A Contrastive Distillation Approach for Incremental Semantic Segmentation in Aerial Images

Edoardo Arnaudo[1,2][0000−0001−9972−599X], Fabio Cermelli[1][0000−0001−7077−697X], Antonio Tavera[1][0000−0002−9013−4007], Claudio Rossi[2][0000−0001−5038−3597], and Barbara Caputo[1][0000−0001−7169−0158]

[1] Politecnico di Torino, Italy {name.surname}@polito.it
[2] LINKS Foundation, Torino, Italy {name.surname}@linksfoundation.com

Abstract. Incremental learning represents a crucial task in aerial image processing, especially given the limited availability of large-scale annotated datasets. A major issue concerning current deep neural architectures is known as catastrophic forgetting, namely the inability to faithfully maintain past knowledge once a new set of data is provided for retraining. Over the years, several techniques have been proposed to mitigate this problem for image classification and object detection. However, only recently the focus has shifted towards more complex downstream tasks such as instance or semantic segmentation. Starting from incremental-class learning for semantic segmentation tasks, our goal is to adapt this strategy to the aerial domain, exploiting a peculiar feature that differentiates it from natural images, namely the orientation. In addition to the standard knowledge distillation approach, we propose a contrastive regularization, where any given input is compared with its augmented version (i.e. flipping and rotations) in order to minimize the difference between the segmentation features produced by both inputs. We show the effectiveness of our solution on the Potsdam dataset, outperforming the incremental baseline in every test.[3]

Keywords: Semantic segmentation · Incremental learning · Aerial images

1 Introduction

Semantic Segmentation represents a key task in aerial image processing, given its wide range of applications, from urban contexts [21], land cover and monitoring [33] or agricultural settings [32]. However, the majority of state-of-art solutions are designed to perform on a static set of categories by means of a full end-to-end training, with no option to integrate new knowledge. Without precautions, deep neural networks tend in fact to forget previously acquired information when a new training set is provided, resulting in poor performance on the old classes.

[3] Code available at: https://github.com/edornd/contrastive-distillation

This phenomenon, known as *catastrophic forgetting* [16], has been addressed and successfully mitigated through a range of different methods [12, 25, 15], mostly considering image classification or object detection. In recent years, a greater deal of effort has been put on specific downstream tasks such as semantic segmentation, with solutions involving representation consistency [9], replay-based methods [29], or knowledge distillation [3]. The problem of incremental learning is extremely relevant also in aerial settings where, despite the growth in resources and data, the scarcity of large-scale annotated aerial datasets remains a crucial drawback for practical applications. In fact, it is often the case that images are collected in the same geographical area [21], or that the data itself is not immediately available, but rather acquired and processed periodically. In this work, we propose to tackle the problem of incremental-class learning (ICL) in the context of semantic segmentation, focusing on aerial imagery. Leveraging on the MiB framework [3], a distillation-based method specifically designed for semantic segmentation tasks, we introduce an additional regularisation based on contrastive distillation, with the aim of exploiting a distinctive feature of such images, namely their invariance to orientation. We explicitly model this feature by comparing the activations produced by the framework on the input and its transformed version, minimising their difference. A first step involves the student network, comparing pairs of augmented inputs, then activations are also compared with the teacher from the previous incremental step, to improve the knowledge distillation. We evaluate our solution on the Potsdam benchmark dataset [21], where it consistently outperforms the robust incremental baseline in every setting. In summary, our contributions can be listed as follows:

- We address the problem of ICL in semantic segmentation of aerial images, providing benchmark results on a popular dataset.
- We propose a new regularization and distillation approach based on contrastive representation learning, addressing the arbitrary orientation of the inputs, one of the key aspects of aerial images.

