Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue



Cite this: *Digital Discovery*, 2025, **4**, 3744

# MaskTerial: a foundation model for automated 2D material flake detection

Jan-Lucas Uslu, *[ab] Alexey Nekrasov,[b] Alexander Hermans,[b] Bernd Beschoten,[a] Bastian Leibe,[b] Lutz Waldecker[a] and Christoph Stampfer[ac]

The detection and classification of exfoliated two-dimensional (2D) material flakes from optical microscope images can be automated using computer vision algorithms. This has the potential to increase the accuracy and objectivity of classification and the efficiency of sample fabrication, and it allows for large-scale data collection. Existing algorithms often exhibit challenges in identifying low-contrast materials and typically require large amounts of training data. Here, we present a deep learning model, called MaskTerial, that uses an instance segmentation network to reliably identify 2D material flakes. The model is extensively pre-trained using a synthetic data generator that generates realistic microscopy images from unlabeled data. This results in a model that can quickly adapt to new materials with as little as 5 to 10 images. Furthermore, an uncertainty estimation model is used to finally classify the predictions based on optical contrast. We evaluate our method on eight different datasets comprising five different 2D materials and demonstrate significant improvements over existing techniques in the detection of low-contrast materials such as hexagonal boron nitride.

Received 17th April 2025
Accepted 20th October 2025

DOI: 10.1039/d5dd00156k

rsc.li/digitaldiscovery



## Introduction

The ability to combine different 2D materials into van der Waals heterostructures has opened up new ways to study fundamental phenomena in solids,[1–6] to tailor material properties[7–12] and to design device structures with improved performance.[13–17] In most research settings, these heterostructures are assembled from individually exfoliated material flakes.[18–20] The identification and selection of suitable 2D material flakes for device fabrication is the first and an integral part of this process.[21] It has traditionally been performed by researchers who scanned large pieces of exfoliation substrates using a microscope.

Automating the detection of exfoliated 2D material flakes using computer vision algorithms has the potential to significantly improve sample preparation efficiency and accelerate the pace of research.[22] For this task, previous work has explored the use of classical machine learning methods, such as support vector machines (SVMs) and *K*-means clustering.[23–25] These methods rely on the discrete nature of the optical contrast values of 2D materials with respect to the substrate material.[26] This discrete nature is a result of their atomic-scale thickness, where each layer of material corresponds to a single, uniform atomic plane. Optical contrast variations arise due to interference effects, where the interaction of light with the material and substrate depends on the exact number of these atomic layers, leading to quantized optical contrasts for each layer count. However, the performance of current detection models typically decreases significantly for materials with low optical contrast and they are sensitive to variations in substrate thickness and lighting conditions.[21] More recently, deep learning approaches using neural networks have been employed to address these limitations.[27–30] Although these methods offer greater versatility, they typically require large amounts of labeled training data, which can be impractical to obtain in a research setting, especially if the yield of exfoliated materials is low.

Recently, the emergence of foundation models in artificial intelligence has transformed numerous fields by providing pre-trained models that can be fine-tuned for diverse tasks with minimal labeled data.[31] Foundation models, such as GPTs[32] for language and vision transformers[33] (ViTs) for image processing, leverage extensive pre-training on large and diverse datasets, enabling them to generalize across domains with limited additional training. These models are often trained on large-scale datasets using self-supervised or unsupervised learning techniques, allowing them to capture broad representations of data. This versatility makes them particularly powerful for tasks where labeled data is scarce or hard to obtain, as they can transfer learned features effectively to new domains.

The success of foundation models in other domains inspires the potential for similar advancements in 2D material flake

[a]*2nd Institute of Physics and JARA-FIT, RWTH Aachen University, 52074 Aachen, Germany*
[b]*Visual Computing Institute, RWTH Aachen University, 52074 Aachen, Germany*
[c]*Peter Grünberg Institute (PGI-9), Forschungszentrum Jülich, 52425 Jülich, Germany*

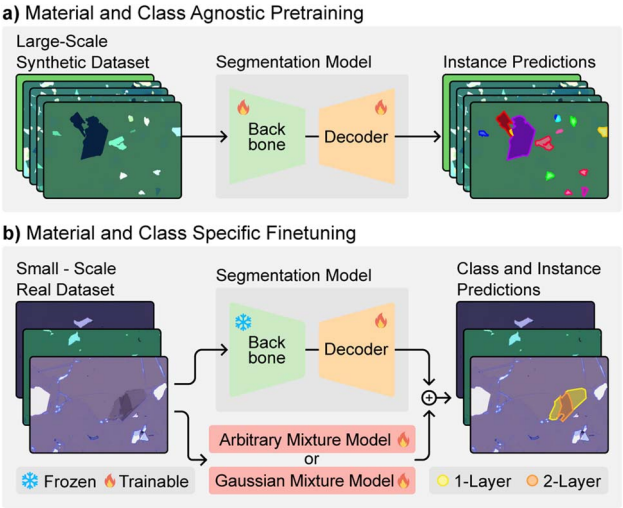


Fig. 1 MaskTerial uses a two step approach to train a robust foundation model for 2D material flake detection. (a) The segmentation model is pre-trained on a large corpus of synthetic data of multiple to learn a good internal representation of the data. The synthetic data contains no information about the material type or thickness of the flakes. This results in a material and class agnostic pre-trained foundation model. During this step, both the backbone and the decoder are trained. (b) After pre-training, a small number of images is used to fine-tune the classification model and the decoder of the foundation model. The classification model can be either the new arbitrary mixture model or any existing Gaussian mixture model from ref. 21.

