	59.08	91.35
	CL	59.60	93.77
	ACL	61.27	92.42
AdDb2	Baseline	63.15	93.97
	CutMix	64.52	92.80
	CL	65.22	95.32
	ACL	66.40	93.23
AdDb3	Baseline	56.20	93.55
	CutMix	57.85	93.59
	CL	58.32	95.54
	ACL	62.29	94.86
AdDb4	Baseline	58.98	92.31
	CutMix	59.84	90.63
	CL	60.63	94.64
	ACL	62.50	92.72
CarDb	Baseline	74.81	94.95
	CutMix	76.10	94.80
	CL	76.73	97.01
	ACL	76.91	98.21

Table 3. The ablation study result on AdDb and CarDb. Compared with CutMix, CL and ACL achieve better performance on AdDb and CarDb Significant values are in [bold].

Dataset	Model	ACC(%)	Top-3-ACC(%)
AdDb-0	vanilla network	66.73	94.67
	Siamese network -binary	67.49	93.84
	Siamese network -difference	66.80	94.97
	Siamese network -contrastive	68.33	95.81
	CutMix	69.28	94.79
	CL	69.86	95.95
	ACL	70.64	95.11
CarDb	vanilla network	74.81	96.71
	Siamese network -binary	75.17	96.47
	Siamese network -difference	74.48	96.26
	Siamese network -contrastive	75.53	96.41
	CutMix	76.10	94.80
	CL	76.73	97.01
	ACL	76.91	98.21

Table 4. Comparison with Siamese networks. CL/ACL takes the absolute dominant Significant values are in [bold].

result comes from three aspects. On one hand, compared with CutMix, the discriminative capability of Siamese CNN is weak. In CutMix/CL/ACL, The spatial feature context with convex combination is learned deeply. However, Siamese network learns the contrastive information on the flat feature by fully connected layers, with very little spatial feature. The former undoubtedly possesses significantly more abundant spatial information than the latter. On the other hand, CutMix produces much more patch combinations and pairing combinations than that of Siamese network. Many patch-pairs with various combinations in CutMix are learned to get much more robust model.

Besides comparing recognition performance, we also give some analysis on model flexibility and memory requirement. About the flexibility of model design and network implementation, Siamese CNN consisting of

Characteristic	Vanilla	Siamese	CutMix	CL	ACL
Performance↑	✓	✓✓	✓✓✓	✓✓✓✓	✓✓✓✓✓
Flexibility(portability)↑	✓✓✓	✓	✓✓✓	✓✓✓	✓✓
Memory requirement↓	✓(1×)	✓✓(2×)	✓(1×)	✓(1×)	✓✓(2×)

Table 5. Comparison of five approaches. CL/ACL is competitive on three evaluation indexes

