Evaluating Deep CNNs for Multi-Label Concept Detection in ROCOv2 Radiology Image Dataset by Team LekshmiscopeVIT

Notebook for the ImageCLEF Lab at CLEF 2025

Aryan Sahni[1,*,†], Rachit Gupta[2,†], Raamigaani Venugopal Reddy[3,†] and Lekshmi Kalinathan[4,†]

[1]*Vellore Institute of Technology,Chennai-600127, India*

Abstract

The "Lekshmiscopevit" team presents a ResNet50-based approach for the Concept Detection Task of the ImageCLEF Medical 2025 challenge, using the Radiology Objects in Context version 2 (ROCOv2) dataset. Our experiments explored multiple deep learning architectures, including InceptionV3, DenseNet, and custom convolutional models, with and without pretrained ImageNet weights. Among these, the ResNet50 model consistently outperformed the others, achieving the highest accuracy in both the validation and the test sets. Training was carried out using 80,091 radiology images, 17,277 images used for validation, and 19,267 for testing. To assess the effect of label space complexity, we also experimented with reducing the number of predicted labels to the top most frequently occurring UMLS CUIs. This label reduction improved model performance by alleviating class imbalance and increasing generalization.

Keywords

Transfer Learning, Deep Learning, Concept detection, ResNet50, Multi-label Classification

1. Task Performed

In the context of the ImageCLEFmedical Caption 2025 challenge [1], we have contributed to the Concept Detection Task, which is the task of detecting clinically relevant UMLS concepts directly from radiological images. The task represents a building block toward automatic image captioning and scene understanding in the medical field. We created and trained Multi-Label Classification models to make predictions for UMLS concepts related to each image in the dataset. The concepts were chosen from a filtered portion of the UMLS 2022AB release, including those with greater frequency and specific semantic types to maintain relevance and feasibility. Our method used the ROCOv2 dataset, which contained a training set of 80,091 radiology images, a validation set of 17,277 images, and a test set of 19,267 images. [2] The forecasted concepts were assessed with set coverage measures, namely precision, recall, and F1-score, depicting the correctness and completeness of the concept sets produced by the models. The experiments were all run on only the official training data, as per the task guidelines, to make it comparable with other participating systems. The codes and trained model can be found in the following GitHub repository: https://github.com/C0okiegranny221/CONCEPT-DETECTION.

2. Main Objectives of the Experiments

The main goal of our experiments was to create a successful deep learning pipeline for multi-label concept detection from radiology images within the ImageCLEFmedical Caption 2025 challenge [1].

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

aryn.sahni22@gmail.com (A. Sahni); rachitgupta988@gmail.com (R. Gupta); venugopalreddy.r2023@vitstudent.ac.in (R. V. Reddy); lekshmi.k@vit.ac.in (L. Kalinathan)

https://github.com/C0okiegranny221 (A. Sahni)

0009-0000-3355-8455 (A. Sahni); 0009-0003-0506-6033 (R. Gupta); 0009-0000-8188-7755 (R. V. Reddy); 0000-0002-7005-742X (L. Kalinathan)

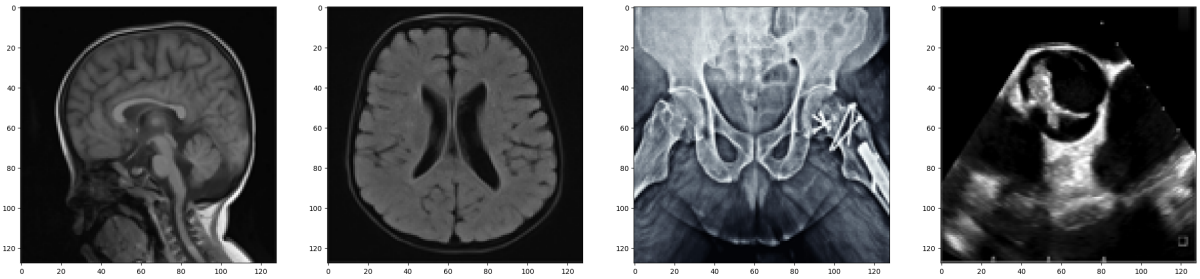

Figure 1: Radiology images from the ROCOv2 dataset [2] ImageCLEFmedical_Caption_2025_test_0, CC-BY (Qi et al.2023), ImageCLEFmedical_Caption_2025_test_1, CC-BY (Qi et al.2023), ImageCLEFmedical_Caption_2025_test_10, CC-BY (Muacevic et al.2024), ImageCLEFmedical_Caption_2025_test_100, CC-BY (Mirhosseini et al.2023)

