A SEMANTIC SEGMENTATION-GUIDED APPROACH FOR GROUND-TO-AERIAL IMAGE MATCHING

Francesco Pro[1], *Nikolaos Dionelis*[2], *Luca Maiano*[1], *Bertrand Le Saux*[2], *Irene Amerini*[1]

[1]Sapienza University of Rome, Italy
[2]European Space Agency (ESA), ESRIN, Φ-lab, Italy

ABSTRACT

Nowadays the accurate geo-localization of ground-view images has an important role across domains as diverse as journalism, forensics analysis, transports, and Earth Observation. This work addresses the problem of matching a query ground-view image with the corresponding satellite image without GPS data. This is done by comparing the features from a ground-view image and a satellite one, innovatively leveraging the corresponding latter's segmentation mask through a three-stream Siamese-like network. The proposed method, Semantic Align Net (SAN), focuses on limited Field-of-View (FoV) and ground panorama images (images with a FoV of 360°). The novelty lies in the fusion of satellite images in combination with their segmentation masks, aimed at ensuring that the model can extract useful features and focus on the significant parts of the images. This work shows how SAN through semantic analysis of images improves the performance on the unlabelled CVUSA dataset for all the tested FoVs.

Index Terms— Earth Observation data, Ground-to-aerial image matching, Semantic segmentation, Data fusion

1. INTRODUCTION

The accurate ground-to-aerial image matching task is a challenging problem to tackle, which raises high interest today. The correct geo-localization of ground view images in Computer Vision plays an important role across different domains, including journalism, forensics analysis, and Earth Observation. With the rapid increase of satellite images covering a huge part of the Earth's surface, the task of determining the geographical location represented in a given ground-view image, becomes a cross-view matching problem between the images taken by a camera placed at the ground level and the aerial images taken by a satellite or aerial vehicle, e.g. drone.

The different point of view between an image captured at ground level and that taken by a satellite poses an alignment problem between the two views. Added to this, the change of context and differences in terms of image quality between the two captured views add another level of complexity. The ground images generally have a limited field of view, which is defined as the observable area captured by an optical device; the lower this value, the lower the quantity of information inside the image that is useful for comparison with an aerial one. A series of AI methods have been exploited in the last years to improve the image-matching problem. This work addresses the problem of matching a query ground-view image with the corresponding satellite image, proposing a method that focuses on both ground panoramas and limited *Field-of-View* (FoV) image data.

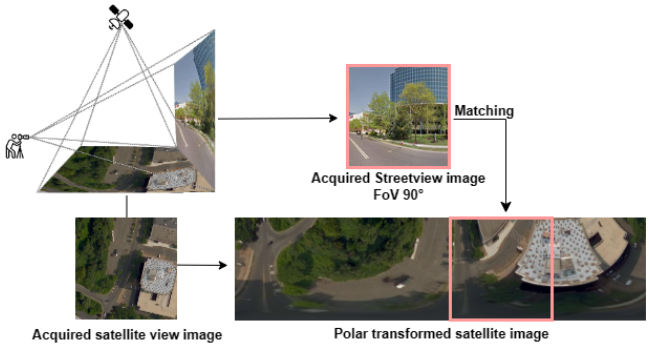

Fig. 1. Example of the ground-to-aerial matching problem. The *query* ground-view image is matched to the polar transformed aerial image.

The novelty of this work is in the introduction of satellite images in combination with their estimated segmentation masks, as additional information, to ensure that the model can focus on the significant parts of the images, such as roads, buildings, and vegetation that can be helpful for the data matching. In particular, our Semantic Align Net (SAN) model is based on a Siamese-like network [1, 2, 3] with three CNN branches extracting features from the ground, the satellite, and the segmentation images. The last two mentioned extracted features are concatenated and, then, correlated with the ones extracted from ground images. The input data to this model are pre-processed, performing normalization, polar transformation, and semantic segmentation on labelled and unlabelled data. The SAN model comprises a triplet loss that is used to train the networks using a similarity function, estimating the orientation angles between the images. Figure 1

shows the points of view from which the two images are captured and the red box represents the identification of the query ground view image inside the satellite one. To overcome the scarcity of labelled aerial images and the difficulties of creating annotations for these images, we have implemented a new method for performing semantic segmentation from unlabelled data (NEOS) in parallel with this work [4]. This is a *Transformer*-based model for semantic segmentation to segment satellite images of the non-annotated CVUSA dataset used for the geo-localization task. This allows us to evaluate our model on this unannotated dataset. We show that the proposed ground-to-aerial matching model, SAN, improves the performance compared to the baseline models on a subset of the unlabelled CVUSA dataset on all the tested FoVs.

