Boundary-aware 3D Building Reconstruction from a Single Overhead Image

Jisan Mahmud True Price Akash Bapat Jan-Michael Frahm
University of North Carolina at Chapel Hill
{jisan,jtprice,akash,jmf}@cs.unc.edu

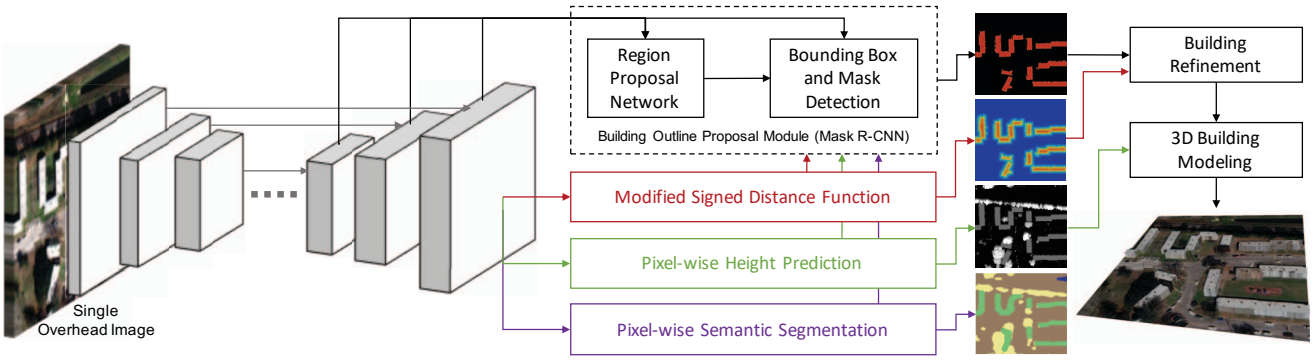

Figure 1: Overview of our method. Given a single overhead image, our multi-task, multi-feature deep network generates 2D building outline proposals, a pixel-wise heightmap, a modified signed distance function (BPSH), and pixel-wise semantic labels. Building outlines are then refined using the BPSH and combined with the height estimates to produce a 3D model.

Abstract

We propose a boundary-aware multi-task deep-learning-based framework for fast 3D building modeling from a single overhead image. Unlike most existing techniques which rely on multiple images for 3D scene modeling, we seek to model the buildings in the scene from a single overhead image by jointly learning a modified signed distance function (SDF) from the building boundaries, a dense heightmap of the scene, and scene semantics. To jointly train for these tasks, we leverage pixel-wise semantic segmentation and normalized digital surface maps (nDSM) as supervision, in addition to labeled building outlines. At test time, buildings in the scene are automatically modeled in 3D using only an input overhead image. We demonstrate an increase in building modeling performance using a multi-feature network architecture that improves building outline detection by considering network features learned for the other jointly learned tasks. We also introduce a novel mechanism for robustly refining instance-specific building outlines using the learned modified SDF. We verify the effectiveness of our method on multiple large-scale satellite and aerial imagery datasets, where we obtain state-of-the-art performance in the 3D building reconstruction task.

1. Introduction

Automated scene understanding and detection from overhead images has long been of interest to the computer vision community [8, 37, 15, 9, 35, 16, 39, 24, 53]. Identifying and modeling 3D buildings from overhead imagery plays an important role in a number of applications. After a major disaster like a hurricane or an earthquake, automated building modeling from overhead imagery, e.g. satellite and aerial images and LIDARs, can provide a vital clue indicating the effect on human settlements and can aid in disaster preparation and assessment. Such modeling can also facilitate urban planning and analysis, digital mapping, overhead surveillance, and city modeling in video games and movies. For this task, satellite and aerial imaging services offer a unique advantage in providing reasonably high-resolution images while remaining relatively cost-effective to capture. In comparison, while aerial or satellite LIDAR can provide highly accurate scene geometry, such scans are often costly to capture and provide low-resolution geometry. Ground-level views are also ineffective: dedicated photography can only efficiently capture a small area, street-view imagery does not necessarily capture buildings from all sides, and city-scale reconstruction requires careful registration of potentially millions of images that still may not completely capture sparsely imaged regions [19, 7]. For large-scale

2575-7075/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPR42600.2020.00052

scene analysis, satellite and aerial imaging provides the best trade-off for overall resolution and spatial coverage.

Many existing methods of reconstruction from satellite or aerial imagery utilize geometric constraints induced by multiple views of the scene and rely on photometric matching [30, 41, 31, 56, 57, 27]. The assumption of similar appearance across images restricts these approaches to use images captured preferably over a short period of time (*e.g.* several days or weeks). In contrast, reconstruction from a single satellite or aerial image does not have this requirement. Single-view approach can provide fast 3D reconstruction while being economic in data capturing. This approach can model sparsely imaged regions, for which one might have just one view of the region. Moreover, obtaining multi-view aerial/satellite images or LIDAR data is typically infeasible for historical remote sensing data [50], making single-image scene understanding and geometry modeling an important problem to solve.

Several methods have proposed to perform building detection and height estimation directly from a single satellite image [29, 40, 55]. This task, however, comes with its own set of challenges. Without the geometric and appearance-matching constraints afforded by having multiple views, modeling — and even detecting — individual buildings is made difficult due to the relatively low ground resolution of the imagery, especially for satellite images.[1] Appearance cues, such as the texture differences between a sidewalk next to a building and the building roof, are often degraded in overhead views. On the other hand, having a prior knowledge of (low-resolution) overhead appearance, and how it relates to semantics and height above the ground, can provide the vital context needed for solving what would otherwise be an ill-posed problem of surface modeling.

