LABEL-FREE NEURAL SEMANTIC IMAGE SYNTHESIS

Jiayi Wang[1] **Kevin Alexander Laube**[1] **Yumeng Li**[1,2] **Jan Hendrik Metzen**[1]
Shin-I Cheng[1] **Julio Borges**[1] **Anna Khoreva**[1]
[1]Bosch Center for Artificial Intelligence [2]University of Mannheim
{jiayi.wang2, kevinalexander.laube, yumeng.li, janhendrik.metzen, shin-i.cheng, julio.borges, anna.khoreva}@de.bosch.com

ABSTRACT

Recent work has shown great progress in integrating spatial conditioning to control large, pre-trained text-to-image diffusion models. Despite these advances, existing methods describe the spatial image content using hand-crafted conditioning inputs, which are either semantically ambiguous (e.g., edges) or require expensive manual annotations (e.g., semantic segmentation). To address these limitations, we propose a new label-free way of conditioning diffusion models to enable fine-grained spatial control. We introduce the concept of *neural semantic image synthesis*, which uses neural layouts extracted from pre-trained foundation models as conditioning. Neural layouts are advantageous as they provide rich descriptions of the desired image, containing both semantics and detailed geometry of the scene. We experimentally show that images synthesized via neural semantic image synthesis achieve similar or superior pixel-level alignment of semantic classes compared to those created using expensive semantic label maps. At the same time, they capture better semantics, instance separation, and object orientation than other label-free conditioning options, such as edges or depth. Moreover, we show that images generated by neural layout conditioning can effectively augment real data in various perception tasks.

1 INTRODUCTION

Controllable image synthesis enables users to specify the desired image content, while relying on a generative model to fill in details that align with the distribution of natural images. This has been popularized by large-scale text-to-image (T2I) diffusion models (mid, 2023; Ramesh et al., 2022; Balaji et al., 2022; Rombach et al., 2022) that express content through natural language. Recent work (Li et al., 2023b; Zhang & Agrawala, 2023; Zhao et al., 2023a; Mou et al., 2023; Qin et al., 2023) introduced additional adapters to integrate spatial conditioning control into the diffusion process for direct image content specification. These methods have shown that it is possible to employ segmentation, edge, depth, and normal maps as well as skeletal poses of a reference image as description of the image's content. Given this variety, it is natural to ask what descriptor is best suited to specify the spatial and semantic contents of scenes. We argue that two properties are key to the general applicability of a descriptor: *richness of semantic and spatial content* and the *ease to obtain descriptor-image pairs* for fine-tuning pre-trained text-to-image (T2I) diffusion models.

Semantic segmentation maps are a popular descriptor choice (Xue et al., 2023; Wang et al., 2022; Saharia et al., 2022), being interpretable high-level abstractions. However, creating them for real images requires costly and tedious pixel-wise manual annotation. Even more so, segmentation maps do not contain full information about the object pose, orientation, or geometry. On the other hand, image *edge* and *depth* can be easily obtained from unlabeled images (e.g., by using pretrained detectors (Xie & Tu, 2015; Ranftl et al., 2022)) to cheaply create descriptor-image pairs (Zhang & Agrawala, 2023). However, they contain limited spatial information and are ambiguous in terms of the object semantics. For example, both "cat" and "blanket" are plausible interpretations for the object boundaries seen in Fig. 2. Similarly for depth maps, the semantics can be misinterpreted. In short, existing conditioning descriptors can not satisfy both desired properties at once.

In this work, we propose a new way of conditioning T2I diffusion models to enable fine-grained spatial control which does not require expensive human annotations. We introduce the concept of *neural semantic image synthesis*, which derives its conditioning from dense neural features extracted from large-scale foundation models (FMs). Recent work Zhou et al. (2022); Oquab et al. (2023); Zhao et al. (2023b); Li et al. (2023a) has shown that these features preserve the semantic content and geometry of the images well, and thus are well-suited for being rich spatial descriptors of the

