ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448422

ZE-FESG: A ZERO-SHOT FEATURE EXTRACTION METHOD BASED ON SEMANTIC GUIDANCE FOR NO-REFERENCE VIDEO QUALITY ASSESSMENT

*Yachun Mi, Yu Li, Yan Shu, Shaohui Liu**

Harbin Institute of Technology, Harbin, China

ABSTRACT

Although the current deep neural network based no-reference video quality assessment (NR-VQA) methods can effectively simulate the human visual system (HVS), their interpretability is getting worse. The current methods only extract the low-level features of space and time of the video and do not consider the impact of high-level semantics. However, the high-level semantic information in the video related to human subjective perception and related to its own quality can be perceived by the HVS. In this work, we design the multidimensional feature extractor (MDFE), which takes the text descriptions related to video quality factors as semantic guidance, and uses the Contrastive Language-Image Pre-training (CLIP) model to perform zero-shot multidimensional feature extraction. Then, we further propose a zero-shot feature extraction method based on semantic guidance (ZE-FESG), which treats the MDFE as a feature extractor and acquires all the semantically corresponding features of the video by sliding over each frame of the video. Extensive experiments show that the proposed ZE-FESG has better interpretability and performance than the current mainstream 2D-CNN based feature extraction methods for NR-VQA. The code will be released on https://github.com/xiao-mi-d/ZE-FESG.

Index Terms— Video Quality Assessment, Semantic Guidance, Zero-shot, Multidimensional Feature Extractor

1. INTRODUCTION

Rapid advances in technology have dramatically lowered the barriers to video production, which has allowed more and more users to shoot and upload videos to online platforms using portable devices. Hence, good no-reference video quality assessment (NR-VQA) algorithms have become essential.

It is well known that the mean opinion score (MOS) for video quality is based on human perception. So both the traditional manual feature-based NR-VQA method and the deep learning-based NR-VQA method aim to simulate the perceptual ability of the human visual system (HVS) to some extent.

*Corresponding author. This study is supported by State Key Laboratory of Communication Content Cognition, People's Daily Online under Grant 2020YFB1406902 and A12003.

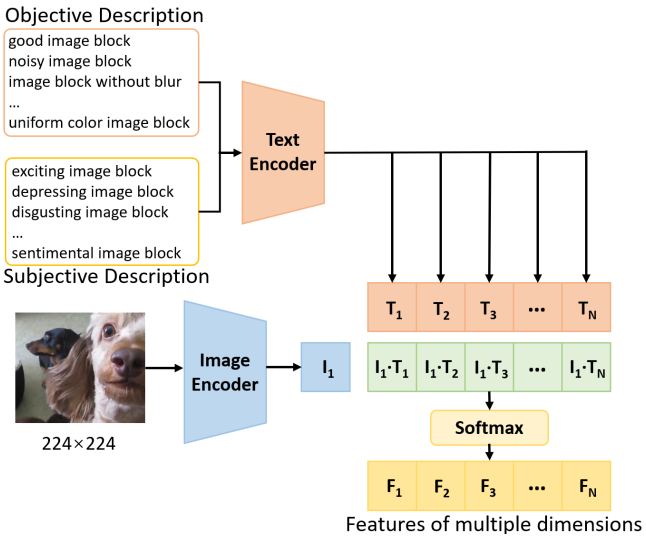

Fig. 1. A summary of our multidimensional feature extractor (MDFE). We use objective and subjective text descriptions as semantic guidance inputs to CLIP [1], and then calculate the similarity value between the image and each text description to obtain multidimensional features related to quality factors.

