KNOWLEDGE TRANSFER FOR LABEL-EFFICIENT MONOCULAR HEIGHT ESTIMATION

Zhitong Xiong [1], *Xiao Xiang Zhu* [1,2]

[1] Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany
[2] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany

ABSTRACT

Estimating height from monocular remote sensing images is one of the most efficient ways for building large-scale 3D city models. However, existing deep learning based methods usually require a large amount of training data, which could be cost-consuming or even not possible to obtain. Towards a label-efficient deep learning model, we propose a new task and dataset for weak-shot monocular height estimation. In this task, only the relative height labels between pairs of a small portion of points are given, which is cheaper and more friendly for humans to annotate. In addition, to enhance the model performance under the sparse and weak-shot supervision, we propose a Transformer-based network for transferring the learned knowledge from a large-scale synthetic dataset to real-world data. Experimental results have shown the effectiveness of the proposed method on a public dataset under the sparse and weak supervision.

Index Terms— Monocular height estimation, relation modeling, transfer learning, weakly-supervised Learning.

1. INTRODUCTION

Geometric information from 3D cities can be used for urban planning and disaster monitoring, which have a close relationship with the life of residents living in cities. Light Detection And Ranging (LiDAR) can actively acquire the Digital Surface Model (DSM) data that contains accurate height information. However, LiDAR is cost consuming, and cannot obtain data in a timely-updated manner. In this context, obtaining the geometric data from monocular remote sensing imagery [1, 2] is essential for a rapid and accurate response to time-critical world events, e.g., natural disasters and damage forecasting [3].

Monocular height estimation (MHE) targets at predicting height data from a single aerial image, which has a broad application potential in practice owing to its fairly simple data acquisition requirements. Considering the representation learning advantages [4, 5], various deep learning-based monocular height estimation methods have been proposed and steady accuracy gains have been achieved. Mou and Zhu [1] designed residual convolutional networks for height estimation and demonstrated its effectiveness on instance segmentation task. Christie [6] proposed to estimate the geocentric pose from monocular oblique images, and promising results have been achieved. Conditional generative adversarial network (cGAN) [7] was proposed to frame height estimation as an image translation task. Kunwar [8] exploited semantic labels as priors to enhance the performance of height estimation on the large-scale Urban Semantic 3D (US3D) dataset [9].

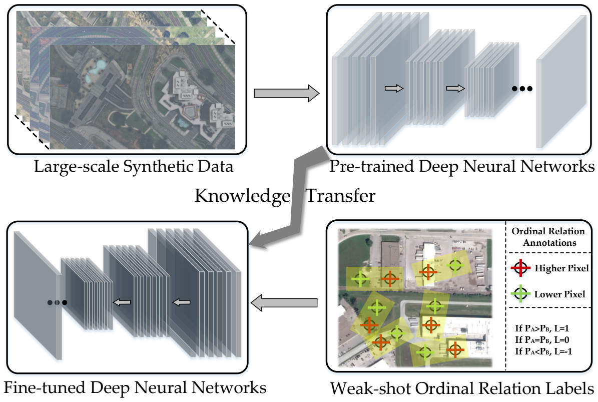
Large-scale Synthetic Data
Pre-trained Deep Neural Networks
Knowledge Transfer
Fine-tuned Deep Neural Networks
Weak-shot Ordinal Relation Labels
Ordinal Relation Annotations
Higher Pixel
Lower Pixel
If $P_A > P_B$, L=1
If $P_A = P_B$, L=0
If $P_A < P_B$, L=-1

Fig. 1. Illustration of the weak shot monocular height estimation task. We aim to transfer the learned knowledge from synthetic dataset to real-world datasets using only sparse ordinal relation annotations.

However, existing deep learning based methods usually rely on the large scale training data. This requirement will be quite challenging for real-world remote sensing applications at a global scale. It is cost-consuming to acquire enough ground truth labels for height estimation, such as using LiDAR or other 3D modeling pipelines. Furthermore, in some emergency situations, we have no time to obtain enough training data. Then, most existing methods will not work due to the label-scarce problem.

