Convolutional transformer network for paranasal anomaly classification in the maxillary sinus

Debayan Bhattacharya[a], Finn Behrendt[b], Lennart Maack[l], Benjamin Tobias Becker[c], Dirk Beyersdorff[d], Elina Petersen[e], Marvin Petersen[f], Bastian Cheng[g], Dennis Eggert[h], Christian Betz[i], Anna Sophie Hoffmann*[j], and Alexander Schlaefer*[k]

[a,b,k,l]Hamburg University of Technology, Hamburg, Germany
[a,c,d,e,f,g,h,i,j] Universitätsklinikum Hamburg-Eppendorf, Hamburg, Germany

ABSTRACT

Large-scale population studies have examined the detection of sinus opacities in cranial MRIs. Deep learning methods, specifically 3D convolutional neural networks (CNNs), have been used to classify these anomalies. However, CNNs have limitations in capturing long-range dependencies across the low and high level features, potentially reducing performance. To address this, we propose an end-to-end pipeline using a novel deep learning network called ConTra-Net. ConTra-Net combines the strengths of CNNs and self-attention mechanisms of transformers to classify paranasal anomalies in the maxillary sinuses. Our approach outperforms 3D CNNs and 3D Vision Transformer (ViT), with relative improvements in F1 score of 11.68% and 53.5%, respectively. Our pipeline with ConTra-Net could serve as an alternative to reduce misdiagnosis rates in classifying paranasal anomalies.

Keywords: Paranasal anomaly, maxillary sinus, anomaly classification, CNN, Transformer, Hybrid Network

1 INTRODUCTION

Paranasal sinus anomalies are a common but clinically significant finding in the radiological assessment of the head and neck area.[1] These anomalies present various treatment challenges[2] and have been the subject of numerous studies to analyze their occurrence and progression in the general population.[3] Accurate diagnosis of paranasal inflammations is crucial for patient care and cost reduction. Clinicians rely on cross-sectional views from CT and MRI to assess these conditions. Misdiagnosis can cause unnecessary concern and costs.[4]

Deep learning is widely used for paranasal pathology screening, including sinusitis classification,[5,6] fungal ball detection,[7] and polyp and cyst detection.[8–10] However, CNNs struggle to capture global cues, while transformers excel in computer vision tasks[11,12] because of learning global cues. However, without pretraining, vision transformers fail to learn meaningful representations.[13] Hence, hybrid models combining CNNs and transformers have been proposed[14–16] that benefit from the inductive bias of CNNs. Our novel ConTra-Net network draws inspiration from TransMed[17] and forms dependencies across multiple CNN feature levels.[14] By combining CNNs inductive bias with self-attention mechanisms, we capture global context and increase representation quality across the low and high level features. ConTra-Net classifies healthy and anomalous maxillary sinuses, offering potential for diverse paranasal anomaly identification. We are leveraging CNNs and Multi-head self-attention (MHSA) block to improve the accuracy and efficiency of diagnosing paranasal anomalies in the maxillary sinus. These anomalies, including polyps and cysts, can be differently located along the sinus walls and present in various shapes, sizes, and contrasts. Therefore we hypothesize that extracting features with long-range dependencies, large receptive fields and invariance with respect to the appearance of the anomaly may be beneficial for the classification task.

Our main contributions can be summarised as follows. First, we propose a hybrid convolutional transformer network (ConTra-Net) for classifying maxillary sinus anomalies. This network combines the inductive biases of a CNN with the ability to capture long-range dependencies among multi-level features, resulting in an improved classification performance. Second, we investigate which combination of interaction between low and high level features leads to the best classification performance. Finally, we investigate the influence of the input volume size on the classification performance. Overall, our study aims to leverage ConTra-Net to improve the accuracy and efficiency of diagnosing paranasal anomalies in the maxillary sinus.

