

Combining Supervised Ground Level Learning and Aerial Unsupervised Learning for Efficient Urban Semantic Segmentation

Youssef Bouaziz^{1,2}^a, Eric Royer¹ and Achref Elouni²

¹*Institut Pascal, Université Clermont Auvergne, Clermont-Ferrand, France*

²*LIMOS, Université Clermont Auvergne, Clermont-Ferrand, France*
{first_name.last_name}@uca.fr

Keywords: Semantic Segmentation, Aerial Imagery, 3D Point Clouds, Label Propagation.

Abstract: Semantic segmentation of aerial imagery is crucial for applications in urban planning, environmental monitoring, and autonomous navigation. However, it remains challenging due to limited annotated data, occlusions, and varied perspectives. We present a novel framework that combines 2D semantic segmentation with 3D point cloud data using a graph-based label propagation technique. By diffusing semantic information from 2D images to 3D points with pixel-to-point and point-to-point connections, our approach ensures consistency between 2D and 3D segmentations. We validate its effectiveness on urban imagery, accurately segmenting moving objects, structures, roads, and vegetation, and thereby overcoming the limitations of scarce annotated datasets. This hybrid method holds significant potential for large-scale, detailed segmentation of aerial imagery in urban development, environmental assessment, and infrastructure management.

1 INTRODUCTION

Semantic segmentation of aerial imagery is crucial in fields such as urban planning, environmental monitoring, and autonomous navigation. By assigning semantic labels to every pixel, valuable insights into spatial patterns and functional elements can be derived, enabling large-scale analysis and decision-making. However, aerial photos pose unique challenges due to varying perspectives, occlusions, and the need to integrate multiple data sources. A significant limitation is the lack of annotated ground-truth data, which impedes training high-performing supervised models and benchmarking new approaches.


The scarcity of comprehensive ground-truth datasets forces reliance on alternative methods that integrate additional data sources. Most available datasets cover limited areas or lack sufficient resolution and detail, reducing their utility for precise segmentation. Consequently, there is a growing need for methods that leverage diverse data types to overcome these constraints and improve segmentation accuracy.

To address these challenges, we present a novel methodology that fuses 2D semantic segmentation with 3D point cloud data. Building on the graph-based label propagation technique in (Mascaro et al.,

2021), we transfer semantic information from 2D images to 3D points. Edges between 2D pixels and 3D points enable label diffusion, incorporating geometric scene information. Once diffused in 3D, these labels are then projected back onto 2D aerial imagery, ensuring geometrically consistent and detailed segmentations. This approach preserves alignment between the 2D and 3D domains, resulting in accurate and refined labels.

To further enhance the quality of 2D segmentations, we integrate MaskFormer (Cheng et al., 2021), an advanced semantic segmentation algorithm, into our pipeline. MaskFormer excels at segmenting complex scenes, making it ideal for producing high-quality 2D segmentations, which are then used as input for the label diffusion process. We segment the ground-level images into five key classes: moving object, structure, road, vegetation, and other. These classes capture the essential elements commonly found in aerial and urban scenes, providing a comprehensive understanding of the environment.

Our methodology overcomes the limitations of traditional approaches by combining powerful 2D segmentation with graph-based label propagation to achieve accurate and detailed 3D semantic segmentation of aerial photos. This hybrid approach addresses the challenge of limited ground-truth data by leveraging the geometric relationships between 2D and

^a <https://orcid.org/0000-0003-3257-6859>

3D data, allowing us to transfer labels efficiently and consistently. The proposed method has wide-ranging potential for applications in urban development, environmental assessment, and infrastructure management, where large-scale and accurate segmentation of aerial imagery is critical.

2 RELATED WORK

Supervised semantic segmentation from ground-level images is a critical task in autonomous driving and remains an active research area. Early deep learning-based approaches, such as fully convolutional networks (FCN) (Simonyan and Zisserman, 2015) and GoogLeNet (Szegedy et al., 2015), paved the way. Subsequent architectures, like Segnet (Badrinarayanan et al., 2017) and HRNet (Yuan et al., 2020), addressed limitations such as high computational cost, achieving impressive results like 85.1% mIoU on CityScapes. UNet (Ronneberger et al., 2015), initially for medical segmentation, and Deeplab (Liang-Chieh et al., 2015; Chen et al., 2018) introduced innovations to improve efficiency and preserve detail.

