

SEMANTIC-AWARE NETWORK FOR AERIAL-TO-GROUND IMAGE SYNTHESIS

Jinhyun Jang Taeyong Song Kwanghoon Sohn*

School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
E-mail: khsohn@yonsei.ac.kr

ABSTRACT

Aerial-to-ground image synthesis is an emerging and challenging problem that aims to synthesize a ground image from an aerial image. Due to the highly different layout and object representation between the aerial and ground images, existing approaches usually fail to transfer the components of the aerial scene into the ground scene. In this paper, we propose a novel framework to explore the challenges by imposing enhanced structural alignment and semantic awareness. We introduce a novel semantic-attentive feature transformation module that allows to reconstruct the complex geographic structures by aligning the aerial feature to the ground layout. Furthermore, we propose semantic-aware loss functions by leveraging a pre-trained segmentation network. The network is enforced to synthesize realistic objects across various classes by separately calculating losses for different classes and balancing them. Extensive experiments including comparisons with previous methods and ablation studies show the effectiveness of the proposed framework both qualitatively and quantitatively. The code is publicly available at <https://github.com/jinhyunj/SANet>.

Index Terms— Aerial-to-ground image synthesis, transformation, semantic segmentation

1. INTRODUCTION

Aerial-to-ground image synthesis aims to predict corresponding ground-view image at a given aerial-view image. It has received significant attention in the computer vision community as it can be applied to various media industries, including wide-area virtual scene generation, 3D simulation, and gaming. However, it is a very challenging task since the aerial and ground images have an extremely different viewpoints, which makes the scene layouts and object representations in the two images completely different.

Recently, there have been attempts [1, 2, 3, 4, 5] to solve the problem by leveraging generative adversarial networks (GANs) [6, 7]. Few methods [1, 2] impose a ground semantic map as a conditional input for the ground image. However, these methods require semantic maps at the testing phase and the synthesized images are strongly conditioned on them, as in example-guided image synthesis methods [8, 9]. Deng *et al.* [3, 4] adopted conditional GANs [7] that use vector representation extracted from aerial image to produce an appropriate ground image. Regmi and Borji [5] proposed two models (X-Fork and X-Seq) that jointly generate ground images and corresponding semantic maps. Although these works have shown plausible results, they do not handle the structural difference

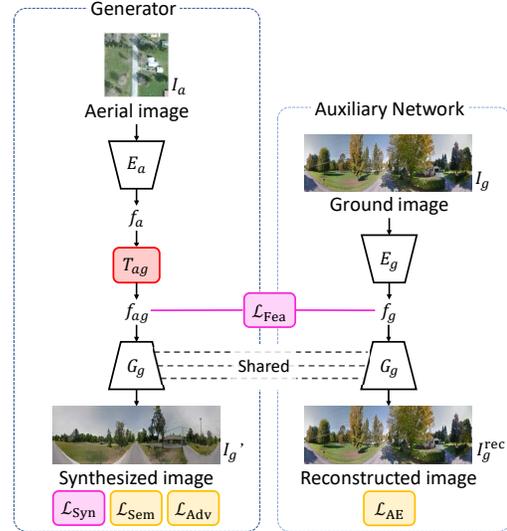


Fig. 1. Overview of our proposed framework. Our aerial-to-ground generator is composed of an aerial encoder, a semantic-attentive feature transformation module, and a ground decoder. It synthesizes a panoramic ground image given an aerial image. The generator is trained using auxiliary ground autoencoder and semantic-aware loss functions.

between the viewpoints or separately consider objects in different semantic classes, resulting in limited performance in difficult scenes which contain multiple objects and complex layout.

Other methods [10, 11, 12] focus on transformation to convert the aerial scene layout into ground perspective. They reduce the geometric difference between two views and mitigate the structural deformation problem. Zhai *et al.* [10] proposed to learn a transformation matrix that turns aerial image into ground-view panorama image. Regmi and Borji [11] applied homography transformation to the aerial images and use them as inputs to synthesize the ground images. These methods adopt coarse alignments of the entire scene layout and often fail to capture detailed transformations, yielding unsatisfactory results. Lu *et al.* [12] proposed a differentiable geo-transformation layer based on orthogonal projection and panoramic rays by using aerial semantic and depth maps. While this method has achieved great success, it is restricted to cases where a large number of ground truth semantic and depth maps are available.

