



Reconstructing the Hemodynamic Response Function via a Bimodal Transformer

Yoni Choukroun¹(✉), Lior Golgher², Pablo Blinder^{3,4}, and Lior Wolf¹

¹ The School of Computer Science, Tel Aviv University, Tel Aviv-Yafo, Israel
choukroun.yoni@gmail.com

² The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

³ Neurobiology, Biochemistry and Biophysics School, Wise Life Science Faculty, Tel Aviv University, Tel Aviv-Yafo, Israel

⁴ The Sagol School for Neuroscience, Tel Aviv University, Tel Aviv-Yafo, Israel

Abstract. The relationship between blood flow and neuronal activity is widely recognized, with blood flow frequently serving as a surrogate for neuronal activity in fMRI studies. At the microscopic level, neuronal activity has been shown to influence blood flow in nearby blood vessels. This study introduces the first predictive model that addresses this issue directly at the explicit neuronal population level. Using in vivo recordings in awake mice, we employ a novel spatiotemporal bimodal transformer architecture to infer current blood flow based on both historical blood flow and ongoing spontaneous neuronal activity. Our findings indicate that incorporating neuronal activity significantly enhances the model's ability to predict blood flow values. Through analysis of the model's behavior, we propose hypotheses regarding the largely unexplored nature of the hemodynamic response to neuronal activity.

Keywords: Hemodynamic Response Function · Bimodal transformers

1 Introduction

The brain consumes copious amounts of energy to sustain its activity, resulting in a skewed energetic budget per mass compared to the rest of the body (about 25% utilized by about 3%, see [2, 16] for an elaborate review of energy utilization). Given this disproportionate need, resources are allocated on a need-basis: active areas signal to the nearby blood vessel to dilate and increase blood flow, bringing a surplus of resources, to that area. This fundamental physiological process is called neurovascular coupling. It is non-trivial to model and different types of neuronal activity have been shown to elicit opposite vascular responses.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43895-0_35.

Neurovascular coupling is a cornerstone of proper brain function and also underpins the ability to observe and study the human brain in action. Imaging methods based on blood oxygenated level dependent (BOLD) approaches rely on it [13], as do methods that are based on rheological properties, such as blood volume and flow speed. Since these methods do not directly measure neuronal activity per-se, but a physiological proxy, i.e. the resulting change in vascular dynamics and oxygen levels, it is of utmost importance to know the precise transform function linking neuronal activity to the observed vascular dynamics. Given the differential response to neuronal activity (see [5] for a timely review), obtaining a cellular and population level hemodynamic response function (HRF) remains an unmet need in this field, that would finally unlock the ability to infer neuronal activity directly from blood flow dynamics [20].

The initial characterization of the hemodynamic response function (HRF) was performed at the *system* level, where *system* refers to large cortical regions encompassing tens of thousands of neurons of different types, without taking into account the fine details of different vascular compartments (see [29] for a succinct review on the original works). At this level, a canonical response function was derived from extensive work on sensory-evoked somatosensory responses. This HRF has become widely accepted and used in the interpretation of BOLD signals. This function consists of three components: an initial dip (its existence and physiological origin are much debated), a prolonged and very pronounced overshoot, followed by a shallower and much shorter undershoot. The initial dip occurs within one second of the sensory stimulus, the overshoot peaks around five seconds later, overshoot and return to baseline level occurs within 15–20 s post stimuli. It should be noted that vascular reactivity is much faster than the collective behavior described by the canonical HRF, with reports showing sensory-evoked vascular responses observed after just 300ms. Recently, more advanced imaging and analysis methods have pushed the formulation of an HRF at the single cell to single blood vessel (capillary) level, pointing to a rather narrow family of possible functions. Importantly, this work also established that the HRF derived at the microscopic level can be partially translated to macroscopic imaging approaches. Nevertheless, single neuron to single vessel responses fail to capture the more complex and varied neuronal population level responses that could be integrated across the extensive vascular network that surrounds them. Here, we exploit a unique dataset, in which neuronal and vascular responses (changes in diameter) were recorded in a volumetric fashion and with relevant temporal resolution, allowing us to establish a novel pipeline to uncover/formulate a many-to-many HRF.