With the thriving of Deep Neural Networks (DNN), DNN-based methods for NR-VQA have obtained excellent performance compared with the traditional methods. Generally, the NR-VQA model consists of two modules: the feature extraction module and the feature aggregation module. The feature extraction module is essential for the whole model, as the quality of the extracted video features determines the performance of the model. However, since the feature extraction modules of current DNN-based NR-VQA methods are extracting high-level abstract visual features of the video, although the performance of NR-VQA methods is improving, the interpretability is declining. Currently, DNN-based NR-VQA models mainly use 2D-CNN [2, 3], 3D-CNN [4], Transformer [5, 6] and their combinations [7, 8, 9, 10] for feature extraction. As shown in table 1. Since the current video quality assessment datasets [11, 12, 13] are labeled only as a score (MOS), this makes these models do not well reflect the video quality related factors such as video content (which affects human subjective thoughts), distortion types (e.g., sharpness, noise, contrast, distortion, etc.), and distortion levels.

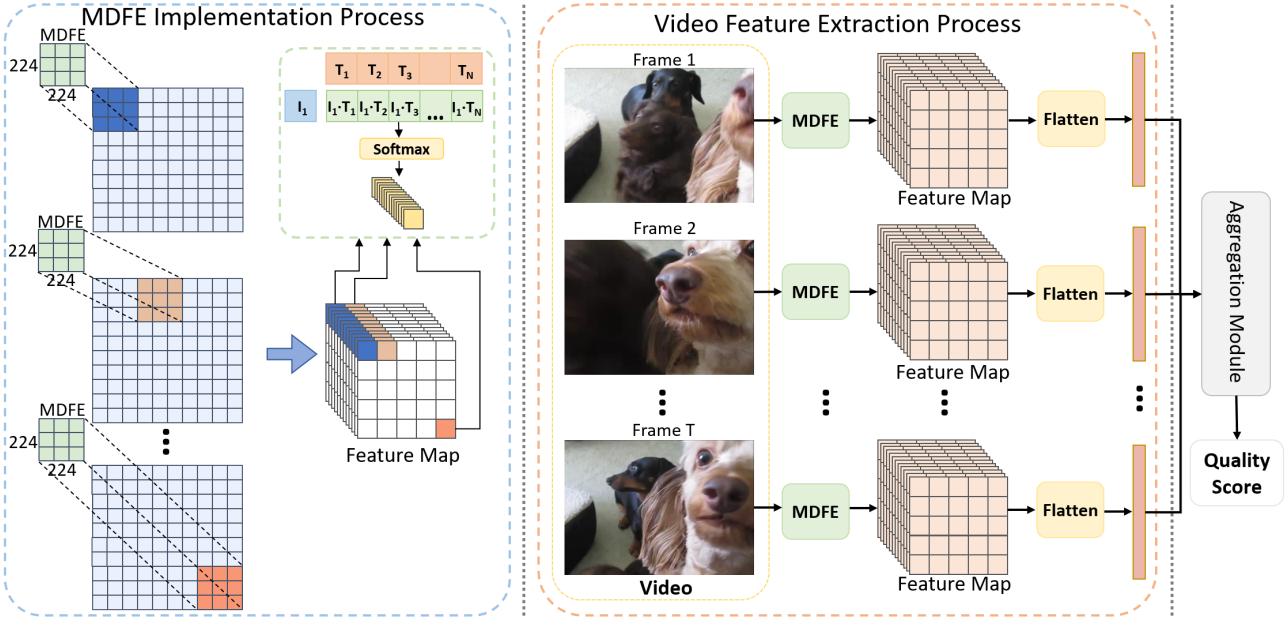

Fig. 2. Overview of the zero-shot feature extraction method based on semantic guidance.

Table 1. Summary of feature extraction models for NR-VQA.

