E^2SAM: A Pipeline for Efficiently Extending SAM's Capability on Cross-Modality Data via Knowledge Inheritance

Sundingkai Su
susundingkai@bupt.edu.cn
Mengqiu Xu
xumengqiu@bupt.edu.cn
Kaixin Chen
chenkaixin@bupt.edu.cn
Ming Wu✉
wuming@bupt.edu.cn
Chuang Zhang
zhangchuang@bupt.edu.cn

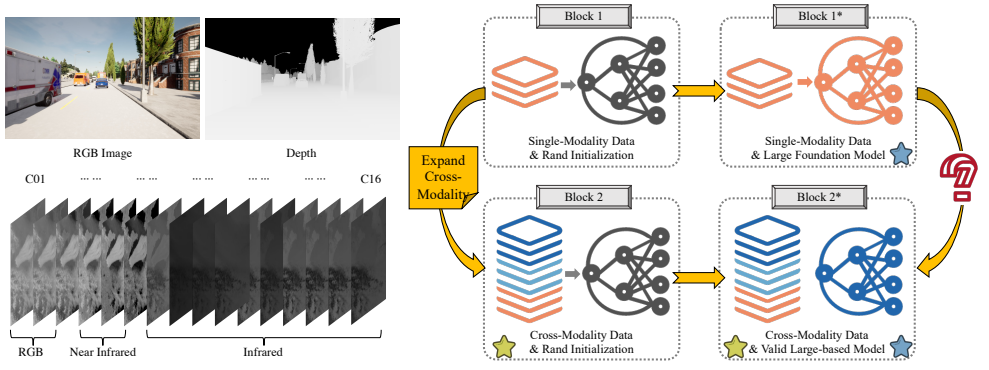
School of Artificial Intelligence,
Beijing University of Posts and
Telecommunications (BUPT),
100876, Beijing, China

Abstract

Segment Anything Model (SAM) has achieved brilliant results on many segmentation datasets due to its strong segmentation capability with visual-grouping perception. However, the limitation of the three-channel input means that it is difficult to apply directly to cross-modality data. Therefore, this paper proposes a pipeline called E^2SAM with the knowledge inheritance stage and downstream fine-tuning stage step by step that can efficiently inherit the capabilities of SAM and extend to both cross-modality data and relevant task-specific application. In order to enable the feature alignment of varying single-modality to cross-modality data, an auxiliary branch with a channel selector and a merge module is designed in the first stage. It is worth noting that we do not need a large amount of additional annotated training data during our pipeline. Furthermore, the strengths of the proposed method are discussed in detail through experiments on generalization performance and resistance to size changes. The experimental results and visualizations on the SFDD-H8 and SHIFT datasets demonstrate the effectiveness of our proposed methods compared to other methods such as random initialization and SAM-based fine-tuning. The code is available at https://github.com/BUPT-PRIS-727/BMVC2023_E2SAM.

1 Introduction

In recent decades, with the continuous development of sensor technology, humans can obtain more and more abundant data through various types of sensors, such as multi-spectral or hyper-spectral data in remote sensing area [9, 28], Lidar or depth data in autonomous driving [14] and CT or X-ray data in medical diagnosis [4], that some samples are as shown



(a) Samples of Cross-Modality Data **(b) Insights of Our Proposed Method**

Figure 1: (a) Some samples of cross-modality data from SHIFT and SFDD-H8 dataset, including depth, near infrared and infrared image other than RGB image. (b) Insights of our proposed method. The arrow from Block 1 to 2 indicates cross-modality data expansion because this ability marked as a yellow star brings better performance under the random initialization setting. Besides, the arrow from Block 1 to 1* indicates the performance gain of foundation models under the same data marked as a blue star. Therefore, the goal of this paper is to utilize both advantages from cross-modality data and large foundation model.

in Fig. 1 (a). These cross-modality data, which are generally considered to be meaningful for data-driven deep learning algorithms, can help algorithms to better recognise and understand scenes, and are therefore of great importance for people to better perceive the world as shown as the arrow from Block 1 to Block 2 in Fig. 1 (b).