2 Related Work

Aerial Semantic Segmentation. Thanks to the recent advancements in deep learning, many semantic segmentation approaches have been proposed over the years [14, 5, 35, 8], focusing mostly on natural images. Most common methods revolve around encoder-decoder, fully-convolutional architectures [6].These techniques have been successfully applied to the field of aerial images in wide range of contexts, such as semantic labelling in urban [1, 8, 18] or agricultural scenarios [32, 18], or land cover tasks [23]. Despite the strong similarities with the natural counterparts, aerial images present some peculiar differences that have been addressed with varying approaches: first, satellite imagery are seldom limited to the visible spectrum and often include additional frequencies [32]. Common solutions to this problem include simpler solutions such as the duplication of input weights [20] or finer multi-modal approaches comprising the fusion of different modalities [23, 33]. Last, a peculiar aspect of aerial and satellite images is represented by the top-down view, in which the orientation becomes arbitrary. In our work we propose to leverage on this peculiar feature, already successfully exploited in classification tasks [24], [31], by applying a contrastive regularization to both the segmentation task and the incremental tasks, to further improve the knowledge distillation between steps.

Incremental Learning. Catastrophic forgetting [16], meaning the inability to remember past knowledge upon learning new information, represents a major issue concerning current deep learning solutions. Several techniques have been proposed to mitigate this issue, with different approaches: replay-based methods [25, 29], exploiting exemplars from old classes parameter isolation parameter-based methods [15], involving a selective pruning so that the weights representing old labels are maintained through the learning steps, and memory-based approaches [34], where important parameters from previous steps are consolidated, forcing the model to maintain a robust representation for old classes. Last, one of the most effective techniques focuses on data and exploits knowledge distillation [12, 3]. The latter is usually carried out with a *teacher-student* approach. Considering Semantic Segmentation on aerial imagery, a first proposal is represented by [29]: here, an hybrid approach comprising both knowledge distillation and additional supporting exemplars is employed. Similarly, in [9] the distillation approach is improved by strengthening the internal representations throughout the learning steps. Compared to image classification, semantic segmentation presents peculiarities that may lead to poor performances when not addressed, such as the presence of a common *background* label. In MiB [3], this issue is tackled by taking into consideration this distributional shift, by means of unbiased losses and regularizations with respect to the background label.

Contrastive Learning. Contrastive learning has become one of the most promising recent techniques in deep learning, closing the gap between supervised and self-supervised settings [17, 7, 10, 2], or even improving the former by learning more robust representations [11]. The objective of Contrastive Representation Learning (CRL) is to cluster together latent representations of similar samples (i.e. *positive examples*), while at the same time increasing the distances between instance representations of different categories *(i.e. negative examples)*. CRL is often applied exploiting pretext tasks (i.e. manually devised tasks solely based on the image itself), including: geometric or color transformations [17, 7], image reconstruction from its parts [19, 28], or cross-modal techniques [13, 30]. These additional tasks can also be paired with more traditional supervised settings such as semantic segmentation, in order to improve the results on the main task [28, 30], deal with low resource datasets [4], or integrate additional modalities [22, 30]. Here, we propose a similar approach where the same inputs are augmented twice, however we exploit the resulting representations as a further regularization to induce further invariance with respect to the applied transformations, during both standard training and knowledge distillation.

3 Methodology

3.1 Problem statement

We address the problem of Incremental-Class Learning (ICL) for Semantic Segmentation on aerial images, where we suppose that different portions of data are provided sequentially, each one with a different set of labels.
First, we can define Semantic Segmentation as a pixel-wise classification, where each pixel x_i composing a generic image $x \in X$ with constant dimensions $H \times W$, is associated with a label $y_i \in Y$ representing its category, or eventually associated with a generic and comprehensive *background* class $b \in Y$. The training can be defined as learning a model f_θ with parameters θ, mapping from the image space X to the pixel-wise label space Y, namely: $f_\theta : X \mapsto \mathbb{R}^{|H \times W \times Y|}$.

Considering now the ICL setting, we require multiple sequential training phases named *learning steps*, in which we provide a different set of data samples and labels every time. Specifically, at each step t, we expand the previous set of labels Y^{t-1} with the additional ground truth Y^t, obtaining a new set of labels $C^t = Y^{t-1} \cup Y^t$. At each phase, we are also provided with a new training set D^t, such that each pixel-wise label y_i belongs to one of the current categories Y^t or the generic background class b. We then train a new model f_θ^t on the whole set of categories C^t, deriving the old labels from the outputs of the previous model $f_\theta^{t-1} : X \mapsto \mathbb{R}^{|H \times W \times Y^{t-1}|}$ and the new labels via standard training, exploiting the dataset for the current step. The final goal is to obtain a single model, able to perform well on both and new classes, namely $f_\theta^t : X \mapsto \mathbb{R}^{|H \times W \times C^t|}$.