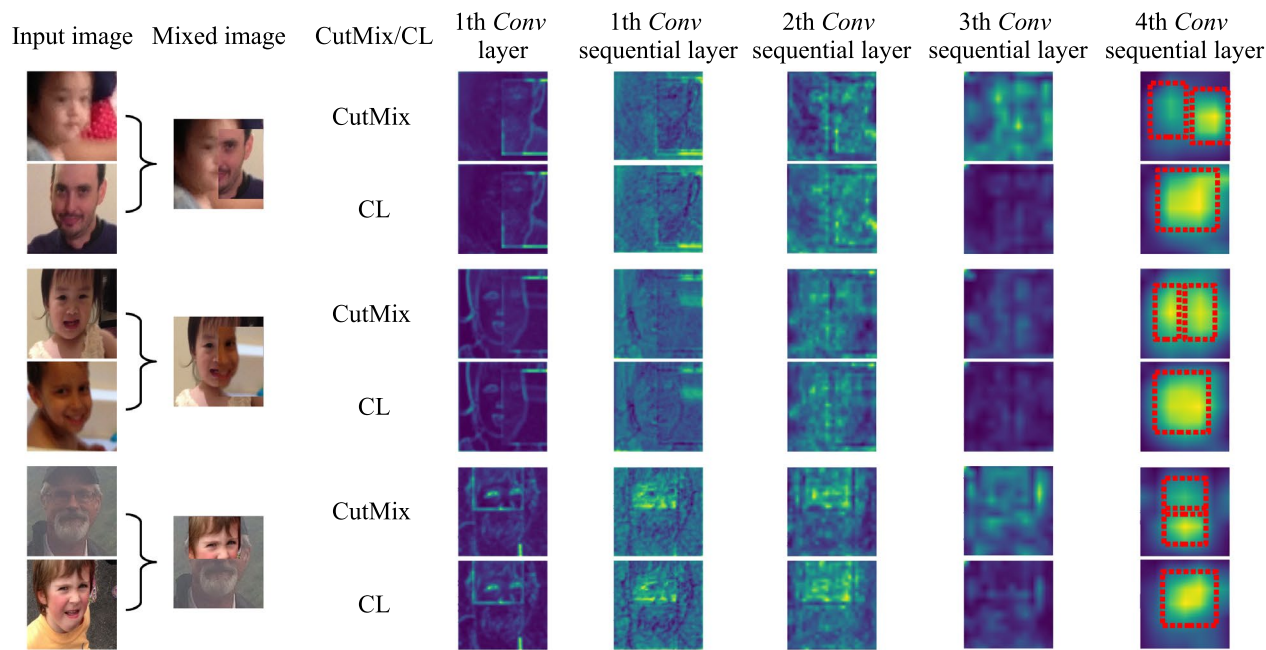


Figure 5. The evolution of feature map. Two separate objects are unexpectedly mixed by the competitive learning task.

two sister networks and two input images is more complex than other four approaches, as shown in Table 5. About memory requirement, Siamese CNN needs roughly twice memory requirement that CL and vanilla CNN need. In general, CL/ACL will be the best choice for IOC among these styles.

An intriguing phenomenon: CutMix VS CL

From Fig. 5, it is interesting to find that two separated areas, marked with red dashed box, are actually mixed in some last convolution layers in CL. For ResNet50, we extract feature map of the first Conv layer and four sequential Conv layers. Two areas are competitively learning in the training process. In CutMix, even at the last Conv layer these two areas are separate from each other. This phenomenon still exists at some front layers in CL, while it disappeared in some last layers. This shows that there are various neuron connections on spatial context between two separated areas, see the red dashed box in Fig. 5. Two images are independent and compete with each other. That is to say, CL implements competitive learning both on the spatial context and the flatted feature. Obviously, we can draw the conclusion that the difference regression term in Eq. 4 certainly plays an important role in the training.

Some analysis on examples

In this section, we explain the role of CL/ACL from ordinal difference prediction and CAM visualization, so as to further show the functionality of competitive learning task.

In Fig. 6, we show six pairs of images, their mixed images and difference groundtruth. After model inference of CL and ACL, their ordinal difference prediction manifests distinct results. The ordinal prediction $\Delta\hat{y}$ is much close to the ground-truth Δy except last two image pairs. This demonstrates that CL and ACL could force network to recognize the difference between two images. The model exactly has learned the difference information, which means adding the second term in Eq. 4 is a valid way to supervise the neural network. In addition, when ACL is implemented, the difference prediction is more closer to the ground-truth in relation to CL.

In Fig. 7, we have selected five cases and shown their probability distribution and gradCAM of CutMix, CL and ACL. By analyzing the gradCAM and probability distribution, we find that in some cases although CutMix and CL can drive the output closer to the ground truth and push the network to focus more on useful information, they still fall short of achieving satisfactory performance. ACL, however, can make model concentrate on as

Category	0	1	4	6	1	4	4	2	7	1	0	5
Input image												
Mixed image												
Difference groundtruth Δy	1		2		3		2		6			5
CL prediction $\Delta \hat{y}$	0.7708		0.1659		4.0142		1.2378		2.4430			1.5829
ACL prediction $\Delta \hat{y}$	0.7608		1.1015		3.1463		1.8153		5.0465			4.3191

Figure 6. Some examples on ordinal difference prediction. The ordinal difference error $|\Delta y - \Delta \hat{y}|$ reflects fine grained information.