2 Materials and Methods

2.1 Dataset

Our population study named Hamburg City Health Study includes MRI images of the head and neck region from participants aged 45 to 74 in the city of Hamburg, Germany. The dataset consists of 299 patients with 174 healthy maxillary sinuses (MS) and 125 MS with anomalies (polyps and cysts) classified as "normal" and "anomaly" classes respectively, confirmed by 2 ENT surgeons and a radiologist. The MRIs of resolution $173 \times 319 \times 319$ voxel were recorded at University Medical Center Hamburg-Eppendorf with FLAIR sequences in NIfTI format. Preprocessing steps were identical to those prescribed by Bhattacharya *et al.*[10] MS volumes of sizes $35 \times 35 \times 35$, $40 \times 40 \times 40$, and $45 \times 45 \times 45$ were extracted from the larger head and neck MRI for the 3D CNN. Our MS volume extraction and dataset samples are shown in figure 1.

Training, validation and test splits: We perform 10-fold patient stratified cross validation set experiments. Altogether, we have 9810 MS volumes in the training set, 1110 MS volumes in the validation set and 1230 MS volumes in the test set. 32% of the MS volumes have anomalies in the training, validation and test sets.

Send correspondence to D.B.)
D.B.: E-mail: debayan.bhattacharya@tuhh.de, * equal contribution

Medical Imaging 2024: Computer-Aided Diagnosis, edited by Weijie Chen, Susan M. Astley, Proc. of SPIE Vol. 12927, 1292717 · © 2024 SPIE · 1605-7422 · doi: 10.1117/12.3005515

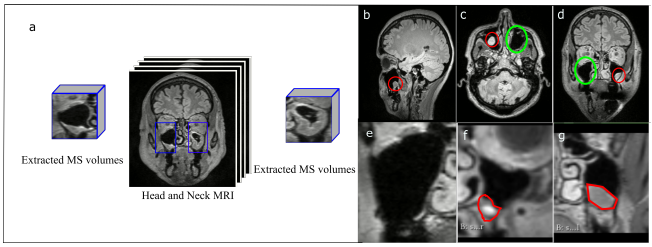

Figure 1. (a) Extracting MS volumes from single head and neck MRI. These MS volumes are passed to the deep learning models. (b) Cyst in the right MS (c) Polyp in the left MS (d) Cyst in the left MS (e) Normal MS of size $35 \times 35 \times 35$ (f) MS of size $40 \times 40 \times 40$ with polyp highlighted in red (g) MS of size $45 \times 45 \times 45$ with cyst highlighted in red

2.2 Convolutional Transformer Network (ConTra-Net)

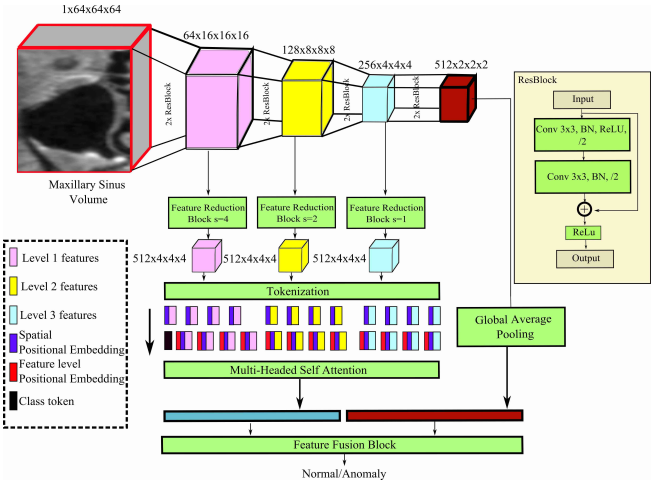

Figure 2. Illustration of ConTra-Net. The s in Feature Reduction Block represents stride of a convolution operation. The pink, yellow and cyan features represent the low, mid and high level features of the CNN.

ConTra-Net improves upon ViT's tokenization process by preserving spatial topology, considering the locality of features through convolutional kernels, and optimizing self-attention. This addresses the drawbacks of destroying spatial topology, insufficient capture of spatial context by feed forward networks, and challenges with optimizing self-attention in ViT. ConTra-Net combines the strengths of CNNs' inductive bias and the ability to capture global cues through the MHSA block. A figure of our proposed ConTra-Net is shown in Figure 2.

