micromachines

MDPI

Article

Extreme Low-Resolution Activity Recognition Using a Super-Resolution-Oriented Generative Adversarial Network

Mingzheng Hou [1,2,†], Song Liu [2,†], Jiliu Zhou [2,*], Yi Zhang [2,*] and Ziliang Feng [1,2]

[1] National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610065, China; houmingzheng@scu.edu.cn (M.H.); fengziliang@scu.edu.cn (Z.F.)
[2] College of Computer Science & College of Software Engineering, Sichuan University, Chengdu 610065, China; 18296875611@163.com
* Correspondence: zhoujl@scu.edu.cn (J.Z.); yzhang@scu.edu.cn (Y.Z.)
† These authors contributed equally to this work.

Abstract: Activity recognition is a fundamental and crucial task in computer vision. Impressive results have been achieved for activity recognition in high-resolution videos, but for extreme low-resolution videos, which capture the action information at a distance and are vital for preserving privacy, the performance of activity recognition algorithms is far from satisfactory. The reason is that extreme low-resolution (e.g., 12 × 16 pixels) images lack adequate scene and appearance information, which is needed for efficient recognition. To address this problem, we propose a super-resolution-driven generative adversarial network for activity recognition. To fully take advantage of the latent information in low-resolution images, a powerful network module is employed to super-resolve the extremely low-resolution images with a large scale factor. Then, a general activity recognition network is applied to analyze the super-resolved video clips. Extensive experiments on two public benchmarks were conducted to evaluate the effectiveness of our proposed method. The results demonstrate that our method outperforms several state-of-the-art low-resolution activity recognition approaches.

Keywords: activity recognition; extreme low-resolution activity recognition; super-resolution; generative network

Citation: Hou, M.; Liu, S.; Zhou, J.; Zhang, Y.; Feng, Z. Extreme Low-Resolution Activity Recognition Using a Super-Resolution-Oriented Generative Adversarial Network. *Micromachines* **2021**, *12*, 670. https://doi.org/10.3390/mi12060670

Academic Editors: Bihan Wen and Zhangyang (Atlas) Wang

Received: 7 May 2021
Accepted: 4 June 2021
Published: 8 June 2021

1. Introduction

The number of videos created by various recording devices has far surpassed what we can process manually. Therefore, it is crucial to develop intelligent video understanding algorithms for various tasks, such as video recommendation and human activity recognition. Many efforts have been made in the field of activity recognition. Typical methods include the two-stream convolution network [1] and C3D [2]. These approaches assume that the provided videos are high-quality and that video regions of human activities are large enough to model spatiotemporal information. However, in certain situations, such as video surveillance in far-field, where a human is usually very far way from the camera, this assumption is invalid as only low-resolution videos are acquired since the ROI (regions-of-interest) can be extremely tiny in the video frames.

Furthermore, some concerns about privacy protection arise. Increasing numbers of cameras, including security and protection system, wearable devices, and even our cellphones, are recording videos at either public or private places. Even worse is that these recording videos are often stored in the cloud. Concerning our privacy, it is risky to store or upload these videos to remote servers for the reason that they can be leaked or stolen. One possible solution is to transmit the videos with the lowest resolution required for recognition or analysis. However, current methods cannot adapt well to these limitations due to severe changes in extracted features, which raise the challenge of effective activity recognition with extreme low-resolution frames.

In response to this problem, many methods have been proposed. Chen et al. [3] introduced a semi-coupled, filter-sharing two-stream network which utilizes high-resolution

videos in the training phase to assist the low-resolution convolutional network (ConvNet) in learning to better distinguish low-resolution features. Based on the observation that training low-resolution videos can benefit from high-resolution data, Xu et al. [4] proposed a fully-coupled two-stream network in which low-resolution videos share all filter parameters with high-resolution videos. Ryoo et al. [5] designed a novel two-stream multi-Siamese network that learns an embedding space shared by low-resolution videos generated with different low-resolution transforms. The aforementioned approaches can be roughly divided into two categories. One is to learn distinguished features for low-resolution data by sharing parameters between high-resolution data and low-resolution videos [3–5]. The other is to extract as much latent information from low-resolution images as possible to improve the recognition rate [5]. Due to the utilization of optical flow for temporal information modeling, the computational costs of these methods are high, which impedes their practical application despite the impressive results these methods achieve. Additionally, in semi-coupled and fully-coupled networks [3,4], high-resolution images are only adopted as auxiliary data to assist training, and latent high-resolution information is not fully explored.

Since video resolution has a critical impact on feature extraction, a direct approach is to enhance the video resolution for activity recognition. Recently, learning-based image/video super-resolution (SR) has been broadly studied, and a great number of methods with state-of-the-art performance have been proposed. We have also noticed that a similar idea has been utilized for other topics in low-resolution scenarios and has achieved encouraging results, such as in face recognition, small object detection, and person re-identification [6–8].

Inspired by this, we propose a super-resolution generative adversarial network for extreme low-resolution activity recognition, which provides a seamless workflow to super-resolve low-resolution images for analyzing human motion. As shown in Figure 1, our approach consists of two modules, namely, a super-resolution module and a spatiotemporal modeling module. Specifically, the super-resolution module can robustly super-resolve high-resolution images from low-resolution images. The spatiotemporal modeling module adopts these generated high-resolution videos as inputs for activity recognition. We must mention that Ugur et al. [9] also proposed a similar method (Prog. DVSR) to ours, which utilizes a progressive generative approach to improve the quality of low-resolution actions followed by a action classifier network. Two main differences exist between both methods: (1) different network structures, including both SR and activity recognition modules are adopted, and (2) Prog. DVSR [9] introduces a weakly trained attention mechanism to help focus on the activity regions in videos, while our approach utilizes long temporal convolution to model the spatiotemporal information in videos.

