MIX-UP SELF-SUPERVISED LEARNING FOR CONTRAST-AGNOSTIC APPLICATIONS

Yichen Zhang[1,*], *Yifang Yin*[2,*], *Ying Zhang*[3,†] *and Roger Zimmermann*[1]

[1]School of Computing, National University of Singapore
[2]Institute for Infocomm Research, A*STAR
[3]School of Computer Science, Northwestern Polytechnical University
{yichenz, dcsrz}@nus.edu.sg, yin_yifang@i2r.a-star.edu.sg, izhangying@nwpu.edu.cn

ABSTRACT

Contrastive self-supervised learning has attracted significant research attention recently. It learns effective visual representations from unlabeled data by embedding augmented views of the same image close to each other while pushing away embeddings of different images. Despite its great success on ImageNet classification, COCO object detection, *etc.*, its performance degrades on contrast-agnostic applications, *e.g.*, medical image classification, where all images are visually similar to each other. This creates difficulties in optimizing the embedding space as the distance between images is rather small. To solve this issue, we present the first mix-up self-supervised learning framework for contrast-agnostic applications. We address the low variance across images based on cross-domain mix-up and build the pretext task based on two synergistic objectives: image reconstruction and transparency prediction. Experimental results on two benchmark datasets validate the effectiveness of our method, where an improvement of 2.5% $\sim$ 7.4% in top-1 accuracy was obtained compared to existing self-supervised learning methods.

***Index Terms*—** Self-supervised learning, contrast-agnostic applications, computer vision, deep learning

1. INTRODUCTION

Self-supervised learning has received increasing research attention from both academia and industry in recent years. It learns from unlabeled data by pretext tasks, and thus reduces the cost for building large and accurately labeled datasets [1].

Several previous studies [2–6] have achieved promising results on self-supervised visual representation learning in the image domain. State-of-the-art methods are mostly based on contrastive learning under the instance discrimination pretext task [7]. These methods aim at increasing the feature similarity between positive pairs while minimizing it between negative pairs. To achieve that, two random augmented views of the same image (*i.e.*, applying random augmentation twice to the same image) are regarded as a positive pair and the augmented views of different images are regarded as negative pairs. Although good performance has been obtained on large-scale benchmark datasets such as ImageNet [8] and COCO [9], challenges still exist in applications where all images are visually similar to each other (see Fig. 1). For these applications, the high similarity among samples becomes an obstacle in optimizing the embedding space for contrastive learning as the visual gap between negative pairs is quite small. Due to this property, we refer to these applications as *contrast-agnostic*. One well-known example of such applications is medical image analysis. Many of them consist of images showing certain body parts captured by specific medical instruments, resulting in visually similar images with low variance across different instances.

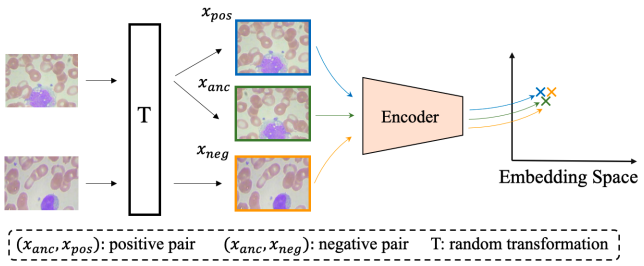

Fig. 1. In contrast-agnostic applications, low variance across images can degrade the performance of contrastive self-supervised learning techniques. The high similarity between images hinders the model to distinguish the positive example x_{pos} and the negative example x_{neg} *w.r.t.* the anchor image x_{anc}, resulting in obtaining less effective representations.

Intuitively, it can be beneficial for representation learning by improving training samples' diversity in these applications. One straightforward solution is to inject random Gaussian noise into the original images. However, Gaussian noise does not contain any semantic information, resulting in minor improvements in representation learning. Instead, as a data augmentation method, mix-up [10] operation has been widely adopted in various applications. It generates synthetic images by mixing different images together. However, based on

∗ Equal contribution.
† Corresponding author.

our experiments, combining mix-up with existing contrastive learning methods, e.g., mixco [5], may only lead to minor improvements in contrast-agnostic applications.