In this paper, we propose a novel framework that imposes enhanced structural alignment with semantic awareness for aerial-to-ground image synthesis. We argue that handling the entire scene at once is insufficient for this complex synthesis problem and therefore, explore semantically different object respectively. To be specific, we introduce a semantic-attentive feature transformation module to align the aerial features into the ground layout. By exploiting at

*Corresponding author

This work was supported by Institute of Information communications Technology Planning & Evaluation (IITP) grand funded by the Korea government(MSIT) (No.2020-0-00056, To create AI systems that act appropriately and effectively in novel situations that occur in open worlds.)

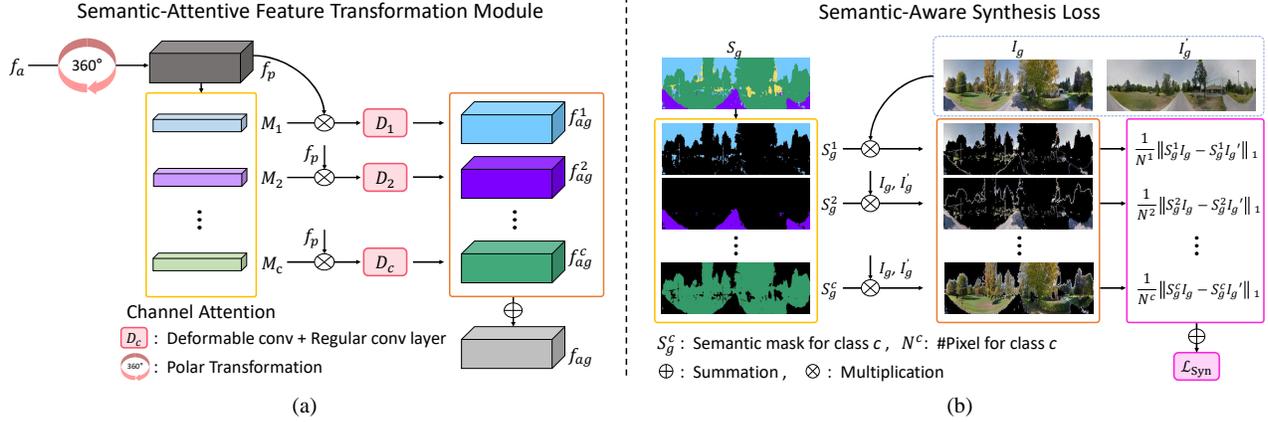


Fig. 2. Illustration of the proposed transformation module and semantic-aware loss function. (a) Our transformation module aligns the structure of the aerial feature to the ground by applying transformation with respect to the semantic classes. **(b)** Semantic awareness of our generator is enforced by a novel loss function which balances the losses across the semantic class.

attention mechanism [13], it separately transforms the features with different semantic representations to achieve better alignment. Furthermore, we present semantic-aware loss functions to handle the objects across the various semantic classes. Being aware that objects with different classes appear in uneven number of pixels and scenes, we balance the losses for different classes by leveraging a pre-trained semantic segmentation network. Experimental results on CVUSA [14] and CVACT [15] datasets demonstrate the effectiveness of the proposed method.

2. PROPOSED METHOD

Fig. 1 illustrates our overall framework. Our goal is to train a deep network that synthesizes a plausible ground panorama image I_g given an aerial image I_a . Based on intuition that the overall layouts and semantics for different objects should be considered for synthesizing realistic images [16], we propose a semantic-attentive feature transformation module and semantic-aware loss functions.

