# BYEL : Bootstrap Your Emotion Latent

**Hyungjun Lee, Hwangyu Lim**
Graduate School of Automotive Engineering, Kookmin University
Seoul
{rhtm13, yooer}@kookmin.ac.kr


**Sejoon Lim**
Department of Automobile and IT Convergence, Kookmin University
Seoul
lim@kookim.ac.kr

## Abstract

With the improved performance of deep learning, the number of studies trying to apply deep learning to human emotion analysis is increasing rapidly. But even with this trend going on, it is still difficult to obtain high-quality images and annotations. For this reason, the Learning from Synthetic Data (LSD) Challenge, which learns from synthetic images and infers from real images, is one of the most interesting areas. In general, Domain Adaptation methods are widely used to address LSD challenges, but there is a limitation that target domains (real images) are still needed. Focusing on these limitations, we propose a framework Bootstrap Your Emotion Latent (BYEL), which uses only synthetic images in training. BYEL is implemented by adding Emotion Classifiers and Emotion Vector Subtraction to the BYOL framework that performs well in Self-Supervised Representation Learning. We train our framework using synthetic images generated from the Aff-wild2 dataset and evaluate it using real images from the Aff-wild2 dataset. The result shows that our framework (0.3084) performs 2.8% higher than the baseline (0.3) on the macro F1 score metric.

## 1 Introduction

Human emotion analysis is one of the most important fields in human-computer interaction. With the development of deep learning and big data analysis, researches on human emotion analysis using these technologies are being actively conducted [1, 2, 3, 4, 5, 6, 7, 8]. In response to this trend, three previous Affective Behavior Analysis in-the-wild (ABAW) competitions were held in conjunction with the IEEE Conference on Face and Gesture Recognition (IEEE FG) 2021, the International Conference on Computer Vision (ICCV) 2021 and the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2022 [9, 10, 11]. The 4th Workshop and Competition on ABAW, held in conjunction with the European Conference on Computer Vision (ECCV) in 2022 comprises two challenges [11]. The first one is Multi-Task-Learning (MTL), which simultaneously predicts Valence-Arousal, Facial Expression, and Action Units. The second one is Learning from Synthetic Data (LSD), which trains with synthetic datasets and infers real datasets.

Due to the successful performance of deep learning, there have been many studies using it to perform human emotion analysis[1, 12, 13]. However, in human emotion analysis using deep learning, a large amount of high-quality facial datasets are required for successful analysis. The problem is that it is difficult to easily utilize such datasets in all studies because the cost of collecting a large number of high-quality images and their corresponding labels is high. Therefore, LSD Challenge, which utilizes synthetic datasets to train a neural network and to apply real datasets to the trained neural network, is one of the most interesting areas.

In this paper, we solve the LSD Challenge of ABAW-4th [11]. A prominent problem to be solved for the LSD Challenge is that the domain of training and inference is different. To solve this problem, Domain Adaptation (DA) techniques are commonly used. DA is a method that increases generalization performance by reducing the domain gap in the feature space of the source and target domains. Traditional DA methods reduce the domain gap in the feature space of the source domain and target domain using the adversarial network [14, 15, 16]. Furthermore, studies have recently been conducted to reduce the gap between source and target domains in feature space using the characteristics of self-supervised learning (SSL) that learn similar representations in feature space without adversarial networks [17, 18]. However, both traditional DA and SSL-based DA have limitations in that both the source domain dataset and target domain dataset are necessary for the training phase. Focusing on these limitations, we propose an SSL-based novel framework that learns the emotional representation of the target domain(real images) using only the source domain(synthetic images). Our contributions are as follows.

- First, we propose the enabled emotion aware Self-Supervised Learning method to learn an invariant features that represent emotion in both the synthetic image and the real image.

- Second, we solve domain adaptation by learning the optimal representation that is also applied in real images using only the synthetic image.

We confirm the efficiency by comparing our contributions with the methods of various cases in 5.4.

## 2 Related Work

### 2.1 Self-Supervised Representation Learning

Recently, studies on methodologies for extracting representations using self-supervision are being actively conducted. MoCo [19] performs contrastive learning based on dictionary look-up. When the key and query representations are derived from the same data, learning is carried out in the direction of increasing the similarity. SimCLR [20] is proposed as an idea to enable learning without an architecture or memory bank. it learns representations to operate as a positive pair of two augmented image pairs.