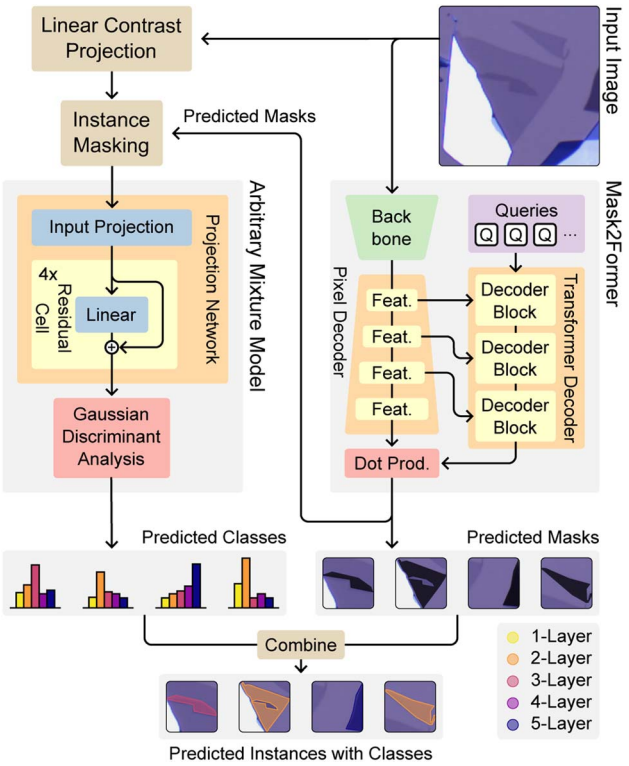


Fig. 2 MaskTerial consists of two models: an instance segmentation model and a classification model. First, the input image is processed by the instance segmentation model returning a set of possible flakes without classifying their thickness class. If possible flakes are found, the image is projected into the contrast space representation following the method described in ref. 21. Afterwards, the masks of the predicted instances are used to extract the contrast values from the transformed image for each of the instances. These are then classified by the classification model to generate probability distributions over the classes of the flakes. The mode of these distributions is used to classify each instance, yielding the final predicted flakes with thickness classes.

detection. By leveraging pre-trained models and domain-specific fine-tuning, foundation models can address key limitations such as the need for large labeled datasets and the challenges posed by low-contrast materials. Building on these principles, we propose a tailored approach to tackle the specific challenges of 2D material flake detection. First, we introduce a deep learning architecture that combines a modified Mask2Former[34] model for instance segmentation with a physics-informed uncertainty estimation head based on the deep deterministic uncertainty (DDU) method.[35] This architecture allows for robust detection and classification of 2D material flakes, even for low-contrast materials such as thin hexagonal boron nitride (hBN). Second, we propose a synthetic data generation pipeline using physical simulations in conjunction with unlabeled data to address the lack of large amounts of labeled data. We show that extensive pre-training of our model using synthetic data (see Fig. 1a) allows it to be fine-tuned with as few as 5 to 10 microscope images per material (see Fig. 1b). Finally, we present eight different datasets covering five different 2D materials used for training and evaluation to validate the performance of our models.

## Model architecture

The MaskTerial architecture combines two deep learning models. The first model, the instance prediction model (Fig. 2 – Mask2-Former), predicts all flakes of interest in the image, regardless of the actual class of the predicted flake. The second model (Fig. 2 – arbitrary mixture model) then takes all the predicted interesting flakes and assigns them classes based on their contrasts, *i.e.* monolayer, bilayer, *etc.* This separation of instance prediction and class prediction has the benefit that, when adding new materials, only the latter model needs to be retrained.

### Instance prediction model

The instance prediction model is based on the Mask2Former[34] architecture (see Fig. 2). It works by first extracting feature representations from the input image using a ResNet50 (ref. 36) backbone. Afterwards, the extracted feature representations are gradually upscaled by a pixel decoder (PD). During upscaling, the features are sequentially fed into the transformer decoder (TD) at multiple levels.

A unique aspect of Mask2Former is its use of learnable query embeddings, introduced by the DETR[37] architecture. We train these queries to act as proxies for potential object instances or specific semantic categories within the image. During the decoding process, these queries interact with the encoded image features *via* cross-attention mechanisms within the TD contextualizing the queries. These contextualized queries are then used to generate segmentation masks of each object by computing the dot product between them and the final feature

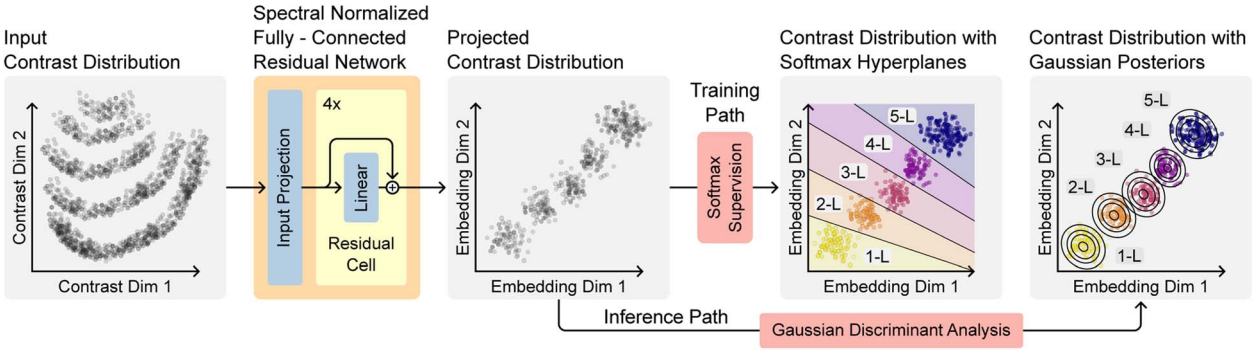


Fig. 3 The arbitrary mixture model (AMM) works by projecting an input distributions onto a distribution of Gaussians. This is achieved by combining a ResNet with spectral normalization techniques. The combination of regularization and residual connections allows the training process to be supervised using a straightforward softmax function and standard cross-entropy loss. After training, a Gaussian is fitted to the embedding representations of each class. During inference, the embedding representations of the input data are evaluated against the learned Gaussians from training to determine class conditional posteriors and thus the probability that any given input contrast belongs to a given thickness class.

representation from the PD. In our case, the segmentation does not classify the instance by layer count, such as monolayer or multilayer; instead, it only identifies interesting objects (*i.e.* a 2D material flake). This improves the detection accuracy for downstream tasks (see Table 2).