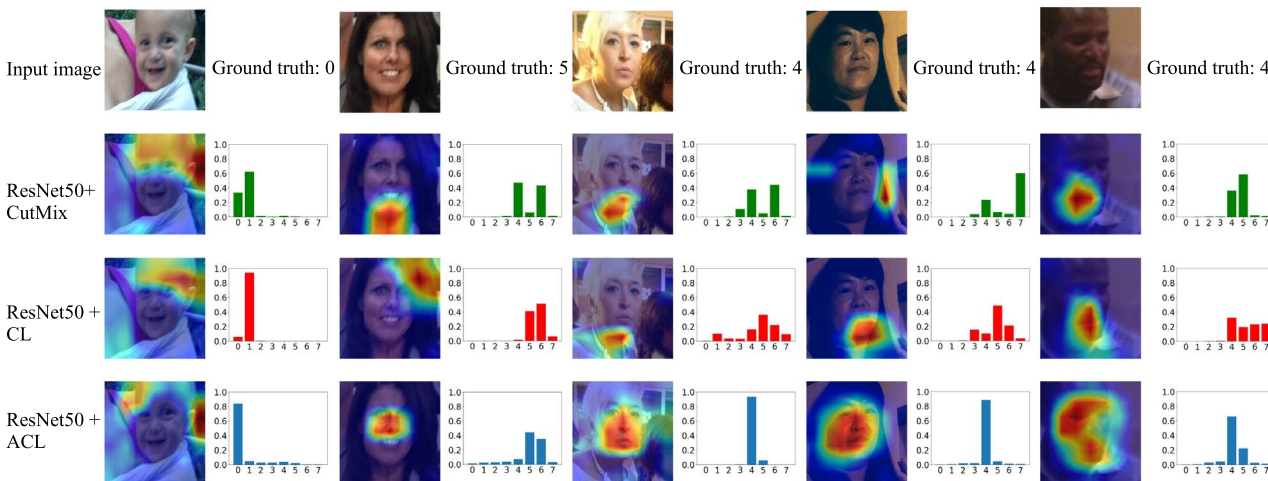


Figure 7. Some examples on prediction and gradCAM in the testing procedure for three approaches. ACL helps learn more discriminative feature in the training.

much important information in the image as possible, such as the whole face of the person, and eventually classify the image correctly. This shows that despite our method is based on CutMix, ACL is able to achieve better results than CutMix and CL by employing augmented enhancement strategy.

Conclusion

In this work, we demonstrate how a simple but effective technique CL/ACL is used for IOC and why it performs well. Mixing policy can help achieve the same function of metric learning that maximizes the inter-class variations and minimizes the intra-class variation. Moreover, it produces much more patch combinations and pairing configurations than that of Siamese network. The ordinal difference is embedded in CL as an accessory, which can further help learn more refined category information and discriminative feature. CL implements competitive learning both on the spatial context and the flat feature. Finally, the dual augmentation and random augmentation are proposed to further enhance the model. From the perspective of application, the CL prototype that we have proposed holds promise for broader application in various general ordinal recognition challenges, including natural language processing and speech recognition. In the future, we desire to study the application of mixing policy on some other areas such as NLP and speech recognition, and also study grid-based learning approaches that analyse how grid size and arrangement of grid influence classification performance. In addition, we will also study how to scientifically balance two losses, i.e., main task and auxiliary task, in the training process.

Data availability

The datasets used and/or analysed during the current study are available in <https://talhassner.github.io/home/projects/Adience/Adience-data.html#frontalized> and <https://web.cs.ucdavis.edu/~yjee/icc2013.html> from the corresponding author on reasonable request.