All existing contrastive learning-based methodologies before BYOL use negative pairs. BYOL [21] achieved excellent performance through a method that does not use negative pairs by using a method that utilizes two networks instead of using a negative pair. In this study, an online network predicts the representation of target network which has same architecture with online network and updates the parameters of the target network using an exponential moving average. As such, the iteratively refining process is bootstrapping.

### 2.2 Human Emotion Analysis

Human Emotion Analysis is rapidly growing as an important study in Human-computer interaction field. In particular, through the Affective Behavior Analysis in-the-wild(ABAW) competition, many methodologies are proposed and their performance has been improved. In the 3rd Workshop and Competition on ABAW, the four challenges i) uni-task Valence-Arousal Estimation, ii) uni-task Expression Classification, iii) uni-task Action Unit Detection, and iv) Multi-Task Learning and evaluation are described with metrics and baseline systems [11].

Many methodologies have been presented through the ABAW challenge. D. Kollias *et al.* [6, 8] exploits convolutional features while modeling the temporal dynamics arising from human behavior through recurrent layers of CNN-RNN from AffwildNet. They perform extensive experiments with CNNs and CNN-RNN architectures using visual and auditory modalities. and show that the network achieves state-of-the-art performance for emotion recognition tasks [5]. According to one study [2], new multi-tasking and holistic frameworks are provided to learn collaboratively, generalize effectively. In this study, multi-task DNNs, being trained on AffWild2 outperform the state-of-the-art for affect recognition over all existing in-the-wild databases. D. Kollias *et al.* [1] present FacebehaviorNet and perform zero- and few-shot learning to the ability to encapsulate all aspects of facial behavior. MoCo [19] is also applied in the field of Human Emotion Analysis. EmoCo [22], an extension of the MoCo framework, removes non-emotional information in the features with the Emotion classfier, and then performs emotion-aware contrastive learning through intra-class normalization in an emotion-specific space.

Also, various new approaches for facial emotion synthesis have been presented. D. Kollias *et al.* [7] propose a novel approach to synthesizing facial effects based on 600,000 frame annotations from the 4DFAB database in terms of valence and arousal. VA-StarGAN [4] applies StarGAN to generate a continuous emotion synthesis image. D. Kollias *et al.* [3] propose a novel approach for synthesizing facial affect. In this study, impact synthesis is implemented by

Figure 1: Problem description of ABAW-4th's LSD(Learning from Synthetic Data) Challenge [11].

fitting a 3D Morphable Model to a neutral image, then transforming the reconstructed face, adding the input effect, and blending the new face and the given effect to the original image.

## 3 Problem Description

ABAW-4th's Learning from Synthetic Data (LSD) Challenge is a task that uses synthetic datasets to train neural networks and classify emotions using trained neural networks in real images. In training phase, we train neural networks $f_\theta$ that classify emotions using $Y_{true} \in \{$Anger, Disgust, Fear, Happiness, Sadness, Surprise $\}$ corresponding to synthetic image $X_{syn} \in \mathbb{R}^{N \times N}$, where $N$ is size of image. Also predicted emotions from $X_{syn}$ are defined as $Y_{pred} \in \{$Anger, Disgust, Fear, Happiness, Sadness, Surprise $\}$. In inference phase, $Y_{pred}$ is obtained using real image $X_{real} \in \mathbb{R}^{N \times N}$. Figure 1 shows our problem description.

## 4 Method

Like previous Self-Supervised Learning frameworks, our method consists of two phases [21, 19, 20]. The first, representation learning is conducted in the pre-training phase, and the second, transfer-learning is performed for emotion classification. We use the Bootstrap Your Emotion Latent (BYEL) framework to do representation learning and then transfer-learning for the emotion classification task. As shown in Figure 2 (a), the BYEL framework performs emotion-aware representation learning on feature extractor $h_\theta$. As shown in Figure 2 (b), $f_\theta$, which consists of pre-trained $h_\theta$ and classifier $c_\theta$, is trained in a supervised learning method in the emotion classification task. The final model, $f_\theta$, is formulated as equation 1, where ∘ is the function composition operator.

$$f_\theta = c_\theta \circ h_\theta (\circ : \textit{function composition operator}) \tag{1}$$

### 4.1 Bootstrap Your Emotion Latent

Inspired by the excellent performance of EmoCo [22] with MoCo [19] applied in the face behavior unit detection task, we apply BYOL [21] to solve LSD tasks. There are several changes in applying BYOL to emotion-aware representaion learning. We add Emotion Classifier $E_\theta$ and Emotion Vector Subtraction.

