Название статьи (авторы статьи): **Are Large-scale Datasets Necessary for Self-Supervised Pre-training**
Автор исследования: Агапова Ольга

Работа написана в 2021 году, засабмичена 20 декабря 2021. У работы через google scholar можно найти две версии, но разницы между субмитами я не нашла.

Alaaeldin El-Nouby – студент ресерчер в FAIR. У него приличный список публикаций (в том числе и в чьей-то команде, например своих супервизоров), но пересечений с текущим исследованием я не нашла, просто можно сказать что в целом он занимается компьютерным зрением и адаптацией к нему известных интересных методов. В данной работе он ссылается на свое же исследование 2021-го года “Xcit: Cross-covariance image transformers”, но только чтобы пояснить, как технически устроено его решение, то есть он использует те же методы, что и в этой работе, max pooling и transposed convolutions в Mask RCNN на этапе object detection.
В целом выглядит как случайная находка, но возможно кто-то из его супервизоров посоветовал ему взять такую тему, например Touvron занимается как раз self-supervised learning и на его публикации в том числе ссылается данная работа (Mathilde Caron, Hugo Touvron, Ishan Misra, Herve´ Jegou, Julien Mairal, Piotr Bojanowski, and Armand ´ Joulin, “Emerging properties in self-supervised vision transformers,” arXiv preprint arXiv:2104.14294, 2021.).

LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference B Graham, A El-Nouby, H Touvron, P Stock, A Joulin, H Jégou, M Douze International Conference on Computer Vision 2021	79	*	2021
Resmlp: Feedforward networks for image classification with data-efficient training H Touvron, P Bojanowski, M Caron, M Cord, A El-Nouby, E Grave, ... arXiv preprint arXiv:2105.03404	62	*	2021
Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction A El-Nouby, S Sharma, H Schulz, D Hjelm, LE Asri, SE Kahou, Y Bengio, ... Proceedings of the IEEE International Conference on Computer Vision, 10304-10312	59	*	2019
XCiT: Cross-Covariance Image Transformers A El-Nouby, H Touvron, M Caron, P Bojanowski, M Douze, A Joulin, ... 35th Conference on Neural Information Processing Systems (NeurIPS 2021)	35	*	2021
Training vision transformers for image retrieval A El-Nouby, N Neverova, I Laptev, H Jégou arXiv preprint arXiv:2102.05644	30		2021
Real-Time End-to-End Action Detection with Two-Stream Networks A Ali, GW Taylor 2018 15th Conference on Computer and Robot Vision (CRV), 31-38	19	*	2018
Skip-Clip: Self-Supervised Spatiotemporal Representation Learning by Future Clip Order Ranking A El-Nouby, S Zhai, GW Taylor, JM Susskind Holistic Video Understanding Workshop ICCV2019	10		2019
Augmenting Convolutional networks with attention-based aggregation H Touvron, M Cord, A El-Nouby, P Bojanowski, A Joulin, G Synnaeve, ... arXiv preprint arXiv:2112.13692	2		2021
Are Large-scale Datasets Necessary for Self-Supervised Pre-training? A El-Nouby, G Izacard, H Touvron, I Laptev, H Jegou, E Grave arXiv preprint arXiv:2112.10740			2021
Spatiotemporal Representation Learning For Human Action Recognition And Localization A Ali			2019