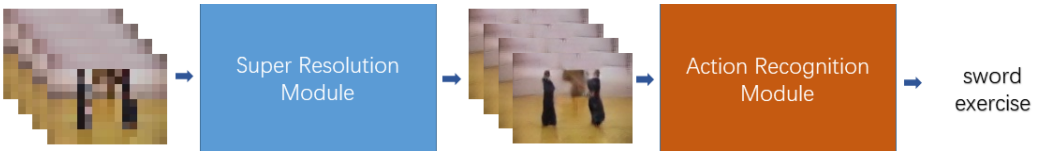


Figure 1. The overview of our approach.

The main contributions of this paper can be summarized as follows.

(1) We propose an extreme low-resolution activity recognition approach aided by a super-resolution generative adversarial network.
(2) A novel training strategy, called long-range temporal convolution, is used in the recognition module to learn action representations over a long temporal range.
(3) Extensive experiments are conducted, which show that the performance of our approach outperforms several state-of-the-art methods by a large margin despite the fact that we use only RGB images as inputs to avoid the extraction of optical flow.

2. Related Work

2.1. General Activity Recognition

The existing research in video activity recognition can be broadly categorized into handcrafted and deep learning-based methods. To represent spatiotemporal information of human motion in videos, various handcrafted-based methods, such as space–time interest points (STIP) [10], histogram of optical flow [11], 3D histogram of gradient [12], and SIFT-3D [13] have been proposed. Presently, an improved dense trajectory [14] has been shown to outperform the handcrafted-based approach. Benefiting from the rapid development of deep learning in computer vision, researchers have started to utilize deep models such as VGG [15] and ResNet [16] to represent spatiotemporal information in video clips or image sequences. Karpathy et al. [17] made the first attempt to deploy deep learning for activity recognition. Later, Simonyan and Zisserman [1] proposed a two-stream ConvNet. The two streams of the ConvNet consist of a spatial stream and a temporal stream, which respectively adopt RGB images and optical flow images as inputs. This network obtained a large-margin recognition rate improvement. To model long-range temporal information, Wang [18] introduced a temporal segment network that obtained a high score on two benchmarks: UCF101 [19] and HMDB51 [20]. While these 2D Conv-based methods have achieved impressive results, they face two difficulties. One difficulty is that they cannot effectively model temporal information in videos although optical flows are adopted as inputs. The other difficulty is that extracting optical flow images is time-consuming. These problems were solved by C3D [2], which applies 3D convolutional filters to model spatiotemporal information from short video clips. Later, Carreira [21] inflated 2D convolutional kernels that successfully leveraged parameters pretrained on ImageNet. Qiu et al. [22] further boosted the performance by decomposing 3D convolutional kernels into 2D convolutional kernels in the spatial domain plus 1D convolutional kernels in the temporal domain.

Generally, promising performance has been achieved by these methods to recognize activity in well-prepared videos. However, there are practical demands for low-resolution activity recognition in some specific fields.

2.2. Low-Resolution Activity Recognition

To address practical problems, several recent approaches [3–5,23] to extreme low-resolution activity have been proposed. These methods can recognize activity to a certain degree in extremely low-resolution (12×16 pixels) videos that even humans cannot identify. The key point of these methods is figuring out how to recover or obtain lost visual information with limited pixels and how to fully utilize the information in high-resolution images. Observing that images downsampled from the same image have different pixels, Ryoo et al. [23] proposed the concept of inverse super-resolution (ISR). This method focused on obtaining more information in low-resolution images generated from a single image by learning an optimal set of image transforms. Additionally, to better learn inherent information obtained from multiple low-resolution images, Ryoo et al. [5] introduced a novel multi-Siamese loss. Ryoo's works are the paradigm for obtaining lost visual information from limited pixels. Another concern is how to utilize high-resolution information. Chen et al. [3] designed a semi-coupled two-stream network in which a low-resolution net shares part filters with a high-resolution net. It employs high-resolution images to assist training. Xu et al. [4] observed that effectively utilizing the information in high-resolution images has a significantly positive impact on the performance improvement of low-resolution recognition. They proposed a fully coupled two-stream network in which high-resolution images are directly adopted as inputs. By utilizing a low-resolution net which shares all convolutional filters with a high-resolution net, the performance of the fully coupled two-stream network is marginally outperformed other methods. In addition, Ugur et al. [9] built a natural low-resolution benchmark TinyVIRAT (https://www.crcv.ucf.edu/tiny-actions-challenge-cvpr2021/, accessed on 3 July 2020) and

proposed a novel method which utilizes a progressive generative approach to improve the quality of low-resolution actions.

Revisiting the approaches [5,23] proposed by Ryoo et al., the significance of recovering or obtaining lost visual information from limited pixels is repeatedly highlighted. From these coupled series methods [3,4], we find that utilizing information in high-resolution images is equally important. However, Ryoo et al. did not leverage information in high-resolution images. A coupled network [3,4] adopts only high-resolution images as inputs to assist in training distinguished features, while low-resolution images are not actually enhanced by the useful information in high-resolution images. Therefore, we introduce a super-resolution module that can simultaneously and effectively recover lost visual information and utilize high-resolution information to enhance low-resolution information.

2.3. Super-Resolution in Other Low-Resolution Recognition Field

On the other hand, many works [6–8] in other low-resolution fields, such as low-resolution face verification, small object detection, and low-resolution person re-identification, employ a super-resolution method to address the low-resolution problem and have all achieved impressive results. Ataer-Cansizoglu et al. [6] proposed a deep learning approach based on identity-preserving super-resolution for very low-resolution face verification. Bai et al. [7] designed an end-to-end multitask generative adversarial network for small object detection. To address the low-resolution and scale mismatching problem in person re-identification, Wang [8] proposed a cascade super-resolution generative adversarial network.

3. The Approach

As shown in Figure 1, in this section we describe, in detail, our approach for extreme low-resolution activity recognition. The basic architecture of our super-resolution module, which adopts a generative adversarial network that can robustly recover images with limited pixels, is discussed first. To utilize information in high-resolution images, we also hold the assumption that high-resolution training videos are available. Then, we introduce the basic architecture of our activity recognition module, which employs a 3D residual convolutional network as a spatiotemporal representation model. Finally, a training strategy, called long-range temporal convolution, will be introduced.