2.1. Network Architecture

Our aerial-to-ground synthesis network, *i.e.*, generator, is composed of an aerial encoder $E_a : I_a \rightarrow f_a$ for mapping aerial image into feature space, a semantic-attentive feature transformation module $T_{ag} : f_a \rightarrow f_{ag}$ for modeling the structural changes of feature, and a decoder $G_g : f_{ag} \rightarrow I'_g$ for synthesizing the ground image. During training, we adopt an auxiliary ground encoder $E_g : I_g \rightarrow f_g$ that maps ground image into feature space, and a pre-trained segmentation network S_{seg} that extracts semantic map from ground images.

Semantic-Attentive Feature Transformation Module. The proposed semantic-attentive transformation module T_{ag} is illustrated in Fig. 2(a). It learns a structural transformation from f_a to f_{ag} , where f_{ag} has structure aligned to I_g . Rather than solely depending on an implicit learning of the transformation [10], we perform initial coarse alignment using polar transformation [17] to generate f_p .

Since objects with different classes are likely to locate in different areas [18] (*e.g.*, sky occupies the upper part of an image whereas road occupies the bottom), we employ semantic-attentive transformation that separately handles alignment of objects in different class. Specifically, we generate channel attentions [13] $\{M_i\}_{i=1}^c$ for c semantic classes and apply them to f_p . For further alignment, we feed them into subsequent warping blocks $\{D_i\}_{i=1}^c$, each of which consists of a deformable convolution layer [19] and a convolution layer. By summing all the attentively aligned features $\{f_{ag}^i\}_{i=1}^c$, the final

semantic-attentive transformed feature f_{ag} is obtained.

Auxiliary Networks. The ground encoder E_g , along with G_g , operate as an autoencoder [20], *i.e.*, E_g extracts feature f_g from I_g and G_g reconstructs the ground image I_g^{rec} . It encourages E_g to extract rich ground-specific feature f_g with sufficient representations to reconstruct ground images. We use f_g as a reference for f_{ag} , thereby guide E_a to extract representative features for synthesizing I'_g , as well as T_{ag} to better model the structural transformations without 3D information [12]. We use separate parameters for E_a and E_g to encourage higher flexibility and capacity in feature extraction [21].

A pre-trained semantic segmentation network S_{seg} takes ground image I_g as an input and outputs a semantic segmentation mask S_g . We use it to improve the semantic-awareness of the generator through semantic-aware losses, presented in the following section.

2.2. Loss functions

Semantic-Aware Synthesis Loss. In order to synthesize plausible ground images, objects across various classes should be considered. Here, we observe uneven number of pixels for different objects in the cross-view image datasets [14, 15] as reported in Fig. 5. This is mainly due to the different sizes of the objects and a prevailing number of scenes without a man-made object. It leads the regular L1 loss between the ground-truth and synthesized images to be dominated by the prevailing objects. Consequently, the model often fails to synthesize the objects across various semantic class.

To alleviate this problem, we balance the losses across the semantic classes by exploiting S_g . Concretely, we compute L1 loss for each class independently using the segmentation mask and average them with their own number of pixels as illustrated in Fig. 2(b). Our novel semantic-aware synthesis loss is defined as

$$\mathcal{L}_{Syn} = \sum_{i=1}^c \frac{w_i}{N_i} \|S_g^i I_g - S_g^i I'_g\|_1, \quad (1)$$

where N_i and S_g^i are the number of pixels and class-mask for class i . We further use w_i as class balancing weight [22, 23] to handle a prevailing number of scenes without a man-made object.

Semantic-Aware Feature Loss. Similar to semantic-aware synthesis loss, we use downsampled S_g and apply separate loss functions for each semantic-attentive branch in the transformation module as

$$\mathcal{L}_{Fea} = \sum_{i=1}^c \frac{w_i}{N_i} \|S_g^i f_g - S_g^i f_{ag}^i\|_1, \quad (2)$$

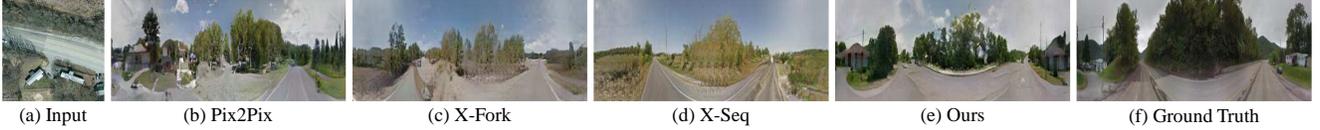


Fig. 3. Qualitative comparison on CVUSA dataset.