We sub-divide the problem of single-view building reconstruction into two sub-problems: (1) detecting 2D outline for each building, and (2) modeling each building's height. To tackle both of these sub-problems together, we propose a multi-task framework that jointly learns four correlated tasks using a deep neural network (Fig. 1):

Task 1: Generate 2D building instance proposals in the form of pixel-wise masks.
Task 2: Predict a modified signed distance function from each building boundary.
Task 3: Predict per-pixel height from the ground (nDSM).
Task 4: Predict pixel-level semantic scene composition.

We propose a technique to resolve overlapping proposals from building detection (Task 1) using learned 2D boundary distance reasoning (Task 2). Notably, we introduce a mixed boundary-label and -distance function, which we call *Boundary Proximity Signed Heatmap* (BPSH) that substantially boosts building outline prediction in Task 1. We propose to learn height regression (Task 3) and semantic segmentation (Task 4) in a joint formulation, which provides additional context for scene understanding. While jointly learning the tasks, we propose a *multi-feature* approach that fuses network features learned for Tasks 2-4 with upstream network features, which serves to improve the instance proposals obtained in Task 1. We demonstrate that our holistic four-part formulation designed to learn generalized feature representations of the scene, along with the novel overlap refinement technique using learned boundary distance reasoning, leads to superior performance in the task of 3D building modeling from a single overhead image.

[1] For instance, Digital Globe's WorldView-3 [4], which is one of the most advanced imaging satellites, captures panchromatic and multi-spectral images with $0.31m$ and $1.24m$ resolution respectively.

2. Related Work

Our multi-task formulation is inspired by several existing works, including approaches in different imaging modalities, and methods that aim to solve a subset of our four tasks. Next, we review works related to estimating 3D geometry from overhead images, including single-view methods; multi-task learning for overhead image understanding; and object instance detection from images.

2.1. Building and ground surface reconstruction

The most common technique for overhead image reconstruction is multi-view stereo [13, 58, 36, 52, 54, 60, 48, 18, 21] using dense image-to-image appearance matching to infer the underlying scene. In [47], Rudner *et al*. use multi-resolution, multi-spectral images from before and after a flood to identify flooded buildings. In contrast to these methods, we target scenarios where reconstruction from a single view is the only viable option.

Historically, techniques for single-view building reconstruction in overhead imagery utilized shadow information from the known pose of the remote camera and the sun-earth relative position. Ok *et al*. [40] use a fuzzy landscape generation approach to model the directional spatial relationship between buildings and their shadows. They detect the building outlines by pruning the non-building regions and using a GrabCut partitioning [46]. Izadi and Saeedi [29] use image primitives such as lines and line intersections, and examine their relationships using a graph search to establish rooftop hypotheses. Height information is then derived from the sun-earth position and shadows. These methods have drawbacks in requiring precise knowledge of the sun-earth relative position, and sun illumination intensity.

A number of works have extended deep-learning approaches for monocular depth estimation from [17, 32] to satellite or aerial domains, sometimes jointly learning an auxiliary task. Wang and Frahm [55] develop a deep framework for parametric building modeling by extending the single-shot multi-box detector (SSD) [34] architecture to 3D space. They predict 2D rectangular building foot-

prints with confidence along each default box and extend the SSD framework by also predicting mean height and orientation of the detected building to generate 3D cuboid building models. This method is limited, however, in its capacity to model non-rectangular buildings. Srivastava *et al*. [50] jointly estimate the nDSM and semantic labeling from monocular satellite images using an encoder-decoder convolutional network. Mou and Zhu [38] propose a similar architecture with skip connections to directly regress height, alone. Mou *et al*. [37] propose spatial relation reasoning for learning semantic segmentation from aerial images. They demonstrate that modeling global relationship between spatial positions and feature maps in networks can provide useful features for segmentation.

2.2. Multi-task learning for overhead imagery

Multi-task CNNs have been shown to boost performance for a variety of correlated tasks compared to single-task architectures [43, 59]. Dai *et al*. [14] design a multi-task network cascade for instance-aware semantic segmentation. Brahmbhatt *et al*. [12] learn convolutional features to predict segmentation between objects and amorphous categories such as ground and water, and utilize this semantic segmentation features at a single stage, for object detection.

In addition to the multi-task ground surface learning methods introduced in the previous subsection, a number of approaches have investigated combining building identification with related tasks for single-view overhead images. Bischke *et al*. [10] and Hui *et al*. [28] jointly learn a binary instance segmentation and a distance function for detecting building outlines from remote sensing images. They show that learning the distance representation guides the network to distinguish between the interior points and boundaries of buildings. Pandey *et al*. [42] train a multi-task CNN to identify indicative factors of urban development and use these features to predict poverty rates across satellite images. Sun *et al*. [51] adopt a similar approach for predicting road topologies, distance functions, and binary masks. In contrast, our approach improves building detection by modeling building outline extraction as an instance detection problem and by learning a novel distance function, which along with the learned scene geometry and scene semantics, provides rich features to the detection task, leading to an overall improvement in the detection performance.

2.3. Object detection

Our method uses Mask R-CNN [25, 6, 33] to generate a set of building outline proposals given an overhead image. Mask R-CNN works by, first, proposing a sparse set of class-agnostic object regions of interest (ROIs) in the image. In the second stage, features are extracted from each of the proposed ROI, and the class of each object is predicted along with its bounding box and mask. Among related work, the precursor R-CNN [22] formulated the classification and localization task in the second stage using a convolutional network, leading to greater accuracy compared to earlier methods. Faster R-CNN [44] formulated both stages with learned sub-networks that utilize the CNN feature map. Mask R-CNN [25] builds on top of this and adds an object mask prediction branch to the classification/localization branch in the second stage. The mask branch predicts one mask for each object category. It also introduces ROIAlign, to avoid any quantization effects when extracting features from the ROIs, allowing for the generation of pixel-wise-accurate masks. Recently, Fu *et al*. [20] demonstrated that object detection prediction can provide good features for semantic segmentation, as well.