To address the issues, we present *MixSSL*, the first *Mix-up Self-Supervised Learning* framework for contrast-agnostic visual representation learning. First, we propose to enrich the diversity of the (target) training samples (*e.g., medical images*) by mixing them with (auxiliary) natural images from ImageNet. This strategy not only increases the variance among training samples, but also improves the model's robustness on unseen instances. Next, we build our pretext task by fulfilling the following two synergistic objectives: image reconstruction and transparency prediction. Reconstruction of the original training sample from the mixed input is a challenging task, which enables our model to extract the key information of the target image effectively. Meanwhile, transparency prediction predicts the mix-up ratio, which serves as an auxiliary guidance by identifying the content of the target image from the mixed input to facilitate reconstruction. Here we summarize the contributions of this paper as follows:

- To the best of our knowledge, we are the first to propose a self-supervised learning framework for contrast-agnostic applications. Cross-domain mix-up is adopted to improve sample diversity and model robustness simultaneously.
- We build our pretext task by jointly performing image reconstruction and transparency prediction. Thus, our model learns to distinguish visual content from different domains and further extract key information from the target domain.
- We perform comprehensive experiments on two contrast-agnostic image datasets. The experimental results show that our method outperforms the state-of-the-art self-supervised learning methods by a large margin on both datasets.

2. RELATED WORK

Since good representations are the foundation of good results [11, 12], a significant number of self-supervised representation learning techniques have been proposed by researchers world-wide in the image domain. Here we roughly divide them into two categories: non-contrastive-based and contrastive-based methods, depending on whether contrastive learning has been adopted for building pretext tasks.

Non-contrastive-based methods. To circumvent the requirement of the labeling process, many pretext tasks have been proposed to train the model in an unsupervised manner. Most of them were inspired by image-level operations and the internal property of data, which provide temporary target for the model training. For example, researchers constructed the training target via predicting the rotation angle [13] or solving the Zigzaw problem [14], whose answer can either be derived from the augmentation process or the original image. Also, other examples include context prediction [15], affine transformation [16] and colorization [17].

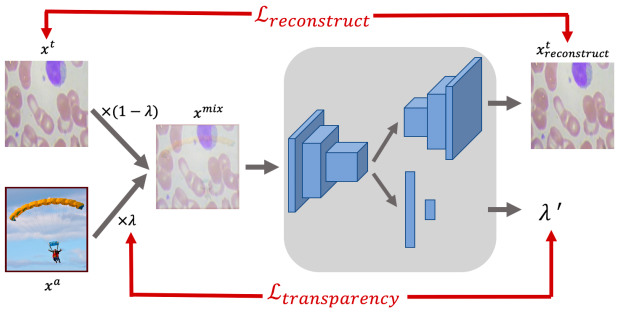

Fig. 2. The proposed method for contrast-agnostic applications.

Contrastive-based methods. As one of the most important branches in self-supervised learning, many studies based on contrastive learning have been developed under the instance discrimination pretext task (matching the encoded views of the same image). They mainly focused on finding a better way to construct or utilize positive or negative pairs for model training. SimCLR [2] adopted the nonlinear projection head between the representation and the loss function. MoCo [3] maintained a key encoder and a corresponding large memory bank (dictionary) to improve stability. Dino [6] adopted a teacher network to facilitate model training. MixCo [5] extended the contrastive learning concept to semi-positives encoded from the mix-up [10] of positive and negative images. Although these methods achieve state-of-the-art results on the ImageNet dataset, they perform unsatisfactorily on contrast-agnostic applications.

3. METHOD

3.1. Overview

Self-supervised learning, especially contrastive self-supervised learning, encounters serious challenges when applied to contrast-agnostic applications, due to the small visual gap among samples. In this section, we introduce *Mix-up Self-Supervised Learning* (*MixSSL*) framework for contrast-agnostic applications. As illustrated in Fig. 2, our method deals with the low-variance challenge by first mixing samples from different domains and then extracting key information of the target domain based on two synergistic objectives.