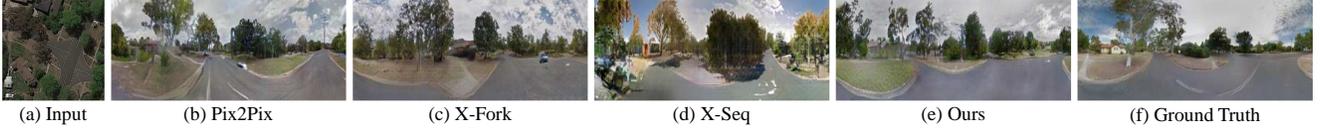


Fig. 4. Qualitative comparison on CVACT dataset.

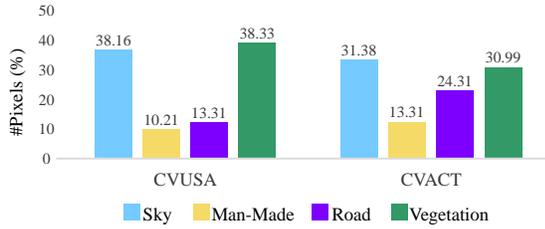


Fig. 5. Class distribution on CVUSA [14] and CVACT [15] datasets. We state the number of pixels for each class. Sky and vegetation are prevailed in the datasets whereas the man-made is few.

so that each channel attention is learned to highlight class-specific features and enables an effective learning of the transformation.

Semantic Consistency Loss. To enforce the semantic consistency in the synthesized image, we constrain the semantic differences between the synthesized image and the ground truth image, defined as

$$\mathcal{L}_{\text{Sem}} = \|S_g - S'_g\|_1, \quad (3)$$

where S'_g is output of S_{seg} with I'_g as input. Different from the previous works [1, 5, 12], the parameters in S_{seg} are fixed when training the generator. It further encourages the synthesized objects to have similar appearance as the ground truth images S_{seg} is trained on.

Ground Image Autoencoding Loss. In order to train E_g to extract ground-representative features f_g , and G_g to synthesize realistic ground images upon the feature, we adopt L1 reconstruction loss between I_g and I_g^{rec} as

$$\mathcal{L}_{\text{AE}} = \|I_g - I_g^{\text{rec}}\|_1. \quad (4)$$

Adversarial Loss. Similar to the previous works for image synthesis [24, 25], we encourage the synthesized images to be indistinguishable from the real images by adopting a discriminator D and applying an adversarial loss as

$$\mathcal{L}_{\text{Adv}} = \mathbb{E}_{I_g} \log D(I_g) + \mathbb{E}_{I'_g} \log(1 - D(I'_g)). \quad (5)$$

Overall Loss. In summary, our full objective is defined as

$$\begin{aligned} \mathcal{L} = & \lambda_{\text{Syn}} \mathcal{L}_{\text{Syn}} + \lambda_{\text{Fea}} \mathcal{L}_{\text{Fea}} + \lambda_{\text{Sem}} \mathcal{L}_{\text{Sem}} \\ & + \lambda_{\text{AE}} \mathcal{L}_{\text{AE}} + \mathcal{L}_{\text{Adv}}, \end{aligned} \quad (6)$$

where λ_{Syn} , λ_{Fea} , λ_{Sem} and λ_{AE} are hyper-parameters that control the weights between the different loss terms.

3. EXPERIMENTS

3.1. Implementation and Experimental Settings

Network Architecture. We adopt the architecture from Zhu *et al.* [25] for our generator and discriminator. Specifically, we com-

pose the encoders and decoder with 4 and 5 residual blocks [26] respectively. We use SegNet [27] architecture for S_{seg} .