Our model needs to combine neuron firing and blood vessel data and employs a multi-modal transformer. There are three types of multi-modal transformers: (i) a multi-modal Transformer where the two modalities are concatenated and separated by the [SEP] token [18,19], and self-attention is used, (ii) co-attention-based model modules that contextualize each modality with the other modality [21,26], and (iii) generative models containing an encoder that uses self-attention on the input and a decoder that uses both the encoded data and data from the decoder’s domain as inputs [3,17,23,30–32,34]. Our model is of

the third type and presents two distinctive properties: pulling from multiple time points and an attention mechanism that is modulated based on distance.

Our results show that the new transformer model can predict the state of blood vessels better than the baseline models. The utility of neuronal data in the prediction is demonstrated by an ablation study. By analyzing the learned model, we obtain insights into the link between neuronal and vascular activities.

2 Data

All procedures were approved by the Ethics Committee of Tel Aviv University for Animal Use and Welfare and followed pertinent Institutional Animal Care and Use Committee (IACUC) and local guidelines. Neuronal activity was monitored in female C57BL/6J transgenic mice expressing Thy1-GCaMP6s. Vascular dynamics were tracked using a Texas Red fluorescent dye, which was conjugated to a large polysaccharide moiety (2 mega Dalton dextran) and retro-orbitally bolus injected under brief isoflurane sedation at the beginning of the imaging day.

425 quasi-linear vascular segments and 50 putative neuronal cell bodies were manually labeled within a volume of $490 \times 500 \times 300 \mu m^3$, which was continuously imaged across two consecutive 1850-second long sessions at an imaging rate of 30.03 volumes per second [9–12]. For neuronal activity estimation, we selected a cuboid volume of interest around each neuronal cell body and summed the fluorescence within it following an axial intensity normalization corresponding to an uneven duty cycle of our varifocal lens.

For vascular diameter estimation we used the Radon transform, [1, 6–9, 22, 24, 28] as its resilience to rotation and poor contrast are particularly useful for our application. Specifically, Gao and Drew have formerly found that thresholding the vascular intensity profile in Radon space is more resilient to noise than other thresholding methods [8]. Based on their observation, we used the time-collapsed imagery to determine a threshold in Radon space, which was then applied separately for each frame in time.

This unique ability to rapidly track neuronal and vascular interactions across a continuous brain volume bears several important advantages [9, 10]. In particular, a greater proportion of the vascular ensemble that reacts to a given neuronal metabolic demand can be accounted for.

3 Method

The HRF learning problem explored in this work is defined as the prediction of current blood flow rates at different vessel segments, given the previous neuronal spikes as well as previous blood flow rates. We propose to design a parameterized deep neural network f_θ for scalar regression of blood flow rates at different vessel segments, such that at a given time t we have

$$f_\theta(S_t, F_t, X_S, X_F) \rightarrow \mathbb{R}^m \quad (1)$$

where the matrix $S_t \in \mathbb{R}^{t_s \times n}$ denotes the n neurons' spikes at the t_s previous samples, while the matrix $F_t \in \mathbb{R}^{t_v \times m}$ denotes the blood flow of the m vessel

segment at the previous t_v time samples. $X_S \in \mathbb{R}^{n \times 3}$ and $X_F \in \mathbb{R}^{m \times 3}$ are the three-dimensional positions of the neurons and vessel segments, respectively.

HRF predictions should satisfy fundamental symmetries and invariance of physiological priors and of experimental bias, such as invariance to rigid spatial transformation (rotation and translation). Therefore, a positional input X_u is transformed to inter-elements Euclidean distances $D_u = \{d_{ij}^u\}_{i,j}$ where $d_{ij}^u = \|(X_u)_i - (X_u)_j\|_2$ for rigid transform invariance.