Nowadays, the emergence of large foundation models [11, 16, 18] has made the field of artificial intelligence flourish rapidly. Among them, Segment Anything Model (SAM) [11] proposed in early April 2023 is more of a milestone with strong segmentation performance with good visual-grouping ability and shape sensitivity. Some works [3, 10, 27] have evaluated SAM on different tasks and fine-tuned with better performance on most diverse datasets. However, due to the limitation of three-channel input and the absence of meteorological remote sensing data in SA-1B, SAM cannot be directly applied to cross modality data with task-specific application as indicated by the arrow from Block 1* to Block 2* in Fig. 1 (b).

There are two approaches to resolve this dilemma. (1) One approach is to realize the adaptation of the model size. For example, it can be achieved by replacing the patch-embedding structure. However, the performance gain is lower because some of the SAM knowledge is discarded. (2) Another approach is to allow cross-modality data as the input of SAM. One naive way is to resample the input dimension through a convolutional module with the cost of compressing data. Moreover, different three-channel-combined data can be regarded as the subset of cross-modality data, and then can be sent to the model one by one in order to integrate at the feature level. However, this method is time-consuming and laborious, and cannot take into account the correlation between different modality data.

Based on the issues mentioned above, our goal in this paper is to obtain a model that can not only allow the cross-modality as input under the premise of not compressing data, but also inherit the feature extraction capabilities of existing large-scale models like SAM for efficient and high-fidelity feature learning, that those are marked as yellow and blue stars in

Fig. 1 (b), respectively. Therefore, we introduce the Knowledge Distillation (KD) spirit [5] to achieve knowledge inheritance as the first stage during the process of encoder initialization weight acquisition. However, most of the current knowledge distillation methods [17, 20, 23] realize the transfer from the teacher model to the student model under the condition of the same modality, thus it is difficult to realize the registration from a single modality to a cross modality. Immediately after, an Auxiliary Branch is presented for feature alignment including the Channel Selector and Merge Module to separate and then integrate features extracted from different modality. In order to obtain better performance, we still adopt the method of fine-tuning on downstream tasks as the second stage, based on the distilled student encoder and randomly initialized decoder. Moreover, we discuss the strengths of the proposed method through experiments on generalization performance and resistance to size changes.

The main contributions of this paper are summarised as follows:

(1) We propose a pipeline named E^2SAM for Efficiently Extending SAM's capability on cross-modality data through knowledge inheritance stage and fine-tuning stage of downstream step by step. It simultaneously allows multiple modalities data as input and has the visual grouping ability of inherited from foundation model such as SAM.

(2) In the knowledge inheritance stage, there is no need to introduce additional data and any annotations since the model can be trained unsupervised. Besides, an Auxiliary Branch with a Channel Selector and a Merge Module is designed in this stage to enable the alignment of efficient features between different modality data.

(3) The experimental results and visualizations on several datasets [7, 21] demonstrate the effectiveness of our proposed method compared to other methods including random initialization and SAM-based fine-tuning.

2 Related Work

Applications for Cross-Modality data. Cross-modality data can be seen everywhere in real-world scenarios [12]. Taking meteorological scenes as an example, the sensors carried by meteorological satellites [8, 9] can simultaneously obtain cross modality data collected varying from visible bands to near-infrared bands, that which are used to comprehensively monitor and forecast meteorological elements. Thus the most obvious and important effort in cross-modality data is valuable and meaningful. In addition, a large number of experiments [7, 13, 22] can prove that the comprehensive use of data from multiple modalities has superior performance compared to using only a single modality. Thus, in this paper, we hope to allow cross-modality data as the input of large foundation model to improve performance.