2. RELATED WORK

The first attempts to this task came from ground-to-ground image-matching. Hays et al. [5] introduced a method aimed to geolocate through the extraction of any type of geographical information from ground-level images. Instead, Chen et al. [6] focused on image relationships through 3D reconstruction and geometric constraints in urban and natural environments. Geometric constraints were also used by Baatz et al. [7] mainly for natural environments. Shan et al. [8] instead contributed to the task's direction, a pipeline with a viewpoint-dependent matching method to handle ground-to-aerial variant viewpoints, while other works [9, 10] model this task as a graph matching problem.

In recent years, a series of methods based on Artificial Intelligence have been introduced that improve performance in this task. In particular, we can group these techniques into the following three main groups:

Generative Adversarial Networks (GAN). This model and architecture is used to synthesize aerial images from ground-view image data, thus reducing the domain gap between the ground and aerial image domains. In the end, this information is combined or fused for the retrieval task [11, 12, 13].

Vision Transformers (ViT). More recent studies introduce ViT exploiting a self-attention mechanism working on temporal sequences of images, trying to model and capture temporal correlations that could help the image matching [14, 15].

Siamese-like networks. This architecture's methods try to learn viewpoint-invariant discriminative characteristics from the images [1, 2, 3, 16]. This is done by passing ground and aerial-view images to CNNs that extract features from them. Our proposed model, SAN, falls into this last category.

3. PROPOSED METHOD

In this work, we propose a Siamese-like network, which we call Semantic Align Net (SAN), that combines features from aerial and ground-view images. Inspired by Shi et al. [2], we apply the polar transformation to the satellite images to bring them into a ground domain format, reducing the gap between the aerial and ground domains. Then, the network is trained to estimate a similarity-matching score between them. Since ground-view images usually have different field-of-views, we compute this score at each possible orientation angle to augment the variability seen during training.

The main contribution of this work is the introduction of the satellite images in combination with their semantic segmentation masks, as additional information, hence making the model able to focus on the important parts of the images, such as on roads, buildings, and vegetation. In this way, the robustness to the image content variation is also increased. In the next paragraphs, we will go into the details of the proposed model through the following steps: (1) Semantic Segmentation, (2) Image Processing, (3) Feature Extraction and Concatenation, and (4) Similarity Matching. Figure 2 presents the proposed architecture.

Semantic segmentation. In the Earth Observation (EO) field, it is very hard to find labelled aerial images, and it is also very difficult to annotate them [17, 18, 19]. The CVUSA dataset used as benchmark for the ground-to-aerial image matching task contains non-segmented aerial images. Therefore, to extract the segmentations from aerial images, we rely on the Non-annotated Earth Observation Semantic Segmentation (NEOS) [4] method. This method is trained on labelled and unlabelled datasets, evaluating the unlabelled CVUSA dataset. NEOS uses a Transformer-based model based on SegFormer [20]. It performs domain adaptation, minimizing the distribution differences of the different domains, so it can work on different datasets having distribution inconsistencies mainly due to acquisition scenes, environment conditions, sensors, and time.

Image pre-processing. Ground-view and aerial images are captured from two completely different points of view, therefore introducing drastic changes of perspective between these two cross-view images. Therefore, in this step, we reduce the differences between the satellite and the ground/street images by cutting and resizing them to a specific shape before passing the images as inputs to the neural networks. This domain gap is reduced by using the polar transformation, by applying the following two formulas:

$$x_i^s = \frac{D_s}{2} - \frac{D_s}{2}\,\frac{(H_v - x_i^t)}{H_v}\cos(\frac{2\pi}{W_v}y_i^t) \tag{1}$$

$$y_i^s = \frac{D_s}{2} + \frac{D_s}{2}\,\frac{(H_v - x_i^t)}{H_v}\sin(\frac{2\pi}{W_v}y_i^t) \tag{2}$$

where $D_s \times D_s$ is the size of an aerial image, and $H_v \times W_v$ denotes the target size of the polar transformation. Here, (x_i^s, y_i^s) is a generic point on the original image, while (x_i^t, y_i^t) is a generic point on the transformed image. To ensure that the scale of the transformed images is consistent, the images in this work are first normalized by dividing by 255, and then centered with the dataset mean (μ) and standardized with the dataset standard deviation (σ).