Thus, the learning problem is refined as $f_\theta : \{S_t, F_t, D_S, D_F, D_{SF}\} \rightarrow \mathbb{R}^m$, where D_S, D_F, D_{SF} represents the Euclidean distance matrix between neurons, vessel segments, and neurons to vessel segments, respectively. We do not include any further auxiliary features or prior in the input.

We model f_θ using a new variant of the Transformer family. The proposed model consists of an encoder and a decoder. The encoder embeds the neurons at *both* spatial and temporal levels. The decoder predicts vessel segment flow by utilizing *both* the past flow values and the spatial information of the vessel segments, along with the neuronal activity via the cross-attention mechanism.

Transformers. The self-attention mechanism introduced by Transformers [30] is based on a trainable associative memory with (key, value) vector pairs, where a query vector $q \in \mathbb{R}^d$ is matched against a set of k key vectors using scaled inner products, as follows

$$A(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (2)$$

where $Q \in \mathbb{R}^{N \times d}$, $K \in \mathbb{R}^{k \times d}$ and $V \in \mathbb{R}^{k \times d}$ represent the packed N queries, k keys and values tensors respectively. Keys, queries and values are obtained using linear transformations of the sequence's elements. A multi-head self-attention layer is defined by extending the self-attention mechanism using h attention *heads*, i.e. h self-attention functions applied to the input, reprojected to values via a $dh \times D$ linear layer.

Neuronal Encoding. To obtain the initial Spatio-Temporal Encoding, for the prediction at time t , we project each neuron to a high d dimensional embedding $\phi_t^s \in \mathbb{R}^{t_s \times n \times d}$ by modulating it with its spike value such that $\phi_t^s = S_t \odot (1_d W^T)$, where $W \in \mathbb{R}^d$ denotes the neuronal encoding. The embedding is modulated by the magnitude of the spike, such that higher neuronal activities are projected farther in the embedding space.

The *temporal encoding* is defined using sinusoidal encoding [30] applied on ϕ and augmented with a learnable embedding such that $\phi_t^s \leftarrow \phi_t^s + p_t \cdot \tilde{p}$ where p_t and \tilde{p} represent the sinusoidal time encoding and the learned vector, respectively. We emphasize the fact that, contrary to traditional transformers, the embedding tensor ϕ_t has an additional spatial dimension such that the tensor is three-dimensional, enabling both spatial and temporal attention.

In order to incorporate the spatial information of the neurons, we propose to insert *spatial encoding* by importing the pairwise information directly into the self-attention layer. For this, we multiply the distance relation by the similarity

tensor as follows

$$A^S(Q, K, D_S) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \odot \psi_S(D_S), \quad (3)$$

with \odot denoting the Hadamard product, and $\psi_S(D_S) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ an element-wise learnable parameterized similarity function. This way, the similarity function scales the self-attention map according to the distance between the elements (in our case the neurons).

Vascular Decoding. The spatio-temporal encoding of the vascular data is similar to the embedding performed by the encoder. The information on each vascular segment is embedded in a high-dimensional vector $\phi_t^F \in \mathbb{R}^{t_v \times m \times d}$ to be further projected by the temporal encoding. The spatial geometric information is incorporated via the pairwise vascular segments' distance matrix D_F via the decoder's self-attention module A^F .

The most important element of the decoder is the cross-attention module, which incorporates neuronal information for vascular prediction. Given the final neuronal embeddings ϕ_t^s , the cross-attention module performs cross-analysis of the neuronal embeddings such that

$$A^{SF}(Q_F, K_S, D_S) = \text{Softmax}\left(\frac{Q_F K_S^T}{\sqrt{d}}\right) \odot \psi_{SF}(D_{SF}), \quad (4)$$

where Q_F and K_S represent the affine transform of ϕ_t^F and ϕ_t^s , respectively. Here also, the (non-square) cross-attention map is modulated by the neuron-vessel distance matrix D_{SF} .