#### 4.1.1 Emotion Classifier.

$E_\theta$ is a matrix with $W_E \in \mathbb{R}^{size\ of\ q_\theta(z) \times C}$, where $C$ is number of emotion class. $W_E$ is a matrix that converts $q_\theta(z)$ into an emotion class. To conduct emotion-aware training as in EmoCo[22], we add Emotion Classifier $E_\theta$ to BYOL framework. The matrix $W_E$ is updated through the, $L_{classify}$, which is the Cross-Entropy of the Emotion Prediction and Emotion Label. As $W_E$ is trained, each column becomes a vector representing the corresponding emotion.

(a) Bootstrap Your Emotion Latent(Pre-training Phase)



(b) Transfer-learning Phase

Figure 2: An illustration of our method.

### 4.1.2 Emotion Vector Subtraction.

Emotion Vector Subtraction is an operation to move $q_\theta(z)$ to the emotion area within the feature space of $q_\theta(z)$. Using $W_E$, we can obtain a prediction vector excluding the emotion information $\overline{q_\theta}(z)$ by subtracting the emotion vector $w_{idx}$ from the $q_\theta(z)$ of $X_{syn}$, like EmoCo [22]. Here, $w_{idx}$ is a column vector of $W_E$ corresponding to the emotion label. In the same way, we subtract the emotion vector from the $z'$ of the target network to obtain the projection vector $\overline{z'}$ excluding the emotion vector $w_{idx}$. The whole process is formulated as equation 2.

$$\overline{q_\theta}(z) = q_\theta(z) - w_{idx}$$
$$\overline{z'} = z' - w_{idx}$$
$$(T : transpose, idx \in \{0, ..., C - 1\})$$
(2)

Figure 2 (a) shows the framework of BYEL, where feature extractor $h_\theta$, decay rate $\tau$, Projection layer $g_\theta$, Prediction layer $q_\theta$ and Augmentation function (t,t') is same as BYOL. The target network ($h_{\theta'}$, $g_{\theta'}$), which is label for representation learning, is not updated by $L_{byol}$ but only through exponential moving average of $h_\theta$, $g_\theta$ like BYOL. This target network update is formulated as an equation 3. In addition, Figure 2 (a) is an example of a situation in which the emotion label is Fear. Here, since the label index of emotion vector corresponding to Fear in $W_E$ is 2, it can be confirmed that $w_{idx}$ is subtracted from $q_\theta(z)$ and $z'$. After subtraction, as in BYOL, BYEL trains $\overline{q_\theta}(z)$ to have the same representation as $\overline{z'}$ so that $h_\theta$ performs emotion-aware representation learning for synthetic image $X_{syn}$.

$$\theta' = \tau * \theta' + (1 - \tau) * \theta (0 \leq \tau \leq 1)$$
(3)

4

### 4.2 Transfer-Learning

After the pre-training phase, we can obtain $h_\theta$ with emotion-aware representation learning. Since $h_\theta$ can extracts emotion representation at $X_{syn}$, $f_\theta$ consists of feature extractor $h_\theta$ and classifier $c_\theta$, which is a one linear layer. As shown in Figure 2 (b), $f_\theta$ is learned in the supervised learning method for the emotion classification task.