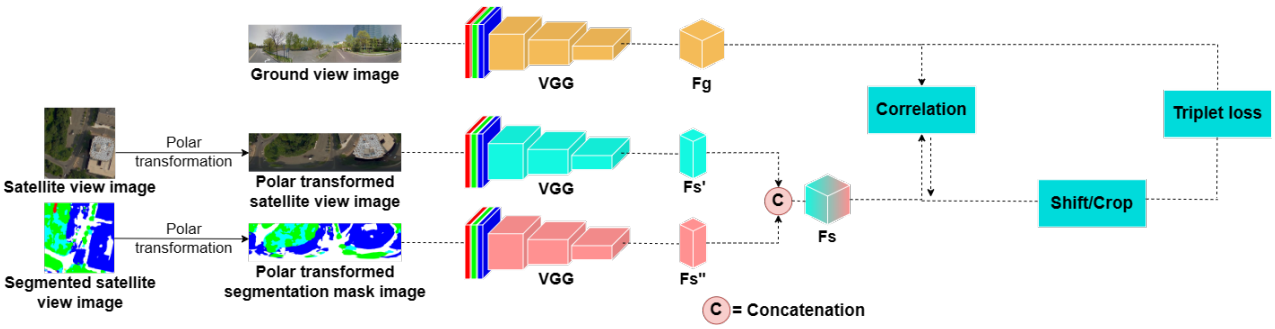

Fig. 2. Our *SAN* network comprises three VGG16 branches extracting features from the (1) ground-view image, the (2) satellite-view image, and (3) its corresponding semantic segmentation mask. The features from the last two branches are correlated and compared with the ones from the ground image to estimate a similarity score.

Feature extraction and concatenation. The images are then passed to the networks. The proposed model, SAN, shown in Figure 2, is composed of three network branches: (1) one for the *ground-view images*, (2) one for the *satellite-view images*, and (3) one for the corresponding *semantic segmentation mask images*. Each branch uses a VGG16 model for extracting features and for generating the corresponding feature maps as output. Each of the networks in Figure 2 is independent from the others, but they all have the same identical structure: (1) 13 Convolutional layers, (2) 3 Max Pooling layers, and (3) 3 Dropout layers. The fully connected and classification layers are truncated from the architecture.

Depending on the chosen FoV, the ground network takes the input images and the output feature maps with variable widths as input. The satellite and segmentation mask network branches take both as input an image of shape (128x512x3) and output feature maps of shape (4x64x8). These two feature maps are concatenated on the third axis, obtaining a resulting shape of (4x64x16). The features are fused to weigh the important features extracted by the satellite images through the ones extracted by the semantic segmentation masks, thus providing a focus on the important elements inside the images.

Similarity matching. In this step, the feature maps extracted from the images are correlated to obtain a similarity score between them and to estimate an orientation angle between polar transformed satellite images and ground-view images.

In general, the relative orientation between the ground-view and the aerial images is not previously known, and it can differ among the images. Therefore, the polar transformation applied to the satellite images aligns them with the ground images up to an unknown azimuth angle. Finding the correct alignment between the aerial and the ground images is crucial for accurate image matching, particularly for limited FoV ground-view images.