3.2 Baseline

As previously mentioned, we adopt MiB as robust incremental baseline [3]. In ICL applied in the context of image classification, a standard approach involves a two-way training, combining a supervised loss on the dataset at the current step D_t with an additional term to maintain the old knowledge. In the case of the selected framework, the latter is carried out through distillation of the old model's outputs. Specifically, the final loss at each learning step becomes:

$$L(\theta^t) = L_{CE}(\theta^t) + \lambda L_{KD}(\theta^t) \tag{1}$$

Where $L_{CE}(\theta^t)$ represents a supervised Cross-Entropy loss, while $L_{KD}(\theta^t)$ represents the Knowledge Distillation term at step t from the previous model $f_{\theta^{t-1}}$, weighted by a factor λ.

As briefly stated in Sec. 2, it is common that two sets of categories, namely Y_i and Y_j share the common background class b, however the semantic regions of the image are assigned to such label is often different in every set. This aspect of semantic segmentation needs to be dealt with during the incremental steps, taking into account that a pixel labeled as background in the dataset D_t might instead belong to one of the previous classes from step 0 to $t-1$. Thus, for each pixel i of a generic image x, the predicted probability $q(i, b)$ for the background class is substituted with:

$$q(i, b) = \sum_{k \in Y^{t-1}} q_x^t(i, k) \tag{2}$$

In other words, the background is not considered as a category on its own, but rather a probability of having an old class *or* actual background.

A similar concept is adopted for the distillation component, where the following distillation loss is applied:

$$L_{KD}^{\theta^t}(x, y) = \frac{1}{N} \sum_{i \in x} \sum_{c \in Y^{t-1}} q_x^{t-1}(i, c) log(q_x^t(i, c)) \tag{3}$$

Where the last term refers to the predicted probabilities for the new model with respect to the old classes. Given that the contribution for the new labels is provided by the Cross-Entropy loss, we require that $q_x^t(i, c) = 0$, $\forall c \in Y^t \setminus \{b\}$. In every other case, the term represents the predicted probability for the new model of having a label c for a pixel i, normalized across old classes as reported in [3]. Again, the distributional shift of the background class needs to be addressed for the incremental learning as well. Consequently, the predicted probability $q_x^t(i, b)$ for this special class is rewritten as:

$$q_x^t(i, b) = \sum_{k \in Y^t} q_x^t(i, k) \tag{4}$$

In other terms, the predicted probability for the background class of the new model is substituted with the probability of having a new class *or* the background. In fact, we expect regions belonging to the new classes to be ignored by the previous model, thus labelled as generic background.

Moreover, excluding similarities among categories, it is extremely likely that predictions for $f_{\theta^{t-1}}$ for the current classes Y^t will fall under the background class. For this reason, we perform the same weight initialization for the final classifier as proposed in [3], so that its outputs for the new classes are uniformly distributed around the background from the very beginning to ease convergence.

3.3 Contrastive Distillation

As stated in Sec. 2, a major difference between natural and aerial images is represented by their orientation: in the former case, the point of view is fundamental to the correct detection of an entity. In fact, in a common scenario we expect to find background and foreground entities in a specific part of the image (e.g. animals in a specific pose, sky on top, ground on the bottom). In the latter case instead, given that the orientation is often arbitrary and simply given by the direction of the observation mean, image rotations around the top-down axis become meaningless for the correct classification or detection.