Recent self-supervised methods like DINO (Caron et al., 2021) demonstrate potential by learning from unlabeled data, offering viable solutions for scenarios with scarce annotations, such as aerial segmentation. Annotated datasets like CityScapes (Cordts et al., 2016) (5000 images) and Mapillary Vistas (Neuhof et al., 2017) (25,000 images) remain essential for training and advancing segmentation models.

For remote sensing tasks, datasets vary significantly in annotations, spectral bands, and resolution (Schmitt et al., 2021). Unlike ground-level datasets, there is no equivalent comprehensive dataset for aerial images. The ISPRS Vaihingen and Potsdam dataset (Rottensteiner et al., 2012) has supported many advancements in urban aerial image segmentation. Recent methods like RS-Dseg (Luo et al., 2024) address challenges by using diffusion models with spatial-channel attention to enhance semantic information extraction, achieving state-of-the-art results on Potsdam. The labor-intensive nature of labeling aerial datasets has led to interest in semi-supervised techniques. These methods use a small set of labeled data to generate pseudo-labels for larger unlabeled datasets, augmenting training. For example, (Desai and Ghose, 2022) successfully trained a land use classification network using just 2% labeled data. To address the lack of labeled data, unsupervised approaches (Ji et al., 2019; Caron et al., 2018; Cho et al., 2021; Hamilton et al., 2022) have been ex-

plored. While their accuracy lags behind supervised and semi-supervised methods, they remain practical when no labeled data is available. For example, on CityScapes, unsupervised methods improved from an mIoU of 7.1 in 2018 (MDC (Caron et al., 2018)) to 21.0 in 2022 (STEGO (Hamilton et al., 2022)).

The fusion of heterogeneous methods and data sources offers promising solutions to segmentation challenges. For instance, (Genova et al., 2021) transfers semantic labels from 2D street-level images to 3D point clouds, bridging modality gaps. Graph-based methods encode semantic relationships across modalities. The scene graph concept (Krishna et al., 2017) has been adapted for semantic segmentation of building interiors (Armeni et al., 2019). Mascaro et al. (Mascaro et al., 2021) introduced "Diffuser," a graph-based label diffusion approach that refines 3D segmentations by leveraging multi-view 2D semantic information. This method avoids 3D training data and integrates effectively with existing 2D segmentation frameworks, broadening its applicability.

3 METHODOLOGY

This section outlines our methodology for semantic segmentation of aerial photos using 2D segmentations and 3D point cloud data. Our approach builds on a label propagation technique (Mascaro et al., 2021) for transferring labels from 2D to 3D domains and incorporates MaskFormer, an open-source segmentation algorithm, to improve 2D segmentation accuracy. By combining these methods, we transfer semantic labels from 2D images to a 3D segmented point cloud, enhancing the accuracy and consistency of labels applied to aerial photos.

The process utilizes two data sources: an aerial image and a set of ground-level images I_k . Ground images are processed with a structure-from-motion algorithm to generate a sparse 3D point cloud, which is georeferenced to associate each point with a pixel in the aerial image (using Meshroom). Ground images are also segmented with a semantic segmentation algorithm trained on manually labeled data, assigning semantic labels to each pixel. Each 3D point in the point cloud is linked to the pixels used in its triangulation, forming associations between ground image pixels, 3D points, and aerial image pixels. These relationships are then encoded into a graph as described in the following paragraphs.

3.1 2D-to-3D Label Diffusion

To summarize, the methodology leverages the output of a 2D semantic segmentation framework to propagate class labels through the point cloud, generating a refined 3D semantic map. The algorithm uses a graph structure comprising nodes that represent both 2D pixels and 3D points. The graph incorporates pixel-to-point and point-to-point edges to enable label diffusion.