To align with the correct orientation angle, for the ground-view and the aerial images, a rotation is needed, when reducing the domain gap between the two views. Aerial images are brought into the ground domain. At this point to find the correct orientation angle, we need a shift on the horizontal axis, both for the aerial and the ground-view images. Using this concept, the correlation between ground and satellite images can be computed along the azimuth angle by fixing the feature map F_s at the bottom and moving the feature map F_g on top of it as a sliding window, executing an inner product on each possible orientation angle. In this way, we obtain a similarity matching score for each orientation angle between the images. For this purpose, we use the correlation operation between the two feature maps: $F_s \star F_g$. Let $F_s \in R^{H \cdot W_s \cdot C}$ and $F_g \in R^{H \cdot W_v \cdot C}$ denote the aerial and ground features respectively, where H and C indicate the height and the channel number of the features. Here, W_s and W_v represent the width of the aerial and ground features respectively. After the polar transformation, we have $W_s \geq W_v$ for all images. The correlation between F_s and F_g is expressed as:

$$[F_s \star F_g](i) = \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W_v} F_s(h, (i+w)\%W_s, c) \cdot F_g(h, w, c) \quad (3)$$

where $F(h, w, c)$ is the feature response at index (h, w, c), and $\%$ denotes the modulo operation. After this computation, the position of the maximum score corresponds to the estimated orientation between ground and aerial images. At the end of this operation, the two final feature maps are used to generate a matrix representing the similarity distance between each compared image. This operation is applied to both matching image pairs and non-matching images. In this way, for matching image pairs, the network is forced to learn similar latent features, while for non-matching images, the similarity score is minimized once the orientation is computed so that the network can learn discriminative features.

4. RESULTS

Implementation details. The three VGG16 networks are structured so that the first ten convolutional layers are pre-

Method	FoV 360°				FoV 180°				FoV 90°				FoV 70°			
	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%
SAN (*our*)	**77,07%**	**92,14%**	**95,62%**	**97,97%**	**48,49%**	**75,53%**	**84,06%**	**91,24%**	**6,23%**	**16,43%**	**24,20%**	**37,07%**	**3,02%**	**9,75%**	**15,44%**	**25,82%**
SCN (*our*)	54,27%	75.67%	82,98%	89,94%	22,21%	45,51%	57,25%	69,84%	1,76%	6,41%	10,02%	16,93%	1,22%	4,11%	6,77%	13,23%
Shi et al. [2]	72,87%	89,62%	92,78%	96,52%	45,28%	72,87%	81,76%	89,57%	6,14%	15,80%	22,93%	35,17%	2,21%	7,31%	12,60%	20,50%

Table 1. Comparison between SAN (proposed) and two baseline methods.

Trained Fov		Tested FoV 360°				Tested FoV 180°				Tested FoV 90°				Tested FoV 70°			
		r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%
360°	SAN (*our*)	**77,07%**	**92,14%**	**95,62%**	**97,97%**	**47,63%**	**75,30%**	**83,43%**	**90,88%**	**18,65%**	**38,92%**	**48,26%**	**60,00%**	**12,28%**	**28,04%**	**36,84%**	**47,40%**
	Shi et al. [2]	72,87%	89,62%	92,78%	96,52%	44,47%	70,74%	79,01%	87,63%	18,10%	36,61%	45.60%	56,48%	11,87%	25,06%	33,00%	43,61%
180°	SAN (*our*)	**67,67%**	**86,95%**	**92,46%**	**96,12%**	**48,49%**	**75,53%**	**84,06%**	**91,24%**	**21,08%**	**43,21%**	**53,50%**	**65,91%**	**14,67%**	**30,84%**	**39,82%**	**52,55%**
	Shi et al. [2]	66,64%	85,06%	91,11%	95,53%	45,28%	72,87%	81,76%	89,57%	20,77%	41,53%	52,42%	64,11%	13,23%	30,20%	39,77%	52,55%
90°	SAN (*our*)	23,07%	**45,96%**	**56,75%**	**70,16%**	**13,72%**	**30,97%**	**42,80%**	**56,88%**	**6,23%**	**16,43%**	**24,20%**	**37,07%**	**3,79%**	**12,10%**	**18,87%**	**29,75%**
	Shi et al. [2]	**23,48%**	43,57%	53,41%	66,59%	12,42%	29,21%	39,19%	54,99%	6,14%	15,80%	22,93%	35,17%	3,61%	11,47%	18,06%	28,04%
70°	SAN (*our*)	**14,67%**	**33,18%**	**42,93%**	**56,52%**	**7,27%**	**21,40%**	**30,43%**	**44,20%**	**3,43%**	**12,19%**	**19,82%**	**31,24%**	**3,02%**	**9,75%**	**15,44%**	**25,82%**
	Shi et al. [2]	11,24%	26,37%	35,85%	49,93%	6,46%	15,58%	23,34%	36,07%	2,84%	9,57%	14,09%	24,74%	2,21%	7,31%	12,60%	20,50%