Feature extraction: The input MS volume $X \in \mathbb{R}^{H \times W \times D}$ is passed through a 3D CNN $F(.)$ with L stages of 3D residual blocks. Formally, it can be expressed as

$$f_l = \mathcal{F}_l(x;\theta) \mid x \in \mathbb{R}^{C_{l-1} \times \frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}} \times \frac{D}{2^{l-1}}}$$

Here, l is the l-th feature level. $f_l \in \mathbb{R}^{C_l \times \frac{H}{2^l} \times \frac{W}{2^l} \times \frac{D}{2^l}}$ represents the feature maps at level l of 3D CNN $\mathcal{F}_l$ (.). C_l represents the channel dimension at the l-th feature level and θ represents parameters of the CNN.

Feature reduction block: In order to allow our proposed ConTra-Net to capture global context while keeping computational costs low, we downsample the resolution and increase the channel dimension of f_l. The feature transformations are performed using 3D depthwise separable convolutions. The depthwise convolution operation spatially downsamples f_l while the pointwise convolution operation increases the channel dimension. Formally, we can express it as

$$\hat{f}_l = \Psi_l(\Upsilon(f_l;k,s);c_{in},c_{out})$$

Here $\Upsilon(.;k,s)$ is the depthwise convolution with kernel size k and stride s. The stride controls the downsample factor of the feature map f_l. $\Psi_l(.;c_{in},c_{out})$ is a 3D Convolution operation of kernel size 1 with input and output channel dimensions c_{in} and c_{out} respectively. Note, $c_{in} = C_l$.

Multi-head self-attention (MHSA) block: The multi-scale features f_l are transformed to $\hat{f}_l \in \mathbb{R}^{c_{out} \times h \times w \times d}$ by the feature reduction block. The $\hat{f}_l$ represent the tokens for the MHSA block. They are flattened into a sequence $x_l \in \mathbb{R}^{N \times c_{out}}$. Here, x_l is

the sequence of the l-th feature level and $N = h \cdot w \cdot d$ is the sequence length. c_{out} represents the embedding dimension. In order to retain the spatial positional information, we add a learnable positional embedding $p_{spatial} \in \mathbb{R}^{N \times c_{out}}$. In other words, we compute $\hat{f}_l^{\,pos} = \hat{f}_l + p_{spatial}$. Then, we concatenate the features $\hat{f}_l^{\,pos}$ arising from different feature levels l into a single large sequence $f_t \in \mathbb{R}^{N_{total} \times c_{out}}$. Here, $N_{total} = N \times (L-1)$ represents the sequence length that spans multiple levels of features. We leave out the features arising from the L-th feature block. To retain the positional information of the low and high level features originating from different parts of the CNN, we further add a learnable positional embedding $p_{level} \in \mathbb{R}^{N_{total} \times c_{out}}$. Formally, $\hat{f}_t = f_t + p_{level}$. Finally, we concatenate the class token $cls \in \mathbb{R}^{1 \times c_{out}}$ to $\hat{f}_t$ such that $\hat{f}_t = (\hat{f}_t \oplus cls) \in \mathbb{R}^{N_{total}+1 \times c_{out}}$. This resulting matrix $\hat{f}_t$ is passed through multiple layers of MHSA blocks and feed forward layers resulting in feature $F_t \in \mathbb{R}^{N_{total}+1 \times c_{out}}$.

Feature Fusion Block The feature fusion block concatenates the MHSA and CNN features and uses the combined feature vector to make class predictions. First, we consider the cls vector from F_t as the representative MHSA feature vector which encodes the global context. Second, we concatenate the CNN feature vector f_L and the MHSA feature vector. The resultant vector is passed through feed forward layers to make the final class prediction. We train ConTra-Net using class weighted cross-entropy loss.

