CEA Comité Español de Automática
XLV JORNADAS AUTOMÁTICA MÁLAGA 2024
UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Segmentación semántica bajo paradigma one-shot learning utilizando SAM y CP-CVV

Duque-Domingo, Jaime.[a,*], Gómez-García-Bermejo, Jaime.[a,b], Zalama, Eduardo.[a,b], Gómez-Ramos, Raúl.[b], Finzi, Alberto.[c]

[a] *ITAP-DISA, Universidad de Valladolid, 47002 Valladolid, España.*
[b] *Centro Tecnológico CARTIF, Boecillo, 47151 Valladolid, España.*
[c] *PRISMA Lab. Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione. Università degli Studi di Napoli "Federico II", Napoli, Italia*

To cite this article: Duque-Domingo, J., Gómez-García-Bermejo, J., Zalama, E., Gómez-Ramos, R., Finzi, A. 2024. Segmentación semántica bajo paradigma one-shot learning utilizando SAM y CP-CVV. Jornadas de Automática, 45. https://doi.org/10.17979/ja-cea.2024.45.10772

Resumen

La detección y segmentación de objetos en escenas complejas se suele llevar a cabo mediante el entrenamiento de modelos de detección y/o segmentación que requieren el etiquetado manual de cientos de imágenes por categoría. Tanto el proceso de etiquetado como el del entrenamiento pueden llegar a ser costosos tanto computacionalmente como a nivel de esfuerzo humano. Las técnicas de segmentación genérica mediante *zero-shot learning* abren la posibilidad a segmentar objetos nunca antes vistos. Sin embargo, estas técnicas no son semánticas y no nos permiten identificar el objeto que se está segmentando. Nosotros proponemos el uso de un método integrado de segmentación genérica y CP-CVV (Class Partitioning and Cross Validation Voting) para detectar y segmentar objetos a partir de una única muestra. Esta técnica permite crear un abanico de posibilidades donde se busca un aprendizaje rápido e incremental y sólo tenemos acceso a una o a un reducido número de imágenes del objeto que deseamos localizar.

Palabras clave: percepción y detección, aprendizaje incremental, one-shot learning, segmentación, integración de modelos

Semantic segmentation under one-shot learning paradigm using SAM and CP-CVV

Abstract

Detection and segmentation of objects in complex scenes is often performed by training detection and/or segmentation models that require manual labelling of hundreds of images per category. Both the labelling and training process can be computationally and human-effort intensive. Generic segmentation techniques using *zero-shot learning* open up the possibility of segmenting previously unseen objects. However, these techniques are not semantic and do not allow us to identify the object being segmented. We propose the use of an integrated method of generic segmentation and CP-CVV (Class Partitioning and Cross Validation Voting) to detect and segment objects from a single sample. This technique allows us to create a range of possibilities where we are looking for fast and incremental learning and we only have access to one or a small number of images of the object we want to locate.

Keywords: perception and sensing, incremental learning, one-shot learning, segmentation, ensemble models

1. Introducción

Cuando queremos detectar objetos en una imagen solemos recurrir al uso de modelos de detección como YOLO (Redmon et al., 2016) o SSD (Liu et al., 2016). Estos modelos suelen requerir cientos de imágenes por categoría durante el entrenamiento para realizar una detección correcta. Por otro lado, cuando nuestro objetivo es la segmentación de objetos solemos utilizar modelos del tipo de Mask R-CNN (He et al., 2017) o diferentes tipos de redes U-NET (Siddique et al.,

*Autor para correspondencia: jaime.duque@uva.es

2021). Nuevamente, para un entrenamiento correcto se suelen necesitar cientos de imágenes. El etiquetado del *ground-truth* de los modelos de segmentación es además más costoso que los de detección ya que tenemos que definir los puntos del borde geométrico que permite delimitar el objeto. En los modelos de detección basta con definir los cuatro puntos del *bounding box*.