Methods	Venus	Feature Extraction
V-MEON[4]	ACMMM'18	C3D
VSFA[2]	ACMMM'19	ResNet-50
GST-VQA[3]	TCSVT'21	VGG-16
PVQ[7]	CVPR'21	PaQ2PiQ + R3D
BVQA[8]	TCSVT'22	2DCNN + 3DCNN
SimpleVQA[9]	ACMMM'22	2DCNN + 3DCNN
FAST-VQA[5]	ECCV'22	Video Swin Transformer
DisCoVQA[6]	TCSVT'23	Video Swin Transformer
MD-VQA[10]	CVPR'23	EfficientNetV2 + R3D

The above problem can be solved by Contrastive Language Image Pre-training (CLIP) [1] model, which is a joint vision and language model trained with massive image-text pairs. Since the 400 million pairs of image-text data used to train CLIP are labeled with natural language descriptions, the perceptual ability of CLIP is highly aligned with human perception. Therefore, the perceptual ability contained in CLIP is worth exploring. Recently, in [14], it has been verified that CLIP is not only very sensitive to obvious semantic features such as visual categorization and action recognition, but also has a good ability to perceive some perceptual aspects of visual information, such as clarity, objective perception (e.g., degree of color, haze, and low light, etc.), and emotional perception (e.g., sadness, pleasure, and calmness, etc.). This shows that CLIP can capture the quality-related factors of an image from multiple dimensions, which provides a new way of thinking for NR-VQA.

In this paper, we propose a zero-shot feature extraction method based on semantic guidance (ZE-FESG) to address the challenges mentioned above. Objective and subjective descriptions related to the quality of the image are used as inputs to CLIP, and the probability values obtained by softmax are used as the acquired feature values corresponding to the text descriptions, as shown in Fig. 1. Then we consider the above CLIP with text descriptions as a small multidimensional feature extractor (MDFE), which extracts a small range of features on each frame of the video and acquires the features of the whole video frame by sliding, as shown in Fig. 2. Finally, we compare our proposed feature extraction method with the current mainstream feature extractors used in NR-VQA under the same aggregation module. Extensive experimental results show that our method not only has better performance but also better interpretability.

2. PROPOSED METHOD

In this section, we first introduce the multidimensional feature extractor (MDFE), as shown in Fig. 1. Then we will elaborate on the proposed zero-shot feature extraction method based on semantic guidance (ZE-FESG), as shown in Fig. 2.

2.1. Multidimensional Feature Extractor (MDFE)

By contrastive learning of images and languages on a large amount of data, CLIP can capture the relationship between human language and visual perception well, and this also gives CLIP the capability to perform visual content quality assessment [14]. Therefore, fully exploiting the perceptual capability of CLIP model can improve its performance in quality assessment tasks. As shown in Fig. 1, we will introduce in detail the objective and subjective descriptions of video quality that we designed to fully exploit the perceptual ability of CLIP, and the process of extracting features by CLIP in the MDFE module.

Design of Objective and Subjective Descriptions of Video Quality. In order to fully exploit the visual perception capability of CLIP, we design as many objective and subjective descriptions related to video quality factors as possible. We design 35 objective descriptions and 17 subjective descriptions, some of which are shown in Fig. 1. Using these text descriptions we will be able to obtain a 52-dimensional feature vector corresponding one-to-one with the semantic information from an image of size 224×224.

Zero-shot CLIP for Feature Extraction. CLIP consists of an visual encoder (E_v) and a text encoder (E_t). The visual encoder has two versions, ResNet [15] and ViT [16], which transform image into feature vectors, and we use VIT-B/32 as the visual encoder in our method. The text encoder is a transformer [17] that takes sentences as input and outputs a feature vector. The formulae are as follows:

$$f_t = E_t(T) \tag{1}$$

$$f_v = E_v(V) \tag{2}$$

where T is the text description and V is an image of size 224×224. Then calculate the cosine similarity f_c between V and T:

$$f_c = \frac{f_v \cdot f_t}{\|f_v\| \, \|f_t\|} \tag{3}$$

Similarity calculation of the image with all the text descriptions will result in the multidimensional feature $F_c = \left\{ f_c^i |_0^{n-1} \right\}$, where n is the number of text descriptions.