### 4.3 Loss

We use three loss functions to train our method. The first is $L_{classify}$ for emotion classification in pre-training phase, transfer-learning phase, the second is $L_{orthogonal}$ to orthogonalize the columns of $W_E$, and the third is $L_{byol}$ in pre-training phase. $L_{classify}$ is formulated as equation 4, where $p$ is the softmax function and $y$ is the ground truth.

$$L_{classify} = -\sum_{c=0}^{C-1} y_c log(p(c))(C : Class\ Number) \tag{4}$$

Inspired by Pointnet's T-Net regularization [23], which helps with stable training of transformation matrix, we use $L_{orthogonal}$ to train $E_\theta$ stably. $L_{orthogonal}$ is formulated as equation 5, where $I$ is identity matrix $\in \mathbb{R}^{C \times C}$ and $\|\cdot\|_1$ is the $L_1$ norm.

$$L_{orthogonal} = \sum_{i=0}^{C-1}\sum_{j=0}^{C-1} \left\| W_E^T * W_E - I \right\|_1 [i][j] \tag{5}$$

$$(C : Class\ Number, I : Identity\ Matrix \in \mathbb{R}^{C \times C})$$

$L_{byol}$ is the same as Mean Square Error with $L_2$ Normalization used by BYOL [21]. $L_{byol}$ is formulated as equation 6, where $\langle \cdot, \cdot \rangle$ is the dot product function and $\|\cdot\|_2$ is the $L_2$ norm.

$$L_{byol} = 2 - 2\frac{\left\langle \overline{q_\theta(z)}, \overline{z'} \right\rangle}{\left\| \overline{q_\theta(z)} \right\|_2 * \left\| \overline{z'} \right\|_2} \tag{6}$$

$L_{byel}$ is obtained by adding $\widetilde{L}_{byol}, \widetilde{L}_{classify}$ obtained by inverting t and t' in Figure 2 (a) to $L_{byol}, L_{classify}, L_{orthogonal}$ as in BYOL. Finally, $L_{byel}$ used in pre-training phase is formulated as equation 7 and Loss used in transfer-learning phase is formulated as equation 4.

$$L_{byel} = L_{byol} + \widetilde{L}_{byol} + L_{classify} + \widetilde{L}_{classify} + L_{orthogonal} \tag{7}$$

## 5 Experiments

### 5.1 Dataset

Like the LSD task dataset in ABAW-4th [11], synthetic images used in method development are all generated from real images used in validation. We can finally get a total of 277,251 synthetic images for training and a total of 4,670 real images for validation. Table 1 shows the detailed distribution of synthetic images and real images. Expression values are $\{0, 1, 2, 3, 4, 5\}$ that correspond to {Anger, Disgust, Fear, Happiness, Sadness, Surprise}.

### 5.2 Settings

In the pre-training phase, we apply LARS [24] optimizer as in BYOL [21] to train the BYEL framework and the $\tau$, augmentation t, projection layer $g_\theta$ and prediction layer $q_\theta$ are the same as BYOL [21], where epoch is 100, learning rate is 0.2, batch size is 256 and weights decay is $1.5 - e^{-6}$. In transfer-learning phase, we apply Adam[25] optimizer to learn the model $f_\theta$ consisting of $h_\theta$ that completed the 100-th epoch learning and 1 linear layer $c_\theta$, where epoch is 100, learning rate is $0.1 - e^{-3}$ and batch size is 256. The size of images $X_{real}, X_{syn} \in \mathbb{R}^{N \times N}$ is all set to $N = 128$. We select a model with the best F1 score across all 6 categories(i.e., macro F1 score) after full learning. All experimental environments are implemented in pytorch [26] 1.9.0.

Table 1: Distribution of datasets by emotion class.

| | Number of Images | |
|---|---|---|
| **Expression** | **Synthetic Image** | **Real Image** |
| **0:Anger** | 18,286 | 804 |
| **1:Disgust** | 15,150 | 252 |
| **2:Fear** | 10,923 | 523 |
| **3:Happiness** | 73,285 | 1,714 |
| **4:Sadness** | 144,631 | 774 |
| **5:Surprise** | 14,976 | 603 |

Table 2: Comparison of macro F1 scores according to methods

| | macro F1 score with unit 0.01($\uparrow$) | |
|---|---|---|
| **Method** | **Validation set** | **Test set** |
| **baseline** | 50.0 | 30.0 |
| **ResNet50 with LSD** | 59.7 | - |
| **BYOL with LSD** | 59.7 | 29.76 |
| **BYEL with LSD** | **62.7** | **30.84** |