Therefore, we explicitly model this orientation bias by introducing an additional regularization, both to the supervised training and the incremental knowledge distillation, using a contrastive-based approach. Specifically, given a generic input image x, we can obtain the output features of the current model $\phi_\theta(x)$ (thus excluding the final classifier). At the same time, given the same image transformed with augmentation T, the model should output a new activation,

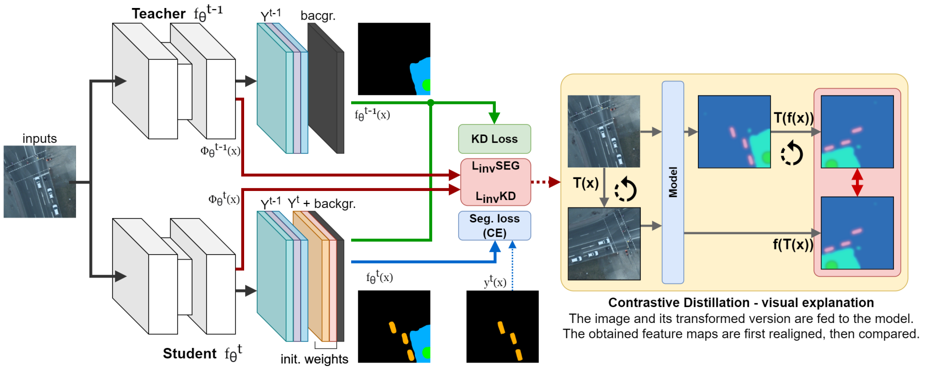

Fig. 1: Overview of the ICL setting on aerial images. For each step t, both the image x and its augmented version $T(x)$ are provided to the old (top) and new (bottom) models. New classes are trained with supervised training on the available ground truth (blue), while old categories are learned through KD (green). Last, features of the augmented inputs are confronted with the augmented features of the normal input, on both distillation and supervised training (red).

namely $\phi_\theta(T(x))$. Given the invariance to rotation, we can assume that both outputs are comparable, minus a transformation, which can be directly applied to the first activation. Formally, we can therefore introduce a regularization term at each learning step t, namely L_{inv}^{SEG} as:

$$L_{inv}^{SEG} = MSE(\phi_{\theta^t}(T(x)),\ T(\phi_{\theta^t}(x)) \tag{5}$$

In other words, the additional term minimizes the differences between *the features of the model on the transformed image and the transformed features of the same model on the original image*, exploiting a Mean Squared Error between the features.

In an ICL setting, we are also interested in transferring the knowledge between $f_{\theta^{t-1}}$ and f_{θ^t}, so that the previous outputs are maintained as unaltered as possible. Together with the standard KD loss from Eq. (3), we can apply the same invariance principle between old and new models. More formally, at each step $t > 0$ we can introduce a further regularization as:

$$L_{inv}^{KD} = MSE(\phi_{\theta^t}(T(x)),\ T(\phi_{\theta^{t-1}}(x)) \tag{6}$$

Simply put, this term minimizes the difference between the features of the new model derived from the transformed image and the transformed features of the old model, obtained from the non-augmented version of the input.

In summary our method comprises three regularizations, therefore the final loss to be minimized can be expressed as:

$$L(\theta^t) = L_{CE}(\theta^t) + \lambda L_{KD}(\theta^t) + \eta L_{inv}^{SEG}(\theta^t) + \rho L_{inv}^{KD}(\theta^t), \tag{7}$$

where the terms λ, η and ρ are scalar factors, weighting the contribution of the additional losses. The overall framework is illustrated in Fig. 1.

4 Results

4.1 Experiments

As described in the previous section, we build our method on top of the MiB framework, which represents a strong baseline for ICL in segmentation tasks. We perform all our experiments on the Potsdam dataset [21], a well known benchmark on aerial imagery providing an urban land cover subdivided into six classes: *impervious surfaces*, *building*, *low vegetation*, *trees*, *cars* and *clutter*. The dataset contains 38 large patches taken from the namesake city, where each patch has a fixed size of 6000 × 6000. Each patch comes with a sampling resolution of $5cm$ and provides five different modalities, namely: red (R), blue (B), green (G), infrared (IR) and a normalized digital surface map (DSM), all encoded as TIFF files. Given our focus on ICL, we only include in our tests inputs composed of RGB and RGBIR, discarding the additional surface map.