Pixel-to-point edges are constructed by creating a subgraph $G^{I_k \rightarrow X}$ for each ground image I_k , where I_k is the k -th image and X represents all 3D points in the scene. Represented as an adjacency matrix, this subgraph facilitates information flow from 2D to 3D. Edges between pixels and points are determined by projecting the 3D points back to the 2D image plane using camera projection matrices from Meshroom. This process links the 2D semantic labels to their corresponding 3D points. The adjacency matrix $G^{I_k \rightarrow X}$ is defined as:

$$G_{ij}^{I_k \rightarrow X} = \begin{cases} 1 & \text{if pixel } p_i \text{ projects to 3D point } x_j \text{ in} \\ & \text{frame } I_k \\ 0 & \text{otherwise} \end{cases}$$

Additionally, point-to-point edges are created by connecting each point to its K nearest neighbors based on Euclidean distance. This step ensures that the subgraph $G^{X \rightarrow X}$ encodes the 3D geometry of the scene point cloud. The adjacency matrix $G^{X \rightarrow X}$ is defined as:

$$G_{ij}^{X \rightarrow X} = \begin{cases} w_{ij} = \exp\left(-\frac{\|x_i - x_j\|_2^2}{0.05}\right) & \text{if point } x_i \text{ is} \\ & \text{among the } K \\ & \text{nearest neighbors of point } x_j \\ 0 & \text{otherwise} \end{cases}$$

The label diffusion graph, denoted as G , combines the pixel-to-point and point-to-point edges. The adjacency matrix G is obtained by concatenating the previously defined adjacency matrices as follows:

$$G = \begin{bmatrix} G^{X \rightarrow X} & G^{I_1 \rightarrow X} & \dots & G^{I_{N_f} \rightarrow X} \\ \mathbf{0} & I^{I_1} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & I^{I_{N_f}} \end{bmatrix}$$

where N_f is the total number of frames, I^{I_k} is the identity matrix of size $N^{I_k} \times N^{I_k}$, and N^{I_k} represents the total number of pixels in image I_k . The identity matrices preserve the structure of the point cloud and ground image pixels in the graph representation during iterations.

To propagate class labels through the graph, a probabilistic transition matrix P is computed by normalizing each row of the adjacency matrix G . The transition matrix P is defined as:

$$P_{ij} = \frac{G_{ij}}{\sum_{k=1}^N G_{ik}}$$

where N is the total number of nodes in the graph.

In order to accumulate the likelihood of each node belonging to each class during the iterative propagation, a label matrix Z is defined. The label matrix Z has dimensions $N \times C$, where N represents the total number of nodes in the graph G , and C represents the number of classes.

The label matrix Z incorporates the semantic labels for both the 3D points and the pixels in each ground image. Z is defined as:

$$Z = \begin{bmatrix} Z^X \\ Z^{I_1} \\ Z^{I_2} \\ \vdots \\ Z^{I_{N_f}} \end{bmatrix}$$

where:

- Z^X is a matrix with dimensions $N^X \times C$, where N^X is the total number of 3D points in the point cloud. It represents the initial semantic labels for the 3D points, initialized to zeros due to the absence of prior knowledge about their class labels at the start of the diffusion process.
- $Z^{I_1}, Z^{I_2}, \dots, Z^{I_{N_f}}$ are the initial semantic labels for the pixels in each ground image I_1, I_2, \dots, I_{N_f} . Each Z^{I_k} is a matrix with dimensions $N^{I_k} \times C$, where each row corresponds to a pixel in image I_k and contains the likelihood of that pixel belonging to different classes, based on the 2D semantic segmentation framework's output.

The label diffusion process iteratively multiplies the transition matrix P with the label matrix Z until convergence or a maximum number of iterations is reached. This iterative operation propagates class probabilities through the graph, capturing contextual information across both 2D and 3D domains. The update equation for the label matrix Z is:

$$Z^{(t+1)} = P \cdot Z^{(t)}$$

where $Z^{(t)}$ represents the label matrix at iteration t . The label diffusion process continues until convergence or the maximum number of iterations is reached.