Of particular interest was finding clinically significant UMLS concepts solely based on visual features, ultimately leading to the facilitation of downstream applications like automatic image captioning and semantic retrieval. To this purpose, we exhaustively tested various convolutional neural network architectures such as ResNet50, DenseNet121, InceptionV3, and custom-designed models under pretrained and randomly initialised weight configurations. These custom models were lightweight CNN models that were made up of dense connections to baseline the tests with the more well-established models later in the test phase. The tests were designed to identify the model architecture with the optimal generalisation performance for unseen medical images. A major experimental aim was to examine the effect of label distribution on model performance by changing the number of concepts employed during training and inference. We explored how limiting predictions to the most common UMLS concepts affected performance metrics like precision, recall, and F1-score. Our highest-performing results were obtained with a ResNet50-based model, and it showed higher accuracy for both the test set and the validation set than other architectures. These results highlight the need for architecture choice and optimisation of label space as key factors in improving visual concept recognition in medical imaging.

3. Approaches Used and Progress Beyond State-of-the-Art

Our strategy towards the concept detection task [1] involved utilising deep convolutional neural networks with effective preprocessing and label encoding techniques suited for multi-label classification. The MultiLabelBinarizer (MLB) was used to encode the UMLS concepts in a binary matrix format, which would allow the model to make multiple concept predictions per image. Input images were preprocessed by Keras ImageDataGenerator, where real-time data augmentation and normalisation were possible, improving the model's generalisation on unseen medical images.

We tried different CNN architectures, such as ResNet50, DenseNet121, and InceptionV3, and found that ResNet50 models with pretrained ImageNet weights performed best overall. Preloading with pretrained weights helped the models transfer low-level feature representations from natural images to medical image data, leading to faster convergence and better accuracy. [3] This transfer learning approach gave a robust initialisation, enabling the network to concentrate on learning domain-specific patterns applicable to radiology. [4]

Compared to conventional concept detection techniques based on handcrafted features or shallow classifiers, our solution achieved a significant improvement by combining deep visual feature learning with multi-label semantic prediction. The use of label frequency analysis and concept space reduction additionally provided performance boosts by concentrating on the most informative clinical concepts. In total, our pipeline demonstrated strong gains on the ROCOv2 validation and test sets [2], demonstrating the power of combining CNN architectures with medical data-specific preprocessing and label optimisation methodologies.

The image size to feed into the neural network was set to a 224 x 224 x 3 image, and the image set was