The obtained cosine similarity feature F_c is then softmax processed to obtain the final extracted feature F:

$$F = \left\{ \frac{e^{f_c^i}}{\sum_{k=0}^{n-1} e^{f_c^k}} |_{i=0}^{n-1} \right\} \tag{4}$$

2.2. Zero-shot Feature Extraction Method based on Semantic Guidance(ZE-FESG)

Restricted by the input limitations of the visual encoder of VIT-B/32, MDFE can only extract a small piece (224×224) of information from the video frame, so we regard MDFE as a feature extraction window, similar to the operation of convolutional kernel, which obtains the features of the whole video frame by sliding. MDFE extracts a 52-dimensional feature vector from each location. Then, it assigns the features that correspond to the same text description to the same feature map. The position of the feature values in the feature map reflects the location of the window when the features are extracted, as shown in Fig. 2 (left). Finally, 52 feature maps will be obtained, where each feature map corresponds to a text description.

As shown in Fig. 2 (right), we perform the above operation for each frame of the video and flatten the obtained feature maps to get the features of the whole video $F_f = \left\{ f_f^j |_{j=0}^{t-1} \right\}$:

$$Map = MDFE(VF) \tag{5}$$

$$f_f = Flatten(Map) \tag{6}$$

Where VF stands for video frame and t stands for the number of video frames.

3. EXPERIMENTS

3.1. Experimental Setup

Model Baselines. Although recent research has focused on methods that combine 2D-CNN and 3D-CNN, as well as methods based on video swin transformer, the goal of this paper is to propose a novel feature extraction method ZE-FESG that is similar to the NR-VQA methods based on 2D-CNN, which extract features from each frame of the video separately. Therefore, we use the feature extraction models based on 2D-CNN that are widely applied in NR-VQA, namely ResNet-50 [15], VGG-16 [18] and EfficientNetV2-S [19], which are pre-trained on ImageNet [20], as the baseline models for comparison, as shown in Table 1. To fairly evaluate the effectiveness of different feature extractors, we adopt the same feature aggregation model proposed in VSFA [2] for all our experiments, which employs GRU [21] for temporal modeling. In particular, in our approach, we do not incorporate 3D-CNN based methods [22, 23] or Transformer methods [16, 24], since it is intuitive to illustrate the effectiveness of our method simply by comparing it with mainstream 2D-CNN approaches [15, 18, 19] for NR-VQA.

Dataset. We train all the models on the KoNViD-1K [12] dataset, which contains 1200 user generated videos targeting natural distortions with a resolution of 960×540. All of these videos have a duration of 8 seconds and a frame rate of 24/25/30 (fps). The dataset is used for training, validation and testing in the ratio of 6:2:2 respectively.

Evaluation Metrics. The Spearman Rank Order Correlation Coefficient (SROCC), the Kendall Rank Order Correlation Coefficient (KROCC), the Pearson Linear Correlation Coefficient (PLCC), and the Root Mean Square Error (RMSE) are used as evaluation Metrics. Higher SROCC, PLCC, KROCC and lower RMSE scores represent models with better performance.

Implementation Details. We employ PyTorch framework and two NVIDIA GeForce RTX 3090 cards to train the model in all experimental implementations. In model training, we use L1 loss and AdamW optimizer. The initial learning rate is set to 1×10^{-4} and adjusted by Linear Warmup.

3.2. Results

Performance Comparison. We compare the mainstream 2D-CNN based feature extraction methods mentioned in Table. 1. The results of the comparison experiments are shown in Table 2. Specifically, we use the 52 text descriptions related to video quality factors introduced in Section 2.1 as a semantic guidance to extract the features of the video and

perform 85 MDFE operations on video frames. The results show that ZE-FESG significantly outperforms the other models when using all 52 text descriptions. When using only objective descriptions, ZE-FESG significantly outperforms ResNet-50 and VGG-16, and is slightly worse than EfficientNetV2. However, when using only subjective descriptions, the performance of ZE-FESG drops dramatically.