3 Experiments, Results and Discussion

3.1 Implementation Details

For our 3D ViT implementation, we use patch sizes of $p = 4$ and embedding dimension of $C = 512$. The depth and number of attention heads of the MHSA are 2 and 4 respectively. For our 3D CNN, we use 3D variant of ResNet50.[18] Each ResNet has $L = 4$ stages of 3D residual blocks. $C_l = \{64, 128, 256, 512\}$ for our experiments. The depth and number of attention heads for our MHSA block is 2 and 4 respectively. Embedding dimension $c_{out} = 512$ for all our experiments. The feature reduction block downsamples the features to a resolution of $h = 4$, $w = 4$, $d = 4$ using strides 4, 2 and 1 for features arising from layers 1, 2 and 3 of the CNN respectively. This results in $N = 64$ and $N_{total} = 192$. With regards to the training configuration, we run our experiments for 100 epochs with a batch size of 16 for all our experiments. The learning rate was set at 0.0001 with a reduction by a factor of 10 if the validation loss did not improve for 5 epochs. Adam optimisation is used to train our deep learning models.

3.2 Classification performance

In our evaluation, we used Area Under Precision Recall Curve (AUPRC), Precision, Recall, and F1 to assess different methods. The positive class in our analysis was the "anomaly" class. As seen in Table 1, ResNet50 was the second best performing method. 3D ViT performed the worst, likely due to lack of pretraining on large scale dataset. Recall was challenging for all models due to the diverse morphological variations of anomalies. However, ConTra-Net achieved the highest AUPRC, Recall, and F1 scores, outperforming ResNet50 by 2.15%, 22.05%, and 11.68% respectively. Our results indicate that the combination of CNN and MHSA features leads to an improved classification performance.

Method	AUPRC	Precision	Recall	F1
ResNet 50	$\mu = 0.93, \sigma = 0.05$	$\mu = \mathbf{0.94}, \sigma = \mathbf{0.08}$	$\mu = 0.68, \sigma = 0.19$	$\mu = 0.77, \sigma = 0.13$
3D ViT	$\mu = 0.69, \sigma = 0.08$	$\mu = 0.69, \sigma = 0.14$	$\mu = 0.52, \sigma = 0.20$	$\mu = 0.56, \sigma = 0.10$
ConTra-Net	$\mu = \mathbf{0.95}, \sigma = \mathbf{0.04}$	$\mu = 0.92, \sigma = 0.10$	$\mu = \mathbf{0.83}, \sigma = \mathbf{0.15}$	$\mu = \mathbf{0.86}, \sigma = \mathbf{0.07}$

Table 1. Table of performance measures for different methods, with mean and standard deviation. Bold values indicate highest values.

3.3 Influence of low, mid and high level features on classification performance

Contra-Net employs a MHSA to enable interaction between low, mid, and high level features (pink, yellow, and cyan cubes in Figure 2). An evaluation of the importance of each feature level and their combinations was conducted, revealing that the combination of low and high-level features led to the greatest enhancement in recall performance (Table 2). These results suggest that the classification performance is most influenced by the interplay between low and high-level features.

Low-level	Mid-level	High-level	AUPRC	Precision	Recall	F1
		✓	$\mu = 0.93, \sigma = 0.05$	$\mu = \mathbf{0.94}, \sigma = \mathbf{0.06}$	$\mu = 0.78, \sigma = 0.11$	$\mu = 0.85, \sigma = 0.06$
	✓		$\mu = 0.94, \sigma = 0.03$	$\mu = 0.93, \sigma = 0.06$	$\mu = 0.78, \sigma = 0.19$	$\mu = 0.83, \sigma = 0.12$
	✓	✓	$\mu = 0.94, \sigma = 0.06$	$\mu = 0.93, \sigma = 0.07$	$\mu = 0.78, \sigma = 0.19$	$\mu = 0.83, \sigma = 0.12$
✓			$\mu = 0.94, \sigma = 0.03$	$\mu = 0.94, \sigma = 0.06$	$\mu = 0.76, \sigma = 0.15$	$\mu = 0.83, \sigma = 0.08$
✓		✓	$\mu = \mathbf{0.95}, \sigma = \mathbf{0.04}$	$\mu = 0.92, \sigma = 0.10$	$\mu = \mathbf{0.83}, \sigma = \mathbf{0.15}$	$\mu = \mathbf{0.86}, \sigma = \mathbf{0.07}$
✓	✓		$\mu = 0.93, \sigma = 0.04$	$\mu = 0.92, \sigma = 0.08$	$\mu = 0.71, \sigma = 0.24$	$\mu = 0.77, \sigma = 0.17$
✓	✓	✓	$\mu = 0.93, \sigma = 0.05$	$\mu = 0.91, \sigma = 0.08$	$\mu = 0.77, \sigma = 0.21$	$\mu = 0.82, \sigma = 0.09$

Table 2. Table of performance measures for different configurations of feature levels. Bold values indicate highest values.