Datasets. We conduct experiments on two commonly used cross-view image datasets, CVUSA [14] and CVACT [15], each containing 35,532/8,884 and 35,532/92,802 train/test image pairs. We apply pre-processing to adjust aerial images into a similar scale and exclude ground image areas with large panoramic distortions. Then we resize the aerial and ground images into 256×256 and 128×512 . For both datasets, we use S_{seg} trained on CVUSA dataset, with the segmentation maps provided by [10] as pseudo label which contains four classes of sky, man-made, road and vegetation.

Training Details. We set the class balancing weights w_i in (1) and (2) as 0.5, 2, 1, and 1 for sky, man-made, road, and vegetation classes, respectively. We aim to handle the scarcity of the man-made class and avoid the network being overfitted to the ground-truth sky representation. We set the loss weights in (6) as $\lambda_{\text{Syn}} = 10$, $\lambda_{\text{Fea}} = 2$, $\lambda_{\text{Sem}} = 2$, and $\lambda_{\text{AE}} = 5$. We use Adam optimizer [28] with momentum parameters 0.5 and 0.999, and fixed learning rate of 0.0002. The network parameters are initialized with normal distribution with zero mean and 0.02 standard deviation. We do not perform any data augmentation and train our network for 30 epochs with batch size of 4. All the experiments are conducted using Pytorch [29] library, on a single NVIDIA RTX 2080Ti X GPU.

Evaluation Protocols. For quantitative evaluations, we follow the protocols presented in [5]. We measure the visual quality of the synthesized images by Peak-Signal-to-Noise Ratio (PSNR), Structural-Similarity Index (SSIM), and Sharpness Difference (SD) with ground-truth image. We also measure the realism and diversity of the synthesized images by Inception Score (IS), Top-k prediction accuracy, and KL divergence. We additionally evaluate the pixel-wise semantic consistency of the synthesized images using mean Intersection-over-Union (mIoU).

3.2. Comparison with State-of-the-Art Methods

We compare the proposed method with Pix2Pix [24], X-Fork [5], and X-Seq [5]. Since these methods handle input and output images of same shapes, we apply few modifications. We change the bottleneck kernel size from (4, 4) to (1, 4) and use unconditional discriminator. We also remove the skip connections in Pix2Pix [24]. Except the above modifications, we follow the original settings.

We present qualitative results in Figs. 3 and 4. Compared to the previous methods [24, 5], we observe that our method shows the most visually plausible results. Specifically, our results show the clearest image appearance and consistent layout with the ground-truth images. It demonstrates that our semantic-attentive feature transformation module successfully aligns the intermediate features

Table 1. Quantitative evaluation of PSNR, SSIM, Sharpness Difference, KL Loss and mIoU.

Methods	CVUSA					CVACT				
	PSNR	SSIM	SD	KL	mIoU	PSNR	SSIM	SD	KL	mIoU
Pix2Pix [24]	19.0631	0.3864	17.8758	4.64±1.18	0.3013	19.5376	0.4022	17.4920	3.64±0.93	0.3048
X-Fork [5]	19.7425	0.4106	18.1640	4.91±1.24	0.2962	20.1629	0.4134	17.7542	3.55±0.90	0.3005
X-Seq [5]	19.6859	0.4292	18.2379	6.42±1.38	0.2944	18.8307	0.4062	17.6511	4.13±1.03	0.2798
Ours	19.6604	0.4363	18.2497	3.66±1.04	0.3068	19.6944	0.4168	17.9001	3.44±0.93	0.3118

Table 2. Quantitative evaluation of inception score and classification accuracy.