3. Our Approach

Given a single satellite or aerial image, we develop a multi-task, multi-feature, and building-boundary-aware deep-learning framework to solve the problem of 3D building modeling. For best performance, we expect overhead images to be captured from on or close to nadir views.

As mentioned in the introduction, we design a deep network to jointly learn four tasks that are jointly trained in an end-to-end fashion (Fig. 2), with shared feature representations serving as the backbone for each individual task prediction. We propose to use feature representations for Tasks 2-4 to provide rich high-level information to Task 1 for learning more robust initial building outline proposals. By design, all four tasks are intertwined and work together to improve contextual information for building identification. By predicting boundaries, recovering ground and building surfaces, and identifying building pixels versus surrounding objects like trees — all within an object detection framework — our method can accurately identify, localize, and ultimately model the buildings in a given image.

We address the first two tasks in the next subsection and then describe how we solve the remaining tasks. Finally, we present our multi-feature learning approach, plus a technique to refine the boundary prediction with instance-level information, both of which improve the final reconstruction.

3.1. Building outline detection

We formulate the estimation of a building outline as a 2D object detection problem. These initial detection proposals (Task 1) are subsequently refined using a novel modified signed distance function (BPSH, Task 2) learned by our network. While the building detection proposal is tasked with identifying building instances, the BPSH learning is designed to sharply learn the boundaries of the instances, especially for buildings close to each other. Our experiments show that learning a shared feature representation for these tasks improves performance for both tasks.

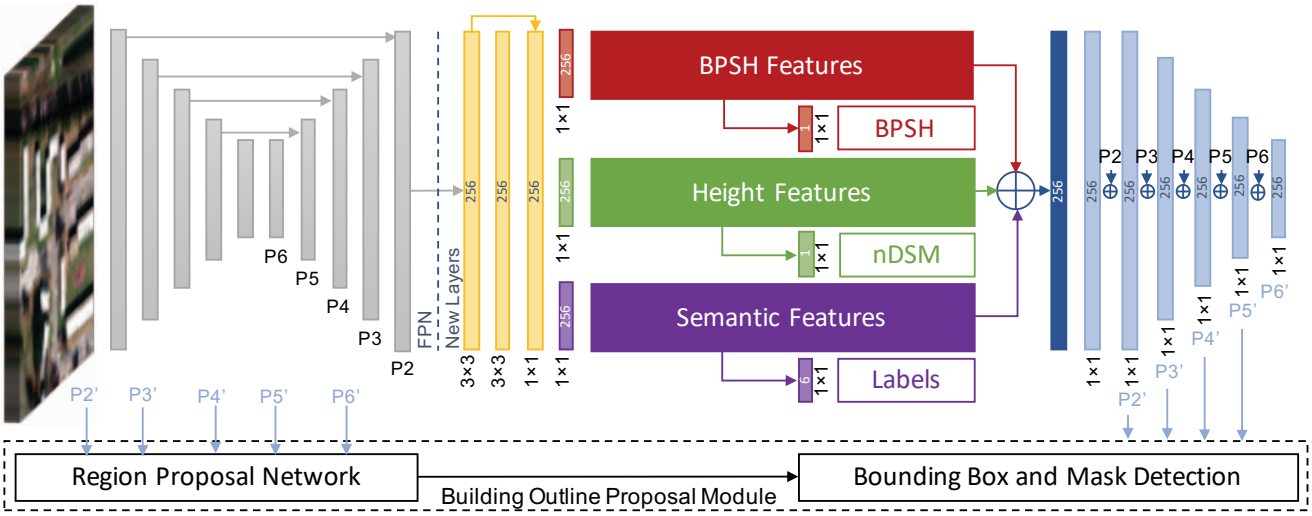

Figure 2: Our final multi-task, multi-feature learning framework. Features from BPSH prediction, nDSM prediction, and semantic segmentation prediction tasks are added to the features of FPN at different scales to aid the building outline proposal.

3.1.1 Building proposal generation

We use the Mask R-CNN [25, 6] framework to generate initial building proposals. In contrast to regular multi-class detection, we are only interested in a single class of objects: buildings. The Feature Pyramid Network (FPN) [33] built on top of ResNet-101 [26] is used as the backbone of Mask R-CNN. FPN uses a top-down architecture with lateral connections to build an in-network feature pyramid from a single-scale input. This creates high-level semantic feature maps with fine details at different scales, each of which is used to generate a set of foreground regions of interest (ROI) proposals using a region proposal network. Features for each ROI are then extracted and used to predict a building-label confidence, bounding box, and a (28×28) building mask. The generated masks with high confidence give us an initial set of 2D building outline proposals.

3.1.2 Signed distance function regression

Mask R-CNN often yields overlapping building instance proposals for buildings positioned close to each other. In practice, however, buildings will rarely overlap in near-nadir overhead images. To resolve true building proposals from the overlapped ones, one naïve approach is to use an extreme non-maximum suppression, removing all proposals that overlap a proposal of higher confidence. However, the low ground resolution of overhead images usually leads to lower-confidence proposals for smaller buildings. As a result, smaller buildings near larger buildings are often suppressed, reducing the overall detection recall drastically.