Recent works [10, 11] have shown that it is possible to train the depth estimation networks using only sparse annotations of relative depth. For human beings, it is difficult to tell the absolute height given a remote sensing image. However, we humans are good at answering questions like, *which pixel is higher between a given pair of locations?*. Thus, to handle

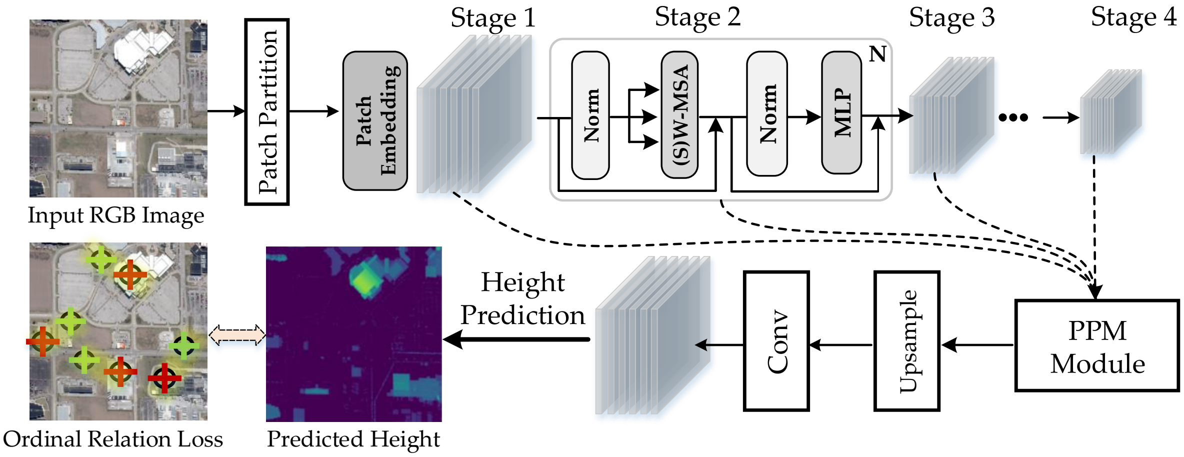

Fig. 2. The whole pipeline of the proposed method. First, multi-scale deep features are extracted by the Swin Transformer. Then, ordinal relation-based loss is used for model training.

the label-scarce problem, we propose to use the sparse and weak annotations, i.e., the relative height relations for training the height estimation model in a weakly-supervised manner.

To enhance the ability for learning relative height relations, we further employ the Transformer-based deep architectures owing to the effective self-attention mechanism in Transformers. Furthermore, to tackle the overfitting problem, we propose to transfer the knowledge learned from a large-scale synthetic dataset, GTAH [12] to real-world data.

2. METHODOLOGY

2.1. Benchmark Dataset Construction

To handle the difficulty in label annotation for monocular height estimation task, we propose to build a benchmark dataset for weak-shot height estimation with only sparse ordinal relation annotations. First, we adopt the GTAH dataset, a large-scale synthetic dataset captured from Grand Theft Auto, for model pre-training. As GTAH contains contains 28,627 height maps obtained under different imaging conditions, such as weather, daytime, shadow, camera height, and pose, it is suitable for training a robust deep model for height estimation. To transfer the learned knowledge from the synthetic dataset to real-world datasets, we further take in the DFC 2019 dataset, as shown in Fig. 1. For this real-world data, only relative height annotations are provided. We aim to design deep learning models to learn to predict the complete height maps only supervised by the sparse relative height annotations.

In practice, we randomly sample 800 pairs of pixels from each ground truth height map, and compute the relative height labels for each pair of locations. As there are 262,144 (512 × 512) pixels in total for each height map, 800 pairs only occupies 0.3% of the whole training labels. Thus, using this sparse and easy to get annotations for model training will greatly reduce the label efforts required for height estimation task.