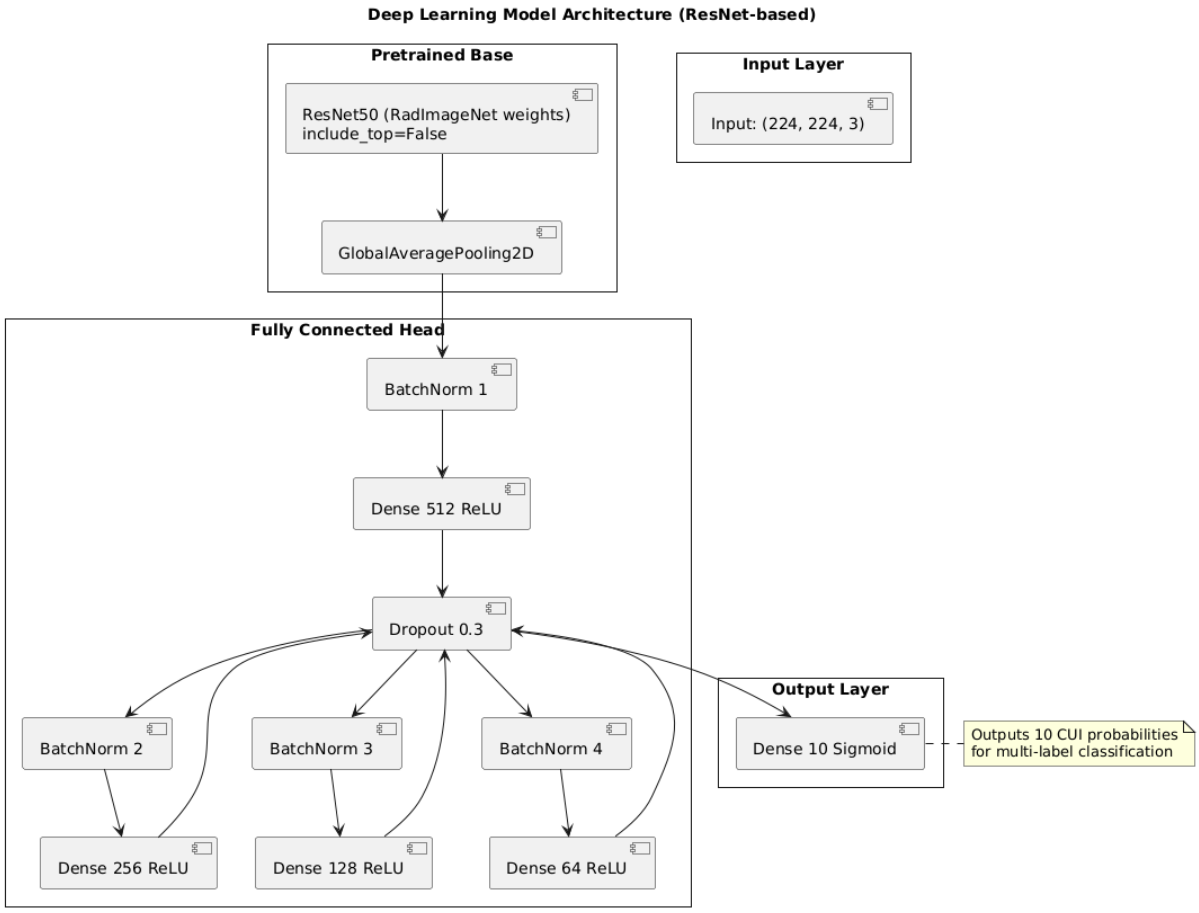

Figure 2: Model architecture with ResNet50 backbone and multi-label output layers.

not shuffled and sent in batches of 128 at a time. Our team used the Adam optimizer, and the learning rate was set to 0.01. We used binary cross-entropy as the loss function in our training and accuracy as the metric. The early stopping function was utilised to avoid overfitting of the model on the training set and was validated against the validation set to a minimum delta of 0.001 and a patience of 3, and the best weights were restored at the end of the training.

4. Resources Used

The experiments were carried out by leveraging a mix of on-campus GPU facilities and cloud-based setups. Much of the training and testing of models was done on GPUs hosted by the high-performance computing facilities of the college, which enabled the necessary computational power for processing the massive ROCOv2 dataset [2]. We also leveraged online environments like Google Colab and Kaggle Notebooks that provided rapid prototyping functionality and working with pretrained models without much hassle. The GPUs used for the training of the models were Tesla T4 with a 30 GB RAM ceiling.

In order to speed up training and increase performance, we utilised pre-trained weights from ImageNet for all deep networks, such as ResNet50, DenseNet121, and InceptionV3 [5]. This application of transfer learning enabled the models to take advantage of learnt features beforehand and thereby decrease training time as well as the likelihood of overfitting, particularly considering the intricacy of radiology image content and the multi-label nature of the task. The integration of heterogeneous computational environments and pre-trained models allowed us to effectively iterate on experiments, tune hyperparameters, and test multiple architectures at scale.