Received: 9 February 2025; Accepted: 4 June 2025

Published online: 02 July 2025

References

- Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (2012).
- He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* (2016).
- Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).
- Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision* (2021).
- Levi, G. & Hassner, T. Age and gender classification using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, Workshop* (2015).
- Yun, S. et al. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE Conference on Computer Vision* 6023–6032 (2019).
- Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: Beyond empirical risk minimization. arXiv preprint [arXiv:1710.09412](https://arxiv.org/abs/1710.09412) (2017).
- Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint [arXiv:2004.10934](https://arxiv.org/abs/2004.10934) (2020).
- Zhang, C., Zhu, C. & Xu, X. Revisiting image ordinal estimation: How to deal with ordinal relationship in deep learning. *J. Electron. Imaging* **28**, 013025 (2019).
- Gao, B.-B., Xing, C., Xie, C.-W., Wu, J. & Geng, X. Deep label distribution learning with label ambiguity. *IEEE Trans. Image Process.* **26**, 2825–2838 (2017).
- Niu, Z., Zhou, M., Wang, L., Gao, X. & Hua, G. Ordinal regression with multiple output CNN for age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition* 4920–4928 (2016).
- Hou, W., Gao, X., Tao, D. & Li, X. Blind image quality assessment via deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* **26**, 1275–1286 (2017).
- Lu, X., Lin, Z., Shen, X., Mech, R. & Wang, J. Z. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *International Conference on Computer Vision*, 990–998 (2015).
- Zhang, Y., Zhao, R., Dong, W., Hu, B.-G. & Ji, Q. Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 7034–7043 (2018).
- Datta, R., Joshi, D., Li, J. & Wang, J. Z. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, 288–301 (2006).
- Datta, R., Joshi, D., Li, J. & Wang, J. Z. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.* **40**, 5 (2008).
- Fu, Y., Guo, G. & Huang, T. S. Age synthesis and estimation via faces: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1955–1976 (2010).
- Han, H., Otto, C. & Jain, A. K. Age estimation from face images: Human vs. machine performance. In *IEEE International Conference on Biometrics (ICB)*, 1–8 (2013).
- Lee, Y.J., Efros, A.A. & Hebert, M. Style-aware mid-level representation for discovering visual connections in space and time. In *International Conference on Computer Vision*, 1857–1864 (2013).
- Yi, D., Lei, Z. & Li, S. Z. Age estimation by multi-scale convolutional network. In *Proceedings of Asian Conference on Computer Vision* (2015).
- Zhu, H., Zhou, Q., Zhang, J. & Wang, J. Z. Facial aging and rejuvenation by conditional multi-adversarial autoencoder with ordinal regression. arXiv preprint [arXiv:1804.02740](https://arxiv.org/abs/1804.02740) (2018).
- Bao, Z. et al. General vs. long-tailed age estimation: An approach to kill two birds with one stone. *IEEE Trans. Image Process.* **32**, 6155–6167 (2023).
- Shou, Y.-W., Cao, X. & Meng, D. Masked contrastive graph representation learning for age estimation. *Pattern Recognit.* **158**, 110974 (2023).
- Bao, Z. et al. Divergence-driven consistency training for semi-supervised facial age estimation. *IEEE Trans. Inf. Forensics Secur.* **18**, 221–232 (2023).
- Chao, Z., Zhu, C., Liu, Y., Wen, H. & Wang, Z. Image ordinal estimation: Classification and regression benefit each other. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (2017).
- Schroff, F., Kalenichenko, D. & Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 815–823 (2015).
- Kong, S., Shen, X., Lin, Z., Mech, R. & Fowlkes, C. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, 662–679 (2016).
- Chen, W., Chen, X., Zhang, J. & Huang, K. Beyond triplet loss: A deep quadruplet network for person re-identification. arXiv preprint [arXiv:1704.01719](https://arxiv.org/abs/1704.01719) (2017).
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
- Wan, L., Zeiler, M., Zhang, S., Le Cun, Y. & Fergus, R. Regularization of neural networks using dropconnect. In *International Conference on Machine Learning*, 1058–1066 (2013).
- Wang, S. & Manning, C. Fast dropout training. In *International Conference on Machine Learning*, 118–126 (2013).
- DeVries, T. & Taylor, G. W. Improved regularization of convolutional neural networks with cutout. arXiv preprint [arXiv:1708.04552](https://arxiv.org/abs/1708.04552) (2017).
- Zhong, Z., Zheng, L., Kang, G., Li, S. & Yang, Y. Random erasing data augmentation. arXiv preprint [arXiv:1708.04896](https://arxiv.org/abs/1708.04896) (2017).
- Inoue, H. Data augmentation by pairing samples for images classification. arXiv preprint [arXiv:1801.02929](https://arxiv.org/abs/1801.02929) (2018).
- Tokozume, Y., Ushiku, Y. & Harada, T. Between-class learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5486–5494 (2018).
- Walawalkar, D., Shen, Z., Liu, Z. & Savvides, M. Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. arXiv preprint [arXiv:2003.13048](https://arxiv.org/abs/2003.13048) (2020).
- Summers, C. & Dinneen, M. J. Improved mixed-example data augmentation. In *Proceedings of Winter Conference on Applications of Computer Vision*, 1262–1270 (2019).