Podemos encontrar técnicas capaces de realizar operaciones como la detección, clasificación o segmentación de objetos basándose en paradigmas como *few-shot learning* (FSL) y, más restrictivamente, *one-shot learning* (OSL). En FSL disponemos de varias imágenes para detectar o clasificar una nueva mientras que en OSL disponemos de una única imagen por categoría. Existen diversos enfoques de OSL, como puede ser el aprendizaje inductivo o transductivo, en función de si el modelo no ha tenido o sí acceso a imágenes de evaluación sin etiquetado durante su entrenamiento.

El problema de la segmentación semántica dentro del paradigma del OSL ha sido uno de los que mayor interés ha tenido en los últimos años (Shaban et al., 2017; Luddecke and Ecker, 2021). La investigación se ha desarrollado utilizando diferentes métodos, como son la creación de clases prototipo (Chen et al., 2021), técnica también utilizada en FSL (Wang et al., 2019; Liu et al., 2020). Otros autores han utilizado redes de grafos piramidales con mecanismos de atención en regiones (Zhang et al., 2019), o la extracción de características ricas del objeto desde tres perspectivas (Zhang et al., 2021): 1) incrustación global para captar las características generales; 2) incrustación de picos para captar la información más discriminativa; 3) incrustación adaptativa para captar las dependencias internas de largo alcance. Finalmente, en SG-One (Zhang et al., 2020) han utilizado redes profundas con estrategias de *pooling* para producir las características que guían la segmentación y comparaciones mediante similaridad del coseno. Todos estos métodos han basado su funcionamiento en modelos completos. Nosotros planteamos partir de la segmentación genérica para conseguir la segmentación semántica mediante un potente sistema de comparación de imágenes como es CP-CVV (Class Partitioning and Cross Validation Voting) (Duque-Domingo et al., 2023). Este es un modelo inductivo que no utiliza imágenes de evaluación durante el entrenamiento. El modelo se implementa mediante un conjunto de redes siamesas que se integran e infieren si dos imágenes pertenecen a una misma categoría. En el artículo de presentación de CP-CVV el modelo se utilizó para clasificación de imágenes en un paradigma OSL.

Por otra parte, en los últimos años se ha producido una revolución con los modelos de segmentación genérica, como SAM (Kirillov et al., 2023). Este modelo trabaja bajo el paradigma del *zero-shot learning* (ZSL), por lo que es capaz de segmentar objetos incluso cuando sean desconocidos para él. Esto se consigue gracias a un costoso entrenamiento que se ha desarrollado con un gran número de imágenes de muchos tipos etiquetadas previamente. Uno de los problemas de este modelo es que no segmenta semánticamente asignando cada región a una clase concreta, por lo que no sabemos a qué se corresponden las regiones delimitadas.

En nuestro método, la conexión SAM CP-CVV la llevamos a cabo en dos etapas. Durante una primera etapa se realiza un proceso de segmentación de todas las posibles regiones que aparecen en una imagen, tanto si son regiones generales como subregiones. En una segunda etapa, para cada una de las máscaras se realizan comparaciones utilizando la técnica CP-CVV con una imagen de referencia después de su alineamiento. A continuación, se presenta el desarrollo de este método y los experimentos desarrollados.

2. Análisis del método

La Figura 1 muestra nuestro sistema durante el proceso de inferencia. Una imagen de entrada y una imagen de referencia de un objeto son introducidas en el *pipeline*. La imagen de referencia es un objeto segmentado que queremos buscar en la imagen de entrada, mientras que esta última es una escena compleja donde el objeto puede aparecer bajo distinta perspectiva e incluso con oclusiones. Mediante SAM extraemos todas las posibles máscaras que existen en la imagen de entrada, algunas de las cuales pueden ser subregiones de otras regiones más grandes. Después de un preprocesamiento previo de las máscaras y de la imagen de referencia que explicaremos a continuación, se realizan las comparativas mediante CP-CVV que permiten decidir cuál de las máscaras es más parecida al objeto de referencia.