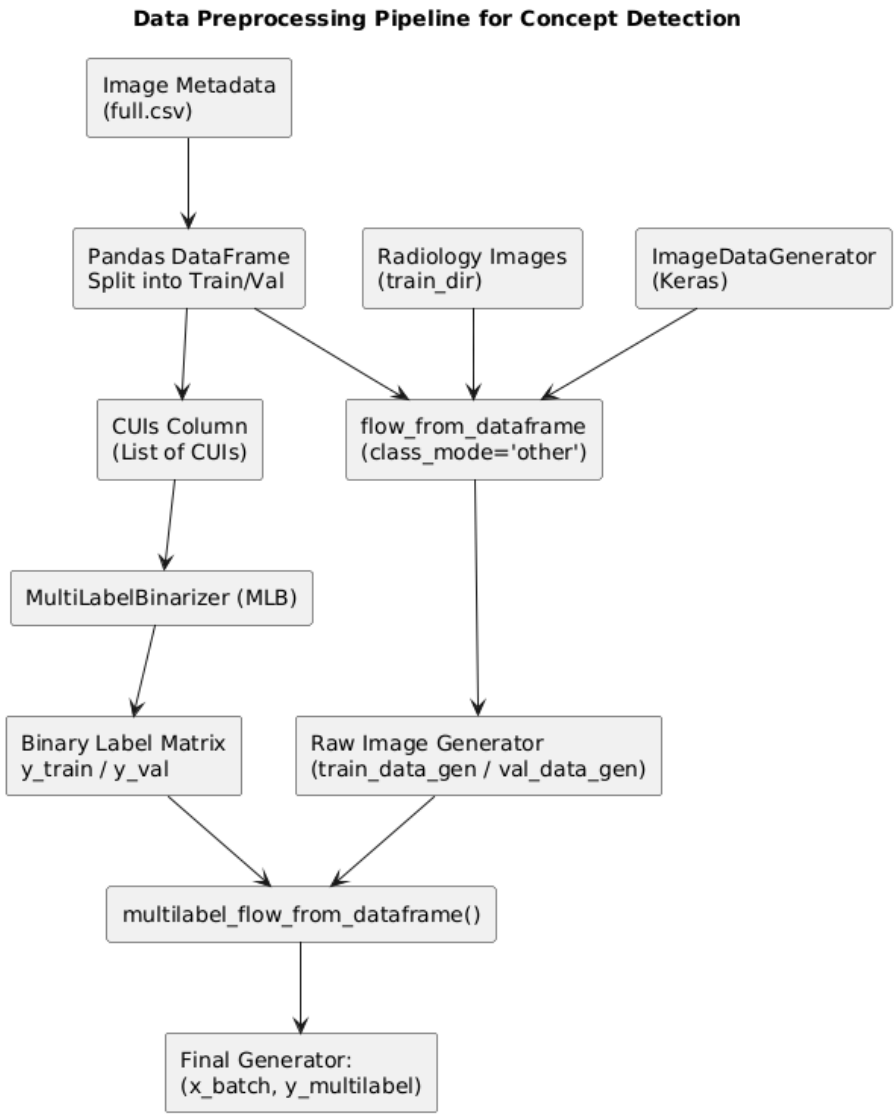

Figure 3: Data preprocessing pipeline from images and labels to batch generator.

5. Results Obtained

Table 1
Comparison of Model Performance

Model	Exact Match Accuracy	Average Jaccard Index	Precision	F1 Score
ResNet50	0.1018	0.3929	0.6579	0.4686
InceptionV3	0.0892	0.3615	0.6311	0.4442
Inception-ResNetV2	0.0955	0.3768	0.6450	0.4579
DenseNet121	0.0990	0.3860	0.6528	0.4635

To compare the performance of various deep learning models for the Concept Detection Task [1], we performed several experiments with pretrained models such as InceptionV3, DenseNet121, and ResNet50. We fine-tuned all models on the ROCOv2 dataset [2] with pretrained ImageNet weights and trained them for an unrestricted number of epochs with early stopping as a callback and the same preprocessing and training pipeline to ensure uniformity.

Among the architectures tested, the ResNet50-based model performed the best. When limiting the

output space to the top 10 most commonly occurring UMLS concepts, the model recorded a Jaccard index of 0.3929 with a corresponding exact match of 0.1018. This outcome captures the power of taking advantage of high-frequency concepts and pretrained feature extractors in a multi-label classification environment for understanding medical images.