Task 2 of our multi-task learning framework tackles this problem. We learn to regress a modified truncated signed

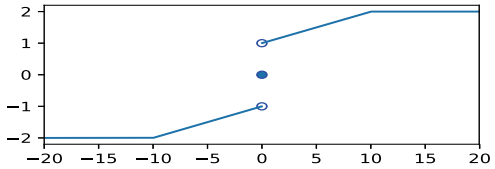

Figure 3: BPSH function. X-axis: signed distance in pixels from the closest building boundary. Y-axis: BPSH score.

distance function – the *boundary proximity signed heatmap* (BPSH) – from individual building boundaries. The BPSH is similar to a traditional 2D truncated signed distance for the building boundaries in the scene, except that it clearly distinguishes the zero level set (Fig. 3). The BPSH of pixel p is positive inside of buildings and negative outside:

$$BPSH(p) = \begin{cases} 0 & \text{if } D_b(p) = 0 \\ 1 + \frac{\min(D_b(p),\tau)}{\tau} & \text{if } p \text{ is inside} \\ -1 - \frac{\min(D_b(p),\tau)}{\tau} & \text{if } p \text{ is outside} \end{cases} \quad (1)$$

Here, τ is the truncation distance, set to 10 px for all of our experiments. $D_b(p)$ is the Euclidean distance to the nearest pixel that lies on a building boundary. The behavior is similar to a ternary labeling function, with the additional context of nearness to the boundary. Thus, learning the BPSH encourages the network to learn the outlines of the buildings.

We create a new sub-network for predicting the BPSH from the spatially largest layer of FPN [33] ($P2$ in Fig. 2). $P2$ is $4\times$ down-sampled compared to the original image dimension; our predicted BPSH has this dimension. From $P2$, we extract a shared feature representation (see supplementary) used for Tasks 2, 3, and 4. Two further 1×1 convolutions finally generate the BPSH prediction.

3.1.3 Overlap refinement using the BPSH

During inference, we generate the BPSH prediction along with the building outline proposals from Mask R-CNN. We extract the BPSH zero level set as all pixels p having BPSH$(p) \in [-0.5, 0.5]$. A score S_i for each proposal is then computed as the sum of its detection confidence c_i and the agreement between its mask and BPSH:

$$S_i = c_i + \max\left(1 - \frac{\lambda}{|M_i|}\sum_{p \in M_i} |D_M(p) - D_B(p)|, 0\right), \quad (2)$$

where M_i is the set of building-labeled pixels in the proposed mask, $D_M(p)$ is the distance of pixel p from the mask's building boundary, and $D_B(p)$ is the distance to the BPSH zero-level set. In our experiments, we use $\lambda = 0.1$. By design, the boundary-agreement term in S_i promotes correctly proposed smaller buildings that overlap with other incorrect larger proposals, since these smaller buildings are likely to have higher conformity to the BPSH zero-level set.

Based on these scores, we run non-maximum suppression (NMS), removing the proposals that overlap other, higher-scored proposals. This NMS retains the buildings that have high prediction confidence and higher conformity to the predicted BPSH. However, while NMS can correctly suppress low-score proposals that overlap with higher-scored correct proposals, the situation can arise where an incorrect proposal overlaps with a lower-scored correct one, but both are removed by NMS. Thus, after NMS, we add back non-overlapping suppressed proposals where both the building outline detector and the BPSH predict a building with high confidence (see supplementary). We find that this step greatly improves the recall of our final detection.

3.2. Building height generation

The second sub-problem we tackle is generating the height of each detected building (Task 3). Our framework predicts the per-pixel height from the ground, known as the normalized digital surface model (nDSM). The nDSM sub-network is similar to the BPSH sub-network and utilizes the same shared feature representation obtained by the three convolutions applied after layer $P2$. We then use two task-specific 1×1 convolutions to generate the height prediction.

3.3. Semantic segmentation

We learn a pixel-wise semantic segmentation for the classes of building, ground, water, high vegetation and low vegetation. As before, two 1×1 convolutions are applied on the shared feature representation to generate the pixel-wise joint class probability distribution. We show that training for semantic segmentation (when such data is available) along with the other three tasks improves the building outline detection; see Sec. 5.2 for an ablation study.

3.4. Multi-feature learning

In the original Mask R-CNN with a FPN backend [33], layers $P6$ down to $P2$ (Fig. 2) are used to generate the region proposals, and the feature maps of $P5$ down to $P2$ are used to generate the second-stage predictions of classification, bounding-box regression, and mask prediction at different scales. Instead of using these layers to generate proposals directly, we combine the high-level features from the three other tasks just before the final prediction layers, and fuse them together at different scales ($P2$ through $P6$) to generate $P2'$, $P3'$, $P4'$, $P5'$, and $P6'$ (Fig. 2). These augmented layers carry rich contextual information about the scene, as well as features from the semantic segmentation, nDSM, and BPSH predictions at different scales. The region proposal network and ROI-specific network in turn can utilize this to generate richer sets of building proposals.

3.5. Instance-level reasoning to improve BPSH

The building outline prediction task is modeled as an object instance detection problem, utilizing instance-level reasoning. On the other hand, the BPSH seeks to sharply learn the building boundaries in a pixel-level reasoning fashion. To boost the final boundary prediction, we propose to fuse the two predicted modalities together in a post-processing stage to induce instance-level reasoning explicitly into the learned BPSH. We use a small skip-connected encoder-decoder network that takes as input the rasterized predictions of building outlines (Task 1), predicted BPSH (Task 2), and original image. The network outputs a refined BPSH that is improved by the context of the rasterized mask. We apply the final predicted BPSH for the overlap refinement task (Sec. 3.1.3) to generate our final set of building predictions. This gives a slight, but notable, increase to the final accuracy. Note that this second network is trained separately from our primary multi-task, multi-feature network.