The spatio-temporal map is of dimensions $A^{SF} \in \mathbb{R}^{t_v \times t_s \times h \times m \times n}$ where h denotes the number of attention heads. Thus, we perform aggregation by averaging over the neuronal time dimension, in order to remain *invariant* to the temporal neuronal embedding and to gather all past neuronal influence on blood flow rates. This way, one can observe that the proposed method is not limited to any spatial or time constraint. The model can be deployed in different spatio-temporal settings at test time, thanks to both the geometric spatial encoding and the Transformer's sequential processing ability. Finally, the output module reprojects the last time vessel embedding into the prediction space.

Architecture and Training. The initial encoding defines the model embedding dimension $d = 64$. The encoder and the decoder are defined as the concatenation of $L = 3$ layers, each composed of self-attention and feed-forward layers interleaved with normalization layers. The decoder also contains N additional cross-attention modules. The output layer is defined by a fully connected layer that projects the last vascular time embedding into the objective dimension m . An illustration of the model is given in Fig. 1.

The dimension of the feed-forward network is four times that of the embedding [30]. It is composed of GEGLU layers [25], with layer normalization set to the pre-layer norm setting, as in [15, 33]. We use an eight-head self-attention module in all experiments. The geometric filtering first augments the distance using

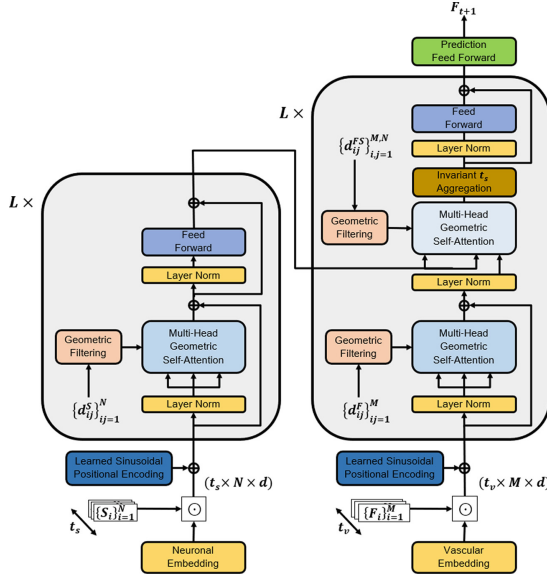


Fig. 1. Illustration of the proposed HRF Transformer architecture. The main differences from the traditional Transformers are the Geometric self-attention modules and the unified spatiotemporal analysis induced by the time aggregation module.

Fourier features [27] and the module is a fully connected neural network with two 50-dimensional hidden layers and GELU non-linearities, expanded to all the heads of the self-attention module. We provide the module with the element-wise inverse of the distance matrix instead of the regular Euclidean matrix, both in order to reduce the dynamic range and since closer elements may have a higher impact.

The training objective is the Mean Squared Error loss

$$\mathcal{L} = \mathbb{E}_t \left(\sum_j^m \|f_\theta(S_t, F_t, D_S, D_F, D_{SF}) - F_{t+1}\|^2 \right) \quad (5)$$

The Adam optimizer [14] is used with 32 samples per minibatch, for 300 epochs. We initialized the learning rate to $5 \cdot 10^{-5}$ coupled with a cosine decay scheduler down to $1 \cdot 10^{-6}$ at the end of the training. The dataset of the first data collection session has been split by 85%, 7.5% and 7.5% for the training, validation, and testing set, respectively. Training time is approximately 20 h for time windows $t_s = t_v = 10$, on an NVIDIA RTX A600. Testing time is approximately 0.25 ms per sample.

4 Experiments

We compare the proposed method, dubbed **H**emodynamic **R**esponse **F**unction Transformer (HRFT), with several popular statistical and machine-learning

Table 1. The prediction errors of the various methods on two test sets. The first is obtained in the same session for which the training samples were collected (top half of the table, and the second in a separate session 30 min later (bottom half).