### Classification model

The second component of MaskTerial is the arbitrary mixture model (AMM), which assigns layer thicknesses to optical contrasts of the flakes. As discussed in our previous work,[21] variations in the oxide thickness in the $Si/SiO_2$ wafers, used to exfoliate the 2D material flakes, lead to non-trivial distributions of these contrasts (see Fig. 3 – input contrast distribution), making them difficult to fit and the detection unreliable. To counteract this, we propose a model which learns a regularized mapping of arbitrary class distributions in the optical contrast space to Gaussians, solving the problem of non-trivial distributions while preserving the interpretability of the contrast distributions (see Fig. 3 – projected contrast distribution). To achieve this, we use an approach for uncertainty estimation in deep learning proposed by Mukhoti *et al.*,[35] who introduced the deep deterministic uncertainty (DDU) method. It addresses the limitations of traditional probabilistic models, such as Bayesian neural networks, which can be computationally expensive and difficult to train.[38]

They proposed to use a combination of spectral normalization and residual connections to constrain the model to learn a smooth and locally linear embedding space. Spectral normalization works by constraining the eigenvalues of each weight matrix by dividing them by their largest eigenvalue during each training step.[39] Residual connections allow the network to learn perturbations around the identity function, which has been shown to improve the stability and convergence of deep networks.[36] Together, these two methods impose a bi-Lipschitz constraint on the model, leading to a robust and sensitive embedding space while preventing feature collapse to a single point.[35,39,40]

During inference, the class probabilities are computed by evaluating the probability density function of each class-conditional Gaussian at the embedding space coordinates of the input feature, giving a probability of the instance belonging to any given class.

The ability to interpret the resulting distributions is an important aspect of training the model in this way. Unlike typical deep learning models that learn arbitrary functions minimizing some objective function, the model can provide uncertainty estimates based on a regularized projection to class-conditional Gaussians.

## Synthetic data generation

Effective training of deep learning models requires a large amount of labeled data to ensure that the model accurately captures the underlying data distribution.[41] However, collecting and annotating real-world data is often challenging and time consuming. To address this issue, we developed a synthetic data generation engine that incorporates physical knowledge and simulations to generate images that closely resemble real-world microscopy images to pre-train the segmentation model.

The image generation process has two main phases, a shape mining phase in which we extract plausible flake shapes from a dataset of unlabeled images (Fig. 4a), and a generation phase, in which we use the extracted shapes to generate new images together with the ground truth masks (Fig. 4b).

We extract shapes from a dataset of around 100 000 unlabeled images of exfoliated graphite from an internal database. These unlabeled images where collected in an automated fashion by exfoliating graphene and scanning the wafers using a motorized microscope. These unlabeled images are only used for the synthetic data generation and not annotated to be used as testing datasets to avoid contamination of the datasets by bleeding information. Since most commonly used 2D materials have a hexagonal crystal structure, we assume that the shapes of their exfoliated flakes will generally be similar to those of graphite.

The images are converted to grayscale, and we then apply a stepped brightness threshold. By setting specific brightness ranges to one and all other values to zero, we create binary masks for different brightness levels. We then use a connected

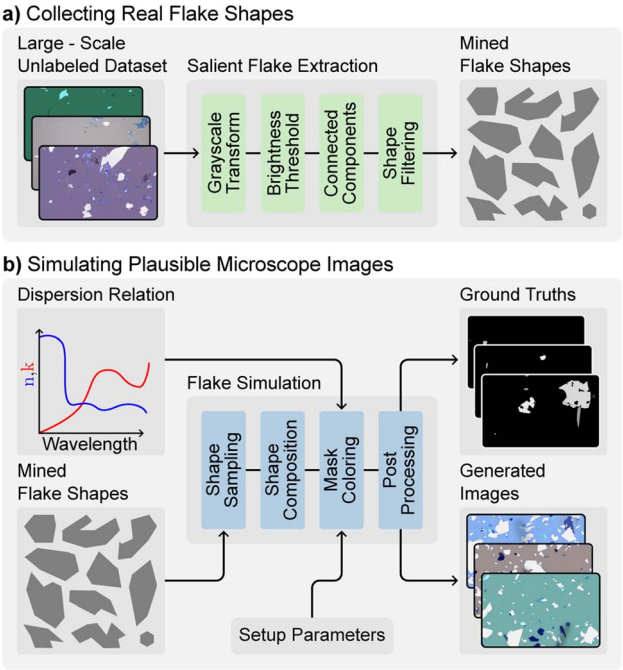


Fig. 4 The workflow for generating synthetic flake images consists of two steps. (a) The process begins by mining real flake shapes from a large unlabeled dataset of exfoliated flake images. This involves first converting images to grayscale, applying brightness thresholds, finding connected components in the resulting binary masks, and finally filtering the detected shapes to keep only high quality flake shapes. (b) The second step uses the previously mined flake shapes to generate plausible synthetic images with associated ground truth masks. This is done by first sampling a set of shapes from the mined shapes and scattering them over an empty placeholder image to create a grayscale ground truth image. Then, using the optical dispersion relation of the target material and the setup parameters such as visible light spectrum, camera activation curve, and substrate thickness, the colors of the material are simulated. Finally, post-processing, such as adding noise, vignetting and shadows, is applied to create the final synthetic image.

components algorithm[42] to extract all connected shapes from the binary masks. Finally, we filter these shapes using an L2 classifier.[21] In total, 35 000 flake shapes were collected.

In the image generation process (Fig. 4b), we sample a random number of shapes (ranging from 1 to 500) and randomly place them on an empty canvas at different angles, sizes, positions, and thicknesses, creating a grayscale image where the pixel values correspond to the number of layers for any given pixel. When shapes overlap, their layer counts are added in the overlapping area, creating a grayscale mask.

The color of the flakes and the background is approximated using a simulation based on the transfer-matrix method (TMM). First, the reflectance of each pixel is calculated taking into account the thickness of the $SiO_2$ layer of the substrate, the dispersion relation of the material and the thickness of the pixel considering the grayscale mask. The color is calculated by integrating the simulated spectral reflectance multiplied by the camera activation curve and the light source spectrum for each RGB channel.