Finally, the likelihood values in the label matrix Z are converted to 3D point labels by assigning each

point the class with the highest accumulated probability. This is achieved by finding the index j^* that maximizes the likelihood in the i -th row of Z :

$$j^* = \arg \max_j Z_{ij}$$

Once the index j^* is determined, the corresponding class label is assigned to the i -th point, making it the most likely class based on the accumulated probabilities. Repeating this for all points in the 3D semantic map yields a refined representation where each point is labeled according to the likelihood values in Z .

In summary, the proposed methodology integrates 2D semantic segmentation results with geometric information from the 3D point cloud, refining semantic labels through a label diffusion process. The 3D label matrix Z^X iteratively accumulates class likelihoods for each node using propagated information from neighbors. The final 3D semantic map offers a more accurate and consistent classification of object classes in the scene.

3.2 Semantic Segmentation of Aerial Photos

The methodology for semantic segmentation of aerial photos involves applying a label diffusion process similar to Section 3.1, but adapted for a 3D-to-2D Label Diffusion process.

This approach transfers labels from a labeled 3D point cloud to 2D unlabeled orthophotos, leveraging the rich semantic information in the 3D point cloud. Additionally, an unsupervised segmentation network (Kim et al., 2020) identifies similarities between orthophoto regions, refining label propagation and aligning semantically similar regions for more accurate segmentation.

The matrices used in this process differ from those in 2D-to-3D diffusion. The label matrix Z' is initialized using the 3D semantic segmentation results from the point cloud (Z^X). It consists of two blocks: the 2D label matrix Z'^O , corresponding to orthophoto pixels, initially set to zeros; and the 3D label matrix Z'^X , containing class probabilities from the segmented point cloud (Z^X). The structure of Z' is:

$$Z' = \begin{bmatrix} Z'^O \\ Z'^X \end{bmatrix}$$

In this representation, Z'^O is a matrix with dimensions $N^O \times C$, where N^O is the total number of pixels in the aerial photos, and C is the number of semantic classes. Similarly, Z'^X has dimensions $N^X \times C$, where N^X is the total number of 3D points in the point

cloud. The label matrix Z' initializes semantic probabilities for 3D points, while Z'^O , representing the 2D orthophoto pixels, is filled with zeros. This provides the starting point for the label diffusion process.

To represent connectivity between orthophoto pixels and 3D points in the graph, we construct the adjacency matrix G' , which includes the pixel-to-pixel adjacency matrix $G'^{O \rightarrow O}$ and the point-to-pixel adjacency matrix $G'^{X \rightarrow O}$.

To establish these connections, we define a sliding window $\mathcal{S}\mathcal{W}$ around each pixel p_i in the aerial photo. For each pixel p_j within this window, we compute a score combining a semantic similarity score (S_{ij}) and a neighborhood score (W_{ij}).

The neighborhood score (W_{ij}) is computed using a Gaussian filter centered at p_i , defined within the sliding window. This filter assigns weights to neighboring pixels based on their spatial proximity to p_i , following a Gaussian distribution. Closer pixels receive higher weights, while farther ones receive lower weights, thus measuring the spatial relationship between p_i and p_j .

To compute the semantic similarity score (S_{ij}), we use an unsupervised segmentation algorithm based on differentiable feature clustering (Kim et al., 2020). The number of classes generated by the algorithm is fixed to a maximum of C , aligning with the classes used in Section 3.1. The segmented aerial photo is utilized to assign semantic similarity scores between p_i and p_j . Pixels belonging to the same class are assigned a score of 1, reflecting strong semantic similarity. For pixels in different classes, a small non-zero score (e.g., $1e-9$) is assigned to allow slow label propagation.

By combining W_{ij} and S_{ij} , we construct the pixel-to-pixel adjacency matrix $G'^{O \rightarrow O}$, encoding spatial and semantic relationships between pixels. It is defined as:

$$G'_{ij}{}^{O \rightarrow O} = W_{ij} \times S_{ij}$$

where $G'_{ij}{}^{O \rightarrow O}$ denotes the element at the i th row and j th column of the matrix, W_{ij} is the neighborhood score between pixels p_i and p_j , and S_{ij} represents their unsupervised semantic similarity.