Methods	CVUSA					CVACT				
	Inception score			Accuracy		Inception score			Accuracy	
	All	Top-1	Top-5	Top-1	Top-5	All	Top-1	Top-5	Top-1	Top-5
Pix2Pix [24]	2.2454	2.0252	2.2045	29.43	67.66	1.7930	1.6808	1.8094	23.48	65.05
X-Fork [5]	2.4556	2.1217	2.4857	29.82	69.99	1.9412	1.7042	1.9686	25.41	67.03
X-Seq [5]	2.2055	2.0558	2.1902	24.83	63.70	2.1648	1.7772	2.1115	19.88	57.39
Ours	2.5367	2.1429	2.5087	34.48	72.58	2.1762	1.8577	2.1293	26.24	63.78
Real Data	3.2930	2.5634	3.2235	-	-	2.4226	2.0046	2.4087	-	-

**Fig. 6.** Qualitative ablation study results on CVUSA dataset.**Table 3.** Quantitative ablation study results on CVUSA dataset.

Setup	CVUSA			
	SSIM	KL	IS (All)	mIoU
w/o T_{ag}	0.4246	5.47±1.21	2.4838	0.2823
w/o $\{\mathcal{L}_{Syn}, \mathcal{L}_{Fea}\}$	0.3884	5.77±1.29	2.4989	0.2790
w/o \mathcal{L}_{Sem}	0.4217	3.85±1.29	2.5501	0.2939
Ours	0.4363	3.66±1.04	2.5367	0.3068

to the ground layout by focusing on every different semantic class. We also observe that the proposed method synthesizes plausible result across various objects, while others failed (*e.g.*, buildings). This confirms that our semantic-aware loss functions allow the network to handle objects across various classes.

Quantitative results are presented in Tables 1 and 2. The proposed method outperforms the previous methods in all the quantitative measures except for PSNR. Although our PSNR results are slightly lower compared to X-Fork [5], we achieve higher visual quality scores (KL and IS), showing that our method generates more realistic images. In addition, we can see that our method results in the highest mIoU score which verifies that our framework generates the most semantically consistent images with the ground truth.

3.3. Ablation Study

To investigate the importance of the key components in our framework, we conduct experiments on our method without T_{ag} , without $\{\mathcal{L}_{Syn}, \mathcal{L}_{Fea}\}$, and without \mathcal{L}_{Sem} on CVUSA dataset. Results are presented in Table 3 and Fig. 6.

For the setup w/o T_{ag} , we only apply polar transformation to f_a

and do not use any loss function for the transformed feature. For the setup w/o $\{\mathcal{L}_{Syn}, \mathcal{L}_{Fea}\}$, we use regular L1 loss for images and features. We observe that the networks with those setups generate the synthesized images that have unclear structure alignment and are not realistic, compared to our full framework. This observation coincides with the quantitative results, indicating that both components largely affect the results and should be applied cooperatively. For the setup w/o \mathcal{L}_{Sem} , both the qualitative and quantitative results show minor difference from our full framework. It demonstrates that our transformation module, together with semantic-aware loss functions sufficiently align the structure and enhance semantic awareness to the network, thereby enforcing the semantic consistency in the synthesized image.

4. CONCLUSION

In this paper, we proposed a novel framework for aerial-to-ground image synthesis through enforcing the semantic awareness of the network. The proposed semantic-aware network contains a novel semantic-attentive feature transformation module and is trained with semantic-aware loss functions. The transformation module is modeled to align the structures of every object into the corresponding ground layout. Semantic awareness is further enhanced by the proposed loss functions designed to independently calculate losses across every semantic. By handling every semantic categories respectively, our approach succeeded in synthesizing plausible ground image from given aerial image. Extensive experimental results demonstrate the effectiveness of the proposed method in terms of realism and semantic consistency of the synthesized images.

This research was supported by the Yonsei University Research Fund of 2021 (2021-22-0001).