Method	6 Hz		15 Hz		30.03 Hz	
	MSE	NRMSE	MSE	NRMSE	MSE	NRMSE
Persistence	24.38	0.220	12.93	0.160	6.923	0.115
Linear	13.65	0.166	9.660	0.139	5.911	0.107
RNN	13.78	0.168	11.38	0.157	10.31	0.147
HRFT-S	13.34	0.165	9.426	0.138	5.782	0.110
HRFT	13.00	0.162	9.370	0.137	5.783	0.106
Persistence	23.11	0.221	11.97	0.160	7.662	0.125
Linear	17.01	0.192	10.26	0.147	6.002	0.111
RNN	15.193	0.182	13.04	0.172	11.87	0.162
HRFT-S	14.63	0.176	9.820	0.147	5.914	0.110
HRFT	14.34	0.173	9.191	0.143	5.908	0.110

models: (i) naive persistence model, which predicts the previous time step’s vascular input, (ii) linear regression, which concatenates all the input (blood flow and neuronal data) from all times stamps before performing the regression, and (iii) a Recurrent Neural Network composed of two stacked GRU [4] layers. All the methods are experimented with using the same inputs and the models have similar capacity (~ 0.5 M parameters).

In order to understand the impact of neuronal information, we also compare our method with the HRFT encoder only applied to the vascular input, referred to as HRFT-S. The only difference between this model and the full HRFT is the cross-attention module. If the neuronal input is irrelevant or the link is too weak to improve the prediction, HRFT is expected not to outperform HRFT-S, which makes the comparison pertinent.

We present both MSE and Normalized Root MSE. Because of computational constraints, we randomly subsample 55 vessel segments among the 425. We trained the model with temporal windows of size $t_s = t_v = 10$, equivalent to 300 ms according to the original data acquisition’s 30.03 Hz sampling rate.

In addition to the original sampling rate, we also present results for prediction based on lower frequencies, in order to check the ability of the models to capture longer-range dependencies. We note that the error in these cases is expected to be larger, since the time gap between the last measurement and the required prediction is larger.

In order to check the generalization abilities of the methods, we test the trained models on a second dataset obtained 30 min after the sampling of the original dataset (that includes training, validation, and the first test set).

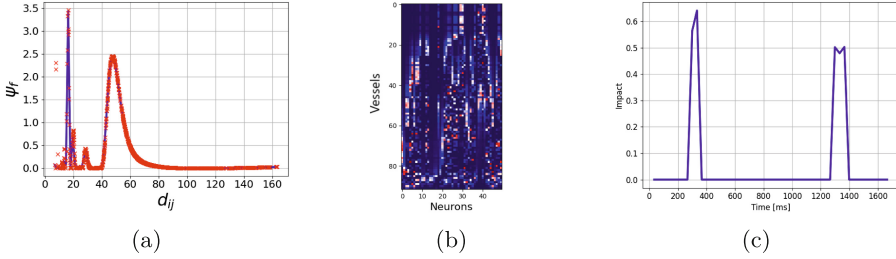


Fig. 2. (a) The learned function ψ_{SF} for the 1st cross-attention layer of the transformer. (b) The magnitude of the self-attention map between every neuron and every vessel at this layer. (c) The impact of the neurons on the vessels (saturated at 90%) for each shift in time as obtained by marginalizing over all 2nd session test samples in the 30.02 Hz dataset. More visualizations of the datasets and the learned features are provided in the Appendix.

The results are presented in Table 1. As can be seen, the HRFT method outperforms all baselines, including the HRFT-S variant, for 6 Hz and 15 Hz. At the original sampling rate, the performance of HRFT and HRFT-S is similar and better than the baselines. This is expected since at this frame rate the history of $t_v = 10$ we employ spans only 300 ms, which is at the limit of the shortest known neurovascular response reported in the literature [29]. It is reassuring that error levels for HRFT remain similar for samples taken 30min after the training set (and the first test set) were collected.