Para mantener una comparación ideal en el modelo CP-CVV, tanto la imagen de referencia como las máscaras obtenidas mediante SAM son preprocesadas. Como se muestra en la Figura 2, cada una de las máscaras es alineada respecto a su eje de inercia. Esto hace que la entrada en la comparación siamesa sea lo más parecida posible. A continuación, las imágenes son recortadas utilizando su *bounding box* y son redimensionadas a un tamaño fijo de entrada del modelo CP-CVV. Además, algunos objetos simétricos pueden quedar rotados en sentido contrario al utilizar su eje de inercia. Por dicho motivo y para facilitar la comparación de la siamesa, se realizan dos comparaciones por máscara, una directa: máscara de referencia/máscara posible; y otra aplicando un *flip horizontal*: máscara de referencia/máscara posible con *flip horizontal*.

2.1. Comparación mediante CP-CVV

El modelo CP-CVV incorpora k redes neuronales siamesas mediante un mecanismo de votación suave/duro (ver Figura 3). A diferencia de la metodología de entrenamiento de las redes siamesas típicas, implica el entrenamiento de cada una de las k redes independientes por separado, utilizando conjuntos distintos de clases de entrenamiento y validación. En CP-CVV, los conjuntos de validación se forman para cada uno de los k slots distribuyendo las n clases entre dichos grupos. Antes de asignar los slots de validación, se aleatoriza el orden de las clases. Este método asegura que las clases potencialmente relacionadas no se agrupen en un mismo entrenamiento. Durante la inferencia, el modelo recibe dos imágenes, lo que significa un par positivo si las imágenes pertenecen a la misma clase y negativo en caso contrario. El par de imágenes se introduce en cada una de las k redes siamesas.

Las redes siamesas que se han utilizado emplean un *backbone* convolucional de tipo ConvNeXt-small (Liu et al., 2022), aunque se pueden utilizar otros tipos. El vector de características de salida de cada backbone se somete a una multiplicación por elementos. Posteriormente, se incorporan tres capas densas, que incluyen *dropout* y normalización. La salida de la red