4. Network Training

We use ground-truth building masks, BPSH maps defined from these masks, ground-truth nDSMs, and ground-truth semantic label maps to train our network. When nDSMs and/or semantic labels are not available (*i.e.* in the SpaceNet dataset), we train only using building masks and BPSH maps. We next detail our loss functions for training.

Overall building estimation loss Our multi-task framework optimizes our four tasks together in an end-to-end fashion. The overall loss function is a combination of the loss functions of individual tasks:

$$Loss = \alpha_1 L_{outline} + \alpha_2 L_{bpsh} + \alpha_3 L_{ndsm} + \alpha_4 L_{sem}. \quad (3)$$

We use $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (1, 2, 3, 2.5)$, which were experimentally chosen by analyzing the training loss to approximately balance the overall loss contributions of the tasks.

Building outline detection loss $L_{outline}$ penalizes the error in our initial 2D building outline proposal. We use the same loss as Mask R-CNN [25] that considers the region proposal loss and ROI-specific classification, regression, and mask prediction loss. For ROI R_i, loss $L^{R_i}_{detection} = L^{R_i}_{cls} + L^{R_i}_{box} + L^{R_i}_{mask}$. Here, L_{cls} is the building-label cross-entropy classification loss, L_{box} penalizes the axis-aligned bounding box using a *smooth* L_1 loss, and L_{mask} is the mean binary cross-entropy loss over the (28×28) prediction window. The latter two losses are defined for candidate ROIs that are assigned a ground-truth building. The final loss, $L_{outline}$, combines $L_{detection}$ and the region-proposal loss [44]. We use "approximate" joint training [44], where the partial gradient of the *ROIAlign* layer is computed while ignoring the gradient w.r.t. ROI coordinates.

BPSH prediction loss L_{bpsh} robustly penalizes the BPSH error $E_{bpsh}(p) = BPSH_{gt}(p) - BPSH_{pred}(p)$ at each output pixel p:

$$L_{bpsh} = \frac{1}{N} \sum_p W_{bpsh}(p) \cdot \text{SmoothL}_1 \left(E_{bpsh}(p) \right). \quad (4)$$

Here, N is the number of pixels in the output image, and W_{bpsh} is a set of per-pixel weights. Each weight is a combination of two different weighting functions:

$$W^{(1)}_{bpsh}(p) = \exp\left(-\frac{BPSH_{gt}(p)^2}{2\sigma^2_{bpsh}} \right), \quad (5)$$

$$W^{(2)}_{bpsh}(p) = \exp\left(-\frac{(d_1(p) + d_2(p))^2}{2\sigma^2_{unet}} \right). \quad (6)$$

The first weighting function gives higher emphasis on the zero level set of the BPSH. We use $\sigma_{bpsh} = 2$ for our experiments. The higher the value of σ_{bpsh}, the lower the emphasis on the zero level set. The second weighting function is inspired by U-Net [45]. Here, $d_1(p)$ denotes the distance (in pixels, in the input image at its original resolution) to the nearest ground-truth building boundary, and $d_2(p)$ is the distance to the border of the second-nearest building's boundary. We set $\sigma_{unet} = 5$. U-Net weighting puts high emphasis on the pixels that are between two building boundaries close to each other. The final BPSH weight is

$$W_{bpsh}(p) = W^{(1)}_{bpsh}(p) + \alpha_{bpsh} \cdot W^{(2)}_{bpsh}(p). \quad (7)$$

Following [45], we set $\alpha_{bpsh} = 10$ for our experiments. This weighting forces the network to learn the building boundaries with high importance, while also emphasizing the pixels that are between two nearby buildings. Our BPSH refinement network (Sec. 3.5) is also trained using this loss.

nDSM prediction loss L_{ndsm} penalizes the height prediction error $E_{ndsm}(p) = NDSM_{gt}(p) - NDSM_{pred}(p)$:

$$L_{ndsm} = \frac{1}{N} \sum_p W_{ndsm}(p) \cdot L'(E_{ndsm}(p)). \quad (8)$$

$W_{ndsm}(p)$ prioritizes building height prediction by up-weighting building pixels. We use a heuristic weight of 5 for ground-truth building pixels and 1 for all other pixels. For L', we initially use the *BerHu* loss [32] for fast convergence, and then switch to the *smooth* L_1 loss for fine-tuning.

Semantic segmentation loss L_{sem} evaluates the average softmax cross-entropy loss for the pixel-wise class predictions. Due to class imbalance for the different semantic labels, the loss at each pixel is weighted by the label's inverse frequency among all pixels in the training set.

5. Experiments

We evaluate our approach on three large-scale satellite datasets: the 2019 IEEE GRSS Data Fusion Contest dataset (GRSS_DFC_2019) [11, 1, 49] containing images, semantic segmentations, and nDSMs (train/test split of 92/16 regions), USSOCOM Urban 3D dataset [23] containing images and nDSMs (130/44 regions), and the SpaceNet Buildings Dataset v2 [5] containing building outlines (7128/1254 images). We also evaluate on two aerial imagery datasets containing images, semantic segmentation, and nDSMs: Potsdam [2] (10/7 regions) and Vaihingen [3] (11/5 regions). A more detailed discussion of these datasets and train-test splits can be found in our supplementary material.