Table 2. Performance comparison of different feature extraction methods. 'o' denotes using only objective description (35), 's' denotes using only subjective description (17), and 'a' denotes using both objective and subjective description.

Method	SROCC↑	KROCC↑	PLCC↑	RMSE↓
VGG-16	0.740	0.542	0.742	0.412
ResNet-50	0.771	0.575	0.775	0.398
EfficientNetV2	0.797	0.599	0.795	0.387
ZE-FESG$_o$	0.782	0.590	0.786	0.401
ZE-FESG$_s$	0.499	0.343	0.531	0.528
ZE-FESG$_a$	**0.814**	**0.622**	**0.826**	**0.375**

By analyzing the results, it can be concluded that the features extracted by CLIP corresponding to the objective descriptions are more representative of the quality of the video. Although the features corresponding to subjective descriptions cannot accurately reflect the feature information of the video, they can be used as a supplement to the objective description features to enrich the extracted video features.

Table 3. Ablation study on text description. 'ob' denotes objective description, 'sub' denotes subjective description; 'f' denotes randomly using the half of the text descriptions, 'a' denotes all descriptions. Perform 85 MDFE operations.

Description	SROCC↑	KROCC↑	PLCC↑	RMSE↓
only-ob-f	0.751	0.550	0.755	0.422
only-ob-a	0.782	0.590	0.786	0.401
only-sub-f	0.358	0.242	0.378	0.576
only-sub-a	0.499	0.343	0.531	0.528
obj-sub-f	0.803	0.602	0.798	0.392
obj-sub-a	**0.814**	**0.622**	**0.826**	**0.375**

Ablation Study. We explore the impact of different numbers of text descriptions on the performance of ZE-FESG, as shown in Table 3. The results show that increasing the number of text descriptions can improve the performance of ZE-FESG. This suggests that more features of the video can be captured using more text descriptions.

Further, we also investigate the effect of the number of MDFE operations on video frames on ZE-FESG. Specifically, we use all 52 text descriptions as semantic guidance, and perform 25, 45, 85 and 165 MDFE operations on each frame of the video respectively. The experimental results are shown in Table 4. The performance of ZE-FESG gradually increases as the number of feature extraction times increases, but when the number reaches 85 times, the performance no longer increases. This indicates that 85 MDFE operations are sufficient to extract almost all the features of the video, and continued MDFE operations yield redundant and repetitive information.

Table 4. Ablation study on the number of times MDFE performs feature extraction on each frame of the video.

Number	SROCC↑	KROCC↑	PLCC↑	RMSE↓
25	0.782	0.596	0.786	0.409
45	0.790	0.604	0.800	0.386
85	0.814	0.622	**0.826**	**0.375**
165	**0.816**	**0.633**	0.811	0.378

Interpretability Study. In order to verify that the features extracted by ZE-FESG have good interpretability, we study the features corresponding to the four text descriptions, as shown in Table 5. Specifically, we add noise or darkness or reduce color or contrast to the original video, respectively. We perform global average pooling on all the feature maps obtained corresponding to a particular description, and use the average of all the values after pooling as the extracted video feature corresponding to that description. The experimental results demonstrate that the video features extracted by ZE-FESG can accurately reflect the changes after performing the operations corresponding to the text descriptions, e.g., after adding noise, the feature values corresponding to the noise will significantly increase. This indicates that the features extracted by ZE-FESG can correspond well to the semantic meaning of the language. At the same time, this also demonstrates the better interpretability of our method.

Table 5. Interpretability study of features extracted by ZE-FESG. '$+$' denotes add, '$-$' denotes reduce.