Every incremental set of labels is assumed to be disjoint from the previous ones. However, given the aerial setting, it is quite common that each image contains many of the available labels. For this reason, we first split the set into disjoint partitions, such that each split only contains a single label. Formally, considering a full dataset $D \subset X \times Y^{|H \times W|}$, we subdivide the available data into $|Y|$ disjoint partitions D_y such that $D_i \cap D_j = \emptyset \ \forall i, j \in Y$ where $i \neq j$, and each partition only contains a set of labels $Y_i = \{i, b\}$, i.e the set of images is unique for each partition and each split only contributes to the whole training with a single label, or a generic background. Every incremental step will then include a variable number of classes, which will in turn require all the partitions corresponding to the involved categories.

4.2 Implementation details

For all the experiments we adopted an encoder-decoder architecture with residual connections, based on the Res-UNet model [8]. Since memory requirements are crucial for the incremental setting, we introduce two optimizations: first, we swap the standard ResNet backbone with an equivalent yet more efficient TResNet with ImageNet pretraining [26]. The latter applies a series of optimizations aimed at maximizing the data throughput on GPU, while at the same time improving the performance over the classical residual architectures. For the experiments concerning four input channels, namely RGBIR, we expand the input layers duplicating the weights of the red channels, with a similar approach to [20]. Second, we apply in-place activated batch normalization also on the decoder, as proposed in [27], further reducing the memory footprint of the architecture.

We train the model for 80 epochs for each step, using AdamW as optimizer with learning rate of 10^{-3} and a cosine annealing scheduler, while reducing to 10^{-4} for the last steps. We adopt a batch size of 8, with effective size equal to 16 given that the pairs generated via contrastive augmentation are also exploited for the supervised training. Given the large size of the inputs, we tile the 6000×6000 images of the Potsdam dataset into patches with size 512×512 with overlap of 12 pixels, which is the minimum amount required to avoid partial tiles while also minimizing the replication of the image content. We perform robust data augmentation as in [8] in every setting, focusing on elastic transformations. Considering the contrastive regularization, we maintain the setting provided in [3]. We set the factors $\eta = \rho = 0.1$ in every test and evaluate as transformation random vertical and horizontal flipping, with rotations by 90-degree angles. In order to monitor the performances, we select 15% of the training set as validation. The final results are reported as F1 scores on the benchmark test set.

4.3 Potsdam dataset

Given the high similarities among image patches and the uniform distribution of the labels among the tiles, the overlapped setting [3, 29], (i.e. images are kept even if they contain future classes), is not complex enough for a robust evaluation of the proposed regularizations. For this reason, we implement the *split* protocol described in Sec. 4.1: we first tile the original patches to obtain fixed-size input images, then we partition the dataset into 5 different disjoint sets, where each one is associated with a single label. Then, we randomly assign each tile to the smallest set among the labels present in the current tile, obtaining a uniform allocation of the data samples among the classes. This configuration can be seen as having 5 different datasets, where each one only contains a single type of annotation. The disjoint splits ensure that the model will work on unseen images at each step, further increasing the robustness of the tests.

We perform tests for two different configurations: first we replicate the testing scenario proposed in [29] where we suppose to receive, for the initial step, the labels for *building* and *trees*, then *impervious surfaces* and *low vegetation*, and as last step *car* (`3-2-1`). Second, we perform a more challenging test with the same order of labels, but provided sequentially (`5S`). For this last configuration, we exclude the *clutter* category, since it is not included in the official benchmarks [21]. Results for both configurations are shown in Tab. 1 and Tab. 2. Given the framework explicitly designed for segmentation, the MiB baseline performs reasonably well, even considering the fully sequential setting. However, the contrastive distillation approach consistently improves the performances in every experiment and every step, as reported in Sec. 4.3, even in the multi-spectral tests. We note that in the simpler `3-2-1` setting the RGB baseline performs on par with the regularized version. We argue because of both the effectiveness of the standard approach and the robust backbone pretrained on RGB images. However, in more challenging scenarios such as `5S`, the contribution of the additional regularization is far more prominent, with a total increase over MiB of around 4%.

Table 1: Class-wise and average results (F1 Score) obtained after 3 incremental steps (`3-2-1`). Double vertical lines indicate label groups for each step.