Formally, given a target contrast-agnostic dataset $X^t = \{x_1^t, x_2^t, \ldots, x_n^t\}$ and an auxiliary dataset $X^a = \{x_1^a, x_2^a, \ldots, x_m^a\}$, our goal is to learn good visual representations for the target dataset X^t with the facilitation of the auxiliary dataset X^a. Since our method is a self-supervised visual representation learning approach, labels from both datasets

are not required.

3.2. MixSSL

Our overall objective is to obtain a well-trained encoder that performs outstandingly at extracting informative visual representations for a target contrast-agnostic application. To achieve this goal, we design our pretext task as a composition of two correlated objectives with cross-domain mix-up applied in data preprocessing. The main objective is the image reconstruction task, which aims to extract the key semantic information from target domain images. The auxiliary objective is the transparency prediction task, which aims to recognize the composition ratio from the mixed samples so as to facilitate target image reconstruction. The overall framework is illustrated in Fig. 2, which consists of a shared encoder for feature extraction, a decoder for image reconstruction and a linear regressor for transparency prediction.

Cross-domain mix-up. The mix-up operation was first proposed in [10] as an image augmentation method. It simultaneously mixes the training samples and their labels as,

$$x = \lambda x_i + (1-\lambda)x_j$$
$$y = \lambda y_i + (1-\lambda)y_j$$

where λ is sampled from the Beta distribution. x and y are samples and labels, respectively. The training of a model with this strategy is capable of reducing the amount of undesirable oscillation when meeting with unseen training samples [10]. We build upon the conventional mix-up strategy and present a cross-domain variant to improve the diversity of our training samples. Our cross-domain mix-up method mixes samples from different domains:

$$x^{mix} = (1-\lambda)x^t + \lambda x^a \tag{1}$$

where x^t and x^a denote two samples from the target dataset and the auxiliary dataset, respectively. The mix-up ratio λ is again sampled from the Beta distribution, varying from batch to batch. Different from [18], where cross-domain mix-up serves as a solution to the domain shift, our cross-domain mix-up method effectively increases the variance among samples in contrast-agnostic applications, but may distract the model from learning specific distribution of the target domain. To solve this issue, we build our pretext task by jointly performing image reconstruction and transparency prediction. Thus, our model learns to separate the content of x^t and x^a from the mixed input x^{mix} and capture the key semantic information of x^t at the same time.

Image reconstruction. As aforementioned, existing contrastive learning methods may confuse the model in contrast-agnostic applications. We alternatively adopt image reconstruction as our main pretext task for self-supervised learning. The reconstruction loss is defined as the distance between the original target image x^t and the reconstructed target image $x^t_{reconstruct}$ from the mixed input x^{mix} as,

$$\mathcal{L}_{reconstruct} = \mathcal{D}(Decoder(Encoder(x^{mix})), x^t) \tag{2}$$

where $\mathcal{D}$ is a distance metric measuring the difference between original and reconstructed images, which is normally implemented as a mean square error (MSE) function. It is worth noting that our proposed method is different from the traditional autoencoder and decoder as we reconstruct an image from its faded version obtained by cross-domain mix-up. The advantages of our method are twofold. First, it motivates the model to separate the mixed input by learning to select the content that is originally from the target image. Second, the key semantic information of target images are extracted by the encoder to perform reconstruction. Facilitated by the transparency prediction task to be introduced in the next section, our model learns more effective visual representations compared to existing contrastive learning based methods.