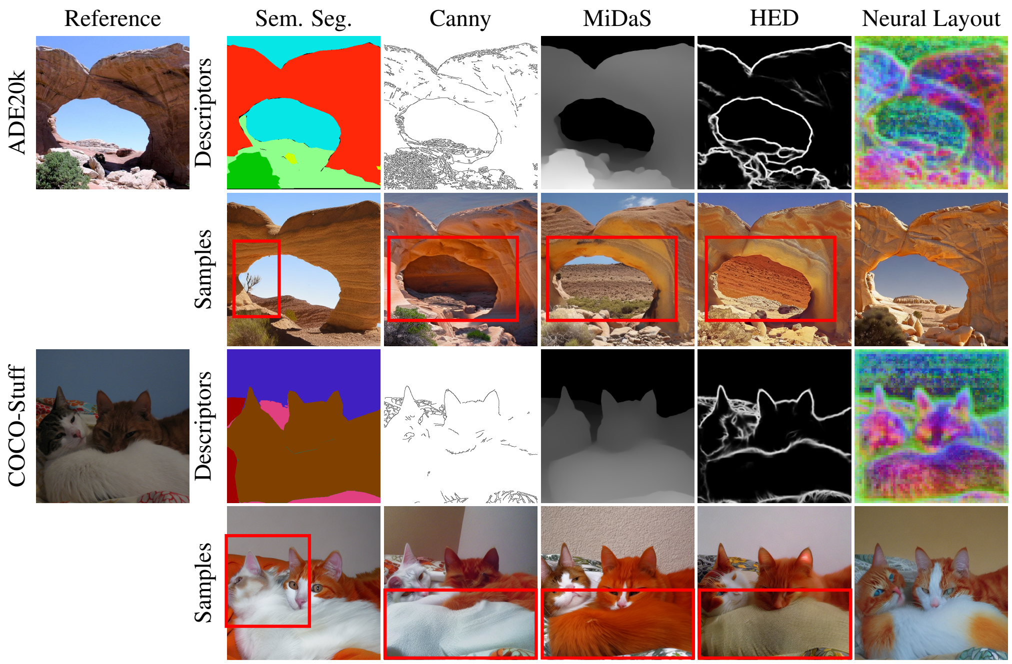

Figure 2: Comparison of images generated by ControlNet with different conditioning types on ADE20k and COCO-Stuff. Neural layouts provide rich description of the desired images, while other inputs contain limited spatial information and are semantically ambiguous.

desired scene. However, these features encode nuisance appearance variations which must be removed to ensure diverse synthesis. Therefore, we introduce an *semantic separation* step using PCA decomposition to extract only the desired information. We refer to these compressed features as a *"neural layout"* (see Fig. 1).

To showcase the benefits of neural layout conditioning, we propose the LUMEN model which stands for **L**abel-free ne**U**ral se**M**antic imag**E** sy**N**thesis. LUMEN builds upon ControlNet (Zhang & Agrawala, 2023) and uses neural layouts extracted from an image's Stable Diffusion features (Rombach et al., 2022) for conditioning (see Fig. 1). We show that images generated by LUMEN achieve similar or superior alignment in semantic layout to the reference image when compared to those created using expensive semantic label maps (see Table 2). In comparison to other label-free conditioning inputs such as edges or depth, images generated with neural layouts capture better the semantics and geometry of the scene (see Fig. 2). Furthermore, we experimentally verify that LUMEN images can serve effectively for data augmentation in perception tasks such as semantic segmentation, depth estimation and object detection (see Table 3).

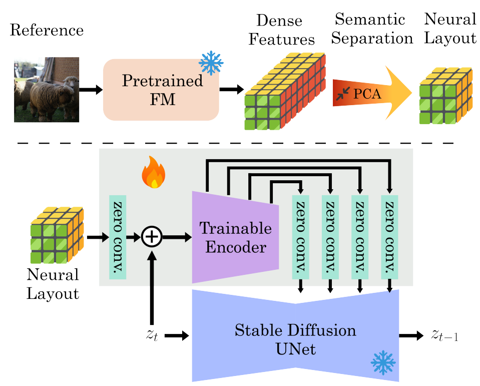

Figure 1: LUMEN uses dense features from foundation models (FMs) to extract neural layouts as conditioning for a ControlNet.

2 Method

In this section, we introduce the concept of *neural semantic image synthesis*. Instead of using ad-hoc conditioning to describe the desired output, neural semantic image synthesis makes use of *neural layouts* derived from the dense features of pretrained foundation models.

Dense Feature Extraction. Modern foundation models make heavy use of self-attention (SA) and cross-attention (CA) (Vaswani et al., 2017). It was noted by several works (Amir et al., 2021; Zhang et al., 2023) that these activation can serve as dense features useful for downstream tasks.

Semantic Separation. Retaining the entire dense feature map f would reveal too detailed information about the reference image $\mathbf{x}_{ref}$. Neural semantic image synthesis would then typically lead to

Original caption: A <u>semi truck</u> is driving down a street.

Reference Sample → banana truck → burning truck + winter + in rainy weather

Figure 3: Image manipulation through prompt editing on the COCO-Stuff validation set. We show an unedited sample, and then show results from either replacing the underlined words in the original caption (→) or appending additional words at its end (+).