Segment Anything Model. The goal of SAM [11] is to establish a large foundation model for image segmentation based on diverse prompt, that its structure consists of an Image Encoder, a Prompt Encoder and a Mask Decoder. A lot of works [10, 19, 31] have evaluated the performance of SAM on multiple modality data through operations such as direct prediction and normalized mask selection. For example, this paper [19] selects 6 commonly remote sensing datasets applied into two applications to measure SAM's ability, and proves that there is still a huge promote for SAM to extend more data types and applications. Although there are also a lot of works based on SAM for secondary development, including combining semantics correlation [31], caption [25], inpainting or generation [30] tasks, aiming to cross-modality data, the methods [7] mainly focus on how to take advantage of the existing SAM model to construct multi-modal datasets, without realizing the integration and expansion of SAM capabilities. Thus, in this paper we want to achieve extending SAM's

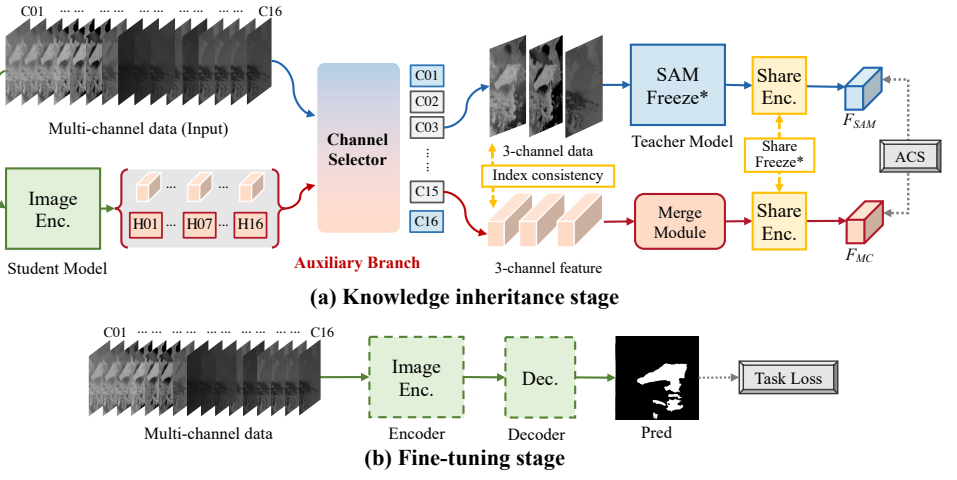


Figure 2: A brief introduction of our proposed method. (a) Knowledge inheritance stage based on a Teacher Model, a Student Model and an Auxiliary Branch marked in blue, green and red color, respectively. It is feasible to achieve alignment through Alignment Constraint selection (ACS) between feature extracted from different modality data. (b) Fine-tuning stage on downstream application through fine-tuning based on distilled Student Model and decoder under Task Loss constraints.

capability via knowledge inheritance on cross-modality.

Knowledge Distillation. Knowledge distillation is first proposed by Hinton *et al.* [6] that these works can transfer knowledge or representations learned by a large teacher model into a smaller student one. On this basis, a large number of response-based [17], feature-based [20], and relation-based [23] knowledge distillation methods have been gradually developed for various applications such as semantic segmentation [6, 29], person ReID [15], medical CT synthesis [4] and so on. However, most current distillation methods focus on feature alignment or global constraints between the same modalities but different models, without considering how to separate and then merge features between different modalities. Thus in our proposed method, an Auxiliary Branch is designed to align representations from different cross-modality data.

3 Methods

In this paper, a pipeline named E^2SAM with knowledge inheritance stage and fine-tuning stage is proposed for efficiently extending SAM's capability on cross-modality data. A brief introduction of our proposed method is shown in Fig. 2.

3.1 Knowledge inheritance stage based on distillation

The core of knowledge inheritance stage is to design multiple branches at the end of the encoding module to assist in training, enabling the model to have the ability inherited from SAM to extract features from more than three channels. During this stage, a teacher model

and a student model are the cornerstones used to represent cross-modality data efficiently. In order to solve feature alignment between the feature F_{SAM} extracted by SAM using single modality data and the feature F_{MC} extracted by student model based on cross modality data, an auxiliary branch is designed in this stage to separate different modality data according to the index output of the Channel Selector and then fuse them in the Merge Module.