2.2. Transformer-based Knowledge Transfer

To model the relative relations for height estimation, we exploit the Transformer-based models to learn the pair-wise relations between different pixels. For the dense prediction tasks, Swin Transformer has advantages in reducing the computation complexity by using the window-based self-attention mechanism [13, 14]. As shown in Fig. 1, the input image is first partitioned evenly into patches. Then multi-head self-attention is computed separately for each patch. Assuming the output feature at layer $l-1$ is x^{l-1}, the formula for the computation of two Swin Transformer layers can be expressed as

$$\begin{aligned} \boldsymbol{x}^{l} &= \text{W-MSA}(\text{LN}(\boldsymbol{x}^{l-1})) + \boldsymbol{x}^{l-1}, \\ \boldsymbol{x}^{l} &= \text{MLP}(\text{LN}(\boldsymbol{x}^{l})) + \boldsymbol{x}^{l}, \\ \boldsymbol{x}^{l+1} &= \text{SW-MSA}(\text{LN}(\boldsymbol{x}^{l})) + \boldsymbol{x}^{l}, \\ \boldsymbol{x}^{l+1} &= \text{MLP}(\text{LN}(\boldsymbol{x}^{l+1})) + \boldsymbol{x}^{l+1}, \end{aligned} \tag{1}$$

where W-MSA denotes the window-based multi-head attention module. The SW-MSA stands for the shifted-window multi-head attention module. In Fig. 1, we have illustrated the whole network architecture of our Swin Transformer-based height estimation model. The PPM module [15] is used to aggregate multi-scale features maps to enhance the learned representations. In this work, we first train the propose deep model on the GTAH dataset. Then, the pre-trained network parameters are transferred to real-world datasets under weak-shot ordinal relation-based supervision.

2.3. Weak-shot Ordinal Relation Modeling

Inspired by the work of Chen et al. [10], we propose to constrain the model training with the ordinal relations. During the training stage, given a Transformner-based deep model pre-trained on the GTAH dataset, we directly predict the height maps $\boldsymbol{y}$ on the real-world dataset. Then, based on the predicted height maps, we compute the loss based on the ordinal relation labels, and enforce the model to learn relative depth:

$$\mathcal{L}_{rh} = \begin{cases} \log\left(1+\exp\left(-\boldsymbol{y}_{i_k}+\boldsymbol{y}_{j_k}\right)\right), & r_k=+1 \\ \log\left(1+\exp\left(\boldsymbol{y}_{i_k}-\boldsymbol{y}_{j_k}\right)\right), & r_k=-1 \\ \left(\boldsymbol{y}_{i_k}-\boldsymbol{y}_{j_k}\right)^2, & r_k=0 \end{cases}, \quad (2)$$

where r_k is the relation label of the k_{th} pair. The relative constraint loss $\mathcal{L}_{rh}$ encourages the predicted depth map to agree with the ground-truth ordinal relations. Note that only the pair-wise relation labels are used here for model training. In this work, $k = 800$ pairs of pixels are randomly sampled for each image.

3. EXPERIMENTS

3.1. Dataset and Implementation Details

In this work, we design a new experimental setting using existing height estimation datasets. First, a large-scale synthetic dataset from [12] is used for pte-training the deep model. Then, a dataset from the Data Fusion Contest (DFC) 2019 [16] is processed to provide ordinal relation-based labels for training. Specifically, 2,200 images are used for training. For each image, 800 pairs of pixels are randomly chosen for generating ordinal relation labels. The left 583 images are used for performance evaluation.