38. Hadsell, R., Chopra, S. & LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1735–1742 (2006).
39. Schroff, F., Kalenichenko, D. & Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 815–823 (2015).
40. Oh Song, H., Xiang, Y., Jegelka, S. & Savarese, S. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4004–4012 (2016).
41. Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, 1857–1865 (2016).
42. Ko, B. & Gu, G. Embedding expansion: Augmentation in embedding space for deep metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 7255–7264 (2020).
43. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E. & Shah, R. Signature verification using a “ siamese ” time delay neural network. *Advances in Neural Information Processing Systems* 737–737 (1994).
44. Chopra, S., Hadsell, R. & LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 539–546 (2005).
45. Weinberger, K. Q. & Saul, L. K. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* **10** (2009).
46. Chechik *et al.* Large scale online learning of image similarity through ranking. *J. Mach. Learn. Res.* **11** (2010).
47. Xuan, H., Souvenir, R. & Pless, R. Deep randomized ensembles for metric learning. In *European Conference on Computer Vision*, 723–734 (2018).
48. Movshovitz-Attias, Y., Toshev, A., Leung, T. K., Ioffe, S. & Singh, S. No fuss distance metric learning using proxies. In *International Conference on Computer Vision*, 360–368 (2017).
49. Zhang, Y., Liu, L., Li, C. *et al.* Quantifying facial age by posterior of age comparisons. arXiv preprint [arXiv:1708.09687](https://arxiv.org/abs/1708.09687) (2017).
50. Eidinger, E., Enbar, R. & Hassner, T. Age and gender estimation of unfiltered faces. *IEEE Trans. Inf. Forensics Secur.* **9**, 2170–2179 (2014).
51. Chen, K., Gong, S., Xiang, T. & Change Loy, C. Cumulative attribute space for age and crowd density estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2467–2474 (2013).
52. Lapuschkin, S., Binder, A., Müller, K.-R. & Samek, W. Understanding and comparing deep neural networks for age and gender classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1629–1638 (2017).
53. Goh, C. K., Liu, Y. & Kong, A. W. K. A constrained deep neural network for ordinal regression. *IEEE Conference on Computer Vision and Pattern Recognition* (2018).
54. Liu, Y., Wang, F. & Kong, W. Probabilistic deep ordinal regression based on gaussian processes. In *IEEE International Conference on Computer Vision (ICCV)* (2019).
55. D’Áz, R. & Marathe, A. Soft labels for ordinal regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).
56. Li, W., Huang, X., Lu, J., Feng, J. & Zhou, J. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *IEEE Conference on Computer Vision and Pattern Recognition* (2021).

Acknowledgements

We thank Ce Zhu for helpful discussions and supervision. This research is supported by Sichuan Science and Technology Program (NO.2023NSFSC0495), Industry-University-Research Innovation Fund for Chinese Universities - Hillstone Project (No.2022HS041), Intelligent Policing Key Laboratory of Sichuan Province (No.ZN-JW2025KFZD001, No.ZNJW2023ZZMS001, ZNJW2022KFMS001).

Author contributions

Chao Zhang: Conceptualization on research idea, Methodology, Experiment, Review and Writing. Chao Feng: Experiment, Formal analysis, Review and Writing. Jianmei Cheng: Formal analysis, Methodology, Supervision, Review and Writing. Keren Fu: Methodology, Review and Writing.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025