3.1. Super-Resolution Module

Similar to most prior works [3,4,23], we assume that in the training phase, we have high-resolution videos. Unlike semi-coupled [3] and fully coupled [4] networks that take high-resolution images as inputs to learn distinguished features, we recover low-resolution images via a generative adversarial network to enhance low-resolution features. Figure 2 shows the general architecture of our super-resolution module.

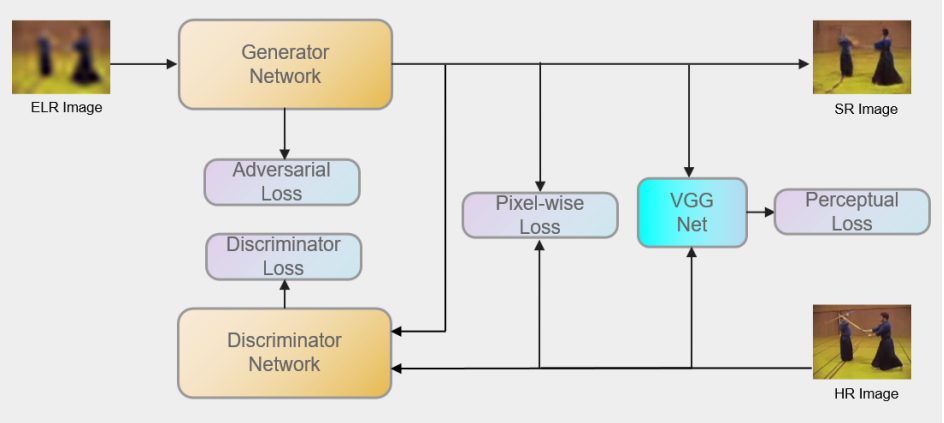


Figure 2. The general architecture of our super-resolution module. ELR denotes extreme low resolution, and SR and HR represent super-resolution and high resolution, respectively.

3.1.1. Generative Adversarial Network

Since generative adversarial network (GAN) [24] was proposed, its strong performance in generating life-like images has impressed us. GANs optimize the generator and discriminator, in turn, via an adversarial process, which enables the generator to achieve an optimal state. The loss of a GAN can be formulated as follows:

$$L_{GAN} = E_{x \sim P_{data}(x)}[\log D_\theta] + E_{z \sim P_z(z)}[\log (1 - D_\theta(G_w(z)))] \quad (1)$$

where x represents the real data, z denotes the random noise, and D_θ and G_w stand for the discriminator and generator, respectively. The adversarial process between the discriminator and generator can be formulated to

$$\arg \min_G \max_D L_{GAN}(G, D) \quad (2)$$

Our goal is simultaneously recovering low-resolution images and obtaining distinguishable features for activity recognition. It is difficult to obtain lost visual information from such limited pixels (12×16). More importantly, unlike other super-resolution tasks [25,26] that are focused on reconstructing images without losing details, we concentrate on recovering lost information that can contribute to recognition. Many studies [2,21,22] have confirmed that capturing the motion of humans to model spatiotemporal information is vital for activity recognition. Therefore, the lost information we want to recover from limited pixels is clear silhouettes of humans and objects, which can be used to model the motion of humans. In summary, the proposed GAN should have the ability to deal with large downscale factors and to roughly restore the outline of humans. Inspired by prior attempts [25–27] in super-resolution with a large scale factor ($\times 8$), we adopt the unique architecture of a generator that can effectively deal with a large scale factor in SDSR [27] and the relativistic discriminator used in ESRGAN [26].

3.1.2. Network Architecture

Our generator consists of a feature extractor and an upsampler. Figure 3 demonstrates the general architecture of our generator, and Figure 4 illustrates, in detail, the architectures of the feature extractor and upsampler. In particular, the feature extractor in the generator we used in SDSR [27] adopts the dense deep back-projection network (D-DBPN) [28] as the backbone and improves the ability to extract features from extreme low-resolution images by utilizing the residual in the residual dense block (RRDB) proposed by Wang et al. [26]. The number of RRDB blocks in our feature extractor is set to 10. To learn effective mapping from extreme low-resolution images to high-resolution images, the unique architecture of the upsampler in SDSR [27] is employed. In the upsampler, the features extracted from extreme low-resolution images are upscaled and downscaled alternatively with deep back-projection layers. Specifically, the extracted features are upscaled three times and downscaled two times using the architecture illustrated in Figure 4b. Borrowing the idea from ESRGAN [27], we adopt the relativistic discriminator [29] to determine whether the high-resolution label is more realistic than the generated image.