5. REFERENCES

- [1] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan, “Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation,” in *CVPR*, 2019, pp. 2417–2426. 1, 3
- [2] Hao Ding, Songsong Wu, Hao Tang, Fei Wu, Guangwei Gao, and Xiao-Yuan Jing, “Cross-view image synthesis with deformable convolution and attention mechanism,” in *PRCV*, 2020, pp. 386–397. 1
- [3] Xueqing Deng, Yi Zhu, and Shawn Newsam, “What is it like down there? generating dense ground-level views and image features from overhead imagery using conditional generative adversarial networks,” in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2018, pp. 43–52. 1
- [4] Xueqing Deng, Yi Zhu, and Shawn Newsam, “Using conditional generative adversarial networks to generate ground-level views from overhead imagery,” in *arXiv preprint arXiv:1902.06923*, 2019. 1
- [5] Krishna Regmi and Ali Borji, “Cross-view image synthesis using conditional gans,” in *CVPR*, 2018, pp. 3501–3510. 1, 3, 4
- [6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” in *NeurIPS*, 2014, pp. 2672–2680. 1
- [7] Mehdi Mirza and Simon Osindero, “Conditional generative adversarial nets,” in *arXiv preprint arXiv:1411.1784*, 2014. 1
- [8] Miao Wang, Guo-Ye Yang, Ruilong Li, Run-Ze Liang, Song-Hai Zhang, Peter M Hall, and Shi-Min Hu, “Example-guided style-consistent image synthesis from semantic labeling,” in *CVPR*, 2019, pp. 1495–1504. 1
- [9] Haitian Zheng, Haofu Liao, Lele Chen, Wei Xiong, Tianlang Chen, and Jiebo Luo, “Example-guided scene image synthesis using masked spatial-channel attention and patch-based self-supervision,” in *arXiv preprint arXiv:1911.12362*, 2019. 1
- [10] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs, “Predicting ground-level scene layout from aerial imagery,” in *CVPR*, 2017, pp. 867–875. 1, 2, 3
- [11] Krishna Regmi and Ali Borji, “Cross-view image synthesis using geometry-guided conditional gans,” *CVIU*, vol. 187, pp. 102788, 2019. 1
- [12] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R Oswald, Marc Pollefeys, and Rongjun Qin, “Geometry-aware satellite-to-ground image synthesis for urban areas,” in *CVPR*, 2020, pp. 859–867. 1, 2, 3
- [13] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, “Cbam: Convolutional block attention module,” in *ECCV*, 2018, pp. 3–19. 2
- [14] Scott Workman, Richard Souvenir, and Nathan Jacobs, “Wide-area image geolocation with aerial reference imagery,” in *ICCV*, 2015, pp. 3961–3969. 2, 3
- [15] Liu Liu and Hongdong Li, “Lending orientation to neural networks for cross-view geo-localization,” in *CVPR*, 2019, pp. 5624–5633. 2, 3
- [16] Bor-Chun Chen and Andrew Kae, “Toward realistic image compositing with adversarial learning,” in *CVPR*, 2019, pp. 8415–8424. 2
- [17] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li, “Spatial-aware feature aggregation for image based cross-view geolocation,” in *NeurIPS*, 2019, pp. 10090–10100. 2
- [18] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang, “Unsupervised domain adaptation for semantic segmentation via class-balanced self-training,” in *ECCV*, 2018, pp. 289–305. 2
- [19] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei, “Deformable convolutional networks,” in *ICCV*, 2017, pp. 764–773. 2
- [20] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, “Learning internal representations by error propagation,” in *Parallel Distributed Processing*, vol. 1, pp. 318–362, 1986. 2
- [21] Sixing Hu and Gim Hee Lee, “Image-based geo-localization using satellite imagery,” *IJCV*, vol. 128, pp. 1–15, 2019. 2
- [22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, “Focal loss for dense object detection,” in *ICCV*, 2017, pp. 2980–2988. 2
- [23] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun, “Learning to reweight examples for robust deep learning,” in *arXiv preprint arXiv:1803.09050*, 2018. 2
- [24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, “Image-to-image translation with conditional adversarial networks,” in *CVPR*, 2017, pp. 1125–1134. 3, 4
- [25] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *ICCV*, 2017, pp. 2223–2232. 3
- [26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778. 3
- [27] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *TPAMI*, vol. 39, no. 12, pp. 2481–2495, 2017. 3
- [28] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015. 3
- [29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *NeurIPS*, pp. 8024–8035, 2019. 3