Method	**Building**	**Tree**	**Clutter**	**Surf.**	**Low veg.**	**Car**	**Avg.**
MiB (RGB)	0.9116	0.8217	0.2766	0.8918	0.7589	0.8500	0.7517
MiB + CD (RGB)	0.9209	0.8085	0.3119	0.9021	0.7619	0.8541	**0.7599**
MiB (RGBIR)	0.8708	0.8062	0.2682	0.8773	0.7414	0.8176	0.7303
MiB + CD (RGBIR)	0.9178	0.8190	0.3128	0.8950	0.7635	0.8515	**0.7598**

Table 2: Class-wise and avg. results (F1 Score) obtained after 5 incremental steps (5S).

Method	**Building**	**Tree**	**Surfaces**	**Low veg.**	**Car**	**Avg.**
MiB (RGB)	0.8451	0.7449	0.7912	0.7011	0.6759	0.7810
MiB + CD (RGB)	0.9015	0.7515	0.8848	0.7313	0.8287	**0.8195**
MiB (RGBIR)	0.8564	0.7007	0.8575	0.6862	0.8228	0.7847
MiB + CD (RGBIR)	0.8770	0.7740	0.8755	0.7343	0.8437	**0.8209**

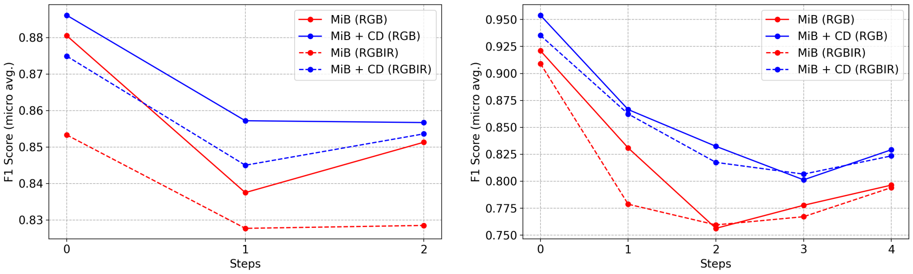

Fig. 2: Micro-averaged F1 scores over the incremental steps in the `3-2-1` configuration (left) and `5S` (right). Blue indicates Contrastive Distillation (CD), dashed lines the RGBIR version.

4.4 Ablation study

In Tab. 3, we report an ablation study highlighting the contribution of our proposals, on the *split 5S* configuration with RGB input. We first start from a simple finetuning (FT): as expected, a new training without considering previous knowledge is detrimental for every step but the last. We then test on the MiB framework that already provides excellent results, with an average increment of more than 60% over the simple FT baseline. Naively introducing the single L_{inv}^{SEG}, (i.e. acting on the current step only) results in better scores for the last class, as expected. However, this negatively affects the performance on previously seen categories, which are not taken into consideration. On the other hand, applying the single L_{inv}^{KD} between current and old model allows for higher scores for previous categories, increasing the average score of 2%, though without obtaining any boost on the labels for the current step. Combining the two regularizations, it is possible to both improve over the current step and increase the performance over old classes, with a significant boost of around 4% over the strong MiB baseline and close to the theoretical upper bound of the *offline* test, representing a static multi-class learning over the whole set at once. As additional test for the distillation capabilities of the regularization, in the second row of Tab. 3 we report results for finetuning, using both the unbiased cross-entropy from [3] and CD, without actual distillation loss. The scores confirm that the additional losses actively contribute in maintaining previous knowledge.

Table 3: Ablation study applied to the 5S setting, as class-wise F1 Scores of the last incremental step and averaged across classes.