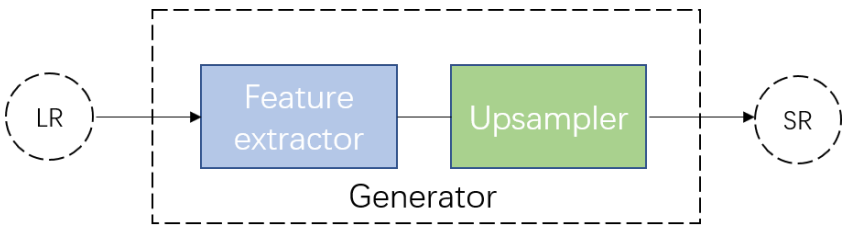
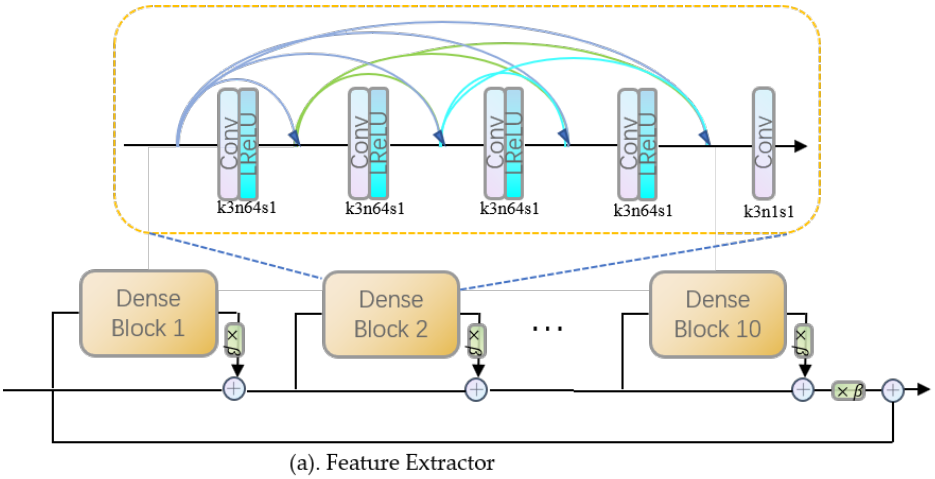
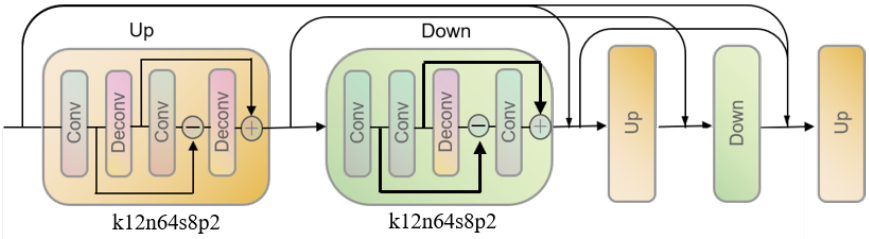


Figure 3. The architecture of our generator.



(a). Feature Extractor



(b). Upsampler

Figure 4. The architecture of the feature extractor and upsampler. Illustration of the (**a**) feature extractor and (**b**) upsampler. β is the residual scaling parameter, which is set to 0.2 k and denotes the kernel size, n represents the number of filters, s is the size of the stride, and p is the size of padding. In (**b**) the conv and deconv share the same numbers of kernel size, features, stride, and padding.

3.1.3. The Loss Function for the Super-Resolution Module

The loss function is critical for the performance of our super-resolution module. Generally, the key component of a GAN's loss is MSE. We additionally introduce the SRGAN adversarial loss and VGG loss to measure the perception similarity between generated images and ground truth. In the following, the details of the MSE loss, adversarial loss, and perception loss-based VGG network are described.

MSE loss. Pixel-wise MSE loss can be computed using the following equation:

$$L_{MSE} = \frac{1}{r^2WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}(I_{x,y}^{HR} - G_\theta(I^{LR})_{x,y})^2 \tag{3}$$

where I^{HR} and I^{LR} respectively represent the high-resolution image and low-resolution image. rW and rH is the size of high-resolution image, where r is the factor of downsampling.

Perception loss. Perception loss is usually used to measure the similarity in feature space, which has proven efficient for SR. Here, we also introduce perception loss to improve the SR performance and a pretrained VGG-19 network is adopted to extract the features from the first 12 convolution layers. We use Φ to represent VGG network extracting features. The perception loss is calculated as follows

$$L_{perc} = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}(\Phi(I^{HR})_{x,y} - \Phi(G_\theta(I^{LR}))_{x,y})^2 \tag{4}$$

where W and H respectively denote the dimensions of the feature maps extracted by VGG network.

Adversarial loss. In addition, adversarial loss is used and it can be calculated as follows:

$$L_{adv} = -\log D_\theta(G_\theta(I^{LR})) \tag{5}$$

where $D_\theta(G_\theta(I^{LR}))$ is adopted to distinguish the super-resolved image $G_\theta(I^{LR})$ from the ground truth image. Finally, the total loss is obtained by combining the MSE loss, perception loss, and adversarial loss as follows:

$$L_{gen} = L_{MSE} + \alpha L_{perc} + \gamma L_{adv} \tag{6}$$

where α and γ are weights trading off the different terms. We set weights $\alpha = 0.006$, $\gamma = 0.001$ in this paper.

3.2. Activity Recognition Module

Formally, we assume that we are given extreme low-resolution videos with L frames. A random frame L_1 is then selected as a temporal start point to generate a video clip $\{L_1, L_2, L_3, ..., L_k\}$. Our goal is to recognize the activity in such extreme low-resolution videos. The process can be represented as follows:

$$score = H(F(G(L_1; w), G(L_2; w), ..., G(L_k; w))) \tag{7}$$

where G is the generator of our super-resolution module, and w represents its parameters. F can be an arbitrary end-to-end activity recognition model. Different from prior works [3–5] that employed a two-stream network adopting optical flow as inputs, a residual 3D convolutional network [30] is selected as F due to its powerful ability to model spatiotemporal information and avoid precomputing optical flow. Based on the output of model F, the probability of each activity class will be computed by the prediction function H. Here, we adopt the softmax function for H.

Specifically, the architecture of activity recognition module is shown in Figure 5 and details of each part are illustrated in Table 1. Our recognizer consists of 5 convolutional parts of which the 1st part includes 64 $7 \times 7 \times 7$ convolutional filters and the remaining parts are composed of ResNeXt blocks. The series of ResNet block have a strong power on extracting feature and can alleviate the problem of gradient vanishing. Figure 6 depicts the block architecture of ResNet series in which ResNeXt is adopted since its group convolution further eases training and improves performance.

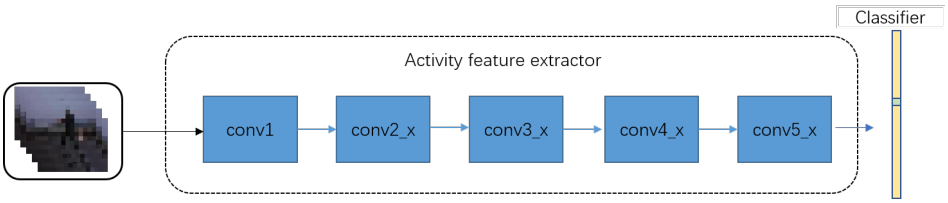


Figure 5. The architecture of our activity recognition module.