To link the 3D points in the point cloud with the corresponding pixels in the aerial photos, a manual alignment of the point cloud with the orthophoto is performed. This ensures that each 3D point is accurately projected onto its respective pixel in the orthophoto.

After alignment, the point-to-pixel adjacency matrix $G'^{X \rightarrow O}$ is constructed. For each pixel in the orthophoto, the matrix assigns a score of 1 if a 3D point

projects onto that pixel, indicating a connection in the graph representation.

By linking the nodes representing 3D points to the pixels in the aerial photos, the point-to-pixel adjacency matrix facilitates the label diffusion process, propagating semantic labels from the 3D point cloud to corresponding pixels. This propagation leverages the graph structure, enabling the transfer of semantic information for the segmentation of aerial photos.

To construct the adjacency matrix G' , we concatenate the pixel-to-pixel adjacency matrix $G^{O \rightarrow O}$ with the point-to-pixel adjacency matrix $G^{X \rightarrow O}$. The identity matrix I^O is added to preserve the structure of the point cloud and the aerial photo in the graph representation during the multiplication of G' with Z' .

$$G' = \begin{bmatrix} G^{O \rightarrow O} & G^{X \rightarrow O} \\ \mathbf{0} & I^O \end{bmatrix}$$

The resulting adjacency matrix G' encodes the connectivity between pixels and 3D points, enabling the propagation of semantic labels from the 3D point cloud to the aerial photo.

The label diffusion process iteratively updates the label matrix Z' by multiplying it with the transition matrix P' , which is derived by normalizing G' as described in Section 3.1.

This 3D-to-2D label diffusion process enhances the semantic segmentation of aerial photos by leveraging the geometric context and rich information from the 3D point cloud. The result is more accurate and detailed class labeling, supporting various applications in aerial image analysis and scene understanding.

4 EXPERIMENTS AND RESULTS

In this section, we evaluate our proposed methodology, including the 2D-to-3D and 3D-to-2D label diffusion processes detailed in Section 3. Our experiments demonstrate the effectiveness of the semantic segmentation pipeline through qualitative and quantitative analyses. Performance is assessed using the Mean Intersection over Union (mIoU) metric against a manually labeled ground-truth orthophoto. We also investigate the impact of key parameters, such as the sliding window size $S\mathcal{W}$, on the label diffusion process.

4.1 Dataset and Experimental Setup

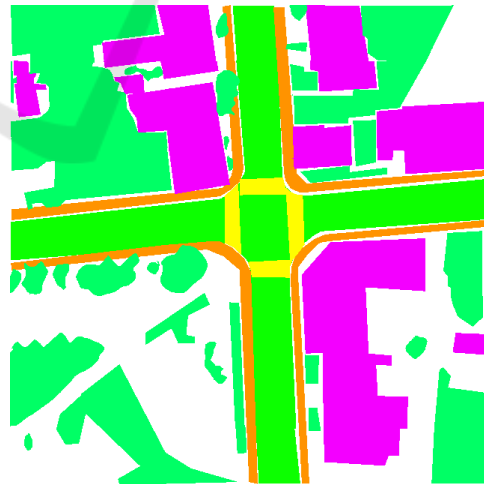
We conducted experiments using ground-level images and a corresponding aerial photo. The ground images were segmented into 66 object categories (Neuhold

et al., 2017), which were remapped into five classes: moving object, structure, road, vegetation, and other. These classes are well-suited for urban environments and cover the key semantic elements of the scene.

The ground photos were processed with Meshroom to generate a 3D point cloud, forming the basis for the 2D-to-3D label diffusion process. The 3D point cloud was then manually aligned with the aerial orthophotos to enable the 3D-to-2D label diffusion process. A manually labeled ground-truth orthophoto was created to compute the Mean Intersection over Union (mIoU), serving as a benchmark for evaluation. Figure 1 shows the manually labeled ground-truth orthophoto used in our analysis.



(a) Orthophoto



(b) Ground-truth orthophoto

Figure 1: Manually labeled ground-truth orthophoto used for mIoU calculation.