Finally, the images are post-processed by adding a layer of random tape residue emulated with simplex noise, random shadows, a vignetting effect, and Gaussian camera noise to closely resemble real images.

We generated about 42 000 synthetic images with ground truth masks each for graphene, chromium triiodide ($CrI_3$), hBN, tantalum disulfide ($TaS_2$), molybdenum diselenide ($MoSe_2$), tungsten disulfide ($WS_2$) and tungsten diselenide ($WSe_2$), resulting in a total of about 300 000 synthetic images, which were used for pre-training.

## New datasets

To fine-tune and evaluate the model, we collected eight new datasets from five different materials. We collected three datasets for exfoliated graphite and two datasets for $WSe_2$ with different substrate thicknesses to measure the robustness of the models. These datasets are the low, medium, and high variance datasets to denote the range of different substrate thicknesses in the training and testing sets. The low variance datasets contain images with substrate thicknesses within ~5 nm of the ~90 nm substrate thicknesses used. The medium and high variance datasets contain images with ranges of ~10 nm and ~20 nm, respectively. In addition, we have collected datasets for hBN, $MoSe_2$, and $WS_2$ with substrate thickness variations of about 10 nm. The training and test images are from independent exfoliation runs to ensure that the test images do not bleed into the training images. Table 1 lists the datasets, the number of images in the train and test sets, and the number of exfoliation runs in the train and test sets.

Table 1 We collected eight datasets from five materials to measure the performance of the models on different materials and substrate variations. We chose a 50/50 train test split to better capture the data distribution in the test set

| Dataset | Train/test images | Annotated classes | Train/test exfoliations |
|---|---|---|---|
| Graphene (low) | 425/1362 | 1–4 layers | 2/10 |
| Graphene (medium) | 357/325 | 1–4 layers | 8/9 |
| Graphene (high) | 438/480 | 1–4 layers | 10/10 |
| $WSe_2$ (low) | 92/420 | 1–3 layers | 2/12 |
| $WSe_2$ | 97/99 | 1–3 layers | 5/5 |
| hBN | 73/62 | 1–3 layers | 2/3 |
| $WS_2$ | 53/94 | 1 layer | 2/2 |
| $MoSe_2$ | 63/97 | 1–2 layers | 7/8 |

## Training

### Instance prediction model

The instance prediction model was trained in two stages. First, we performed extensive pre-training using the 300 000 simulated images on 8 NVIDIA V100 GPUs (see Fig. 1a). We trained for 90 000 iterations with a batch size of 56 with images cropped to a resolution of 1024 × 1024. We used the AdamW[43] optimizer with a learning rate of $10^{-4}$, a weight decay of $5 \times 10^{-2}$. The learning rate scheduler we used was a simple linear decay scheduler. Finally, we enabled gradient clipping throughout the model and clipped them to $10^{-2}$. The pre-training took about 52 hours.

For further fine-tuning, we used the pre-trained instance prediction model as a base while freezing the parameters of the backbone (see Fig. 1b). This training was performed on a single NVIDIA V100 for 500 iterations with a batch size of 24 with images cropped to a resolution of 512 × 512. We used the same parameters as for pre-training, except that we changed the learning rate to $10^{-5}$. Using this setup each fine-tune takes about 5 to 7 minutes. All fine-tuning experiments used the same hyperparameters and only the real images from mechanical exfoliation where used for fine-tuning and no synthetic images where used after the pre-training step.

### Classification model

For training the AMM, we first extracted contrast values from all annotated flakes to obtain the contrast distribution. To ensure robust model training, we implemented a multi-stage denoising pipeline to remove outliers and artifacts that could compromise classification performance. The denoising process targets three primary sources of contamination: (i) edge pixels from flakes where mask erosion was insufficient, (ii) imaging artifacts such as dust particles, and (iii) transition pixels between regions of different thickness. We used a two-step approach: first applying a $k$-nearest neighbor ($k = 25$) classifier across all classes to reassign mislabeled points based on their local neighborhood. This is followed by DBSCAN clustering.[44] The DBSCAN algorithm was configured with conservative parameters ($\varepsilon = 0.1$, min_samples = 10% of class size) to preserve the core distribution structure while removing only clear outliers, thereby minimizing any impact on downstream classification performance. Following denoising, we normalized the contrast distribution using $z$-score normalization (zero mean, unit variance) and applied balanced sampling to ensure equal representation across all thickness classes, effectively addressing class imbalance issues inherent in the dataset.

For training, we used the Adam optimizer[45] with a learning rate of 0.01, a batch size of 10 000, and 5000 iterations with a dropout probability of 10%. For the loss function, we used the standard cross entropy. The network used an embedding dimension of 16, a depth of 4 and a spectral coefficient of 0.5. Training on a CPU takes about 5 minutes.

## Evaluation

We use the Average Precision at 50% IoU (AP50) as the evaluation metric. The AP50 is the area under the precision–recall curve at a threshold where the Intersection over Union (IoU) between