Table 1. The architecture of activity feature extractor. F is the number of feature channels corresponding in Figure 6, and N is the number of blocks in each layer.

Part	Output Size	F	N
conv1	$8 \times 56 \times 56$	64	conv $7 \times 7 \times 7$, 64
conv2_x	$8 \times 56 \times 56$	128	8
conv3_x	$4 \times 28 \times 28$	256	24
conv4_x	$2 \times 14 \times 14$	512	36
conv5_x	$1 \times 7 \times 7$	1024	3

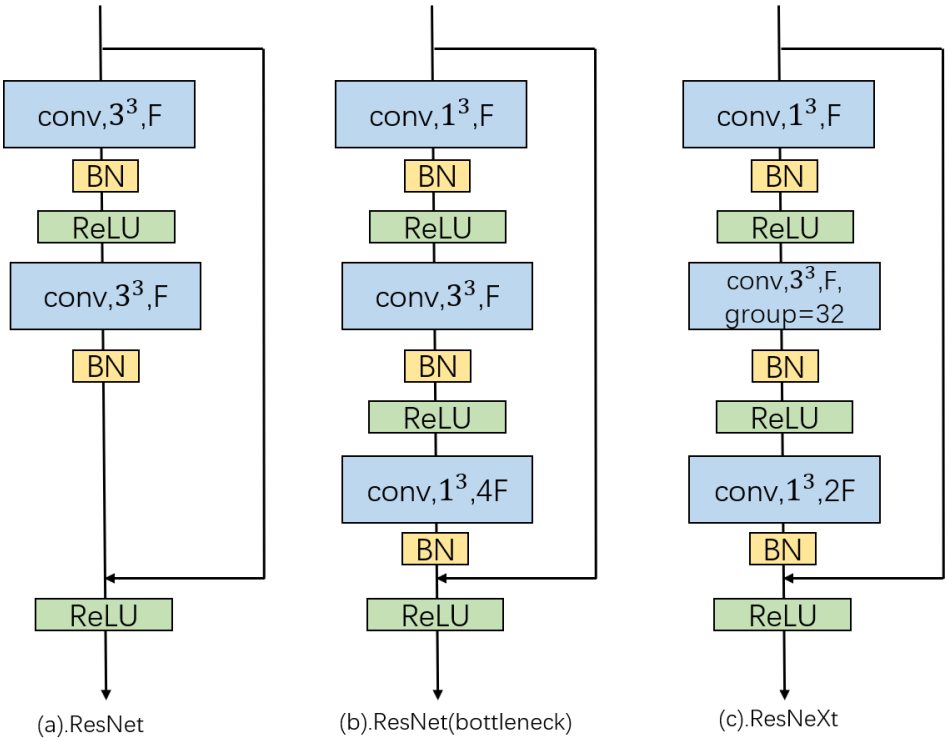


Figure 6. Block of each architecture. conv and 3^3 denote a $3 \times 3 \times 3$ convolutional filter, while F and group is the number of feature channels and groups of group convolution, respectively.

Formally, we use cross entropy loss to train the activity recognition module, and the loss function can be given as follows:

$$l_{clf} = -\sum_{c=0}^{C-1} y_c \log F(G(L_1, ..., L_k)) \tag{8}$$

where $L_1, ..., L_k$ are the low-resolution video frames, C and y_c are the number of action class and the labels of action, respectively.

3.3. Training Strategy

3.3.1. Data Augmentation

Prior works on low-resolution recognition mainly performed experiments on two standard datasets, i.e., HMDB51 [20] and UCF101 [19], which have only 3.7 k and 0.2 k training videos, respectively. The scale of the two datasets is truly small. Since the similarity of adjacent frames in videos is extraordinarily high, it makes no sense to use all of the frames in a video for training our super-resolution module. Compared with other similar low-resolution tasks [7,8] using GANs, such as SOD-MTGAN [7], which has 80k training images, the amount of data our super-resolution module can use is much smaller. It is risky to train with such a limited amount of data as it can easily cause overfitting.

Motivated by the practice in [1], data augmentation is employed for training our proposed GAN. In the learning phase of the GAN for HMDB51, the UCF101 dataset is introduced. Different from modifying the architecture of our network, we directly merge two datasets. Specifically, we first divide the two datasets into training and test sets according to the official partition file. Then, the training sets of HMDB51 and UCF101 are merged to train the super-resolution module.

3.3.2. Long-Range Temporal Convolutions

Previous works for high-resolution activity recognition with CNN architectures, such as C3D [2] and R2 + 1D [22], typically learned activity representations at the level of a

few video frames and thus failed to model longer-range temporal information. Despite this minor flaw, these methods have a powerful performance due to the abundant spatial information in high-resolution videos. However, for extreme low-resolution videos, the spatial information of a single frame is limited. Following the idea of [31], we use long-term temporal convolutions to model spatial–temporal information over a longer range to better learn low-resolution video representations. Specifically, the number of input frames is typically 16. We boost the number to 64, which can cover a more complete temporal extent to operate spatial–temporal convolutions.

4. Experiments

4.1. Dataset

The HMDB51 [20] and Dogcentric [32] datasets have been popularly used for extreme low-resolution recognition evaluation in previous works [3–5,23]. We choose the HMDB51 dataset to make a direct comparison between our approach and previous works. The UCF101 [19] dataset is chosen instead of Dogcentric for the following reasons. On the one hand, our goal is to recognize reliable human, not dog, activities at distances and to preserve human privacy in extreme low-resolution videos. The videos in Dogcentric are taken from the dog's viewpoint and record the dog's activities, such as turning the dog's head to the right/left and playing with a ball. On the other hand, UCF101 contains various videos ranging from videos in which humans near the camera to videos in which humans are poorly visible in the wild, which fits our goal effectively. All these factors make UCF101 a more reasonable and challenging dataset for extremely low-resolution activity recognition.