Transparency prediction. Recall that we construct training samples by mixing images from different domains as shown in Fig. 2. The mix-up operation combines the target and the auxiliary images by linear interpolation with weight λ. We subsequently introduce an auxiliary pretext task, termed transparency prediction, to predict the weight λ based on the mixed input sample. It is beneficial to jointly optimize image reconstruction and transparency prediction as the two tasks are intrinsically correlated with each other. Here we model the transparency prediction task as a regression problem and predict λ based on multi-layer perception (MLP). The loss function is given by,

$$\lambda' = MLP(Encoder(x^{mix})) \tag{3}$$
$$\mathcal{L}_{transparency} = |\lambda' - \lambda| \tag{4}$$

where $|\lambda' - \lambda|$ is the absolute difference between the predicted transparency λ' and the ground-truth transparency λ.

Overall loss function. In summary, the above two objectives encourage the encoder to extract well-trained visual representations from the target images based on self-supervised learning. The overall loss is defined as a linear combination of $\mathcal{L}_{reconstruct}$ and $\mathcal{L}_{transparency}$ as:

$$\mathcal{L}_{pretrain} = \mathcal{L}_{reconstruct} + \gamma\mathcal{L}_{transparency} \tag{5}$$

where γ is a hyper-parameter balancing the impact of the two losses.

4. EXPERIMENTS

4.1. Experimental setup

We adopt two benchmark datasets to evaluate our proposed mix-up self-supervised learning framework, including the blood cell dataset and the cervix dataset. They can both be regarded as contrast-agnostic applications due to the high visual

Pre-trained model	Mix	Top-1 Accuracy (%)				
		Eosinophil	Lymphocyte	Monocyte	Neutrophil	Average
Rand	×	76.40	99.19	75.02	94.55	86.29 (-)
SimCLR [2]	×	80.26	99.84	83.39	91.99	88.86 (+1.94%)
Dino [6]	×	87.60	100.00	74.68	94.39	89.22 (+2.93%)
MoCo [3]	×	83.79	100.00	77.10	**96.15**	89.26 (+2.97%)
Rotation [13]	×	87.48	100.00	75.48	94.07	89.26 (+2.97%)
ImageNet-self	×	85.07	100.00	78.71	93.91	89.43 (+3.14%)
Mixup+SimCLR	√	**88.28**	100.00	75.02	93.59	89.22 (+2.93%)
MixCo [5]	√	86.19	100.00	77.90	94.71	89.71 (+3.42%)
MixSSL (Ours)	√	87.69	**100.00**	**85.81**	95.35	**92.19** (+5.90%)

Table 1. Performance comparison between our framework and state-of-the-art methods on the blood cell dataset.

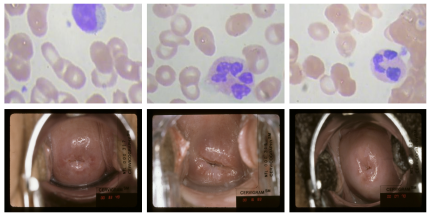

Fig. 3. Data samples from two datasets, including the blood cell dataset (first row) and the cervix dataset (second row).

similarity across samples. Examples from the two datasets are shown in Fig. 3.

Blood cell dataset [19] is a public dataset consisting of four types of blood cell images. It contains nearly 10,000 images for training and nearly 2,500 images for validation. We train our model on the training set and report the results on the validation set.

Cervix dataset consists of four related datasets: NHS [20], ALTS [21], CVT [22] and Biopsy [23]. A total of nearly 20,000 cervical images are included. Following previous work [24, 25], we split out a validation set from the NHS dataset for evaluation and assign binary labels to them according to their original labels. More details about the cervix data description can be found in supplementary material.

We compare our framework with five state-of-the-art self-supervised learning methods [2, 3, 5, 6, 13] on two datasets.

4.2. Implementation details

Following previous work [3], we adopt ResNet-50 [26] as our encoder and ResNet-Generator from CycleGAN [27] as our decoder. The linear regressor contains two linear layers with 128 and 1 units, respectively. All methods including ours are first initialized with an ImageNet pre-trained model Dino [6]. Next, they are trained to learn visual representations on the contrast-agnostic datasets based on pretext tasks, and finally fine-tuned for the downstream classification task with class labels. The parameter γ in the loss function is empirically set to 1 throughout the experiments. For the auxiliary dataset to perform cross-domain mix-up, we adopt ImageNette [28], which is a subset of the ImageNet consisting of 10 classes only. For more details about data preprocessing and training please refer to supplementary material.