To accomplish this task, we pretrain the multi-channel model with the same number of auxiliary branch as the number of channels. The input to the teacher model is random selection of three channels from three-channel data, while the input to the student model is the entire channel data. The features from the teacher model correspond to the outputs of the three auxiliary branch of the student model, and we combine the outputs of these three branches as a three-channel feature. During optimisation, this synthesised feature is used to calculate the similarity to the output features of the teacher model.

3.1.1 Auxiliary branch

Merge Module. In order to transfer capacity from RGB-pretrained model to multi-channels input network, we divide the feature map of student encoder output M_t into c branches which is same size as the teacher network output $M_s \in \mathbb{R}^{B\times H\times W\times D}$, and then combine them into a single tensor through two Fully-Connection (FC) layers.

$$M_{branch} = M_s \cdot W_1 \ , \tag{1}$$

$$M_t = Norm1(Concat(M^i_{branch}, M^j_{branch}, M^k_{branch})W_2) \tag{2}$$

where the fully-connection layer are learnable parameter matrix $W_1 \in \mathbb{R}^{D\times cd_k}$, $W_2 \in \mathbb{R}^{3d_k\times D}$, $M^i_{branch}, M^j_{branch}, M^k_{branch} \in \mathbb{R}^{B\times H\times W\times d_k}$ where i, j, k is the index selected by the Channel Selector. In this work we set $d_k = D/3$, each $M^i_{branch} i.e.$ is independent and corresponds to a specific channel, and we utilise these branches to network optimization to improve the overall feature extraction ability.

Share Encoder. Once we get M_s and M_t, we can begin calculating the loss and optimizing the network. In our work, we use the convolutional layer after the transformer blocks of the teacher model to constrain the features. The output features of the student model and the teacher model are jointly passed through this shared layer and the resulting tensors are used to calculate the loss. The input of the network's computed loss function is as follows:

$$M'_t = Norm2(Conv1(M_t)) \ , \tag{3}$$

$$M'_s = Norm2(Conv1(M_s)) \ , \tag{4}$$

$$Loss = Criteron(M'_s, M'_t) \ , \tag{5}$$

where the parameters of the convolutional layer $Conv1$ are copied from the teacher model and participate in back-propagation but do not update the parameters.

3.1.2 Alignment constraint selection

Our proposed pipeline E^2SAM aims to align the features M'_s extracted from each channel based on SAM with the features M'_t extracted from the student model. In practice, the alignment constraint between two features can be reduced using Mean Squared Error (MSE) or Maximum Mean Discrepancy (MMD) [1] namely.

MSE calculates the distance between two features in numerical terms:

$$\mathcal{L}_{mse} = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2 \ , \tag{6}$$

where y and y' are an element in M_s' and M_t', $n = H \times W \times D$.

Besides, MMD calculates the difference between the two features in the division, where $k(*)$ is a Gaussian kernel,

$$\mathcal{L}_{mmd} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(y_i, y_j) - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(y_i, y_j') + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(y_i', y_j') \ . \tag{7}$$

3.2 Fine-tuning stage

To make the model usable in downstream tasks, a suitable decoding module must be designed or selected to complete the fine-tuning of the downstream tasks after the model's encoding module has been learned. If the size of the dataset is too far from the input of the model, it will be a great waste of computational resources. For the transformer model, the different size of the image input only requires changing the position embedding layer to match the input of the network. This also means that our method can be adapted to different tasks and datasets of different sizes, and the number of model parameters can be changed flexibly with different tasks to save computational resources.

4 Experiments

4.1 Experimental Setups

Datasets and evaluation metrics. The SFDD-H8 [7, 26] dataset for sea fog detection task is a cross-modality data collected from Himiwari-8/9 meteorological satellite, that it includes Visibility (VIS), Near-Infrared (NIR) and Infrared (IR) bands. The training and test dataset contain 1128 and 680 samples at a resolution of 1024×1024, respectively. Since sea fog areas need to be paid more attention in the image, Critical Success Index (CSI), the mean Intersection-over-Union (mIoU), mean Accuracy (mAcc) and all Accuracy (allAcc) are selected as the evaluation metrics according to existed works [7, 32].