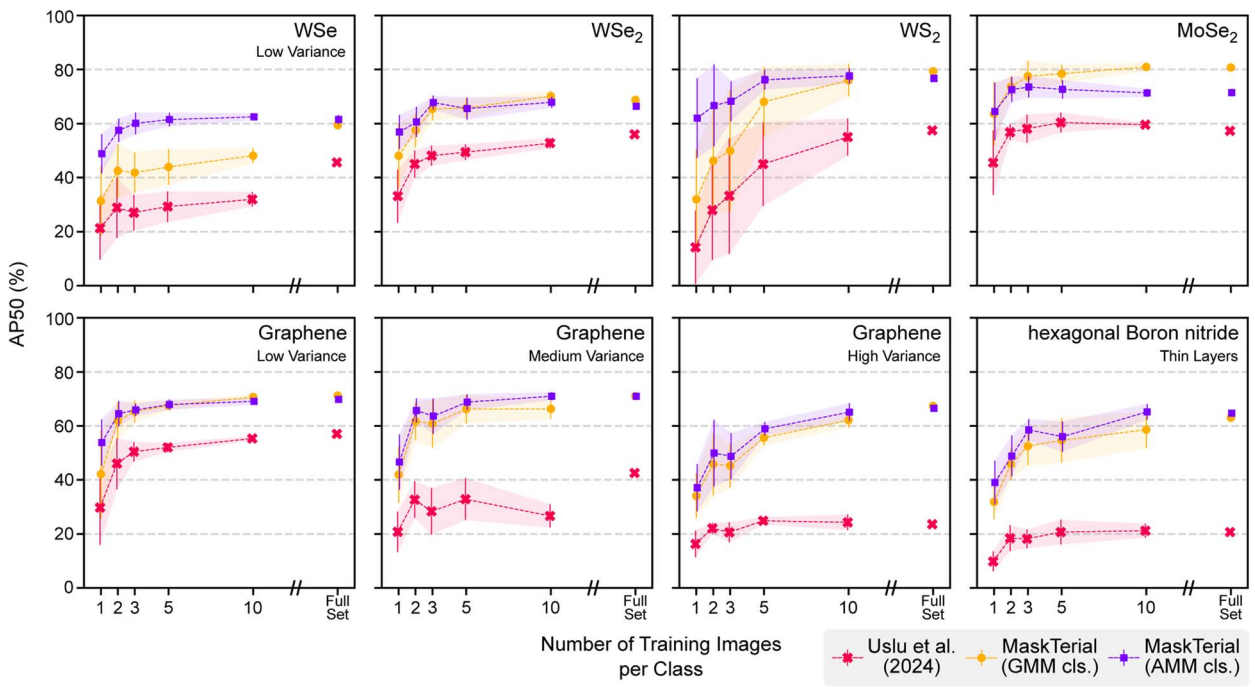


Fig. 5 The model has been evaluated on eight different datasets when trained with different amounts of images per class. MaskTerial outperforms the baseline model from ref. 21 for all training thresholds and for all materials. An interesting find is that MaskTerial outperforms the fully trained baseline model with as little as two images per class for all materials. Furthermore MaskTerial seems to saturate after as little as two images per class for most materials making further training unnecessary.

Table 2 The table shows the impact of synthetic pre-training and classification model choice on the average AP50 scores of various model configurations. The deep learning instance prediction models without synthetic pre-training struggle significantly with detection, as shown by their low AP50 scores. Although using AMM as the classification model has a smaller impact on the AP50 than pre-training, it contributes to more stable and consistent results across datasets. The results are averaged across all datasets, with a threshold of 10 images per class

| Model | Synthetic pre-training | Classification model | Average AP50 |
|---|---|---|---|
| GMM only | ✗ | GMM | 40.8 ± 15.5 |
| AMM only | ✗ | AMM | 43.8 ± 12.2 |
| MaskTerial | ✗ | Mask2Former | 3.6 ± 3.5 |
| | ✗ | GMM | 2.7 ± 2.8 |
| | ✗ | AMM | 2.5 ± 2.4 |
| | ✓ | Mask2Former | 35.2 ± 8.1 |
| | ✓ | GMM | 66.8 ± 10.6 |
| | ✓ | AMM | **68.9 ± 4.9** |

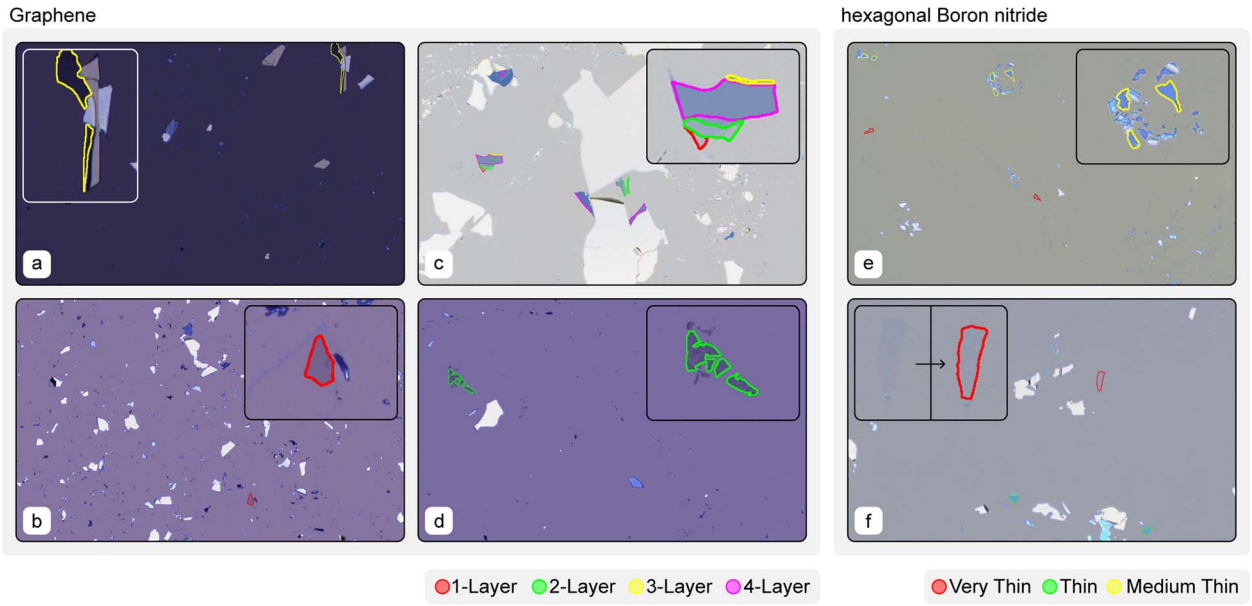


Fig. 6 Example images of the MaskTerial-AMM model's ability to generalize across a wide range of imaging conditions and materials as well as its limitations. (a) Correct prediction of a three-layer flakes under dim lighting conditions. (b) Precise segmentation of flakes in cluttered images. (c) A flake consisting of multiple thicknesses that is segmented into multiple pieces. It also shows one of the model's limitations, namely its tendency to group similar classes into a single flake. The two-layer prediction shows that the model grouped instances from one-layer and two-layer together. (d) For closely clustered flakes, the model tends to group them into a single prediction, whereas multiple cleanly separated instances are expected. (e) The model has a tendency to miss very small flakes. (f) Demonstration of the model's ability to segment and correctly classify low-contrast hBN flakes.

predicted and ground truth boxes is at least 50%. This effectively measures the models performance in detecting instances of interest while minimizing false positives and is used as one of the default metrics when it comes to instance segmentation models. The model was trained with varying numbers of training images to evaluate its performance and ability to handle few-shot learning tasks across different materials.