## 5.3 Metric

We use the evaluation metric F1 score across all 6 categories(i.e., macro F1 score) according to the LSD task evaluation metric proposed in ABAW-4th [11]. F1 score is defined as the harmonic mean of recall and precision and is formulated as equation 8. Finally, the F1 score across all 6 categories (i.e., macro F1 score) is formulated as equation 9. The closer the macro F1 score is to 1, the better the performance.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$
$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{8}$$
$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$P_{LSD} = \frac{\sum_{c=0}^{5} F_1^c}{6} \tag{9}$$

## 5.4 Results

We demonstrate the effectiveness of our method through comparison with the baseline presented in ABAW-4th [11]. A baseline model is set to a transfer-learning model of ResNet50 [27] pre-trained with ImageNet. ResNet50 with LSD is a case where ResNet50 is trained using the LSD dataset. BYOL with LSD is a case of training in the LSD dataset using the BYOL [21] framework and then transfer-learning. BYEL with LSD is our method. Table 2 summarizes results. We also prove that our method is more effective than other methods.

### 5.4.1 Ablation Study.

We analyze the relationship between pre-training epoch and performance through the performance comparison of $f_\theta^e = c_\theta \circ h_\theta^e$ according to the training epoch of pre-training. $h_\theta^e$ represents the situation in which training has been completed using the BYEL framework for as many as $e$ epochs. $f_\theta^e$ is a transfer-learned model using $h_\theta^e$. In Table 3, it can be confirmed that the larger the pre-training epoch, the higher the performance.

Table 3: Comparison of macro F1 scores in ablation study

| | macro F1 score with unit 0.01($\uparrow$) | |
|---|---|---|
| **Method** | **Validation set** | **Test set** |
| **baseline** | 50.0 | 30.0 |
| $f_\theta^{45}$ | 56.9 | - |
| $f_\theta^{90}$ | 59.3 | - |
| $f_\theta^{100}$ | **62.7** | **30.84** |

# 6 Conclusion

In this paper, inspired by EmoCo, we propose an emotion-aware representaion learning framework applying BYOL. This framework shows generalization performance in real images using only synthetic images for training. In section 5.4, we demonstrate the effectiveness of our method. However, it does not show a very large performance difference compared to other methods. Therefore, we recognize these limitations, and in future research, we will apply the Test-Time Adaptation method to further advance.

# References

[1] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021.

[2] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021.

[3] Dimitrios Kollias, Shiyang Cheng, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, 128(5):1455–1484, 2020.

[4] Dimitrios Kollias and Stefanos Zafeiriou. Va-stargan: Continuous affect generation. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 227–238. Springer, 2020.

[5] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.

[6] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019.

[7] Dimitrios Kollias, Shiyang Cheng, Maja Pantic, and Stefanos Zafeiriou. Photorealistic facial synthesis in the dimensional affect space. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[8] Dimitrios Kollias, Mihalis A Nicolaou, Irene Kotsia, Guoying Zhao, and Stefanos Zafeiriou. Recognition of affect in the wild using deep neural networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1972–1979. IEEE, 2017.

[9] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 637–643. IEEE, 2020.

[10] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3652–3660, October 2021.

[11] Dimitrios Kollias. Abaw: Learning from synthetic data & multi-task learning challenges. *arXiv preprint arXiv:2207.01138*, 2022.

[12] Geesung Oh, Euiseok Jeong, and Sejoon Lim. Causal affect prediction model using a past facial image sequence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3550–3556, 2021.

[13] Euiseok Jeong, Geesung Oh, and Sejoon Lim. Multitask emotion recognition model with knowledge distillation and task discriminator. *arXiv preprint arXiv:2203.13072*, 2022.

[14] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[15] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[16] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

[17] Pallavi Jain, Bianca Schoen-Phelan, and Robert Ross. Self-supervised learning for invariant representations from multi-spectral and sar images. *arXiv preprint arXiv:2205.02049*, 2022.

[18] Hiroyasu Akada, Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Self-supervised learning of domain invariant features for depth estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3377–3387, 2022.

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[22] Xuran Sun, Jiabei Zeng, and Shiguang Shan. Emotion-aware contrastive learning for facial action unit detection. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021.

[23] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[24] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.