In order to evaluate the generalization and robustness of our proposed pipeline, the cross-modality dataset SHIFT [21] with domain discrepancy is considered since it is the largest synthetic dataset for autonomous driving and provides the most inclusive set of annotations and conditions. We selected 3040, 1468, 1492 samples and 1268, 596, 640 samples from the three sub-datasets of Daytime, Dawn/dusk, and Night conditions as the training set and test set according to the proportion of the original dataset. In SHIFT dataset, the mean Intersection-over-Union (mIoU), mean Accuracy (mAcc) and all Accuracy (allAcc) are used as main metrics in our experiments.

Implementation Details. All experiments are based on ViT-base model structure with patch size 16 and learning rate of $1e^{-4}$. The models at the first stage were trained for 50 epochs using batch size as 1 on 2 NVIDIA RTX 3090s while the models at the second stage were fine-tuning 100 epochs using batch size as 2 on 1 NVIDIA RTX 3090.

Data		Model	CSI	mIou	mAcc	allAcc
SFDD-H8	C3*-1024	Rand	34.66	65.72	72.63	96.83
		SAM	57.76 (+23.10)	77.88 (+12.16)	86.27 (+13.64)	98.06 (+1.23)
	C16-1024	Rand	52.18	74.86	85.40	74.86
		SAM-ReSam	59.78 (+7.60)	78.88 (+4.02)	**89.69 (+3.42)**	98.04 (+0.43)
		SAM-RePE	56.96 (+4.15)	77.46 (+2.78)	85.79 (+0.39)	98.02 (+0.41)
		Ours	**61.57 (+9.39)**	**79.91 (+5.05)**	87.52 (+2.21)	**98.29 (+0.66)**
	C3*-512	SAM	**60.26**	78.81	**90.06**	98.00
	C16-512	Ours	59.92	**79.01**	87.78	**98.15**
SHIFT	C3-1024	Rand	–	63.72	73.92	94.85
	C3-1024	SAM	–	73.15 (+9.43)	82.31 (+8.39)	96.56 (+1.71)
	C4-1024	Rand	–	78.17	86.44	98.04
	C4-1024	Ours	–	**81.78 (+3.62)**	**89.25 (+2.81)**	**98.59 (+0.55)**

Table 1: The experimental results of our pipeline and other methods based on SFDD-H8 and SHIFT dataset, where C* indicates that we have taken a variety of three-channel and displayed the average values of all evaluation metrics.

4.2 Quantitative and Qualitative Results

To demonstrate the effectiveness of our proposed method, we design two sets of controlled experiments by comparing the input of different modalities and the advantages of inheriting the powerful capabilities of SAM. The quantitative results are as shown in Table 1, where 'Rand' indicates random initialization and 'SAM-' means performing different operations based on SAM [□] as the pre-trained model.

(1) To compare the effects of different modality as inputs, experiments are designed varying the combination of different three-channel data from multi-modal data with the same random initialization operation, as shown in the Line 'C3*-1024 Rand' and 'C16-1024 Rand' in Table 1. It can be seen that due to the introduction of more modality data, the sea fog recognition task produces a significant performance gain of 17.52 CSI points under the same experimental conditions of model training.

(2) To demonstrate the effectiveness of our method for extending the capabilities of SAMs, we first design a set of experiments by using the SAM model as initial weights on three-channel data compared with random initialization, as shown in the 2rd and 3th row of the Table 1. The experimental results prove that the powerful segmentation ability of SAM is valuable for inheritance and extension.