4.2 Qualitative Analysis

4.2.1 2D Semantic Segmentation Results

Figure 2 illustrates the 2D semantic segmentation results of the ground-level photos. Using MaskFormer, the model effectively labeled pixels into the predefined 66 object categories, providing the input for the subsequent 2D-to-3D label diffusion process.

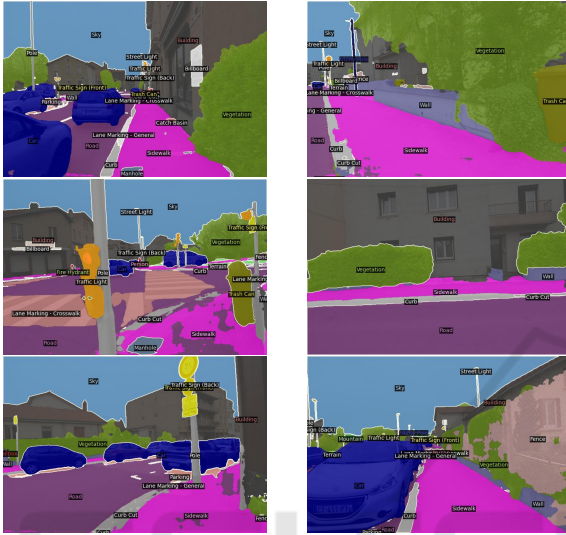


Figure 2: Examples of segmented ground photos.

4.2.2 3D Point Cloud Generation and Labeling

Initially, the ground images were processed with Meshroom to generate an unlabeled 3D point cloud, as shown in Figure 3. This reconstruction captures the scene's geometric structure and serves as the basis for the 2D-to-3D label diffusion process, though it lacks semantic information.

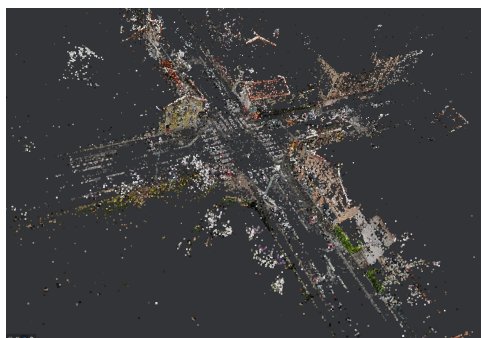


Figure 3: Unlabeled 3D point cloud generated from ground images using Meshroom.

Following the 2D-to-3D label diffusion, the unlabeled 3D point cloud was enriched with semantic labels, as shown in Figure 4. Each point was assigned one of five class labels based on the diffusion pro-

cess, demonstrating the successful transfer of labels from 2D images to 3D space using pixel-to-point and point-to-point constraints.

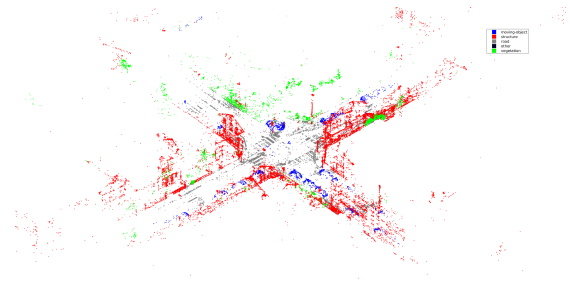


Figure 4: Labeled 3D point cloud after applying the 2D-to-3D label diffusion process.

4.2.3 Orthophoto Segmentation

After aligning the 3D point cloud with the aerial orthophotos, we performed the 3D-to-2D label diffusion to transfer the labels to the 2D orthophoto. As a first step, the orthophoto underwent unsupervised segmentation, shown in Figure 5, which structured the diffusion process by integrating similarity constraints based on feature clustering.

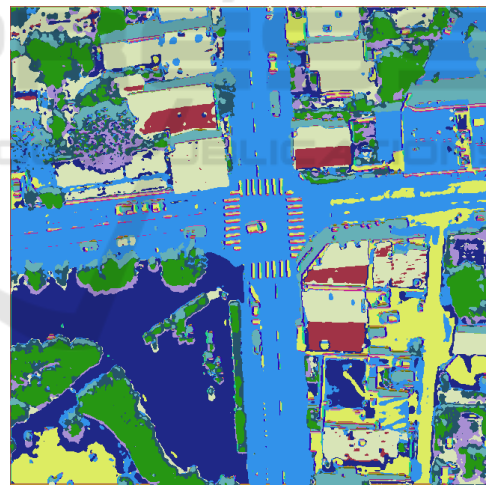


Figure 5: Result of unsupervised segmentation on the orthophoto.