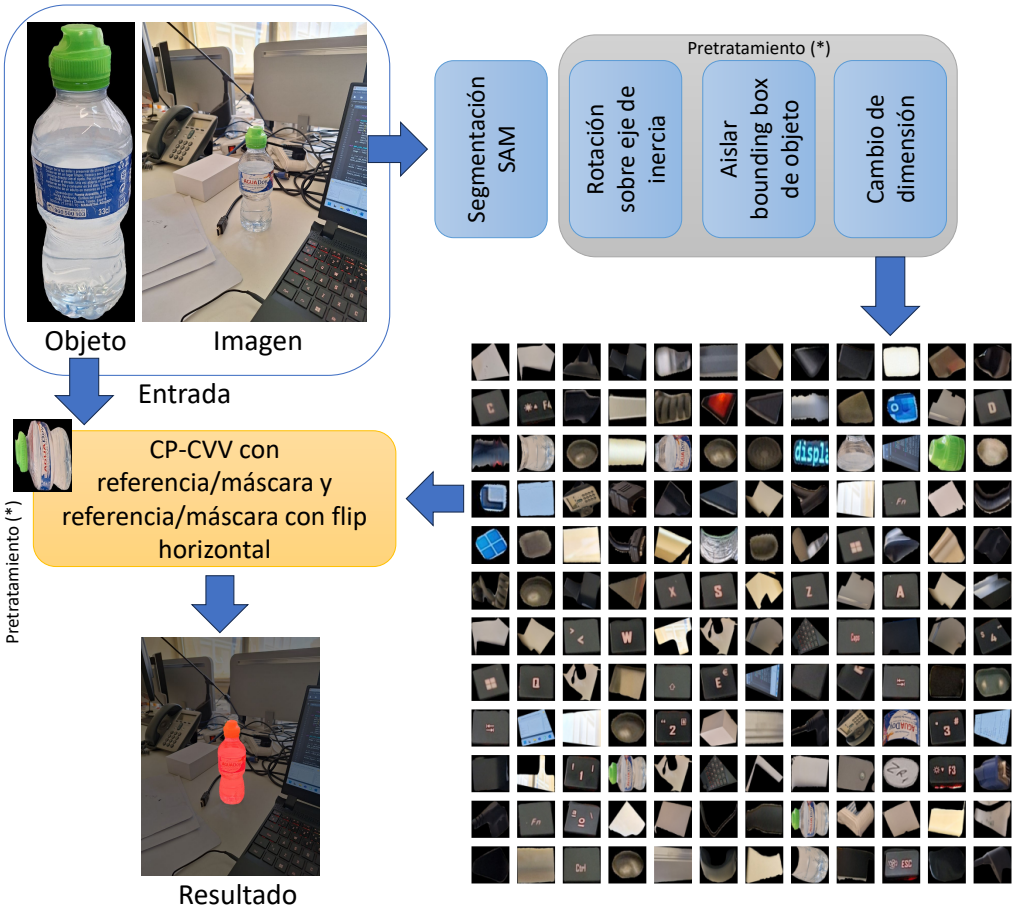

Figura 1: Esquema del modelo de segmentación OSL.

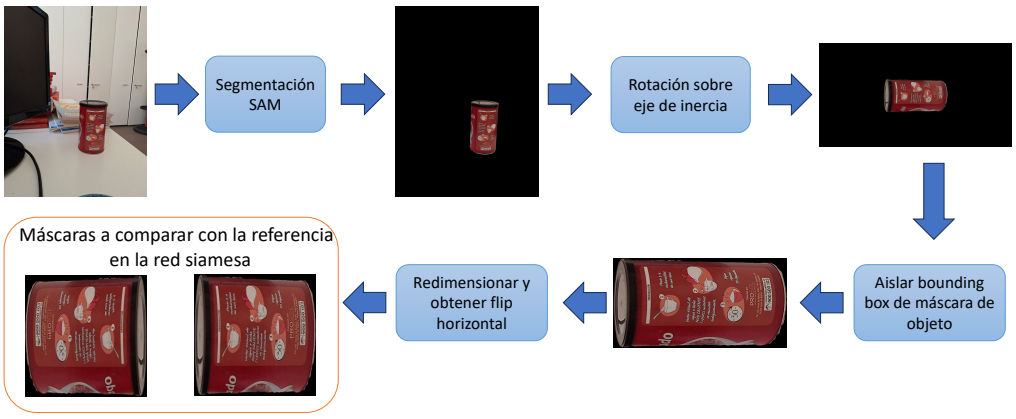

Figura 2: Alineamiento de las máscaras.

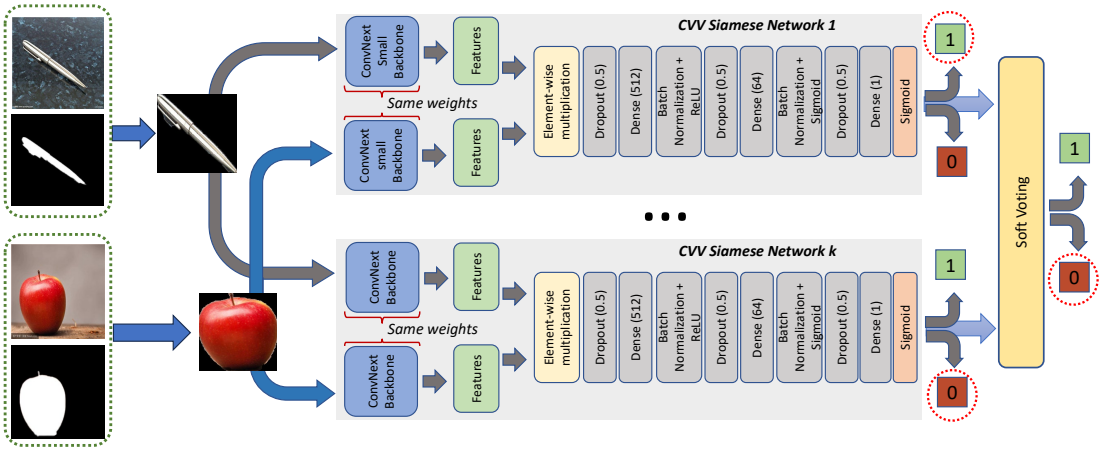

Figura 3: Modelo CP-CVV.