3.4 Volume size

The size of the extracted MS volume is crucial. If it is too small, pathology may be missed or the sinuses may be only partially extracted. If it is too large, irrelevant anatomical information can hinder paranasal anomaly classification. We evaluated Contra-Net performance on MS volumes of sizes $35 \times 35 \times 35$, $40 \times 40 \times 40$, and $45 \times 45 \times 45$ voxels. Results in table 3 showed that AUPRC of $35 \times 35 \times 35$ and $40 \times 40 \times 40$ were same. AUPRC decreased for $45 \times 45 \times 45$ compared to $35 \times 35 \times 35$ and $40 \times 40 \times 40$, indicating limited impact on model performance beyond $40 \times 40 \times 40$. Our results indicate that including additional surrounding structures in larger volumes negatively affected paranasal anomaly classification. We thereby conclude that careful selection of volume size is vital for optimization.

Volume Size	AUPRC	Precision	Recall	F1
$35 \times 35 \times 35$	$\mu = \mathbf{0.95}, \sigma = \mathbf{0.04}$	$\mu = 0.92, \sigma = 0.10$	$\mu = \mathbf{0.83}, \sigma = \mathbf{0.15}$	$\mu = \mathbf{0.86}, \sigma = \mathbf{0.07}$
$40 \times 40 \times 40$	$\mu = \mathbf{0.95}, \sigma = \mathbf{0.04}$	$\mu = \mathbf{0.93}, \sigma = \mathbf{0.06}$	$\mu = 0.79, \sigma = 0.12$	$\mu = 0.85, \sigma = 0.09$
$45 \times 45 \times 45$	$\mu = 0.93, \sigma = 0.03$	$\mu = 0.90, \sigma = 0.09$	$\mu = 0.78, \sigma = 0.10$	$\mu = 0.83, \sigma = 0.06$

Table 3. Influence of volume size on classification performance. Bold values indicate highest values.

4 Conclusion

In this study, we introduced a novel hybrid CNN transformer, ConTra-Net, for the task of paranasal anomaly classification in the maxillary sinus. We observed that the recall metric has high standard deviation illustrating the difficulty of generalizing to unseen anomaly morphologies.We compared ConTra-Net to ResNet50 and 3D ViT, and also performed an ablation study to investigate the influence of low, mid and high level features towards the classification performance. ConTra-Net outperformed ResNet50 in AUPRC, Recall and F1, suggesting that learning global features using MHSA may be beneficial for this task. Interaction of low and high level features through MHSA proved to be the most beneficial towards paranasal anomaly classification. A limitations to our work is the need for further improvement in the F1 score of ConTra-Net in order to make it applicable to real-world clinical scenarios. Despite this limitation, our results provide a promising deep learning solution for paranasal anomaly classification in the maxillary sinus.

REFERENCES

[1] Wilson, R., Kuan Kok, H., Fortescue-Webb, D., Doody, O., Buckley, O., and Torreggiani, W. C., "Prevalence and seasonal variation of incidental mri paranasal inflammatory changes in an asymptomatic irish population," *Irish medical journal* **110**(9), 641 (2017).

[2] Hansen, A. G., Helvik, A.-S., Nordgård, S., Bugten, V., Stovner, L. J., Håberg, A. K., Gårseth, M., and Eggesbø, H. B., "Incidental findings in mri of the paranasal sinuses in adults: a population-based study (hunt mri)," *BMC ear, nose, and throat disorders* **14**(1), 13 (2014).

[3] Tarp, B., Fiirgaard, B., Christensen, T., Jensen, J. J., and Black, F. T., "The prevalence and significance of incidental paranasal sinus abnormalities on mri," *Rhinology* **38**(1), 33–38 (2000).