Type	Noisy	Dark	Colorful	Contrast
Original	0.0289	0.0523	0.0783	0.0657
$+noisy$	**0.2103**	0.0604	0.0645	0.0704
$+dark$	0.0482	**0.1601**	0.0626	0.0475
$-colorful$	0.0274	0.0633	**0.0112**	0.0599
$-contrast$	0.0318	0.0564	0.0651	**0.0175**

4. CONCLUSION

In this paper, we propose ZE-FESG, a zero-shot feature extraction method based on semantic guidance for NR-VQA. We first design MDFE, which takes 52 objective and subjective text descriptions related to video quality factors as semantic guidance, and then performs zero-shot multidimensional feature extraction on video frames using CLIP. We consider the MDFE as a feature extraction window, which acquires features throughout the video frame by sliding over it. The MDFE performs sliding operation on each frame of the video to obtain the features of the whole video. Experimental results show that our proposed ZE-FESG has better performance and interpretability. We believe that the feature extraction method based on semantic guidance proposed in this paper will offer some insights for developing the next generation of NR-VQA models.

5. REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. on Machine Learning (ICML)*. PMLR, 2021, pp. 8748–8763.

[2] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proc. ACM Multimedia Conf. (MM)*, 2019, pp. 2351–2359.

[3] B. Chen, L. Zhu, G. Li, F. Lu, H. Fan, and S. Wang, "Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment," *IEEE Trans. Circuits Syst. Video Technol*, vol. 32, no. 4, pp. 1903–1916, 2021.

[4] W. Liu, Z. Duanmu, and Z. Wang, "End-to-end blind quality assessment of compressed videos using deep neural networks.," in *Proc. ACM Multimedia Conf. (MM)*, 2018, pp. 546–554.

[5] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin, "Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling," in *Proc. European Conf. on Computer Vision. (ECCV)*. Springer, 2022, pp. 538–554.

[6] H. Wu, C. Chen, L. Liao, J. Hou, W. Sun, Q. Yan, and W. Lin, "Discovqa: Temporal distortion-content transformers for video quality assessment," *IEEE Trans. Circuits Syst. Video Technol*, vol. 33, no. 9, pp. 4840–4854, 2023.

[7] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-vq: 'patching up' the video quality problem," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2021, pp. 14019–14029.

[8] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang, "Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception," *IEEE Trans. Circuits Syst. Video Technol*, vol. 32, no. 9, pp. 5944–5958, 2022.

[9] W. Sun, X. Min, W. Lu, and G. Zhai, "A deep learning based no-reference quality assessment model for ugc videos," in *Proc. ACM Multimedia Conf. (MM)*, 2022, pp. 856–865.

[10] Z. Zhang, W. Wu, W. Sun, D. Tu, W. Lu, X. Min, Y. Chen, and G. Zhai, "MD-VQA: Multi-dimensional quality assessment for ugc live videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2023, pp. 1746–1755.

[11] Z. Sinno and A. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. Image Process*, vol. 28, no. 2, pp. 612–627, 2018.

[12] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The konstanz natural video database (konvid-1k)," in *Proc. IEEE. Conf. Qual. Multimedia Exper. (QoMEX)*, 2017, pp. 1–6.

[13] Y. Wang, S. Inguva, and B. Adsumilli, "Youtube ugc dataset for video compression research," in *Proc. IEEE. Int. Conf. Multimedia Signal Process. (MMSP)*, 2019, pp. 1–5.

[14] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proc. AAAI Conf. on Artificial Intelligence. (AAAI)*, 2023, vol. 37, pp. 2555–2563.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2016, pp. 770–778.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. Annual Conf. on Neural Information Process. Syst. (NeurIPS)*, vol. 30, 2017.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *Proc. Int. Conf. on Machine Learning (ICML)*. PMLR, 2021, pp. 10096–10106.

[20] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255.

[21] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. IEEE Int. Conf. on Computer Vision. (ICCV)*, December 2015, pp. 4489–4497.

[23] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," in *Proc. IEEE Int. Conf. on Computer Vision. (ICCV)*, Oct 2017, pp. 3154–3160.

[24] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2022, pp. 3202–3211.