3.2. Evaluation Methods

Two metrics are used for performance evaluation, including Root Mean Squared Error (RMSE), and Multi-scale Gradient Error (MSGE). RMSE, defined as RMSE $= \sqrt{\Sigma\left(y_i-\hat{y}_i\right)^2/n}$, is a widely-used metric for regression tasks. RMSE is more sensitive to large height values. Similar to [12], we also use the MSGE to measure the correctness of relative relations between different pixels. MSGE is defined as

$$\text{MSGE} = \frac{1}{M}\sum_{k=1}^{K}\sum_{i=1}^{M}\left(\left|\nabla_x R_i^k\right|+\left|\nabla_y R_i^k\right|\right). \quad (3)$$

Since the relative height relation cannot provide the absolute height information, in this paper, we also evaluate the proposed method using the normalized height maps.

3.3. Experimental Results

For performance evaluation, we have conducted experiments on the DFC 2019 dataset to study the effectiveness of the proposed ordinal relation (OR) based model training method. Specifically, five different experiments are conducted: 1) directly using the pre-trained model on GTAH for performance evaluation on the DFC 2019 dataset; 2) model trained using complete training set of DFC 2019; 3) model pre-trained on ImageNet and finetuned using $k = 800$ OR labels; 4) model pre-trained on GTAH and finetuned using $k = 400$ OR labels; 5) model pre-trained on GTAH and finetuned using $k = 800$ OR labels.

Table 1. Experimental results on the DFC 2019 dataset in the weak-shot supervised setting. The best results are in bold.

DFC 19	Absolute Height		Normalized Height	
	RMSE	MSGE	RMSE	MSGE
GTAH	5.812	3.625	1.681	1.454
DFC 19	2.116	3.031	1.452	1.085
ImageNet,OR(k =800)	3.229	2.871	5.523	2.783
GTAH,OR (k =400)	2.973	2.887	1.679	1.368
GTAH,OR (k =800)	**2.900**	**2.865**	**1.462**	**1.295**

The experimental results are presented in Table 1. From this table, it can be observed that the performance is limited when we directly transfer the GTAH pre-trained model to DFC 2019 dataset. However, the results make sense because there are large domain shifts between these two datasets regarding the height distribution and city styles. Using the traditional pixel-wise training labels can result in a satisfactory performance on DFC 2019 dataset.

When OR-based labels are used for the model training, the results can be clearly improved, for example, RMSE is reduced from 5.812 to 2.900. This indicates that although only 0.3% ordinal relation data is used for training, the results can be significantly improved. Furthermore, we also compare the performance of GTAH pre-trained and ImageNet pre-trained models under the OR-based training setting. We can observe that using GTAH for pre-training can improve the height estimation performance by a large margin. In addition, we also observe that using more OR labels can result in better performance. Some qualitative height estimation examples are presented in Fig. 3.

4. CONCLUSION

In this paper, we study the height estimation task in a weak-shot setting. Different from existing deep learning based methods, the proposed model in this work only uses annotations with relative height between pairs of random points. This type of annotation is cheaper and more friendly for humans to annotate. To effectively train the deep model with sparse and weak-shot labels, we design a Transformer-based deep model to transfer the learned knowledge from a large-scale synthetic dataset to real-world datasets. Both qualitative and quantitative experimental results have demonstrated the effectiveness of the proposed method.