		COCO-Stuff					ADE20K				
Conditioning	Label-free	mIoU ↑	SI Depth ↓	FID ↓	LPIPS ↑	TIFA ↑	mIoU ↑	SI Depth ↓	FID ↓	LPIPS ↑	TIFA ↑
Edges (Canny)	✓	44.4	24.7	13.2	0.48	0.84	35.1	25.0	19.1	0.49	0.86
Edges (HED)	✗	49.3	21.4	12.1	0.39	0.74	41.8	**21.1**	17.4	0.37	0.73
Depth (MiDaS)	✗	45.3	24.0	14.3	0.53	**0.88**	34.0	22.1	21.2	0.52	**0.87**
Sem. Seg. (GT)	✗	43.3	28.8	15.3	**0.65**	**0.88**	35.1	27.1	22.6	**0.63**	0.85
Neural Layout	✓	**52.9**	**21.1**	**11.8**	0.36	0.66	**45.7**	**21.1**	**16.1**	0.33	0.67

Table 2: Comparison of different ControlNet conditioning. Neural layout outperforms all other options in terms of image quality (FID), as well as semantic and spatial alignment (mIoU, SI).

samples that are highly similar to $\mathbf{x}_{ref}$, lacking diversity. To prevent this, it is preferable to separate semantic and geometric features from those that encode appearance details. Based on existing works Oquab et al. (2023), we hypothesize that the principal directions of variation in the dense features should at least partially correspond to what humans intuitively understand as spatial and semantic image content. Thus, we implement Principal Component Analysis (PCA) to obtain a linear projector that can remove nuisance variations. To obtain the neural layout $\mathbf{c}_i$ as conditioning, we retain only the information in the top N PCA components. In practice, we perform PCA with $N = 20$ on a random sample of $40,000$ feature vectors extracted from images in the training set.

Foundation Model Backbones. After a thorough ablation, shown in Table 1, we select Stable Diffusion as the default foundation model backbone. Following Zhang et al. (2023), we extract the intermediate activations from layer 2, 5, and 8 of SD's U-Net and upsampled them to match the resolution of layer 8. All activations are then concatenated across the channel dimension. According to Zhang et al. (2023), SD features have a strong sense of spatial layout, which makes them a promising candidate for neural layouts.

3 Experiments

Evaluation Metrics. We measure the image synthesis quality of our method using FID (Heusel et al., 2017) for perceptual quality, average LPIPS (Zhang et al., 2018) between generated samples for diversity, and TIFA (Hu et al., 2023) for text controllability. We additionally evaluate how well each conditioning captures the semantic composition and geometry of the scene. For alignment with semantic layouts, we use mIoU between ground truth segmentation label and those predicted by a pretrained segmenter. However, since mIoU does not contain 3D information, we use the scale-invariant depth error (SI depth) (Eigen et al., 2014) as a metric for geometric consistency.

3.1 Neural Layout Design Space

We explore the design space of neural layouts on the diverse COCO-Stuff dataset (Caesar et al., 2018) to determine how to best extract descriptive semantic and spatial information from a given reference image. In Table 1, we compare the quality of image generated from DINO, DINOv2, CLIP, and SD features. We observe that SD features provide the best perceptual image quality and also retain the semantic content best, and DINOv2 is a close second. Although CLIP conditioning can generate more varied images, this diversity is due to the weak semantic and spatial constraints imposed during synthesis. Since CLIP is trained with an image-level objective, it is less suitable to capture precise pixel-level information without further processing. Therefore, we choose to base our neural layout on SD features.

Feat.	mIoU ↑	SI Depth ↓	FID ↓	LPIPS ↑	TIFA ↑
CLIP	41.7	25.7	15.6	**0.58**	**0.84**
DINO	49.5	22.2	12.8	0.41	0.77
DINOv2	51.1	22.0	12.8	0.42	0.78
SD	**51.4**	**21.5**	**12.2**	0.42	0.79

Table 1: Comparison of different FMs for extracting neural layouts on COCO-Stuff.

3.2 Comparison to Existing Conditioning

We evaluate the effects that different conditioning have on image synthesis using the challenging COCO-Stuff (Caesar et al., 2018) and ADE20k (Zhou et al., 2017) datasets. As Table 2 shows, neural layout as the condition results in images that best preserve the semantic content, outperforming others in terms of both mIoU and SI Depth while achieving better image quality. Surprisingly, Sem. Seg. achieves only similar or worse results in terms of mIoU compared to all other conditioning. We believe this is due to the large number of difficult semantic classes in COCO-Stuff and ADE20k with sometimes semantically ambiguous labels, and due to the long tail distribution on rare classes. We also see that HED edges performs the best among the existing conditions in terms of FID, mIoU, and SI Depth, as it well encodes the object boundaries with additional image details being captured by soft edges. However, the semantic class of the object within the boundary can be ambiguous, resulting in a lower mIoU (see Table 2).

Note that we observe again the trade-off between information content constraining the image and the diversity and editability of the results. Canny edge and depth have low semantic content and Sem. Seg. does not constrain appearance or geometry, consequently, they often achieved the best LPIPS and TIFA at the cost of worse alignment, geometry or image quality. We also observe that despite the lower TIFA, LUMEN still responds well to a variety of out-of-distribution prompt edits (see Fig. 3). Therefore, text-prompting creates additional variations for data synthesis.