Method	**Building**	**Tree**	**Imp. surf.**	**Low veg.**	**Car**	**Average**
FT	0.0	0.0	0.0	0.0	0.8708	0.1742
FT, Unb. CE + CD	0.6118	0.4927	0.6924	0.2909	0.5275	0.5231
MiB	0.8491	0.7625	0.8480	0.6751	0.7703	0.7810
MiB + L_{inv}^{SEG}	0.8178	0.7452	0.8514	0.6781	0.8186	0.7822
MiB + L_{inv}^{KD}	0.9079	0.7522	0.8815	0.7011	0.7895	0.8064
MiB + L_{inv}^{SEG} + L_{inv}^{KD}	0.9015	0.7515	0.8848	0.7313	0.8287	**0.8196**
Offline	0.9510	0.8535	0.9063	0.8415	0.8942	0.8893

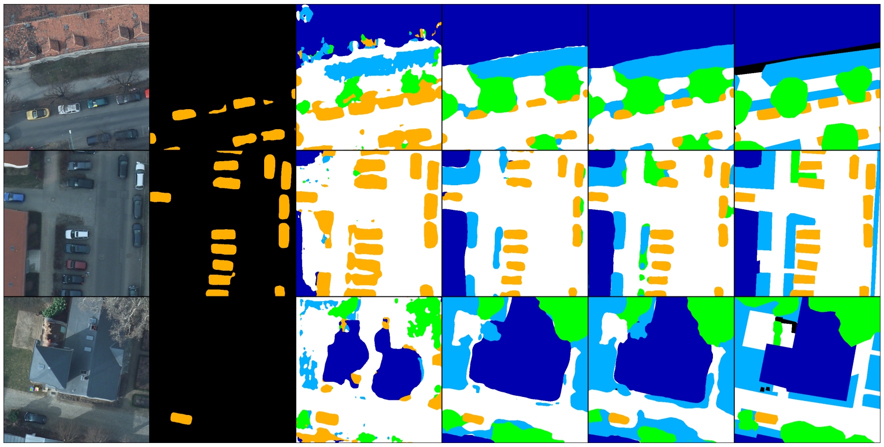

Fig. 3: From left to right: input, finetuning (FT), Finetuning with unbiased CE and Contrastive Distillation (FT + CD), Modelling the Background (MiB), MiB with Contrastive Distillation (MiB + CD), ground truth.

5 Conclusions

We addressed the problem of incremental learning in the context of semantic segmentation of aerial imagery, proposing a new regularization based on contrastive distillation to explicitly model the orientation invariance of such top-down images. In our experiments, we first provide benchmark results for the current state-of-the-art technique on natural images, already displaying excellent performances. We then demonstrate the effectiveness of our simple additional solution leveraging on the same framework, that consistently outperforms the strong baseline leading to a more stable sequential training. Nevertheless, incremental learning remains a challenging problem, especially considering different data sources and domains. Future works could provide more insight on this technique with additional datasets and explore more diverse scenarios, where datasets not only come with different annotations, but also from different domains.

Acknowledgements This work was developed in the context of the Horizon 2020 projects SHELTER (grant agreement n.821282) and SAFERS (grant agreement n.869353).

References

1. Audebert, N., Le Saux, B., Lefèvre, S.: Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks. ISPRS Journ. Phot. Rem. Sens. **140**, 20–32 (2018)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882 (2020)
3. Cermelli, F., Mancini, M., Rota Bulò, S., Ricci, E., Caputo, B.: Modeling the background for incremental learning in semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (June 2020)
4. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. In: Adv. Neural Inform. Process. Syst. (2020)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Eur. Conf. Comput. Vis. (September 2018)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Int. Conf. Mach. Learn. pp. 1597–1607. PMLR (2020)
8. Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journ. Phot. Rem. Sens. **162**, 94–114 (2020)
9. Feng, Y., Sun, X., Diao, W., Li, J., Gao, X., Fu, K.: Continual learning with structured inheritance for semantic segmentation in aerial imagery. IEEE Trans. Geo. Rem. Sens. (2021)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9729–9738 (2020)
11. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Adv. Neural Inform. Process. Syst. vol. 33, pp. 18661–18673 (2020)
12. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Trans. Pattern Anal. Mach. Intell. **40**(12), 2935–2947 (2017)
13. Loghmani, M.R., Robbiano, L., Planamente, M., Park, K., Caputo, B., Vincze, M.: Unsupervised domain adaptation through inter-modal rotation for rgb-d object recognition. IEEE Robot. and Autom. Lett. **5**(4), 6631–6638 (2020). https://doi.org/10.1109/LRA.2020.3007092
14. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3431–3440 (2015)
15. Mallya, A., Lazebnik, S.: Packnet: Adding multiple tasks to a single network by iterative pruning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7765–7773 (06 2018). https://doi.org/10.1109/CVPR.2018.00810
16. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psych. Learn. Motiv., vol. 24, pp. 109–165. Elsevier (1989)

17. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: IEEE Conf. Comput. Vis. Pattern Recog. (June 2020)
18. Nogueira, K., Dalla Mura, M., Chanussot, J., Schwartz, W.R., dos Santos, J.A.: Learning to semantically segment high-resolution remote sensing images. In: Int. Conf. Pattern Recog. pp. 3566–3571 (2016)
19. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Eur. Conf. Comput. Vis. pp. 69–84 (2016)
20. Pan, B., Shi, Z., Xu, X., Shi, T., Zhang, N., Zhu, X.: Coinnet: Copy initialization network for multispectral imagery semantic segmentation. IEEE Geos. Rem. Sens. Lett. **16**(5), 816–820 (2019). https://doi.org/10.1109/LGRS.2018.2880756
21. society for photogrammetry, I., remote sensing: Potsdam dataset (2018)
22. Pielawski, N., Wetzer, E., Öfverstedt, J., Lu, J., Wählby, C., Lindblad, J., Sladoje, N.: Comir: Contrastive multimodal image representation for registration. In: Adv. Neural Inform. Process. Syst. vol. 33, pp. 18433–18444 (2020)
23. Piramanayagam, S., Saber, E., Schwartzkopf, W., Koehler, F.W.: Supervised classification of multisensor remotely sensed images using a deep learning framework. Rem. Sens. **10**(9) (2018). https://doi.org/10.3390/rs10091429
24. Qi, K., Yang, C., Hu, C., Shen, Y., Shen, S., Wu, H.: Rotation invariance regularization for remote sensing image scene classification with convolutional neural networks. Rem. Sens. **13**(4) (2021). https://doi.org/10.3390/rs13040569
25. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2001–2010 (2017)
26. Ridnik, T., Lawen, H., Noy, A., Friedman, I.: Tresnet: High performance gpu-dedicated architecture. Winter Conf. App. Comp. Vis. pp. 1399–1408 (2021)
27. Rota Bulò, S., Porzi, L., Kontschieder, P.: In-place activated batchnorm for memory-optimized training of dnns. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018)
28. Singh, S., Batra, A., Pang, G., Torresani, L., Basu, S., Paluri, M., Jawahar, C.: Self-supervised feature learning for semantic segmentation of overhead imagery. In: Brit. Mach. Vis. Conf. vol. 1, p. 4 (2018)
29. Tasar, O., Tarabalka, Y., Alliez, P.: Incremental learning for semantic segmentation of large-scale remote sensing data. IEEE Journ. Select. Topics App. Earth Observ. Rem. Sens. **12**(9), 3524–3537 (2019)
30. Valada, A., Mohan, R., Burgard, W.: Self-supervised model adaptation for multimodal semantic segmentation. Int. J. Comput. Vis. **128**(5), 1239–1285 (2020)
31. Wang, G., Wang, X., Fan, B., Pan, C.: Feature extraction by rotation-invariant matrix representation for object detection in aerial image. IEEE Geos. Rem. Sens. Lett. **14**(6), 851–855 (2017). https://doi.org/10.1109/LGRS.2017.2683495
32. Yang, S., Yu, S., Zhao, B., Wang, Y.: Reducing the feature divergence of rgb and near-infrared images using switchable normalization. In: IEEE Conf. Comput. Vis. Pattern Recog. Work. pp. 206–211 (jun 2020). https://doi.org/10.1109/CVPRW50498.2020.00031
33. Yuan, Q., Shafri, H.Z.M., Alias, A.H., Hashim, S.J.b.: Multiscale semantic feature optimization and fusion network for building extraction using high-resolution aerial images and lidar data. Rem. Sens. **13**(13) (2021). https://doi.org/10.3390/rs13132473
34. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: Int. Conf. Mach. Learn. ICML'17, vol. 70, p. 3987–3995 (2017)
35. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conf. Comput. Vis. Pattern Recog. (July 2017)