4.3. Comparison to the state-of-the-art

We compare our framework with five state-of-the-art self-supervised learning methods on both the blood cell dataset and the cervix dataset, and report the results in Tables 1 and 2. We also report the results obtained by using random initialization (Rand) or self-supervised pre-trained weights on ImageNet (ImageNet-self) as baselines for comparison.

Table 1 reports both the per-class and average top-1 accuracy, together with the average performance gain against the baseline Rand on the blood cell dataset. We observe that our method outperforms all the other methods in terms of the average accuracy. For the most difficult class, *i.e.*, the Monocyte, our method obtains a promising result of 85.81%, while its competitors only achieves top-1 accuracy of less than 80%. Compared to the second best solution [5], our method outperforms it by 1.5% in Eosinophil, 7.91% in Monocyte, 0.64% in Neutrophil, and 2.53% in average accuracy.

Table 2 shows the result comparison on the cervix dataset. In addition to the top-1 accuracy, we also report the precision, recall, F1 score, and ROC-AUC since the cervix cancer detection is a binary classification problem. ROC-AUC is an essential metric for medical applications, which reveals the overall model capability. As can be seen, our method obtains the best result in terms of F1 score, ROC-AUC, and accuracy. Compared to the second best solution [5], our model achieves an improvement of 9.49% in recall, 4.40% in F1 score, 2.14% in ROC-AUC, and 2.84% in accuracy.

Additionally, we observe that self-supervised models trained on the ImageNet (ImageNet-self) outperforms most of the self-supervised models trained on medical images. One possible reason is that existing contrastive-based approaches suffer from the low-variance challenge among medical images, which makes them less effective in visual representation learning for such contrast-agnostic applications. Our method,

Pre-trained model	Mix	Metrics (%)				
		Precision	Recall	F1 score	ROC-AUC	Accuracy
Rand	×	63.32	68.51	65.88	75.70	70.45 (-)
SimCLR [2]	×	71.82	83.69	77.21	82.26	79.55 (+9.10)
Dino [6]	×	75.31	87.72	81.06	84.47	82.95 (+12.50)
MoCo [3]	×	73.12	**93.26**	81.91	85.90	82.95 (+12.50)
Rotation [13]	×	81.46	78.17	79.71	86.12	83.52 (+13.07)
ImageNet-self	×	79.72	80.86	80.35	86.21	83.52 (+13.07)
Mixup+SimCLR	✓	75.31	79.55	77.36	83.91	80.68 (+10.23)
MixCo [5]	✓	**81.72**	79.53	80.61	86.92	84.09 (+13.64)
MixSSL (Ours)	✓	81.27	89.02	**85.01**	**89.06**	**86.93** (+16.48)

Table 2. Performance comparison between our framework and state-of-the-art methods on the cervix dataset.

Dataset	$\mathcal{R}$	$\mathcal{T}$	$\mathcal{C}$	$\mathcal{A}$	Accuracy (%)
Bloodcell dataset		✓			89.26
	✓				89.63
	✓	✓			**92.19**
	✓		✓		89.22
	✓			✓	90.39
Cervix dataset		✓			82.95
	✓				84.09
	✓	✓			**86.93**
	✓		✓		81.25
	✓			✓	82.96

Table 3. Performance comparison between different objectives. $\mathcal{R}$: image reconstruction; $\mathcal{T}$: transparency prediction; $\mathcal{C}$: contrastive learning; $\mathcal{A}$: auxiliary label prediction.

on the other hand, addresses the low-variance challenge by applying cross-domain mix-up assisted with two complementary objectives, leading to a better performance when being fine-tuned for downstream tasks.

4.4. Ablation studies

In this section, we conduct ablation studies to investigate the usage of mix-up operation and the design of pretext task.