Although other models like DenseNet121 and InceptionV3 were competitive in their performance, they could not equate the accuracy levels that ResNet50 attained under equivalent training conditions. These results indicate that ResNet50 is especially best suited to the visual representation requirements of the task of concept detection, particularly when used in tandem with label frequency filtering to decrease the output space complexity.

6. Result Analysis

Even though they are well known for excellent feature extraction strengths, InceptionV3 and DenseNet121 performed poorly compared to ResNet50 under the multi-label concept detection task based on ROCOv2 radiology images [2]. Various architectural and optimisation-related aspects probably attributed to this result [5]. ResNet50 proved to be the best model for this task based on its fast convergence, strong residual learning mechanism, and ability to generalise under a low-label and domain-specific environment. [6] The inferior performance of alternative architectures and the test-validation gap are explained by architectural mismatch with task requirements, distributional shift sensitivity, and transfer learning constraints from non-medical domains. Future research might delve into longer training lengths, pretraining for the medical domain, and curriculum-based label additions for further optimisation.

6.1. Architectural Complexity vs. Task Requirements

InceptionV3, with its deeply modular structure consisting of multiple convolutional filters of different sizes in parallel (e.g., 1×1, 3×3, 5×5), is optimised to learn multi-scale features. [7] While useful in general natural image classification applications, this multiplicity of feature scales could have introduced redundancy in learning features for medical images, where fine-grained domain-specific patterns predominate and might not correspond well to general-purpose, multiresolution filters. In addition, numerous parallel branches add computational and memory requirements, which may slow down convergence in a short number of training epochs.

DenseNet is based on dense connectivity, where every layer takes inputs from all earlier layers. This design promotes feature reuse and prevents vanishing gradients. [8] But practically, this dense connectivity can create overfitting on the unnecessary fine-grained details in high-resolution radiology images [2], especially when the training is conducted for a few epochs and without domain-specific pretraining. The nature of DenseNet to retain many low-level features can also result in information dilution in subsequent layers, which could be undesirable in an application where semantic abstraction and concept-level recognition are more important.

By contrast, ResNet50's residual connections allow stable gradient flow and speed up convergence through the ability to learn identity mappings when deeper transformations are not required. [9] This aspect is especially beneficial for transfer learning with pretrained weights, allowing for efficient adaptation to the target domain without excessive degradation. ResNet50 thereby balances depth, simplicity, and transferability [5] and is more suitable for medical concept detection with sparse label space and high intra-class similarity.

6.2. Impact of Label Space Restriction

Another important factor in the observed performance differences is the decision to restrict the output label space to the top 10 most frequent UMLS concepts. This substantially reduced label sparsity and class imbalance, which typically plague multi-label classification tasks. Models with higher capacity (e.g., DenseNet) may require larger label diversity to showcase their full representational power. Conversely,

ResNet50 took advantage of the decreased complexity of labels to allow it to converge better in the lower-dimensional label space.

6.3. Gap between Validation and Test Set Performance

Although the performance in validation went up to 0.3929, in the unseen test set it was significantly lower achieveing a F1 score of 0.1494 and a secondary F1 score of 0.2298.There could have been several reasons why this gap ensued. First the Dataset Distribution Shift ,the ROCOv2 dataset [2]has been reported to have a wide range of imaging modalities and diagnostic scenarios. The validation set was drawn from the same distribution as training data, while the test set includes totally unseen images, potentially including underrepresented modalities, resolutions, or clinical conditions. Such domain shift can cause reduced generalization. The model could also have been privy to Overfitting to Most Frequent Patterns as the model was trained and tested on most frequent 10 concepts, which could result in overfitting to those dominant patterns. If the test set has a slightly different frequency distribution or a greater level of label noise, the model would not be able to adapt and hence will perform lower accuracy. Batch normalization layers found in all utilized models are batch statistics sensitive. At training time, they learn to match the training/validation batch distribution. At inference time over the test set (particularly when conducted in small batches), statistics mismatch causes suboptimal scaling of activations and performance degradation.