We also provide confusion matrices in the supplementary information to compare the MaskTerial model with an AMM head against the GMM from ref. 21 for different confidence- and size thresholds.

### Quantitative results

We compare the Mask2Former instance segmentation model with an AMM classification head (MT-AMM) against the Gaussian mixture model (GMM) of ref. 21 and the Mask2Former instance segmentation model with a GMM classification head (MT-GMM). The models were evaluated for both the few-shot and full-data tasks with our eight datasets. All models were trained and evaluated ten times in different data subsets to obtain metrics on their performance (see Fig. 5).

The results show that the MT-AMM and MT-GMM outperform the GMM baseline by a large margin of at least 10% on all datasets and for any number of training examples. The performance increase is particularly strong for materials with low optical contrast, such as thin layers of hBN, and highly varying substrate thicknesses, with metrics improving by up to 40%.

For almost all materials, we see diminishing returns for the metrics even with more training examples, indicating that the model has already learned the distributions from only 2 to 5

example images. The MT-AMM is particularly strong when in the low data regime, it outperforms the MT-GMM in this regime while also providing more stable performance (see Table 2). When using more data, the MT-GMM starts to match the performance of the MT-AMM.

Some qualitative results are shown in Fig. 6. The MT-AMM is able to predict flakes in images which are very dark (a), cluttered (b and e) as well as flakes which are stuck together (c and d). The model is particularly good at detecting very low contrast flakes such as few-layer hBN (f).

### Ablations

To determine which contributions most improved the model's performance, we conducted ablation studies. Specifically, we evaluated the impact of synthetic pre-training and the choice of classification model (GMM *vs.* AMM *vs.* Mask2Former) on the detection metrics.

Table 2 highlights the effect of these components: models without synthetic pre-training achieve very low AP50 scores. Having the Mask2Former model predict the classes itself also reduces performance, resulting in unstable and inconsistent detection across datasets. These results show the critical role of synthetic pre-training and the stability advantage provided by AMM.

## Conclusion

In this paper, we have presented a deep learning architecture, paired with a synthetic data generator, that improves upon existing algorithms for the detection of 2D material flakes in microscopy images, particularly for materials with low optical contrast (see Fig. 6f). Our model combines an instance prediction model with an uncertainty estimation model to make decisions based on physical features. We have shown that our model significantly outperforms current state-of-the-art methods and can be trained on as few as 5 to 10 images per class in a few minutes. Furthermore the model is able to be run on low-end consumer grade GPUs, such as a Nvidia RTX 3060 Ti, with low inference times of about $\approx$280 ms per image for full HD images of resolution 1920 pixels by 1200 pixels. Better High-end GPUs such as the Nvidia RTX 4080 Super are able to push the inference times down to $\approx$120 ms per image.

The strength of our approach is the use of physical inductive biases in the model architecture. By incorporating physical knowledge into the decision-making process, our model provides interpretable predictions that can be validated for further downstream processing, such as stacking of different 2D flakes into van der Waals heterostructures. In addition, our few-shot learning capability allows for the detection of difficult to exfoliate and detect materials, which is a significant advancement over existing methods that require large annotated datasets.

We have also introduced a synthetic data generator that mimics the true distributions of microscopy images. This generator allows us to create large datasets for pre-training deep learning models, reducing the need for extensive data collection and annotation. Our study shows that pre-training with synthetic data significantly improves the performance of our instance detection model, highlighting the potential of this approach in the 2D materials community.

We also published the code together with implementation details, demos, guides, training scripts and evaluation scripts to GitHub,[46] so researchers can finetune the model to their materials.

Despite the strengths of our approach, there are some limitations to consider. First, the model is less efficient at detecting small instances with an area less than 200 pixels (see Fig. 6d). This is a common challenge for deep learning models without specialized layers and techniques. Second, when predicting instances that are close together, the model tends to combine them into a single prediction, leading to instance misclassification (see Fig. 6c).

In summary, our novel deep learning architecture and synthetic data generator represent a significant step forward in the automated detection of 2D materials in microscopy images. By exploiting physical inductive biases and few-shot learning capabilities, our models enable the detection of rare materials and provide interpretable predictions. Although there are some limitations to our approach, we believe that our contributions lay the foundations for future research in this area and have the potential to have a major impact on the field of 2D materials science.

## Conflicts of interest

There are no conflicts of interest to declare.

## Data availability

All code, datasets, and model weights generated and analysed during this study are deposited in Zenodo and are publicly available at: **https://doi.org/10.5281/zenodo.14415557**. The development repositories for the code are also available on GitHub:

- Model code: **https://github.com/Jaluus/MaskTerial**.
- Synthetic data generator code: **https://github.com/Jaluus/MaskTerial-DataGen**.

Supplementary information is available. See DOI: **https://doi.org/10.1039/d5dd00156k**.

## Acknowledgements

This project has received funding from the European Research Council (ERC) under grant agreement no. 820254 and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – Cluster of Excellence Matter and Light for Quantum Computing (ML4Q) EXC 2004/1 – 390534769. A. N. acknowledges funding by the BMBF project "WestAI" (grant no. 01IS22094D).

## References

1 K. S. Novoselov, D. Jiang, F. Schedin, T. J. Booth, V. V. Khotkevich, S. V. Morozov and A. K. Geim, Two-dimensional atomic crystals, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 10451.