Mix-up operation. Recall that we apply mix-up to make training samples more diverse in our framework. Then a question naturally arises: can mix-up also improve the performance of existing contrastive self-supervised methods? To find the answer, we compare SimCLR and MoCo with their mix-up variants Mixup+SimCLR and MixCo in Tables 1 and 2. The results show that the mix-up operation only leads to minor improvements in the SimCLR and MoCo frameworks. Our proposed method outperforms Mixup+SimCLR and MixCo by 2.48% $\sim$ 6.25% in terms of the top-1 accuracy, which indicates that our method leverages the mix-up operation to deal with the challenges in a more effective way. Reconstructing the original images from the mixed input where only partial information is accessible is a challenging task. By training based on this objective, our encoder has stronger generalization capability and is able to extract informative representations of target images better compared to other methods.

Pretext task. In addition to the image reconstruction ($\mathcal{R}$) and the transparency prediction ($\mathcal{T}$) adopted in our framework, we also evaluate the effectiveness of two alternatives tasks, namely contrastive learning ($\mathcal{C}$) and auxiliary label prediction ($\mathcal{A}$). Results are reported in Table 3. As it shows, both $\mathcal{R}$ and $\mathcal{T}$ are essential in our proposed framework, missing either of them results in performance degradation of more than 2% in terms of accuracy. The experimental results also indicate that the image reconstruction task $\mathcal{R}$ is more important for visual representation learning of medical images. The transparency prediction task $\mathcal{T}$ can be considered as an auxiliary objective to facilitate the reconstruction as both of them are intrinsically correlated with each other.

Next, we compare our transparency prediction task $\mathcal{T}$ with its alternatives $\mathcal{C}$ and $\mathcal{A}$. $\mathcal{C}$ utilizes the contrastive loss that narrows the gap between two augmented views, while $\mathcal{A}$ leverages auxiliary label information as an instruction. By jointly optimizing with the reconstruction loss, our transparency prediction task $\mathcal{T}$ outperforms $\mathcal{C}$ and $\mathcal{A}$ by 2.97% and 1.80% on the blood cell dataset and by 5.68% and 3.97% on the cervix dataset. Task $\mathcal{C}$ optimizes the feature space based on mixed samples without distinguishing the content from different domains. Task A is a supervised task where labels of the auxiliary dataset are required. Our proposed transparency prediction task $\mathcal{T}$ addresses the drawbacks of $\mathcal{C}$ and $\mathcal{A}$, which has been shown to be more effective in visual representation learning for contrast-agnostic applications.

5. CONCLUSION

In this paper, we propose a novel self-supervised visual representation learning framework for contrast-agnostic applications. Our method deals with the low-variance challenge by applying cross-domain mix-up to diversify training samples. Furthermore, we build our pretext task based on image re-

construction and transparency prediction to learn representations in the target domain. Extensive experiments have been conducted on two benchmark datasets. Results show that our method outperforms existing contrastive self-supervised learning methods by a large margin.

6. ACKNOWLEDGEMENT

This research was supported by Singapore Ministry of Education Academic Research Fund Tier 1 under MOE's official grant number T1 251RES2029.

7. REFERENCES

[1] Qingyang Hu, Qinming He, Hao Huang, Kevin Chiew, and Zhenguang Liu, "Learning from crowds under experts' supervision," in *PAKDD*, 2014, vol. 8443, pp. 200–211. 1

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1597–1607. 1, 2, 4, 5

[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020, pp. 9729–9738. 1, 2, 4, 5

[4] Zhenguang Liu, Zepeng Wang, Luming Zhang, Rajiv Ratn Shah, Yingjie Xia, Yi Yang, and Xuelong Li, "Fastshrinkage: Perceptually-aware retargeting toward mobile platforms," in *ACM Multimedia*, 2017, pp. 501–509. 1