[4] Ma, Z. and Yang, X., "Research on misdiagnosis of space occupying lesions in unilateral nasal sinus," *Lin chuang er bi yan hou tou jing wai ke za zhi = Journal of clinical otorhinolaryngology, head, and neck surgery* **26**(2), 59–61 (2012).

[5] Jeon, Y., Lee, K., Sunwoo, L., Choi, D., Oh, D. Y., Lee, K. J., Kim, Y., Kim, J.-W., Cho, S. J., Baik, S. H., Yoo, R.-E., Bae, Y. J., Choi, B. S., Jung, C., and Kim, J. H., "Deep learning for diagnosis of paranasal sinusitis using multi-view radiographs," *Diagnostics (Basel, Switzerland)* **11**(2) (2021).

[6] Kim, Y., Lee, K. J., Sunwoo, L., Choi, D., Nam, C.-M., Cho, J., Kim, J., Bae, Y. J., Yoo, R.-E., Choi, B. S., Jung, C., and Kim, J. H., "Deep learning in diagnosis of maxillary sinusitis using conventional radiography," *Investigative radiology* **54**(1), 7–15 (2019).

[7] Kim, K.-S., Kim, B. K., Chung, M. J., Cho, H. B., Cho, B. H., and Jung, Y. G., "Detection of maxillary sinus fungal ball via 3-d cnn-based artificial intelligence: Fully automated system and clinical validation," *PLOS ONE* **17**, 1–19 (02 2022).

[8] Bhattacharya, D., Behrendt, F., Becker, B. T., Beyersdorff, D., Petersen, E., Petersen, M., Cheng, B., Eggert, D., Betz, C., Hoffmann, A. S., and Schlaefer, A., "Unsupervised anomaly detection of paranasal anomalies in the maxillary sinus," in [*Medical Imaging 2023: Computer-Aided Diagnosis*], Iftekharuddin, K. M. and Chen, W., eds., **12465**, 124651B, International Society for Optics and Photonics, SPIE (2023).

[9] Bhattacharya, D., Becker, B. T., Behrendt, F., Bengs, M., Beyersdorff, D., Eggert, D., Petersen, E., Jansen, F., Petersen, M., Cheng, B., Betz, C., Schlaefer, A., and Hoffmann, A. S., "Supervised contrastive learning to classify paranasal anomalies in the maxillary sinus," in [*Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*], Wang, L., Dou, Q., Fletcher, P. T., Speidel, S., and Li, S., eds., 429–438, Springer Nature Switzerland, Cham (2022).

[10] Bhattacharya, D., Behrendt, F., Becker, B. T., Beyersdorff, D., Petersen, E., Petersen, M., Cheng, B., Eggert, D., Betz, C., Hoffmann, A. S., and Schlaefer, A., "Multiple instance ensembling for paranasal anomaly classification in the maxillary sinus," (2023).

[11] Gao, W., Wan, F., Pan, X., Peng, Z., Tian, Q., Han, Z., Zhou, B., and Ye, Q., "Ts-cam: Token semantic coupled attention map for weakly supervised object localization," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* , 2866–2875 (2021).

[12] Yu, S., Ma, K., Bi, Q., Bian, C., Ning, M., He, N., Li, Y., Liu, H., and Zheng, Y., "Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification," in [*Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*], de Bruijne, M., Cattin, P. C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., and Essert, C., eds., 45–54, Springer International Publishing, Cham (2021).

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N., "An image is worth 16x16 words: Transformers for image recognition at scale," in [*International Conference on Learning Representations*], (2021).

[14] Jang, J. and Hwang, D., "M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer," in [*2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 20686–20697 (2022).

[15] Dai, Y., Gao, Y., and Liu, F., "Transmed: Transformers advance multi-modal medical image classification," *Diagnostics* **11**(8) (2021).

[16] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y., "Transunet: Transformers make strong encoders for medical image segmentation," *CoRR* **abs/2102.04306** (2021).

[17] Dai, Y., Gao, Y., and Liu, F., "TransMed: Transformers advance Multi-Modal medical image classification," *Diagnostics (Basel)* **11** (July 2021).

[18] Hara, K., Kataoka, H., and Satoh, Y., "Learning spatio-temporal features with 3d residual networks for action recognition," (2017).