Specifically, HMDB51 consists of 6766 video clips that are collected from movies and web videos with 51 activity categories. UCF101 is a popular video dataset containing 13,320 video clips belonging to 101 activity classes. The resolution of the above two datasets is 240 $\times$ 320 pixels. To simulate an extremely low-resolution dataset, we resize these videos to 12 $\times$ 16 pixels with average downsampling and then resize the 12 $\times$ 16 videos back to their original size using bicubic interpolation. Several corresponding low- and high-resolution frames are shown in Figure 7. Then, these datasets are split into two parts via the provided train/test split files.

Figure 7. Visualization of some corresponding low-, super-, and high-resolution images of HMDB51. The left column shows low-resolution images; the middle column shows super-resolved images; the right column shows high-resolution images.

4.2. Implementation Details

Our training process consists of two stages: (1) training the super-resolution module and recovering super-resolution frames from low-resolution video and (2) training the recognition module with the recovered frames of each video as inputs.

For the super-resolution module, we train our GAN on the HMDB51 and UCF101 datasets at low resolution from scratch. As discussed before, we use simulated low-resolution data as inputs and high-resolution data as labels. Adam [33] is adopted to

optimize the network parameters with a learning rate of 10^{-3} and a weight decay of 10^{-5}. The whole process stops at 300 epochs, with the batch size set to 60.

For the recognition module, we follow [30]. Using their available pretrained model, we finetune it on the HMDB51 and UCF101 datasets at low resolution. We adopt 16/64 frames as inputs, respectively. Stochastic gradient descent [34] is employed to optimize the network parameters with a learning rate of 10^{-3} and a weight decay of 10^{-5}. All our experiments are implemented in PyTorch on Ubuntu with two Nvidia 1080Ti GPUs.

4.3. Ablation Studies

Influence of the Super-Resolution Module. Table 2 (the 1st row vs. 2nd row and 3rd row vs. 4th row) compares the performance of our model with/without the super-resolution module. From Table 2, it is observed that without long-range temporal convolutions but with the enhancement of our super-resolution module, the performance of our model outperforms other methods without the super-resolution module by a small margin (i.e., 0.6% accuracy on HMDB51 and 1% accuracy on UCF101). After long-range temporal convolutions are introduced, the influence of the super-resolution module increases. The performance of our model with the super-resolution module outperforms other methods without that module by a sizable margin (i.e., 1.2% in accuracy on HMDB51 and 1.6% on UCF101). Figure 7 shows the corresponding low- and high-resolution frames and super-resolved frames recovered from the super-resolution module. These results demonstrate that the super-resolution module can effectively help increase the accuracy of low-resolution activity recognition. With long-range temporal convolutions, the lost information recovered from low-resolution frames can be more fully explored. In addition, as shown in Table 3, our approach obtains considerable performance on TinyVIRAT dataset which makes a margin of about 1% comparing with baseline model(the 1st row vs. 2nd row and 3rd row vs. 4th row).

Influence of Long-Range Temporal Convolutions. From Table 2 (1st row vs. 3rd row and 2nd vs. 4th row), we can see that the accuracy drops by 7.5% and 8.1%, respectively, without long-range temporal convolutions, and from Table 3 (1st row vs. 3rd row and 2nd vs. 4th row), we can see that the accuracy drops by 5.2% and 5.1%, respectively, without long-range temporal convolutions. The reason is that without long-range temporal convolution, we can only model a limited amount of the temporal information which is important for activity recognition. To effectively learn spatial–temporal information in low-resolution videos, we use long-range temporal convolutions to train our network.

Evaluation of Our Method. As shown in Figure 8, the confusion matrices illustrate that the performance of our proposed model with the super-resolution module and long-range temporal convolutions is visually more remarkable than that of our baseline method. Figure 8b shows that the recognition accuracy of most activities is considerably high. However, several actions, such as 'hit', 'jump', and 'shoot bow' are misrecognized as 'swing baseball', 'catch', and 'laugh'. This is because these actions have similar subactions and lose too much information in the extreme low-resolution videos, which is demonstrated in Figure 9.

Table 2. Performance of our proposed method with/without two mechanisms on HMDB51 and UCF101.

Super-Resolution Module	Long-Range Temporal Convolutions	Accuracy	
		HMDB51	UCf101
×	×	45.8%	65.3%
✓	×	46.6%	66.2%
×	✓	53.3%	69.6%
✓	✓	54.6%	71.1%

Table 3. Performance of our proposed method with/without two mechanisms on TinyVIRAT.

Super-Resolution Module	Long-Range Temporal Convoluton Module	F1 Score TinyVIRAT
×	×	73.89%
✓	×	74.11%
×	✓	79.02%
✓	✓	79.77%

4.4. State-of-the-Art Comparison

We compare our proposed method with several state-of-the-art low-resolution activity recognition models [3–5,35] on the challenging 12 × 16 HMDB51 dataset. Table 4 lists the performance, modalities and number of input frames, from which we conclude that our method outperforms all other state-of-the-art methods on the HMDB51 dataset. More importantly, in the case where only 16-RGB frames are used as input, our method still obtains better performance than the second-best low-resolution recognition model by approximately 1.5%. If we follow the set of input frames of previous works, our method outperforms the second best model by a large margin. Moreover, we make a comparison on UCF101 dataset between our proposed method and DVSR. Table 5 shows the comparing result from which we can see our approach outperforms DVSR [9] by a considerable margin of accuracy. This clearly demonstrates the effectiveness of our method on low-resolution activity recognition.

Table 4. The performance of the proposed method and other state-of-the-art methods on the 12 × 16 HMDB51 dataset.