2 A. K. Geim and K. S. Novoselov, The rise of graphene, *Nat. Mater.*, 2007, **6**, 183.
3 K. S. Novoselov, A. Mishchenko, A. Carvalho and A. H. C. Neto, 2D materials and van der Waals heterostructures, *Science*, 2016, **353**, 6298.
4 D. Zhong, K. L. Seyler, X. Linpeng, R. Cheng, N. Sivadas, B. Huang, E. Schmidgall, T. Taniguchi, K. Watanabe, M. A. McGuire, W. Yao, D. Xiao, K.-M. C. Fu and X. Xu, Van der Waals engineering of ferromagnetic semiconductor heterostructures for spin and valleytronics, *Sci. Adv.*, 2017, **3**, e1603113.
5 Y. Cao, V. Fatemi, S. Fang, K. Watanabe, T. Taniguchi, E. Kaxiras and P. Jarillo-Herrero, Unconventional superconductivity in magic-angle graphene superlattices, *Nature*, 2018, **556**, 43.
6 R. Niu, Z. Li, X. Han, Z. Qu, D. Ding, Z. Wang, Q. Liu, T. Liu, C. Han, K. Watanabe, T. Taniguchi, M. Wu, Q. Ren, X. Wang, J. Hong, J. Mao, Z. Han, K. Liu, Z. Gan and J. Lu, Giant ferroelectric polarization in a bilayer graphene heterostructure, *Nat. Commun.*, 2022, **13**, 6241.
7 L. Britnell, R. M. Ribeiro, A. Eckmann, R. Jalil, B. D. Belle, A. Mishchenko, Y.-J. Kim, R. V. Gorbachev, T. Georgiou, S. V. Morozov, A. N. Grigorenko, A. K. Geim, C. Casiraghi, A. H. C. Neto and K. S. Novoselov, Strong Light-Matter Interactions in Heterostructures of Atomically Thin Films, *Science*, 2013, **340**, 1311.
8 A. C. Ferrari, F. Bonaccorso, V. Fal'ko, K. S. Novoselov, S. Roche, P. Bøggild, S. Borini, F. H. L. Koppens, V. Palermo, N. Pugno, J. A. Garrido, R. Sordan, A. Bianco, L. Ballerini, M. Prato, E. Lidorikis, J. Kivioja, C. Marinelli, T. Ryhänen, A. Morpurgo, J. N. Coleman, V. Nicolosi, L. Colombo, A. Fert, M. Garcia-Hernandez, A. Bachtold, G. F. Schneider, F. Guinea, C. Dekker, M. Barbone, Z. Sun, C. Galiotis, A. N. Grigorenko, G. Konstantatos, A. Kis, M. Katsnelson, L. Vandersypen, A. Loiseau, V. Morandi, D. Neumaier, E. Treossi, V. Pellegrini, M. Polini, A. Tredicucci, G. M. Williams, B. Hee Hong, J.-H. Ahn, J. Min Kim, H. Zirath, B. J. van Wees, H. van der Zant, L. Occhipinti, A. Di Matteo, I. A. Kinloch, T. Seyller, E. Quesnel, X. Feng, K. Teo, N. Rupesinghe, P. Hakonen, S. R. T. Neil, Q. Tannock, T. Löfwander and J. Kinaret, Science and technology roadmap for graphene, related two-dimensional crystals, and hybrid systems, *Nanoscale*, 2015, **7**, 4598.
9 C. Androulidakis, K. Zhang, M. Robertson and S. Tawfick, Tailoring the mechanical properties of 2D materials and heterostructures, *2D Mater.*, 2018, **5**, 032005.
10 H. H. Fang, B. Han, C. Robert, M. A. Semina, D. Lagarde, E. Courtade, T. Taniguchi, K. Watanabe, T. Amand, B. Urbaszek, M. M. Glazov and X. Marie, Control of the Exciton Radiative Lifetime in van der Waals Heterostructures, *Phys. Rev. Lett.*, 2019, **123**, 067401.
11 D. Tebbe, M. Schütte, K. Watanabe, T. Taniguchi, C. Stampfer, B. Beschoten and L. Waldecker, Tailoring the dielectric screening in $WS_2$–graphene heterostructures, *npj 2D Mater. Appl.*, 2023, **7**, 29.
12 F. Volmer, M. Ersfeld, P. E. Faria Junior, L. Waldecker, B. Parashar, L. Rathmann, S. Dubey, I. Cojocariu, V. Feyer, K. Watanabe, T. Taniguchi, C. M. Schneider, L. Plucinski, C. Stampfer, J. Fabian and B. Beschoten, Twist angle dependent interlayer transfer of valley polarization from excitons to free charge carriers in $WSe_2/MoSe_2$ heterobilayers, *npj 2D Mater. Appl.*, 2023, **7**, 1.
13 F. H. L. Koppens, T. Mueller, P. Avouris, A. C. Ferrari, M. S. Vitiello and M. Polini, Photodetectors based on graphene, other two-dimensional materials and hybrid systems, *Nat. Nanotechnol.*, 2014, **9**, 780.
14 L. Mennel, J. Symonowicz, S. Wachter, D. K. Polyushkin, A. J. Molina-Mendoza and T. Mueller, Ultrafast machine vision with 2D material neural network image sensors, *Nature*, 2020, **579**, 62.
15 J. F. Sierra, J. Fabian, R. K. Kawakami, S. Roche and S. O. Valenzuela, Van der Waals heterostructures for spintronics and opto-spintronics, *Nat. Nanotechnol.*, 2021, **16**, 856.
16 M. C. Lemme, D. Akinwande, C. Huyghebaert and C. Stampfer, 2D materials for future heterogeneous electronics, *Nat. Commun.*, 2022, **13**, 1392.
17 E. Icking, D. Emmerich, K. Watanabe, T. Taniguchi, B. Beschoten, M. C. Lemme, J. Knoch and C. Stampfer, Ultrasteep Slope Cryogenic FETs Based on Bilayer Graphene, *Nano Lett.*, 2024, **24**, 11454.
18 M. Yi and Z. Shen, A review on mechanical exfoliation for the scalable production of graphene, *J. Mater. Chem. A*, 2015, **3**, 11700.
19 F. Pizzocchero, L. Gammelgaard, B. S. Jessen, J. M. Caridad, L. Wang, J. Hone, P. Bøggild and T. J. Booth, The hot pick-up technique for batch assembly of van der Waals heterostructures, *Nat. Commun.*, 2016, **7**, 1.
20 R. Frisenda, E. Navarro-Moratalla, P. Gant, D. P. De Lara, P. Jarillo-Herrero, R. V. Gorbachev and A. Castellanos-Gomez, Recent progress in the assembly of nanodevices and van der Waals heterostructures by deterministic placement of 2D materials, *Chem. Soc. Rev.*, 2018, **47**, 53.
21 J.-L. Uslu, T. Ouaj, D. Tebbe, A. Nekrasov, J. H. Bertram, M. Schütte, K. Watanabe, T. Taniguchi, B. Beschoten, L. Waldecker and C. Stampfer, An open-source robust machine learning platform for real-time detection and classification of 2D material flakes, *Mach. Learn.: Sci. Technol.*, 2024, **5**, 015027.
22 B. Ryu, L. Wang, H. Pu, M. K. Y. Chan and J. Chen, Understanding, discovery, and synthesis of 2D materials enabled by machine learning, *Chem. Soc. Rev.*, 2022, **51**, 1899.
23 S. Masubuchi, M. Morimoto, S. Morikawa, M. Onodera, Y. Asakawa, K. Watanabe, T. Taniguchi and T. Machida, Autonomous robotic searching and assembly of two-dimensional crystals to build van der Waals superlattices, *Nat. Commun.*, 2018, **9**, 1413.
24 X. Lin, Z. Si, W. Fu, J. Yang, S. Guo, Y. Cao, J. Zhang, X. Wang, P. Liu, K. Jiang and W. Zhao, Intelligent identification of two-dimensional nanostructures by machine-learning optical microscopy, *Nano Res.*, 2018, **11**, 6316.
25 Y. Li, Y. Kong, J. Peng, C. Yu, Z. Li, P. Li, Y. Liu, C.-F. Gao and R. Wu, Rapid identification of two-dimensional materials via machine learning assisted optic microscopy, *J. Materiomics*, 2019, **5**, 413.