siamesa se aproxima a un valor de 0 o 1, dependiendo de si el par se clasifica como positivo o negativo. En la integración de varios clasificadores mediante votación dura, se cuentan los pares clasificados como positivos y negativos, y la decisión final se basa en el recuento mayoritario. En la votación suave, los valores de salida de los distintos clasificadores se promedian. Si el resultado promedio supera $\frac{1}{2}$, el par se clasifica como positivo; de lo contrario, se clasifica como negativo.

Consideremos que P_c representa el resultado acumulativo obtenido al sumar las salidas sigmoidales ponderadas de diferentes redes siamesas para una clase de prueba c. Así, P_c indica el valor asignado a una imagen de prueba que pertenece a una categoría determinada, evaluada seleccionando una imagen aleatoria por clase. Sea p_{ci} la salida sigmoidea del clasificador i con una imagen de la categoría c y w_i su ponderación. Para la votación suave, P_c se calcula según la Ecuación (1).

$$P_c(x) = \sum_{i=1}^{k} w_i \cdot p_{ci}(x) \tag{1}$$

Para identificar la categoría más similar, seleccionamos la más cercana a 0. Esto se logra mediante el uso de la función $arg_c min$, como se representa en la Ecuación (2). Esta función produce la clase ganadora.

$$C_{soft}(x) = arg_c min\,[P_c(x)] \tag{2}$$

3. Experimentación

Aunque se ha utilizado un modelo preentrenado de SAM basado en un backbone ViT, capaz de segmentar regiones de objetos desconocidos, para CP-CVV se ha tenido que entrenar el modelo utilizando un conjunto de imágenes diferentes de las que se quieren utilizar durante la inferencia. Para ello se ha seleccionado una base de datos de imágenes con objetos y sus segmentaciones llamada FSS-1000 (Li et al., 2020). Este conjunto de datos contiene 1,000 tipos diferentes de objetos.

Para el entrenamiento se han dividido las 1,000 clases en 100 de prueba y 900 de entrenamiento, distribuyendo estas últimas según el planteamiento mencionado anteriormente. Utilizando $k = 5$ slots, cada red siamesa se entrenó con 720 clases de entrenamiento y 180 clases de validación. El entrenamiento se llevó a cabo con técnicas de *data augmentation*, aplicando cambios de perspectiva, rotación, traslación e iluminación. Como nuestro objetivo es la comparación de máscaras, se han segmentado los objetos de las imágenes previamente utilizando los datos suministrados en FSS-1000.

Los resultados de la clasificación se evaluaron utilizando los datos de prueba. Para cada imagen de prueba, se seleccionaron imágenes aleatorias tanto de su categoría como de otras categorías. El accuracy de la clasificación de un objeto de la base de datos FSS-1000 en su clase correcta fue de un 92.30 % para CP-CVV frente a 86.04 % cuando sólo utilizamos una única red siamesa.

Una vez que el modelo CP-CVV estaba entrenado, se han realizado diversos experimentos con objetos e imágenes reales. En la Figura 4 se pueden observar algunos de los resultados de la detección y segmentación. La selección correcta de la máscara devuelta por el método CP-CVV ofrece una confianza de entre el 98 y 99 %. Ante ligeras variaciones de perspectiva, como podemos ver en el caso del teléfono, el método es capaz de obtener correctamente la segmentación. En aproximadamente un 10 % de las imágenes tratadas, SAM no compuso correctamente las máscaras esperadas, dividiendo por ejemplo el objeto en varias máscaras sin generar una que agrupara todas. Se observa que esto ocurre especialmente ante objetos severamente ocluidos.

También se han realizado experimentos ante casos de oclusión ligera, como se muestra en la Figura 3. Se puede observar cómo la segmentación se realiza correctamente para la búsqueda del teléfono.