In addition, we design two comparative experiments on cross-modality data, that which are SAM-ReSam through resampling the input dimension using a convolutional module and SAM-RePE by replacing the Patch-Embedding structure, respectively. Specifically, SAM-ReSam uses a trainable convolutional module to downscale multi-channel data to three channels. SAM-RePE loads the SAM's pretrained weights but modifies the first layer to adjust multi-channel inputs. Compared these aboved methods, our proposed pipeline achieves the best performance and is illustrated the effectiveness of our method.

The qualitative results through visualization of RGB images, Ground Truth (GT) and different methods' predictions based on SFDD-H8 dataset as shown in Fig 3. Compared with the random initialization methods, the SAM-based models have better recognition performance for fog areas due to the more fitting edges and significantly reduces misjudgment areas. The qualitative results of the SHIFT dataset will be found in the Supplementary I.

Besides, for the performance of the knowledge inheritance stage, we select the distillation

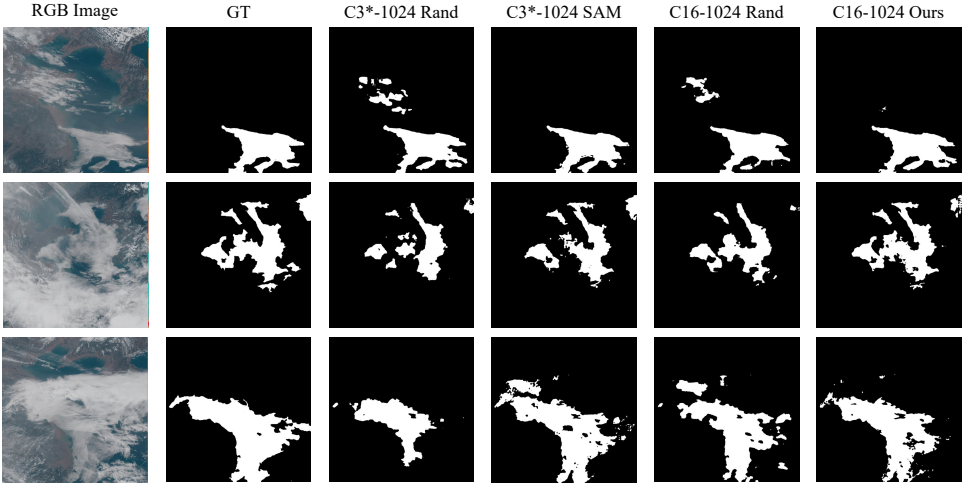


Figure 3: The qualitative results through visualization of RGB images, Ground Truth (GT) and different methods' predictions based on SFDD-H8 dataset.

model every 20 epochs and use the prediction results of fine-tuning on downstream task to evaluate the inheritance quality, which is shown in the Supplementary II. We can find that fine-tuning performance on downstream tasks is better as the knowledge inherited from SAM is more complete.

In practice, the alignment constraint in the knowledge inheritance stage can be chosen according to the dataset. MSE is suitable for larger datasets or datasets with many unlabelled data. It allows the model to converge more quickly. Conversely, MMD is more appropriate when the data is small or there is no unlabelled data. In our work, we have found that MSE is more responsive to the learning progress of the distillation task than MMD, and the lower the MSE, the better the performance in downstream tasks. The details of different alignment constraint selection experiments are supplied in the Supplementary III.

Moreover, we attempt to reduce the size of the cross-modality data input to the Student Model in knowledge distillation stage, in order to extend the application to more abundant datasets and tasks. We have performed the above experiments on the SFDD-H8 dataset with the resolution as 512x512 px. It can be found that although the recognition performance of the model is slightly reduced, the performance of our method is basically the same as that of directly using SAM fine-tuning. Thus, inheriting the segmentation ability of SAM can effectively resist the impact of size changes.

In addition, we visualize the ratio of different channel selection on SFDD-H8 dataset during training process as shown in Fig 4 (a), where the abscissa represents the average probability of being selected and the ordinate represents the number of iterations. Although the selection ratio of different channels fluctuates to a certain extent in the early stage of training, the probability of selection of different channels is basically equal throughout the training process, and they are all maintained around the value of 1/16.