26 P. Blake, E. W. Hill, A. H. Castro Neto, K. S. Novoselov, D. Jiang, R. Yang, T. J. Booth and A. K. Geim, Making graphene visible, *Appl. Phys. Lett.*, 2007, **91**, 063124.
27 Y. Saito, K. Shin, K. Terayama, S. Desai, M. Onga, Y. Nakagawa, Y. M. Itahashi, Y. Iwasa, M. Yamada and K. Tsuda, Deep-learning-based quality filtering of mechanically exfoliated 2D crystals, *npj Comput. Mater.*, 2019, **5**, 124.
28 B. Han, Y. Lin, Y. Yang, N. Mao, W. Li, H. Wang, K. Yasuda, X. Wang, V. Fatemi, L. Zhou, J. I.-J. Wang, Q. Ma, Y. Cao, D. Rodan-Legrain, Y.-Q. Bie, E. Navarro-Moratalla, D. Klein, D. MacNeill, S. Wu, H. Kitadai, X. Ling, P. Jarillo-Herrero, J. Kong, J. Yin and T. Palacios, Deep-Learning-Enabled Fast Optical Identification and Characterization of 2D Materials, *Adv. Mater.*, 2020, **32**, 2000953.
29 S. Masubuchi, E. Watanabe, Y. Seo, S. Okazaki, T. Sasagawa, K. Watanabe, T. Taniguchi and T. Machida, Deep-learning-based image segmentation integrated with optical microscopy for automatically searching for two-dimensional materials, *npj 2D Mater. Appl.*, 2020, **4**, 3.
30 F. Ramezani, S. Parvez, J. P. Fix, A. Battaglin, S. Whyte, N. J. Borys and B. M. Whitaker, Automatic detection of multilayer hexagonal boron nitride in optical images using deep learning-based computer vision, *Sci. Rep.*, 2023, **13**, 1595.
31 R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. A. Creel, J. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. E. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. F. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. P. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. F. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. H. Roohani, C. Ruiz, J. Ryan, C. R'e, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. P. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. A. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, On the Opportunities and Risks of Foundation Models, *arXiv*, 2021, DOI: **10.48550/arXiv.2108.07258**.
32 T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, Language Models are Few-Shot Learners, in *Advances in Neural Information Processing Systems (NeurIPS)*, Curran Associates, Inc., 2020, vol. 33, pp. 1877–1901.
33 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in *International Conference on Learning Representations (ICLR)*, 2021.
34 B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, Masked-attention Mask Transformer for Universal Image Segmentation, in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1280–1289.
35 J. Mukhoti, A. Kirsch, J. van Amersfoort, P. H. Torr, and Y. Gal, Deep Deterministic Uncertainty: A New Simple Baseline, in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 24384–24394.
36 K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition, in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
37 N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, End-to-End Object Detection with Transformers, in *European Conference on Computer Vision (ECCV)*, Springer International Publishing, Cham, 2020, pp. 213–229.
38 R. Chandra, K. Jain, R. V. Deo and S. Cripps, Langevin-gradient parallel tempering for Bayesian neural learning, *Neurocomputing*, 2019, **359**, 315.
39 T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, Spectral Normalization for Generative Adversarial Networks, in *International Conference on Learning Representations (ICLR)*, 2018.
40 J. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax Weiss, and B. Lakshminarayanan, Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness, in *Advances in Neural Information Processing Systems (NeurIPS)*, Curran Associates, Inc., 2020, vol. 33, pp. 7498–7512.
41 Y. LeCun, Y. Bengio and G. E. Hinton, Deep learning, *Nature*, 2015, **521**, 436.
42 F. Bolelli, S. Allegretti, L. Baraldi and C. Grana, Spaghetti Labeling: Directed Acyclic Graphs for Block-Based Connected Components Labeling, *IEEE Trans. Image Process.*, 2020, **29**, 1999.
43 I. Loshchilov and F. Hutter, Decoupled Weight Decay Regularization, in *International Conference on Learning Representations (ICLR)*, 2019.
44 M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, AAAI Press, 1996, pp. 226–231.
45 D. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, in *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.
46 The GitHub repository for MaskTerial, **https://github.com/Jaluus/MaskTerial**.