7. Perspectives for Future Work

While the existing method concentrated on utilizing ResNet50 with pre-initialized ImageNet weights and decreasing the label domain to common ideas, there are a variety of promising avenues that can be improved upon to increase performance and extend applicability. Further research can delve into self-supervised pretraining on medical image datasets, e.g., MIMIC-CXR or CheXpert, to improve feature representations towards the domain specific meaning of radiology. In addition, multi-modal learning by blending image features with metadata or subset caption text can potentially allow more context-sensitive predictions. Another novel avenue is the utilization of graph neural networks (GNNs) to capture co-occurring and hierarchical semantic relationships between UMLS concepts, enabling the model to draw on semantic dependencies in classification. A transformer-based approach has also shown to to be better performing in cases of multilabel classification [10]. In addition, uncertainty-aware learning with Bayesian deep learning may alleviate label noise and dataset bias, particularly in the long tail of infrequent concepts. Finally, adding visual grounding or attention maps to identify concept-related regions within the image would make the system more interpretable for clinical users, paving the way for hybrid AI-human diagnostic pipelines.

Acknowledgments

This research was supported by the Department of Science and Technology (DST), India, under the Fund for Improvement of S&T Infrastructure in Universities and Higher Educational Institutions (FIST) Program [Grant No. SR/FST/ET-I/2022/1079], along with a matching grant from VIT University. The authors express their sincere gratitude to DST-FIST and the VIT management for their financial assistance and the infrastructural support provided for this work.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 Turbo, QuillBot, and Grammarly in order to: Grammar and spelling check. Further, the author(s) used GPT-4 Turbo for rephrasing sentences or paragraphs to improve clarity, conciseness, or style. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] H. Damm, T. M. G. Pakull, H. Becker, B. Bracke, B. Eryilmaz, L. Bloch, R. Brüngel, C. S. Schmidt, J. Rückert, O. Pelka, H. Schäfer, A. Idrissi-Yaghir, A. Ben Abacha, A. García Seco de Herrera, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2025 – Medical Concept Detection and Interpretable Caption Generation, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[2] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. Ben Abacha, A. García Seco de Herrera, H. Müller, P. Horn, F. Nensa, C. M. Friedrich, ROCOv2: Radiology Objects in Context Version 2, an Updated Multimodal Image Dataset, Scientific Data 11 (2024). doi:10.1038/s41597-024-03496-6.

[3] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, T. Ganslandt, Transfer learning for medical image classification: a literature review, BMC medical imaging 22 (2022) 69. doi:10.1186/s12880-022-00793-7.

[4] L. Torrey, J. Shavlik, Transfer learning, in: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, IGI global, 2010, pp. 242–264. doi:10.4018/978-1-60566-766-9.ch011.

[5] S. R. Shah, S. Qadri, H. Bibi, S. M. W. Shah, M. I. Sharif, F. Marinello, Comparing inception V3, VGG 16, VGG 19, CNN, and ResNet 50: A case study on early detection of a rice disease, Agronomy 13 (2023) 1633. doi:10.3390/agronomy13061633.

[6] D. Lu, Q. Weng, A survey of image classification methods and techniques for improving classification performance, International journal of Remote sensing 28 (2007) 823–870. doi:10.1080/01431160600746456.

[7] N. Sharma, V. Jain, A. Mishra, An analysis of convolutional neural networks for image classification, Procedia computer science 132 (2018) 377–384. doi:10.1016/j.procs.2018.05.198.

[8] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, S. Yan, HCP: A flexible CNN framework for multi-label image classification, IEEE transactions on pattern analysis and machine intelligence 38 (2015) 1901–1907. doi:10.1109/tpami.2015.2491929.

[9] S. Mascarenhas, M. Agarwal, A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification, in: 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), volume 1, 2021, pp. 96–99. doi:10.1109/CENTCON52345.2021.9687944.

[10] J. Lanchantin, T. Wang, V. Ordonez, Y. Qi, General multi-label image classification with transformers, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 16478–16488. doi:10.1109/cvpr46437.2021.01621.