Methods	Modalities	Input Frames	Accuracy
pLRN + Tennet [35]	RGB	-	21.7%
ISR [23]	RGB	-	28.68%
Semi-Coupled [3]	RGB + Optical flow	64	29.2%
Multi-Siamese [5]	RGB + Optical flow	64	37.7%
DVSR [9]	RGB	16	41.63%
Fully-Coupled [4]	RGB + Optical flow	64	44.96%
Ours	RGB	16	46.4%
Ours	RGB	64	54.4%

Table 5. The performance of the proposed method and other state-of-the-art methods on the UCF101 dataset. LRTC denotes long-range temporal convolutons.

Method	Input Size	Accuracy
Bicubic I3D	14 × 14	14.1%
DVSR	14 × 14	68.2%
Prog.DVSR	14 × 14	70.6%
Ours	12 × 16	66.2%
Ours + LRTC.	12 × 16	71.1%

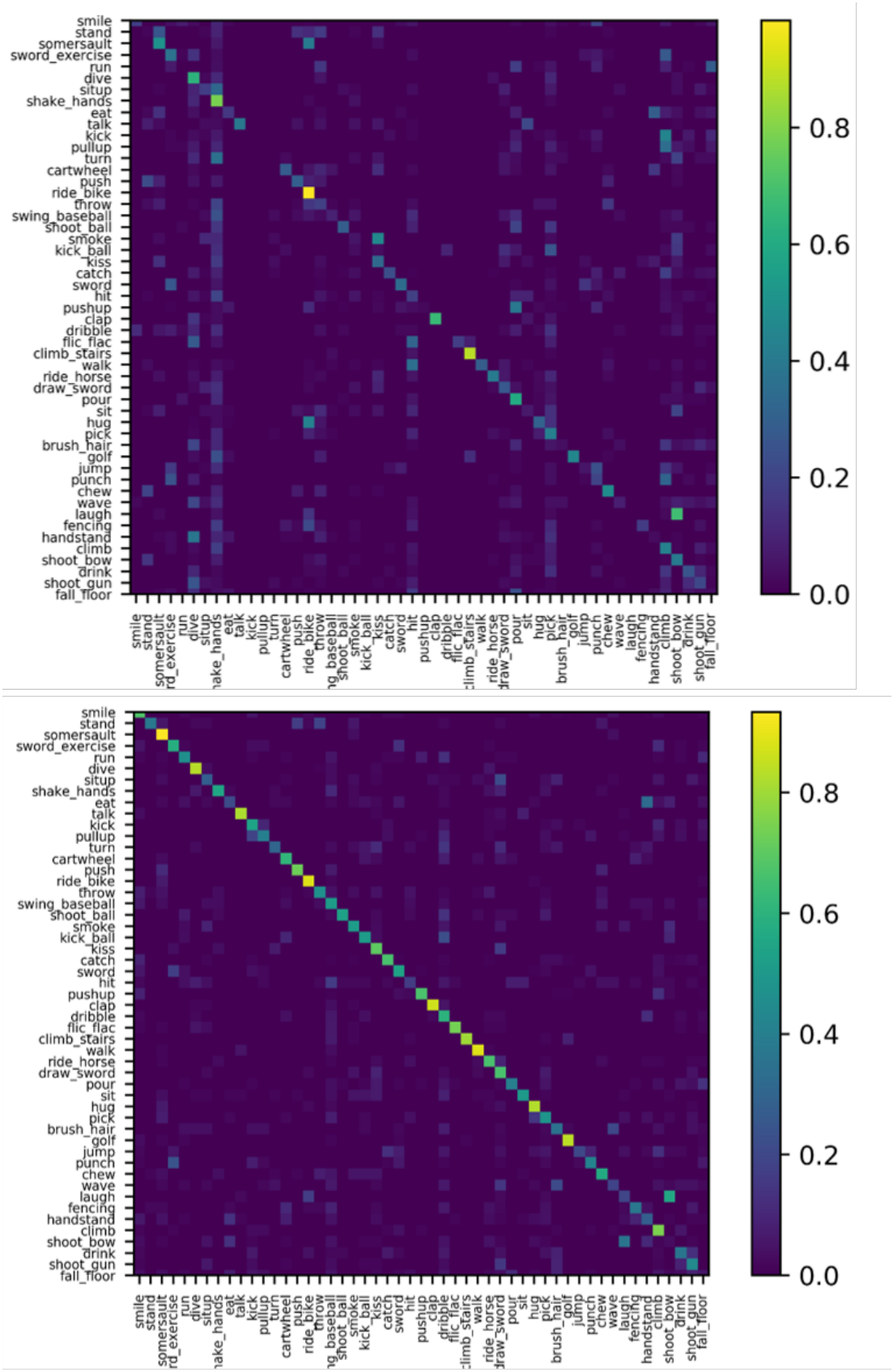


Figure 8. Confusion matrix on the 12 × 16 HMDB51 dataset. The X-axis denotes the predicted labels, and the y-axis presents the ground truth labels. (**a**) The result of our baseline model without the super-resolution module and long-range temporal convolutions. (**b**) The result of our proposed model with a super-resolution module and long-range temporal convolutions.

(a). Ground truth (b). Misrecognized results

Figure 9. Snapshots of parts of misrecognized activities. (**a**) is the ground truth label, and (**b**) is the misrecognized activity.

5. Discussion and Conclusions

We must mention that in this paper, our goal of using GAN is to generate super-resolution images from low-resolution images to help recognition. It is true that we can use more advanced variants of GAN to obtain better super-resolution performance, but we restrict our choice to SDSR based on two factors: (1) the basic idea of this manuscript is to propose a framework for extreme low-resolution activity recognition, not a new SR method; and (2) for activity recognition, it is not necessary to recover all the details but general silhouettes of humans and objects. It must also be mentioned that different SR modules may further improve the subsequent recognition performance, and this is planned for our future work.