To gain insights into the HRF, we examine the HRFT model. The learned distance function ψ_{SF} of the 1st cross attention layer is depicted in Fig. 2(a) (other layers are similar). The plot shows the learned function in blue and the actual samples in red. Evidently, this prior on the attention is monotonically decreasing with the distance between the neuron and the blood vessel. Panel (b) shows the cross-attention in the same layer. We note that some neurons have little influence, and the rest of the attention is scattered relatively uniformly. Panel (c) considers the derivative of the prediction vector F_{t+1} by each of the neuron data, summed over all test samples of the 2nd session at 6 Hz, and all neurons and vessels. There are two negative peaks (contractions) that occur at 333 ms and 1333 ms, which is remarkably consistent with current knowledge [29]. There is also a dilation effect at 666 ms. The 15 Hz data with $t_v = 10$ captures shifts of 0–700 ms and the 300 ms peak is clearly visible in that model as well.

5 Conclusions and Future Work

We present the first local HRF model. While for the baseline methods, the performance is at the same level with and without neuronal data (omitted from the tables), the transformer we present supports an improved prediction capability using neuronal firing rates (ablation) and also gives rise to interesting insights regarding the behavior of the hemodynamic response function.

Limitations. Our main goal is to verify the ability to model HRF by showing that using neuronal data helps predict blood flow beyond the history of the latter. The next challenge is to scale the model in order to be able to model more vessels (without subsampling) and longer historical sequences (larger t_v, t_s). With transformers being used for very long sequences, this is a limitation of our resources and not of our method.

Acknowledgements. The authors thank David Kain for conducting the mouse surgery. This project has received funding from the ISRAEL SCIENCE FOUNDATION (grant No. 2923/20) within the Israel Precision Medicine Partnership program. It was also supported by a grant from the Tel Aviv University Center for AI and Data Science (TAD). It was also supported by the European Research Council, grant No 639416, and the Israel Science Foundation, grant No 2342/21. The contribution of the first author is part of a PhD thesis research conducted at Tel Aviv University.

References

1. Asl, M.E., Koohbanani, N.A., Frangi, A.F., Gooya, A.: Tracking and diameter estimation of retinal vessels using gaussian process and radon transform. *J. Med. Imaging* **4**(3), 034006 (2017)
2. Buxton, R.B.: Thermodynamic limitations on brain oxygen metabolism: physiological implications. *bioRxiv*, pp. 2023–01 (2023)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. *arXiv preprint [arXiv:2005.12872](https://arxiv.org/abs/2005.12872)* (2020)
4. Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches. In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111 (2014)
5. Drew, P.J.: Vascular and neural basis of the bold signal. *Current Opin. Neurobiol.* **58**, 61–69 (2019). <https://doi.org/10.1016/J.CONB.2019.06.004>
6. Drew, P.J., Blinder, P., Cauwenberghs, G., Shih, A.Y., Kleinfeld, D.: Rapid determination of particle velocity from space-time images using the radon transform. *J. Comput. Neurosci.* **29**(1), 5–11 (2010)
7. Fazlollahi, A., et al.: Efficient machine learning framework for computer-aided detection of cerebral microbleeds using the radon transform. In: *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pp. 113–116. IEEE (2014)
8. Gao, Y.R., Drew, P.J.: Determination of vessel cross-sectional area by thresholding in radon space. *J. Cereb. Blood Flow Metab.* **34**(7), 1180–1187 (2014)
9. Golgher, L.: Rapid volumetric imaging of numerous neuro-vascular interactions in awake mammalian brain. Ph.D. thesis, Sagol School of Neuroscience, Tel Aviv University (2022)
10. Gur, S., Wolf, L., Golgher, L., Blinder, P.: Microvascular dynamics from 4d microscopy using temporal segmentation. In: *Pacific Symposium on Biocomputing 2020*, pp. 331–342. World Scientific (2019)
11. Har-Gil, H., et al.: Pysight: plug and play photon counting for fast continuous volumetric intravital microscopy. *Optica* **5**(9), 1104–1112 (2018)
12. Har-Gil, H., Golgher, L., Kain, D., Blinder, P.: Versatile software and hardware combo enabling photon counting acquisition and real-time display for multiplexing, 2d and continuous 3d two-photon imaging applications. *Neurophotonics* **9**(3), 031920 (2022)