5.1. Evaluation

Table 1 shows results for the SpaceNet and GRSS_DFC_2019 datasets, comparing our proposed method versus the state-of-the-art methods of Wang and Frahm [55], Mou and Zhu [38], and Srivastava *et al.* [50], each of which performs a subset of our four network tasks. We use the GRSS_DFC_2019 dataset to evaluate height, 2D building outlines, and semantic labels. We evaluate 2D building outlines for the SpaceNet dataset, which does not have height or semantic data. For 2D building outline evaluation, we compute the F1 score at an intersection over union (IoU) threshold of 0.5. Building height errors are evaluated using mean absolute error (MAE) in meters and root-mean-squared error (RMSE) in meters; non-building ground-truth pixels are not considered in this metric. In addition to evaluating pixel-wise height regression, we also evaluate median building height regression to account for small misalignments between the overhead imagery and ground-truth labels. Semantic segmentation is evaluated using F1 for the dominant classes of building, ground, and vegetation. In all cases, we demonstrate that our proposed learning framework leads to better accuracy for 2D building outline detection, height regression, and semantic segmentation, often with substantial gains over the state of the art. See Fig. 4 and our supplementary material for a qualitative comparison of the different methods.

Wang and Frahm's [55] method predicts building outlines and heights according to cuboid models. However,

Method	SpaceNet	GRSS_DFC_2019							
	Bldg. Outline	Bldg. Outline	Median Height		Pixelwise Height		Semantic Segmentation (F1)		
	F1	F1	MAE	RMSE	MAE	RMSE	Building	Ground	Tree
Ours	**68.87**	**68.34**	**1.85**	**2.79**	**3.34**	**5.02**	**94.2**	**95.2**	**81.0**
Wang & Frahm [55]	61.60	57.86	1.89	2.94	-	-	-	-	-
Mou & Zhu [38]	-	-	2.26	3.19	3.62	5.40	-	-	-
Srivastava *et al*. [50]	-	-	2.45	3.59	3.74	5.85	76.8	92.6	76.6

Table 1: Our method achieves higher F1 scores for 2D building outline detection in single-view satellite image datasets versus state-of-the-art methods, indicating its superior performance. We also achieve lower MAE and RMSE in median and pixelwise building height prediction, and we show superior performance in class-wise F1 scores for semantic segmentation. Building median height is evaluated to account for small misalignments between the images and ground-truth labels.

Method	Outline	Median Height		Pixelwise Height	
	F1	MAE	RMSE	MAE	RMSE
Ours	**82.89**	**1.05**	2.25	**2.34**	**6.15**
[55]	69.98	1.06	2.35	-	-
[38]	-	**1.05**	**2.24**	2.35	6.62
[50]	-	1.31	2.64	2.90	7.70

Table 2: Results on the Urban 3D dataset. Our method achieves the best performance in building outline detection and comparable performance in the height prediction.

	SpaceNet	GRSS_DFC_2019
	Outline F1	Outline F1
MRCNN	65.0	63.3
+ TSDF	65.6	62.7
+ BPSH	**66.4**	64.3
+ Sem. Seg.	-	65.1
+ nDSM	-	66.9
+ Multi-feat.	-	**67.5**

Table 3: Ablative results for our multi-task method, with each row adding an additional network component to the previous row. "+ TSDF" is a comparison to "+ BPSH". The "Mask R-CNN" result does not include overlap refinement.

we found that this approach did not generalize well in our datasets, since many of the buildings do not have a rectangular footprint as required by their method, leading to poor detections. To improve their method's competitive performance, we removed the building orientation regression output by their approach. We also changed their detection method to Faster R-CNN [44] with ROIAlign and FPN, giving it a more competitive performance than SSD [34]. Tables 2 and 4 demonstrate the performance of our proposed model against these methods on the Urban 3D dataset and the Potsdam and Vaihingen datasets, respectively. In all cases, we obtain building reconstruction with higher outline detection and height prediction accuracy. Our auxiliary task compares favorably to the state-of-the-art semantic segmentation results on the Potsdam and Vaihingen datasets. We especially demonstrate the superior performance in the F1 score for building detection against [55].

3D Modeling. We generate 3D models using the extracted building outlines and the median nDSM height per building. Fig. 4 shows these models for various inputs. Additional results can be found in our supplementary material.

5.2. Ablation

To evaluate the overall effectiveness of our framework, we perform an ablation analysis of our network with different sub-tasks activated. Table 3 shows these results, starting from a baseline Mask R-CNN building detection network and successively adding BPSH regression, semantic segmentation, height regression, and our expanded multi-feature architecture. It can be observed that as new tasks are added to the multitasking framework, the final building outline detection accuracy consistently improves. We observe the best performance using the multi-feature architecture, which supports our hypothesis that the features related to semantic segmentation, height, and BPSH prediction at different scales provide a richer set of features for the building outline proposal task.

Table 3 (second row) shows the result of learning truncated signed distance function (TSDF) instead of the BPSH. We used a TSDF cutoff of 10 px, linearly scaled to the range [-1, 1]. We found the BPSH to outperform the TSDF when trained for a similar number of epochs. We hypothesize that this is due to the ternary behavior of the BPSH: cost in misidentifying a building boundary is higher for the BPSH than the TSDF, and thus the TSDF network is not inclined to converge as fast to finely localize building outlines. When trained for a much longer period of time, we noticed a similar performance between these two techniques. See the supplementary materials for more ablative analysis.