4.3 Quantitative Analysis

Comparing our results with other state-of-the-art methods presents challenges due to the scarcity of works focusing on orthophoto semantic segmentation. Moreover, our approach relies on orthophotos linked to ground-level images, making direct comparisons with existing methodologies difficult. Nonetheless, our results, measured against the manually labeled ground-truth orthophoto, highlight the robustness and

effectiveness of our method in urban environments.

To evaluate the impact of local constraints, we tested our 3D-to-2D label diffusion algorithm with different sliding window sizes ($S\mathcal{W}$): 9×9 , 25×25 , and 65×65 pixels. The sliding window determines the neighboring pixels considered during label propagation in the orthophoto.

Figure 6 shows the mIoU progression across label diffusion iterations, calculated against the manually labeled ground-truth orthophoto. The results indicate

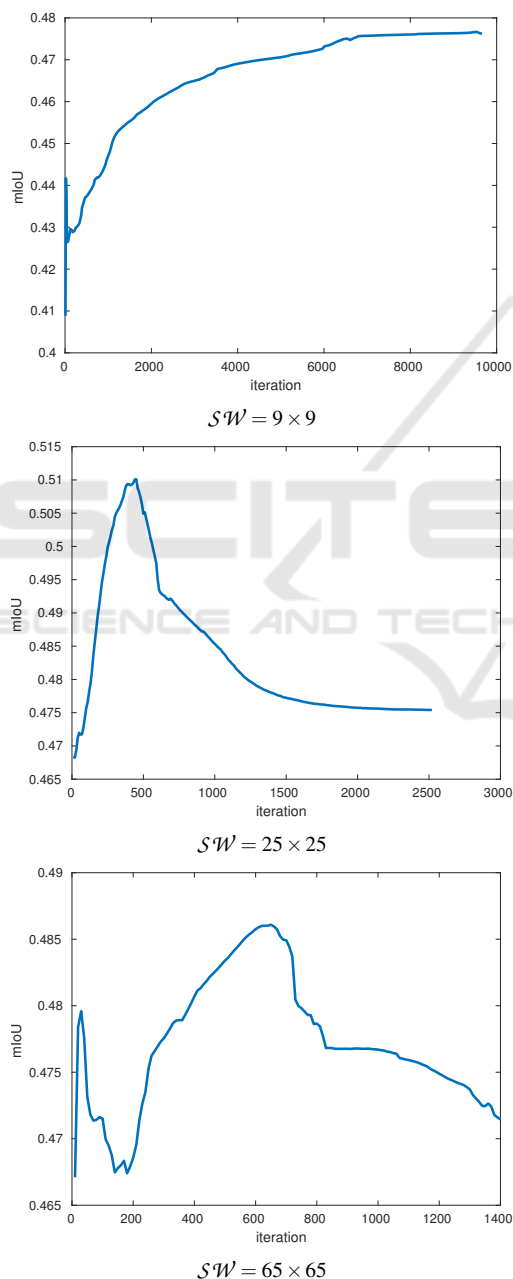


Figure 6: Evolution of mIoU across iterations with different sliding window sizes.

that a sliding window size of $S\mathcal{W} = 25 \times 25$ achieves the best balance, avoiding the limitations of too few constraints or excessive noise from larger windows.

The best segmentation result was achieved after 500 iterations with a 25×25 sliding window. Figure 7 illustrates the final segmented orthophoto. This configuration balanced local and global context, yielding the highest mIoU score of 0.51.