Model		All	Daytime	Dawn/dusk	Night
All-C4	Rand	78.17	78.00	79.41	77.20
All-C4	Ours	81.78 (+3.62)	81.58 (+3.58)	82.51 (+3.10)	81.12 (+3.92)
Daytime-C4	Rand	76.03	73.50	63.62	55.05
Daytime-C4	Ours	76.57 (+0.54)	77.86 (+4.18)	77.52 (+13.90)	73.50 (+18.45)
Dawn/dusk-C4	Rand	61.82	60.25	63.78	62.88
Dawn/dusk-C4	Ours	73.58 (+11.76)	73.01 (+12.76)	74.24 (+10.46)	73.61 (+10.73)
Night-C4	Rand	52.88	43.63	59.36	64.58
Night-C4	Ours	72.79 (+19.91)	71.49 (+27.86)	73.01 (+13.38)	74.21 (+9.63)

Table 2: Generalization performance experimental results based on SHIFT dataset. The text in blue color indicate the results directly test on other sub-dataset.

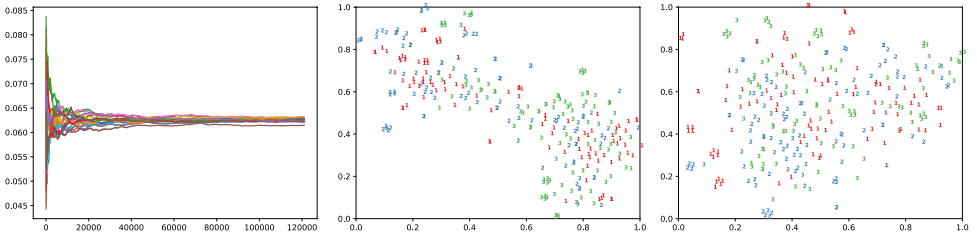


Figure 4: (a) The ratio of different channel selected on SFDD-H8 dataset during training process, that which is shown the probability of different channels being selected is nearly equal. (b) t-SNE visualizations for the aligned feature F_{stu} from rand initialization model and our pipeline, where red, blue and green numbers represent daytime, dawn/dusk and night samples, respectively.

4.3 Experiments for generalization ability

In order to explore the advantages of the ability based on SAM extending, we attempt to evaluate the model generalization performance on the SHIFT dataset with domain differences constructed for unsupervised domain adaptation task, mainly selecting three sub-datasets: daytime, dawn/dusk, and night. We leverage both models with random initialization and our proposed pipeline, trained on a single sub-dataset and directly tested on other datasets. There are the results of experiments in Table 2. The experimental results show that compared with the random initialization method, we can better adapt to the data of various distributions through the knowledge distillation of SAM.

Besides, the t-SNE [24] visualizations for the aligned feature F_{stu} between rand initialization model and our proposed pipeline are as shown in Fig 4 (b1) and (b2). The red points represent daytime data in SHIFT while the blue points represent night samples. We can find that the distribution of data points of different colors in b2 is more confusing than that in b1. Thus, we can draw a conclusion the model from our proposed pipeline has better generalization performance in dealing with data with domain differences.

5 Conclusion

In this paper, in order to realize the capability expansion of SAM based on cross-modality data, we propose a universal two-stages pipeline, which is the knowledge inheritance stage for inheriting SAM capability, and the fine-tuning stage for better downstream adaptation. An auxiliary branch including a Channel Selector and Merge Module is designed to separate different cross-modality to achieve feature alignment. It is worth mentioning that we do not need to lead into additional data and labels during the knowledge inheritance process, reducing the cost of collecting and annotating data. Through meticulous experiments and visualization results, it can be demonstrated that our method can efficiently inherit the ability of SAM on cross-modality data without compressing data.

In future work, our method also can be applied not only cross-modality data, but also various computer vision tasks that which need fusion of diverse data, such as change detection, temporal optical flow and other tasks.