Conviene mencionar que este método puede buscar y segmentar diversos objetos en la imagen a partir de una única imagen de referencia, pero los objetos buscados deben tener suficiente grado de relación. La búsqueda produce una segmentación semántica si consideramos que tenemos un catálogo de objetos de los que disponemos de una única imagen. El método tarda en procesar la imagen unos pocos segundos por lo que no puede ser utilizado actualmente en aplicaciones que requieren un proceso rápido de imágenes. Sin embargo, se ha empezado a utilizar en la segmentación inicial de los objetos que aparecen en un escenario por parte de un robot. A diferencia de otros modelos que requieren entrenamiento con

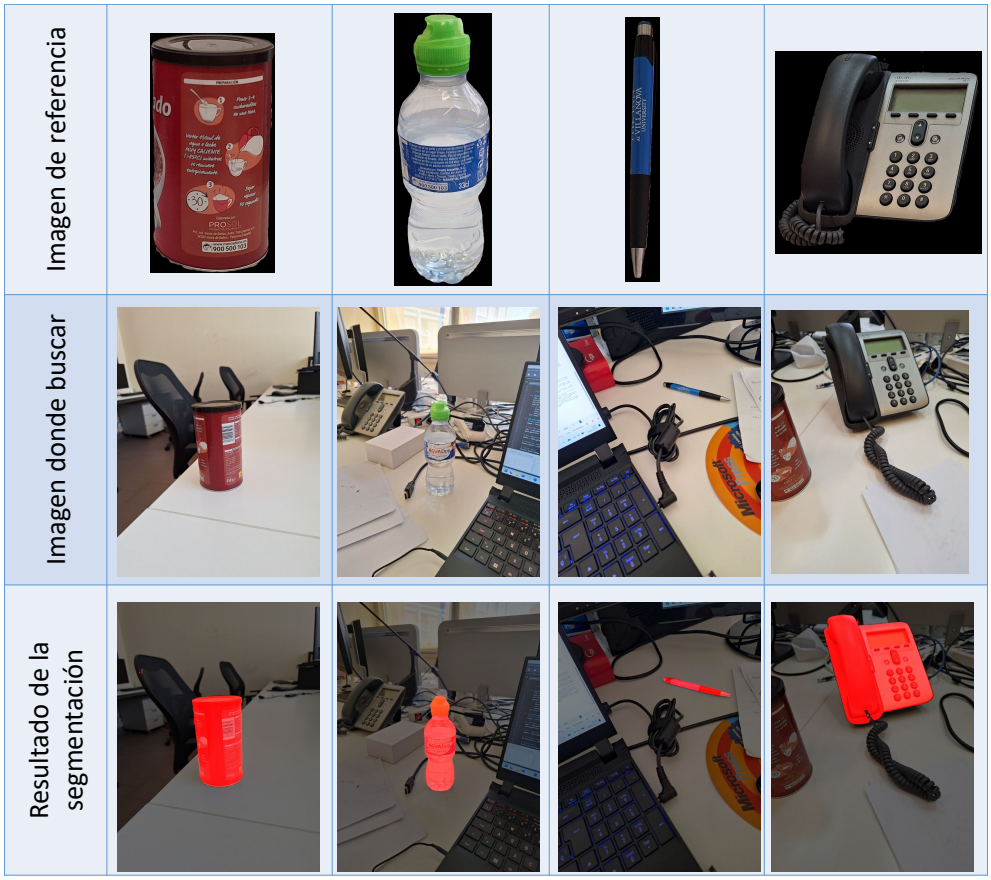

Figura 4: Algunos resultados de la segmentación semántica con SAM+CP_CVV.

(a) Imagen de referencia (b) Imagen donde buscar (c) Resultado

Figura 5: Segmentación en el caso de oclusión ligera.

nuevos objetos, este modelo empieza a segmentar un objeto nuevo con únicamente una fotografía.

4. Conclusiones

Se ha presentado en el artículo un método que permite la segmentación de objetos utilizando técnicas de *one-shot learning* y métodos de segmentación genéricos no semánticos. Este método permite segmentar objetos a partir de una única muestra y ha demostrado ser una alternativa a los métodos que requieren de etiquetado previo de numerosas imágenes y de costosos entrenamientos.