13. Kim, S.G., Ogawa, S.: Biophysical and physiological origins of blood oxygenation level-dependent fMRI signals. *J. Cereb. Blood Flow Metab. Off. J. Int. Soc. Cereb. Blood Flow Metab.* **32**, 1188–206 (2012). <https://doi.org/10.1038/jcbfm.2012.23>
14. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. *arXiv* (2014)
15. Klein, G., Kim, Y., et al.: Open-source toolkit for neural machine translation. In: *ACL* (2017)
16. Levy, W.B., Calvert, V.G.: Communication consumes 35 times more energy than computation in the human cortex, but both costs are needed to predict synapse number. *Proc. Natl. Acad. Sci.* **118**(18), e2008173118 (2021)
17. Lewis, M., et al.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019)
18. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: VisualBERT: a simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019)
19. Li, X., et al.: Oscar: object-semantics aligned pre-training for vision-language tasks. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) *ECCV 2020. LNCS*, vol. 12375, pp. 121–137. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58577-8_8
20. Logothetis, N.K.: What we can do and what we cannot do with fMRI. *Nature*. **453**, 869–78 (2008). <https://doi.org/10.1038/nature06976>
21. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: *Advances in Neural Information Processing Systems*, pp. 13–23 (2019)
22. Mookiah, M.R.K., et al.: A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification. *Med. Image Anal.* **68**, 101905 (2020)
23. Paul, M., Danelljan, M., Van Gool, L., Timofte, R.: Local memory attention for fast video semantic segmentation. *arXiv preprint arXiv:2101.01715* (2021)
24. Pourreza, R., Banaee, T., Pourreza, H., Kakhki, R.D.: A radon transform based approach for extraction of blood vessels in conjunctival images. In: Gelbukh, A., Morales, E.F. (eds.) *MICAI 2008. LNCS (LNAI)*, vol. 5317, pp. 948–956. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88636-5_89
25. Shazeer, N.: GLU variants improve transformer. [arXiv:2002.05202](https://arxiv.org/abs/2002.05202) (2020)
26. Tan, H., Bansal, M.: LXMERT: learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019)
27. Tancik, M., et al.: Fourier features let networks learn high frequency functions in low dimensional domains. *Adv. Neural. Inf. Process. Syst.* **33**, 7537–7547 (2020)
28. Tavakoli, M., Mehdizadeh, A., Pourreza, R., Pourreza, H.R., Banaee, T., Toosi, M.B.: Radon transform technique for linear structures detection: application to vessel detection in fluorescein angiography fundus images. In: *2011 IEEE Nuclear Science Symposium Conference Record*, pp. 3051–3056. IEEE (2011)
29. Uludağ, K., Blinder, P.: Linking brain vascular physiology to hemodynamic response in ultra-high field MRI. *NeuroImage*. **168**, 279–295 (2018). <https://doi.org/10.1016/j.neuroimage.2017.02.063>
30. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: *NeurIPS* (2017)
31. Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Max-deeplab: end-to-end panoptic segmentation with mask transformers. *arXiv preprint arXiv:2012.00759* (2020)

32. Wang, Y., et al.: End-to-end video instance segmentation with transformers. arXiv preprint [arXiv:2011.14503](#) (2020)
33. Xiong, R., et al.: On layer normalization in the transformer architecture. [arXiv:2002.04745](#) (2020)
34. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: ICLR (2021)