6. Conclusion

We presented a multi-task, multi-feature learning formulation for 3D building modeling from a single overhead image. Unlike the existing multi-tasking-learning-based formulations for building footprint detection, we additionally utilize scene geometry and semantics learning to robustly

Dataset	Method	Bldg. Outline	Median Height		Pixelwise Height		Semantic Segmentation (F1)		
		F1	MAE	RMSE	MAE	RMSE	Building	Impervious	Tree
Potsdam	Ours	**71.98**	**1.86**	**2.75**	**2.55**	**3.73**	**97.31**	**92.09**	80.36
	Wang & Frahm [55]	57.97	1.96	2.88	-	-	-	-	-
	Mou & Zhu [38]	-	2.57	3.48	3.32	4.26	-	-	-
	Srivastava *et al.* [50]	-	2.63	3.59	3.6	4.67	93.93	87.27	76.16
	Mou *et al.* [37]	-	-	-	-	-	94.70	91.33	**83.47**
Vaihingen	Ours	**72.85**	**1.10**	**1.51**	**1.43**	**1.93**	**97.33**	**92.11**	87.56
	Wang & Frahm [55]	60.70	1.17	1.55	-	-	-	-	-
	Mou & Zhu [38]	-	1.34	1.80	1.74	2.30	-	-	-
	Srivastava *et al.* [50]	-	1.57	2.05	2.02	2.59	95.56	88.98	88.09
	Mou *et al.* [37]	-	-	-	-	-	94.97	91.47	**88.57**

Table 4: For the Potsdam and Vaihingen aerial datasets, our method achieves state-of-the-art performance in building detection and height regression. We also demonstrate that the auxiliary semantic segmentation task learned by our method has competitive performance for different semantic object categories, and much superior performance for building segmentation.

Figure 4: Comparison of building mask predictions for [55], [50], and our method. Right: Our 3D reconstruction.

detect building footprints. Our multi-feature formulation demonstrates that high-level features from these learned tasks provide rich information to the detector, improving the detection performance. Our boundary-aware approach to BPSH prediction, as well as our overlap and BPSH refinement techniques, also boost performance substantially.

Several avenues of future work arise from the method presented here. The BPSH with its ternary structure has potential use in general applications that require accurate boundary prediction. Our approach to overlap refinement can also be extended in other detection modalities where targets should not overlap each other, such as overhead crowd counting and clustered object detection. Finally, we anticipate that mixed object-detection and 3D-reconstruction frameworks will continue to show mutual benefits, especially for single-view reconstruction tasks.

Acknowledgement: The authors would like to thank the Johns Hopkins University Applied Physics Laboratory and IARPA for providing some data used in this study, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee for organizing the Data Fusion Contest. Satellite imagery courtesy of DigitalGlobe.

References

[1] 2019 IEEE GRSS Data Fusion Contest. http://www.grss-ieee.org/community/technical-committees/data-fusion. Accessed on: 2019-03-15. 6

[2] 2d semantic labeling contest - potsdam. http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html. Last modified: 2019-10-23, Accessed on: 2019-10-23. 6

[3] 2d semantic labeling contest - vaihingen. http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html. Last modified: 2019-10-23, Accessed on: 2019-10-23. 6

[4] DigitalGlobe. https://www.digitalglobe.com. Accessed on: 2019-03-10. 2

[5] SpaceNet on Amazon Web Services (AWS). datasets. the spacenet catalog. https://spacenetchallenge.github.io/datasets/datasetHomePage.html. Last modified: 2018-04-30, Accessed on: 2019-03-15. 6

[6] Waleed Abdulla. Mask R-CNN for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017. 3, 4

[7] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011. 1

[8] Seyed Majid Azimi, Corentin Henry, Lars Sommer, Arne Schumann, and Eleonora Vig. Skyscapes fine-grained semantic understanding of aerial scenes. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1

[9] Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David DeWitt. Roadtracer: Automatic extraction of road networks from aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[10] Benjamin Bischke, Patrick Helber, Joachim Folz, Damian Borth, and Andreas Dengel. Multi-task learning for segmentation of building footprints with deep neural networks. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1480–1484. IEEE, 2019. 3

[11] Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D. Hager, and Myron Z. Brown. Semantic stereo for incidental satellite images. *CoRR*, abs/1811.08739, 2018. 6

[12] Samarth Brahmbhatt, Henrik I Christensen, and James Hays. Stuffnet: Using stuff to improve object detection. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 934–943. IEEE, 2017. 3

[13] Randi Cabezas, Julian Straub, and John W Fisher. Semantically-aware aerial reconstruction from multi-modal data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2156–2164, 2015. 2

[14] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3158, 2016. 3

[15] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[16] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1

[17] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2650–2658, 2015. 2

[18] Gabriele Facciolo, Carlo De Franchis, and Enric Meinhardt-Llopis. Automatic 3D reconstruction from multi-date satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 57–66, 2017. 2

[19] Jan-Michael Frahm, Pierre Fite Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, and Svetlana Lazebnik. Building rome on a cloudless day. In *ECCV*, 2010. 1

[20] Cheng-Yang Fu, Tamara L. Berg, and Alexander C. Berg. Imp: Instance mask projection for high accuracy semantic segmentation of things. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 3

[21] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. DSSD: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017. 2

[22] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 580–587, 2014. 3

[23] Hirsh Goldberg, Myron Brown, and Sean Wang. A benchmark for building footprint classification using orthorectified rgb imagery and digital surface models from commercial satellites. In *Proceedings of IEEE Applied Imagery Pattern Recognition Workshop 2017*, 2017. 6

[24] Sergey Golovanov, Rauf Kurbanov, Aleksey Artamonov, Alex Davydow, and Sergey Nikolenko. Building detection from satellite imagery using a composite loss function. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 1