Table 2. Generalization tests comparing SAN (proposed) with the method by *Shi et al.* [2].

trained on the ImageNet dataset. The following three layers, instead, have been randomly initialized. The first seven layers are frozen, while the others are trainable. The networks have been trained for 30 epochs using the Adam optimizer, with a starting learning rate of value 10^{-5}. Although other networks may fit our case, we will explore this study in the future.

Every experiment reported in this work is evaluated using the *top-K recall* evaluation metric (in Table 1 and Table 2 is indicated with the notation r@1, r@5, r@10, and r@1%), which measures how often a ground-view image is correctly matched with the corresponding aerial image, in terms of distance, among the first K predicted output results. We also release our code for reproducibility[1].

Datasets. We train the semantic segmentation model [4] on the: (a) *Potsdam* [21], and (b) *Vaihingen* [22] datasets. We also train and evaluate both our proposed ground-to-aerial image matching SAN method and [4] on the: (c) *CVUSA* [23] dataset. The proposed method uses a filtered version of it [24] The used ground view images have a FoV value of 360°, so as to work with different values. They are randomly cropped to have the desired width. For all the experiments, we train our model, SAN, and the baselines on 6647 triplets of images and test on 2215 images from [24]. Here, a triplet comprises a ground-view image, a satellite-view image, and the corresponding segmentation mask, where all three represent the same location.

Baselines. We compare the performance of our model against two baselines. (1) The first is the state-of-the-art method introduced by Shi et al. [2]. (2) The second, called Semantic Channel Net (SCN), is a two-branch network that we created as a baseline model to compare it with our SAN model. The first branch is for the ground-view images and the other one is obtained by concatenating the aerial images with their segmentation maks. Differently from our proposed Semantic Align Net (SAN) model, the concatenation is done at the image level and not at the features level.

Experiments. For a reliable comparison, in this work, we use the same FoV values used by Shi et al. [2]: 360°, 180°, 90°, 70°. A different model is trained for each FoV value mentioned above. Based on this, we present two different types of experiments. In the first experiment, which is reported in Table 1, we validate our model by comparing its performance with the two baseline methods introduced before. All three methods are trained and tested with the same CVUSA subset, using the same images split between the training set and the test dataset. According to the results, the proposed method improves the performance for all the FoV values. From Table 1, we can observe a decrease in performance with lower FoV values, and this is mainly because of the fact that the lower the FoV value, the less information in the image.

Our second experiment, reported in Table 2, is a generalization test in which a model trained on a specific FoV value is tested on other FoV values, i.e. on FoV values on which it was not trained. This experiment allows us to see how a model behaves with images with a different FoV from the ones it was trained on. Our model outperforms the one by Shi et al. [2] in all the experiments, thus showing that semantic segmentation introduces useful information to the model, which makes it more robust in all the FoV values.

5. CONCLUSION

This work examines geolocation without using GPS data, and to this end, we model the problem as a cross-view ground-to-aerial image matching task. Our main idea is to integrate the aerial view with its semantic segmentation mask in a three branches Siamese-like network. Our experiments show that this additional stream allows the network to focus on the correct parts of the aerial image and obtain more accurate image matching scores. Our proposed model, SAN, outperforms the other baseline models in all the experiments in this work. Our future work will extend our experiments on the entire CVUSA dataset and will introduce the semantic segmentation masks for the ground-view image data.

[1] https://github.com/pro1944191/SemanticAlignNet

6. REFERENCES

[1] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li, "Optimal Feature Transport for Cross-View Image Geo-Localization," 2019.

[2] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li, "Where am I looking at? Joint Location and Orientation Estimation by Cross-View Matching," 2020.

[3] Royston Rodrigues and Masahiro Tani, "Global Assists Local: Effective Aerial Representations for Field of View Constrained Image Geo-Localization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022, pp. 3871–3879.