6 Acknowledgement

This work was supported by National Natural Science Foundation of China (NSFC), under the Grant number 62076093.

References

[1] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[2] Junzhang Chen and Xiangzhi Bai. Learning to" segment anything" in thermal infrared images through knowledge distillation with a large scale dataset satir. *arXiv preprint arXiv:2304.07969*, 2023.

[3] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. *arXiv preprint arXiv:2304.09148*, 2023.

[4] Chaowei Fang, Liang Wang, Dingwen Zhang, Jun Xu, Yixuan Yuan, and Junwei Han. Incremental cross-view mutual distillation for self-supervised medical ct synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20677–20686, 2022.

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[6] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8479–8488, June 2022.

[7] Bin Huang, Ming Wu, Shuyue Sun, Wei Zhao, Zhanbei Cui, and Cheng Lv. Sea fog monitoring method based on deep learning satellite multi-channel image fusion (in chinese). *Meteorological Science and Technology*, 49(6):823–829, 2021.

[8] Bin Huang, Luming Xiao, Wen Feng, Mengqiu Xu, Ming Wu, and Xiang Fang. Domain adaptation on multiple cloud recognition from different types of meteorological satellite. *Frontiers in Earth Science*, page 1132, 2022.

[9] Yixiang Huang, Ming Wu, Jun Guo, Chuang Zhang, and Mengqiu Xu. A correlation context-driven method for sea fog detection in meteorological satellite imagery. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[10] Wei Ji, Jingjing Li, Qi Bi, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *arXiv preprint arXiv:2304.05750*, 2023.

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[12] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[13] Sumin Lee, Sangmin Woo, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Modality mixer for multi-modal action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3298–3307, 2023.

[14] Haisong Liu, Tao Lu, Yihui Xu, Jia Liu, Wenjie Li, and Lijun Chen. Camliflow: bidirectional camera-lidar fusion for joint optical flow and scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5801, 2022.

[15] Yichen Lu, Mei Wang, and Weihong Deng. Augmented geometric distillation for data-free incremental person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7329–7338, June 2022.

[16] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[17] Mary Phuong and Christoph H Lampert. Distillation-based training for multi-exit architectures. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1355–1364, 2019.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[19] Simiao Ren, Francesco Luzi, Saad Lahrichi, Kaleb Kassaw, Leslie M Collins, Kyle Bradbury, and Jordan M Malof. Segment anything, from space? *arXiv preprint arXiv:2304.13000*, 2023.

[20] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[21] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21371–21382, 2022.

[22] Paolo Testolina, Francesco Barbato, Umberto Michieli, Marco Giordani, Pietro Zanuttigh, and Michele Zorzi. Selma: Semantic large-scale multimodal acquisitions in variable weather, daytime and viewpoints. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[23] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019.

[24] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[25] Teng Wang, Jinrui Zhang, Junjie Fei, Yixiao Ge, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, Shanshan Zhao, Ying Shan, et al. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*, 2023.

[26] Zhaoqing Wang, Xiangyu Kong, Zhanbei Cui, Ming Wu, Chuang Zhang, MingMing Gong, and Tongliang Liu. Vecnet: A spectral and multi-scale spatial fusion deep network for pixel-level cloud type classification in himawari-8 imagery. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 4083–4086. IEEE, 2021.

[27] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[28] Sibo Wu, Mengqiu Xu, Ming Wu, Chuang Zhang, and Hua Shen. Identify, guess and reconstruct: Three principles for cloud removal task. In *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2022.

[29] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12319–12328, June 2022.

[30] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.

[31] Jielu Zhang, Zhongliang Zhou, Gengchen Mai, Lan Mu, Mengxuan Hu, and Sheng Li. Text2seg: Remote sensing image semantic segmentation via text-guided visual foundation models. *arXiv preprint arXiv:2304.10597*, 2023.

[32] Yuan Zhou, Keran Chen, and Xiaofeng Li. Dual-branch neural network for sea fog detection in geostationary ocean color imager. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022.