In this paper, we propose a super-resolution generative network-based method to recognize activities in extreme low-resolution videos. Our method consists of two modules, namely, a super-resolution module and an activity recognition module. The proposed super-resolution module generates super-resolution frames from low-resolution frames, which can recover lost information to improve recognition. The recognition module adopts the recovered frames as inputs and predicts the category of the activity in the low-resolution videos. Extensive experiments on the HMDB51 and UCF101 datasets demonstrate that our method improves the state-of-the-art accuracy performance compared to other methods.

In our future work, more network architectures for both super-resolution and activity recognition will be evaluated. In addition, more datasets with multiple levels of resolution will be included to evaluate the robustness of the proposed model.

Author Contributions: Conceptualization, M.H. and S.L.; methodology, M.H. and S.L.; software, S.L.; validation, M.H., S.L. and Y.Z.; formal analysis, S.L. and Y.Z.; investigation, J.Z., Y.Z. and Z.F.; resources, J.Z., Y.Z. and Z.F.; data curation, S.L.; writing—original draft preparation, M.H. and S.L.; writing—review and editing, Y.Z.; visualization, S.L.; supervision, J.Z., Y.Z. and Z.F.; project administration, Y.Z.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Key R&D Program of China grant number 2018YFC0830300 and LAIW (AI in LAW) Advanced Deployed Discipline of Sichuan University, China.

Conflicts of Interest: We declare no conflict of interest.

References

1. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *arXiv* **2014**, arXiv:1406.2199.
2. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (CVPR), Santiago, Chile, 13–16 December 2015; pp. 4489–4497.
3. Chen, J.; Wu, J.; Konrad, J.; Ishwar, P. Semi-coupled two-stream fusion convnets for action recognition at extremely low resolutions. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rose, CA, USA, 27–29 March 2017; pp. 139–147.

4. Xu, M.; Sharghi, A.; Chen, X.; Crandall, D.J. Fully-coupled two-stream spatiotemporal networks for extremely low resolution action recognition. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, CA, USA, 12–14 March, 2018; pp. 1607–1615.
5. Ryoo, M.; Kim, K.; Yang, H. Extreme low resolution activity recognition with multi-siamese embedding learning. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New Orleans, LA, USA, 2–7 February 2018; pp. 7315–7322.
6. Ataer-Cansizoglu, E.; Jones, M.; Zhang, Z.; Sullivan, A. Verification of very low-resolution faces using an identity-preserving deep face super-resolution network. *arXiv* **2019**, arXiv:1903.10974.
7. Bai, Y.; Zhang, Y.; Ding, M.; Ghanem, B. Sod-mtgan: Small object detection via multi-task generative adversarial network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 206–221.
8. Wang, Z.; Ye, M.; Yang, F.; Bai, X.; Satoh, S. Cascaded SR-GAN for scale-adaptive low resolution person re-identification. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, 13–19 July 2018; pp. 3891–3897.
9. Demir, U.; Rawat, Y.S.; Shah, M. Tinyvirat: Low-resolution video action recognition. In Proceedings of the International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 7387–7394.
10. Laptev, I. On space-time interest points. *Int. J. Comput. Vis.* **2005**, *64*, 107–123. [CrossRef]
11. Chaudhry, R.; Ravichandran, A.; Hager, G.; Vidal, R. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami beach, FL, USA, 20–21 June 2009; pp. 1932–1939.
12. Klaser, A.; Marszałek, M.; Schmid, C. A spatio-temporal descriptor based on 3D-gradients. In Proceedings of the British Machine Vision Conference (BMVC), Leeds, UK, 3–9 September 2008; pp. 1–10.
13. Scovanner, P.; Ali, S.; Shah, M. A 3-dimensional sift descriptor and its application to action recognition. In Proceedings of the ACM international conference on Multimedia, Augsburg, Germany, 24–29 September 2007; pp. 357–360.
14. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 3–6 December 2013; pp. 3551–3558.
15. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
17. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 1725–1732.
18. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2740–2755. [CrossRef] [PubMed]
19. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
20. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the International Conference on Computer Vision (ICCV), Bercelona, Spain, 6–13 November 2011; pp. 2556–2563.
21. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
22. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6450–6459.
23. Ryoo, M.; Rothrock, B.; Fleming, C.; Yang, H.J. Privacy-preserving human activity recognition from extreme low resolution. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), San Francisco, CA, USA, 4–9 February 2017; pp. 4255–4262.
24. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *arXiv* **2014**, arXiv:1406.2661.
25. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
26. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change Loy, C. ESRGAN: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September, 2018.
27. Huang, Y.; Lu, Z.; Shao, Z.; Ran, M.; Zhou, J.; Fang, L.; Zhang, Y. Simultaneous denoising and super-resolution of optical coherence tomography images based on generative adversarial network. *Opt. Express* **2019**, *27*, 12289–12307. [CrossRef] [PubMed]
28. Haris, M.; Shakhnarovich, G.; Ukita, N. Deep back-projection networks for super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 19–21 June 2018, pp. 1664–1673.
29. Jolicoeur-Martineau, A. The relativistic discriminator: A key element missing from standard GAN. *arXiv*, **2018**, arXiv:1807.00734.
30. Hara, K.; Kataoka, H.; Satoh, Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 19–21 June 2018; pp. 6546–6555.

31. Varol, G.; Laptev, I.; Schmid, C. Long-term temporal convolutions for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1510–1517. [CrossRef] [PubMed]
32. Iwashita, Y.; Takamine, A.; Kurazume, R.; Ryoo, M.S. First-person animal activity recognition from egocentric videos. In Proceedings of the International Conference on Pattern Recognition (ICPR), Stockholm, Sweden, 24–28 August 2014; pp. 4310–4315.
33. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
34. Polyak, B.T.; Juditsky, A.B. Acceleration of stochastic approximation by averaging. *SIAM J. Control. Optim.* **1992**, *30*, 838–855. [CrossRef]
35. Yu, T.; Wang, L.; Guo, C.; Gu, H.; Xiang, S.; Pan, C. Pseudo low rank video representation. *Pattern Recognit.* **2019**, *85*, 50–59. [CrossRef]