[4] Nikolaos Dionelis, Francesco Pro, Luca Maiano, Irene Amerini, and Bertrand Le Saux, "Learning from Unlabelled data with Transformers: Domain Adaptation for Semantic Segmentation of High Resolution Aerial Images," *IGARSS*, 2024.

[5] James Hays and Alexei A. Efros, "IM2GPS: estimating geographic information from a single image," in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[6] David M. Chen, Georges Baatz, Kevin Köser, Sam S. Tsai, Ramakrishna Vedantham, Timo Pylvänäinen, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, Bernd Girod, and Radek Grzeszczuk, "City-scale landmark identification on mobile devices," in *CVPR 2011*, 2011, pp. 737–744.

[7] Georges Baatz, Olivier Saurer, Kevin Köser, and Marc Pollefeys, "Large Scale Visual Geo-Localization of Images in Mountainous Terrain," 10 2012, vol. 7573, pp. 517–530.

[8] Qi Shan, Changchang Wu, Brian Curless, Yasutaka Furukawa, Carlos Hernandez, and Steven M. Seitz, "Accurate Geo-Registration by Ground-to-Aerial Image Matching," in *2014 2nd International Conference on 3D Vision*, 2014, vol. 1, pp. 525–532.

[9] Tania Sari Bonaventura, Luca Maiano, Lorenzo Papa, and Irene Amerini, "An Automated Ground-to-Aerial Viewpoint Localization for Content Verification," in *2023 24th International Conference on Digital Signal Processing (DSP)*, 2023, pp. 1–5.

[10] Sebastiano Verde, Thiago Resek, Simone Milani, and Anderson Rocha, "Ground-to-aerial viewpoint localization via landmark graphs matching," *IEEE Signal Processing Letters*, vol. 27, pp. 1490–1494, 2020.

[11] Krishna Regmi and Mubarak Shah, "Bridging the Domain Gap for Ground-to-Aerial Image Matching," 2019.

[12] Xueqing Deng, Yi Zhu, and Shawn Newsam, "Using Conditional Generative Adversarial Networks to Generate Ground-Level Views From Overhead Imagery," 2019.

[13] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé, "Coming Down to Earth: Satellite-to-Street View Synthesis for Geo-Localization," 2021.

[14] Xiaoyang Tian, Jie Shao, Deqiang Ouyang, Anjie Zhu, and Feiyu Chen, "SMDT: Cross-View Geo-Localization with Image Alignment and Transformer," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 2022, pp. 1–6.

[15] Xiaohan Zhang, Waqas Sultani, and Safwan Wshah, "Cross-View Image Sequence Geo-localization," 2022.

[16] Sixing Hu, Mengdan Feng, Rang M. H. Nguyen, and Gim Hee Lee, "CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7258–7267.

[17] Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch, Nicolas Audebert, and Sébastien Lefèvre, "Semi-Supervised Semantic Segmentation in Earth Observation: The MiniFrance Suite, Dataset Analysis and Multi-task Network Study," 2020.

[18] Ying Chen, Xu Ouyang, Kaiyue Zhu, and Gady Agam, "Semantic Segmentation in Aerial Images Using Class-Aware Unsupervised Domain Adaptation," 11 2021, pp. 9–16.

[19] Nicolas Audebert, Alexandre Boulch, Hicham Randrianarivo, Bertrand Le Saux, Marin Ferecatu, Sébastien Lefèvre, and Renaud Marlet, "Deep learning for urban remote sensing," in *2017 Joint Urban Remote Sensing Event (JURSE)*, 2017, pp. 1–4.

[20] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," 2021.

[21] ISPRS, "Semantic Labelling Challenge Potsdam Dataset, International Society for Photogrammetry and Remote Sensing (ISPRS), German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF)," 2016.

[22] ISPRS, "Semantic Labelling Challenge Vaihingen Dataset, International Society for Photogrammetry and

Remote Sensing (ISPRS), German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF)," 2015.

[23] Scott Workman, Richard Souvenir, and Nathan Jacobs, "Wide-Area Image Geolocalization with Aerial Reference Imagery," 2015.

[24] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs, "Predicting Ground-Level Scene Layout from Aerial Imagery," 2016.