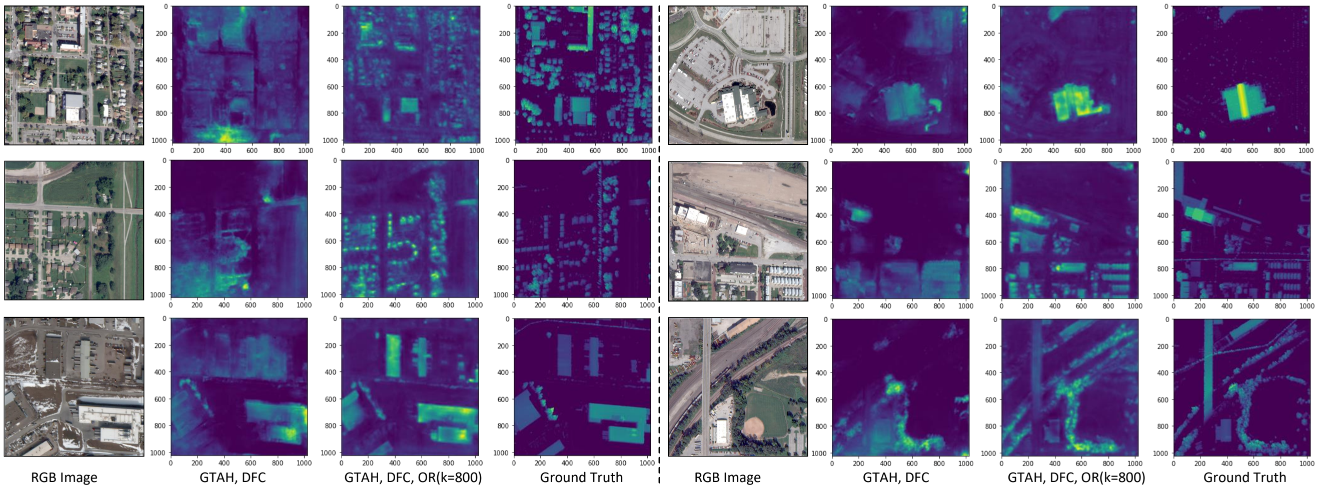

Fig. 3. Visualization of some examples on the DFC 2019 dataset. The columns from left to right are 1) the input image, 2) height maps predicted using pre-trained model on GTAH, 3) height maps predicted using pre-trained model and ordinal relation training, and 4) the ground truth height maps.

5. REFERENCES

[1] Lichao Mou and Xiao Xiang Zhu, "Im2height: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," *arXiv preprint arXiv:1802.10249*, 2018.

[2] Zhitong Xiong, Sining Chen, Yilei Shi, and Xiao Xiang Zhu, "Disentangled latent transformer for interpretable monocular height estimation," 2022.

[3] Ranganath R Navalgund, V Jayaraman, and PS Roy, "Remote sensing applications: an overview," *current science*, pp. 1747–1766, 2007.

[4] Zhitong Xiong, Yuan Yuan, and Qi Wang, "Ai-net: Attention inception neural networks for hyperspectral image classification," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 2647–2650.

[5] Zhitong Xiong, Yuan Yuan, Nianhui Guo, and Qi Wang, "Variational context-deformable convnets for indoor scene parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3992–4002.

[6] Gordon Christie, Rodrigo Rene Rai Munoz Abujder, Kevin Foster, Shea Hagstrom, Gregory D Hager, and Myron Z Brown, "Learning geocentric object pose in oblique monocular images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14512–14520.

[7] Pedram Ghamisi and Naoto Yokoya, "Img2dsm: Height simulation from single imagery using conditional generative adversarial net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 794–798, 2018.

[8] Saket Kunwar, "U-net ensemble for semantic and height estimation using coarse-map initialization," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 4959–4962.

[9] Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D Hager, and Myron Brown, "Semantic stereo for incidental satellite images," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1524–1532.

[10] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng, "Single-image depth perception in the wild," *Advances in neural information processing systems*, vol. 29, pp. 730–738, 2016.

[11] Jae-Han Lee and Chang-Su Kim, "Monocular depth estimation using relative depth maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9729–9738.

[12] Zhitong Xiong, Wei Huang, Jingtao Hu, Yilei Shi, Qi Wang, and Xiao Xiang Zhu, "THE benchmark: Transferable representation learning for monocular height estimation," 2021.

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.

[14] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," *arXiv preprint arXiv:2107.00652*, 2021.

[15] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 418–434.

[16] Bertrand Le Saux, Naoto Yokoya, Ronny Hänsch, and Myron Brown, "Data fusion contest 2019 (dfc2019)," 2019.