Entre las limitaciones del método podemos mencionar el coste computacional del uso de SAM y CP-CVV. Mientras que el primero es costoso cuando se obtienen todas las máscaras posibles de una imagen, planteamos su aceleración utilizando versiones más rápidas como Fast SAM (Zhang et al., 2023). Por otro lado, CP-CVV se puede acelerar reduciendo el número de slots o la dimensión de las imágenes de entrada de las redes siamesas.

El método responde correctamente ante cambios de perspectiva de los objetos, aunque en caso necesario es posible añadir unas pocas imágenes por objeto desde distintas perspectivas. El sistema puede funcionar directamente bajo el paradigma de *few-shot learning* con simplemente añadir algunas imágenes.

La utilidad de la presente investigación es múltiple. Cualquier sistema de aprendizaje incremental que requiera de la segmentación de objetos puede utilizar esta tecnología para aprender nuevos objetos sólo con verlos una vez. Así, por ejemplo, su uso en robótica permitirá tareas como que el robot coja un objeto que sólo ha visto una vez.

Agradecimientos

Esta investigación ha sido realizada en el marco del proyecto ROSOGAR PID2021-123020OB-I00 financiado por MCIN/AEI/10.13039/501100011033/FEDER, UE, y el proyecto EIAROB financiado por la Consejería de Familia de la Junta de Castilla y León - Next Generation EU.

Referencias

Chen, T., Xie, G.-S., Yao, Y., Wang, Q., Shen, F., Tang, Z., Zhang, J., 2021. Semantically meaningful class prototype learning for one-shot image segmentation. IEEE Transactions on Multimedia 24, 968–980.

Duque-Domingo, J., Aparicio, R. M., Rodrigo, L. M. G., 2023. One shot learning with class partitioning and cross validation voting (cp-cvv). Pattern Recognition 143, 109797.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al., 2023. Segment anything. arXiv preprint arXiv:2304.02643.

Li, X., Wei, T., Chen, Y. P., Tai, Y.-W., Tang, C.-K., 2020. Fss-1000: A 1000-class dataset for few-shot segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2869–2878.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. C., 2016. Ssd: Single shot multibox detector. In: European conference on computer vision. Springer, pp. 21–37.

Liu, Y., Zhang, X., Zhang, S., He, X., 2020. Part-aware prototype network for few-shot semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. Springer, pp. 142–158.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. arXiv preprint arXiv:2201.03545.

Luddecke, T., Ecker, A., 2021. The role of data for one-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2653–2658.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788.

Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B., 2017. One-shot learning for semantic segmentation. arXiv preprint arXiv:1709.03410.

Siddique, N., Paheding, S., Elkin, C. P., Devabhaktuni, V., 2021. U-net and its variants for medical image segmentation: A review of theory and applications. Ieee Access 9, 82031–82057.

Wang, K., Liew, J. H., Zou, Y., Zhou, D., Feng, J., 2019. Panet: Few-shot image semantic segmentation with prototype alignment. In: proceedings of the IEEE/CVF international conference on computer vision. pp. 9197–9206.

Zhang, C., Han, D., Qiao, Y., Kim, J. U., Bae, S.-H., Lee, S., Hong, C. S., 2023. Faster segment anything: Towards lightweight sam for mobile applications. arXiv preprint arXiv:2306.14289.

Zhang, C., Lin, G., Liu, F., Guo, J., Wu, Q., Yao, R., 2019. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9587–9595.

Zhang, X., Wei, Y., Li, Z., Yan, C., Yang, Y., 2021. Rich embedding features for one-shot semantic segmentation. IEEE Transactions on Neural Networks and Learning Systems 33 (11), 6484–6493.

Zhang, X., Wei, Y., Yang, Y., Huang, T. S., 2020. Sg-one: Similarity guidance network for one-shot semantic segmentation. IEEE transactions on cybernetics 50 (9), 3855–3865.