[5] Sungnyun Kim, Gihun Lee, Sangmin Bae, and Se-Young Yun, "Mixco: Mix-up contrastive learning for visual representation," *arXiv preprint arXiv:2010.06300*, 2020. 1, 2, 4, 5

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, "Emerging properties in self-supervised vision transformers," *arXiv preprint arXiv:2104.14294*, 2021. 1, 2, 4, 5

[7] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *CVPR*, 2018, pp. 3733–3742. 1

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255. 1

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755. 1

[10] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017. 1, 2, 3

[11] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Shouling Ji, Qi Liu, Shijian Lu, and Li Cheng, "Investigating pose representations and motion contexts modeling for 3d motion prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pp. 1–16, 2022. 2

[12] Yifang Yin, Harsh Shrivastava, Ying Zhang, Zhenguang Liu, Rajiv Ratn Shah, and Roger Zimmermann, "Enhanced audio tagging via multi-to single-modal teacher-student mutual learning," in *AAAI*, 2021, vol. 35, pp. 10709–10717. 2

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018. 2, 4, 5

[14] Mehdi Noroozi and Paolo Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *ECCV*, 2016, pp. 69–84. 2

[15] Carl Doersch, Abhinav Gupta, and Alexei A Efros, "Unsupervised visual representation learning by context prediction," in *ICCV*, 2015, pp. 1422–1430. 2

[16] David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi, "Self-supervised learning of geometrically stable features through probabilistic introspection," in *CVPR*, 2018, pp. 3637–3645. 2

[17] Richard Zhang, Phillip Isola, and Alexei A Efros, "Colorful image colorization," in *ECCV*, 2016, pp. 649–666. 2

[18] Chuanchen Luo, Chunfeng Song, and Zhaoxiang Zhang, "Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup," in *European Conference on Computer Vision*. Springer, 2020, pp. 224–241. 3

[19] "Blood cell dataset.," [Online], 2017, Available: https://www.kaggle.com/paultimothymooney/blood-cells. 4

[20] Rolando Herrero, Allan Hildesheim, Concepcion Bratti, Mark E Sherman, Martha Hutchinson, Jorge Morales, Ileana Balmaceda, Mitchell D Greenberg, Mario Alfaro, Robert D Burk, et al., "Population-based study of human papillomavirus infection and cervical neoplasia in rural costa rica," *Journal of the National Cancer Institute*, vol. 92, no. 6, pp. 464–474, 2000. 4

[21] The ASCUS-LSIL Triage Study ALTS Group, "A randomized trial on the management of low-grade squamous intraepithelial lesion cytology interpretations," *American journal of obstetrics and gynecology*, vol. 188, no. 6, pp. 1393–1400, 2003. 4

[22] Allan Hildesheim, Sholom Wacholder, Gregory Catteau, Frank Struyf, Gary Dubin, Rolando Herrero, CVT Group, et al., "Efficacy of the hpv-16/18 vaccine: final according to protocol results from the blinded phase of the randomized costa rica hpv-16/18 vaccine trial," *Vaccine*, vol. 32, no. 39, pp. 5087–5097, 2014. 4

[23] Mark H Stoler, Brigitte M Ronnett, Nancy E Joste, William C Hunt, Jack Cuzick, and Cosette M Wheeler, "The interpretive variability of cervical biopsies and its relationship to hpv status," *The American journal of surgical pathology*, vol. 39, no. 6, pp. 729, 2015. 4

[24] Ying Zhang, Yifang Yin, Zhenguang Liu, and Roger Zimmermann, "A spatial regulated patch-wise approach for cervical dysplasia diagnosis," in *AAAI*, 2021, vol. 35, pp. 733–740. 4

[25] Ying Zhang, Yonit Zall, Ronen Nissim, Roger Zimmermann, et al., "Evaluation of a new dataset for visual detection of cervical precancerous lesions," *Expert Systems with Applications*, vol. 190, pp. 116048, 2022. 4

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778. 4

[27] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232. 4

[28] "Imagenette dataset.," [Online], 2019, Available: https://github.com/fastai/imagenette. 4