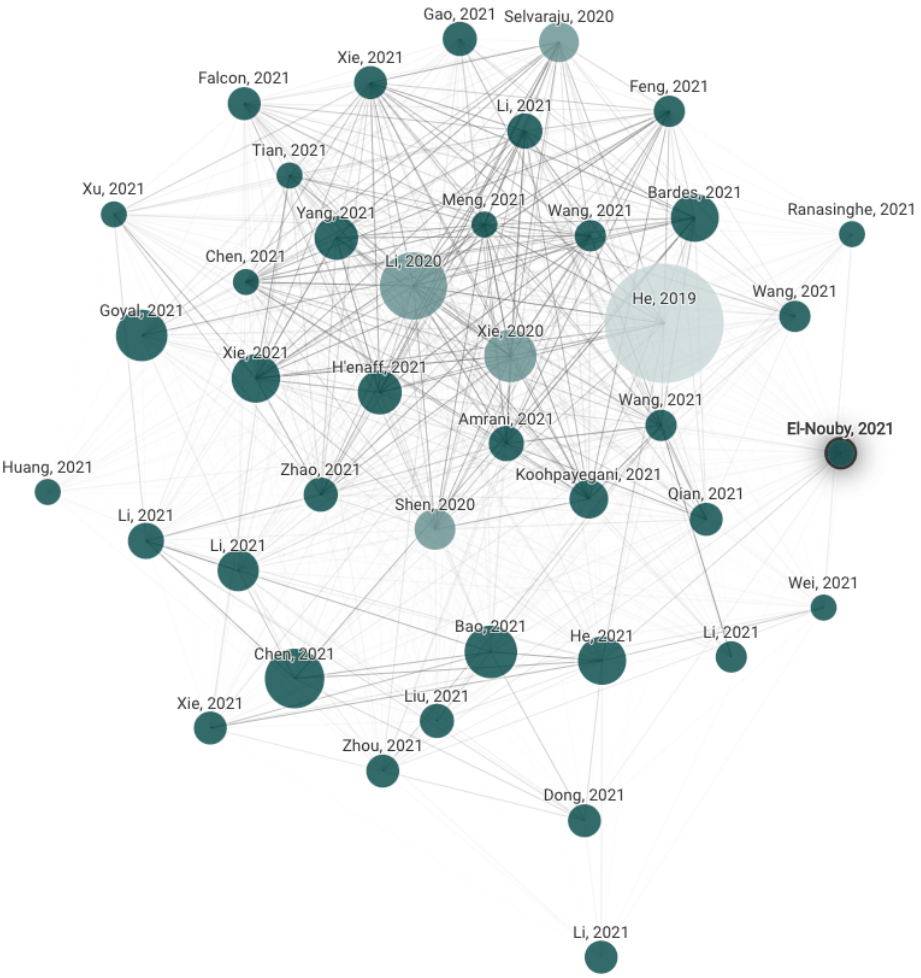
Трудно сказать, какая из цитируемых работ оказала наибольшее влияние на данную работу, но выделить стоит следующие:

1. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in International Conference on Learning Representations, 2021
2. Hangbo Bao, Li Dong, and Furu Wei, “Beit: Bert pre-training of image transformers,” arXiv preprint arXiv:2106.08254, 2021.

На первую из этих работ ссылаются как раз вчерашние конкуренты (о них дальше), она описывает именно общий подход к обучению. А вторая фокусируется именно на MiM+ImageNet, поэтому можно

сказать, что оказывает большое влияние, и в идею исследования El-Nouby ставит ответы на следующие вопросы, оставшиеся после исследования Bao et al:

- Как именно предложенные методы предобучения зависят от размера предобучающей выборки, и в частности, необходимы ли для предобучения миллионы изображений?
- Устойчивы ли предложенные методы к различным типам обучающих изображений, например, не object centric изображений?



Цитируется работа двумя исследованиями, опубликованными уже в 2022 году.
[Context Autoencoder for Self-Supervised Representation Learning]
[Corrupted Image Modeling for Self-Supervised Visual Pre-Training]

Обе эти работы можно рассматривать не только как логическое продолжение данного исследования, но и как конкуренты. Они обе вышли ВЧЕРА)), первая рассматривает тот же подход к self-supervised learning – masked image modeling, и их решение в ходе работы сравнивается с подходом в нашем исследовании El-Nouby, а вторая представляет подход Corrupted Image Modeling. И то и другое, как я поняла, ставит целью пред-обучение на “испорченных” данных.