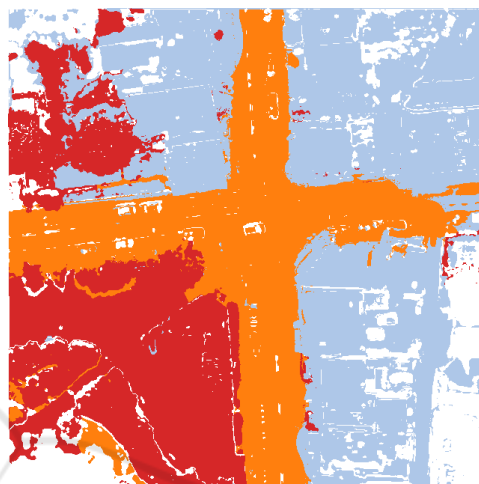


Figure 7: Final semantic segmentation of the orthophoto with $S\mathcal{W} = 25 \times 25$.

5 CONCLUSION

In this paper, we introduced a novel methodology for semantic segmentation of aerial photos by leveraging 2D segmentations and 3D point cloud data. Our approach employs a graph-based label diffusion algorithm to propagate semantic labels from 2D images to a 3D point cloud and subsequently transfer them to aerial photos. The "Meshroom" algorithm was used for 3D reconstruction, providing precise camera poses and accurate geometry from ground-level photos.

By integrating 2D and 3D spatial information, our method achieves accurate and detailed segmentation of aerial photos, effectively capturing intricate scene details. Experimental results validate the proposed framework's effectiveness in urban environments. The optimal configuration ($S\mathcal{W} = 25 \times 25$) highlighted the importance of balancing neighborhood constraints to minimize noise and enhance accuracy.

Future work will focus on improving the unsupervised segmentation stage to better align semantic regions within orthophotos, further enhancing segmentation performance.

REFERENCES

- Armeni, I., He, Z.-Y., Gwak, J., Zamir, A. R., Fischer, M., Malik, J., and Savarese, S. (2019). 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE International Conference on Computer Vision*.
- Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495.
- Caron, M., Bojanowski, P., Joulin, A., and Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*.
- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Hartwig, A. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pages 801–818.
- Cheng, B., Schwing, A. G., and Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation.
- Cho, J. H., Mall, U., Bala, K., and Hariharan, B. (2021). Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16789–16799.
- Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Desai, S. M. and Ghose, D. (2022). Active learning for improved semi-supervised semantic segmentation in satellite images. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1485–1495.
- Genova, K., Yin, X., Kundu, A., Pantofaru, C., Cole, F., Sud, A., Brewington, B., Shucker, B., and Funkhouser, T. (2021). Learning 3d semantic segmentation with only 2d image supervision. In *2021 International Conference on 3D Vision (3DV)*, pages 361–372.
- Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., and Freeman, W. T. (2022). Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*.
- Ji, X., Henriques, J. F., and Vedaldi, A. (2019). Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Kim, W., Kanezaki, A., and Tanaka, M. (2020). Unsupervised learning of image segmentation based on differentiable feature clustering. *IEEE Transactions on Image Processing*, 29:8055–8068.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalan-tidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, (123):32–73.
- Liang-Chieh, C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. (2015). Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In *International Conference on Learning Representations*, San Diego, United States.
- Luo, Z., Pan, J., Hu, Y., Deng, L., Li, Y., Qi, C., and Wang, X. (2024). Rs-dseg: semantic segmentation of high-resolution remote sensing images based on a diffusion model component with unsupervised pretraining. *Scientific Reports*, 14(1):18609.
- Mascaro, R., Teixeira, L., and Chli, M. (2021). Diffuser: Multi-view 2d-to-3d label diffusion for semantic scene segmentation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13589–13595.
- Neuhof, G., Ollmann, T., Rota Bulò, S., and Kotschieder, P. (2017). The mapillary vistas dataset for semantic understanding of street scenes. In *International Conference on Computer Vision (ICCV)*.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing.
- Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Bénitez, S., and Breitkopf, U. (2012). The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, I-3.
- Schmitt, M., Ahmadi, S., and Hänsch, R. (2021). There is no data like more data – current status of machine learning datasets in remote sensing. In *International Geoscience and Remote Sensing Symposium*.
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.
- Yuan, Y., Chen, X., and Wang, J. (2020). Object-contextual representations for semantic segmentation.