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3, 4, 6

[26] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. 4

[27] Benjamin Hepp, Matthias Nießner, and Otmar Hilliges. Plan3d: Viewpoint and trajectory optimization for aerial multi-view stereo reconstruction. *ACM Transactions on Graphics (TOG)*, 38(1):4, 2018. 2

[28] Jian Hui, Mengkun Du, Xin Ye, Qiming Qin, and Juan Sui. Effective building extraction from high-resolution remote sensing images with multitask driven deep neural network. *IEEE Geoscience and Remote Sensing Letters*, 2018. 3

[29] Mohammad Izadi and Parvaneh Saeedi. Three-dimensional polygonal building model estimation from single satellite images. *IEEE Transactions on Geoscience and Remote Sensing - IEEE TRANS GEOSCI REMOT SEN*, 50:2254–2272, 06 2012. 2

[30] Jin Ryong Kim and Jocimar Prates Muller. 3d reconstruction from very high resolution satellite stereo and its application to object identification. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 34, 01 2002. 2

[31] Georg Kuschk. Model-free dense stereo reconstruction for creating realistic 3d city models. *Joint Urban Remote Sensing Event 2013*, pages 202–205, 2013. 2

[32] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *Fourth International Conference on 3D Vision (3DV)*, pages 239–248. IEEE, 2016. 2, 6

[33] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 3, 4, 5

[34] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision*, 2016. 2, 7

[35] Gellert Mattyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[36] Z. M. Moratto, M. J. Broxton, R. A. Beyer, M. Lundy, and K. Husmann. Ames Stereo Pipeline, NASA's Open Source Automated Stereogrammetry Software. In *Lunar and Planetary Science Conference*, volume 41 of *Lunar and Planetary Inst. Technical Report*, page 2364, Mar. 2010. 2

[37] Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12416–12425, 2019. 1, 3, 8

[38] 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. 3, 6, 7, 8

[39] Yoni Nachmany and Hamed Alemohammad. Detecting roads from satellite imagery in the developing world. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 1

[40] Ali Ozgun Ok, Caglar Senaras, and Baris Yuksel. Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 51(3):1701–1717, 2013. 2

[41] Özge Can Özcanli, Yi Dong, Joseph L. Mundy, Helen F. Webb, Riad I. Hammoud, and Victor Tom. Automatic geolocation correction of satellite imagery. *International Journal of Computer Vision*, 116:263–277, 2015. 2

[42] Shailesh M Pandey, Tushar Agarwal, and Narayanan C Krishnan. Multi-task deep learning for predicting poverty from satellite images. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 3

[43] Anastasia Pentina, Viktoriia Sharmanska, and Christoph H Lampert. Curriculum learning of multiple tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5492–5500, 2015. 3

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 3, 6, 7

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 6

[46] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "GrabCut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, Aug. 2004. 2

[47] Tim G. J. Rudner, Marc Rußwurm, Jakub Fil, Ramona Pelich, Benjamin Bischke, Veronika Kopacková, and Piotr Bilinski. Multi^{3}net: Segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. *CoRR*, abs/1812.01756, 2018. 2

[48] Ewelina Rupnik, Marc Pierrot-Deseilligny, and Arthur Delorme. 3D reconstruction from multi-view VHR-satellite images in MicMac. *ISPRS Journal of Photogrammetry and Remote Sensing*, 139:201–211, 2018. 2

[49] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *arXiv preprint arXiv:1906.07789*, 2019. 6

[50] Shivangi Srivastava, Michele Volpi, and Devis Tuia. Joint height estimation and semantic labeling of monocular aerial images with cnns. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5173–5176. IEEE, 2017. 2, 3, 6, 7, 8

[51] Tao Sun, Zehui Chen, Wenxiang Yang, and Yin Wang. Stacked u-nets with multi-output for road extraction. In *CVPR Workshops*, pages 202–206, 2018. 3

[52] C Vincent Tao and Yong Hu. 3D reconstruction methods. *Photogrammetric Engineering & Remote Sensing*, 68(7):705–714, 2002. 2

[53] Vivek Verma, Rakesh Kumar, and Stephen Hsu. 3d building detection and modeling from aerial lidar data. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2213–2220. IEEE, 2006. 1

[54] Ke Wang and Jan-Michael Frahm. Fast and accurate satellite multi-view stereo using edge-aware interpolation. In *2017 International Conference on 3D Vision (3DV)*, pages 365–373. IEEE, 2017. 2

[55] Ke Wang and Jan-Michael Frahm. Single view parametric building reconstruction from satellite imagery. In *2017 International Conference on 3D Vision (3DV)*, pages 603–611. IEEE, 2017. 2, 6, 7, 8

[56] Ke Wang, Craig Stutts, Enrique Dunn, and Jan-Michael Frahm. Efficient joint stereo estimation and land usage classification for multiview satellite data. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016. 2

[57] Bin Wu, Xian Sun, Qichang Wu, Menglong Yan, Hongqi Wang, and Kun Fu. Building reconstruction from high-resolution multiview aerial imagery. *IEEE Geoscience and Remote Sensing Letters*, 12(4):855–859, 2014. 2

[58] Xiuchuan Xie, Tao Yang, Jing Li, Qiang Ren, and Yanning Zhang. Fast and seamless large-scale aerial 3d reconstruction using graph framework. In *Proceedings of the 2018 International Conference on Image and Graphics Processing*, pages 126–130. ACM, 2018. 2

[59] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. Rotating your face using multi-task deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 676–684, 2015. 3

[60] Enliang Zheng, Ke Wang, Enrique Dunn, and Jan-Michael Frahm. Minimal solvers for 3d geometry from satellite imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 738–746, 2015. 2