3.3 Downstream Applications

Training Data for Multiple Tasks. As neural layout specifies both semantic and spatial concepts in an image, the same synthesized data can reuse all annotations to train downstream networks for different tasks. We experimented with this capability by synthesizing data using the 2975 training images from Cityscapes (Cordts et al., 2016) as reference, and reuse the semantic segmentation labels, the depth disparity maps, and 3D bounding boxes of vehicles (Gählert et al., 2020) for training.

Using this, SegFormer (Xie et al., 2021) is trained for semantic segmentation and TaskPrompter (Ye & Xu, 2023) for predicting depth and 3D detection of vehicle. The results are shown in Table 3 and the exact setup is detailed in the supplementary materials. As 3D annotation is available, we follow prior work (Ye & Xu, 2023) and report root-mean-square error (RMSE) of the estimated disparity and the mean detection score (mDS) (Gählert et al., 2020) to evaluate the quality of the 3D tasks. Neural layout performs better or equal to existing conditioning on all tasks simultaneously. It also improves upon the mIoU and mDS compared to a baseline that uses only real data. This demonstrates that neural layout is a more universal conditioning and the data generated using it can be applied across different tasks.

Method	SegFormer mIoU ↑	TaskPrompter (3D) RMSE ↓	TaskPrompter (3D) mDS ↑
Baseline	67.90	4.78	0.19
Edges (Canny)	67.08	5.26	0.16
Edges (HED)	67.40	<u>4.96</u>	<u>0.19</u>
Depth (MiDaS)	68.00	<u>4.96</u>	0.17
SemSeg (GT)	<u>68.48</u>	4.99	**0.20**
Neural Layout	**68.54**	**4.89**	**0.20**

Table 3: Using generated data for training multiple downstream tasks on Cityscapes.

Domain Generalization. By choosing to use ControlNet as our backbone generator, we can use the text prompt to control the domain (mainly appearance) of the synthesized images while using neural layout to control the semantic and geometric content. We use this to perform domain generalization experiments from the daytime only Cityscapes to ACDC (Sakaridis et al., 2021), containing adverse weather and lighting conditions. As shown in Table 4, we verified that images generated by LUMEN can significantly improve the model's generalization ability upon the baseline, which is trained only on Cityscapes. We compare against other diffusion-based methods PnP-Diffusion (Tumanyan et al., 2023), FreestyleNet (Xue et al., 2023), as well as ControlNet (Zhang & Agrawala, 2023) with Sem. Seg. conditioning. The prompt editing of PnP-Diffusion cannot generalize well to the image domain of Cityscapes, leading to little benefits for domain generalization. Both ControlNet with semantic segmentation and FreestyleNet require manual annotation to train the image generator, in contrast to our label-free LUMEN. Yet, our method outperforms FreestyleNet and is overall on par with ControlNet using Sem. Seg.

Method	CS	Rain	Fog	Snow	Night	Avg.
Baseline (CS)	67.9	50.2	60.5	48.9	28.6	47.0
PnP-Diffusion	67.8	50.6	63.5	50.4	30.3	48.7
FreestyleNet	**69.7**	52.7	**69.0**	54.3	32.9	52.2
ControlNet [Sem. Seg.]	68.3	**55.3**	67.3	<u>55.4</u>	<u>34.6</u>	**53.2**
LUMEN (Ours)	68.5	<u>53.4</u>	<u>67.4</u>	**55.6**	**35.1**	<u>52.9</u>

Table 4: Quantitative comparison of synthetic data augmentation techniques for domain generalization from Cityscapes (train) to ACDC (unseen).

4 Conclusion

We introduced the concept of neural semantic image synthesis and established LUMEN as a strong label-free baseline that can simultaneous specify semantic and spatial concepts of the outputs.

REFERENCES

Midjourney. https://www.midjourney.com/, 2023.

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 2014.

Nils Gählert, Nicolas Jourdan, Marius Cordts, Uwe Franke, and Joachim Denzler. Cityscapes 3d: Dataset and benchmark for 9 dof vehicle detection. *arXiv preprint arXiv:2006.07864*, 2020.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023.

Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *CVPR*, 2023a.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023b.

Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Can Qin, Ning Yu, Chen Xing, Shu Zhang, Zeyuan Chen, Stefano Ermon, Yun Fu, Caiming Xiong, and Ran Xu. Gluegen: Plug and play multi-modal encoders for x-to-image generation. In *ICCV*, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Chitwan Saharia, Chris A. Lee, David James Fleet, Huiwen Chang, Jonathan Ho, Mohammad Norouzi, Tim Salimans, and William Chan. Palette: Image-to-image diffusion models. In *SIGGRAPH*, 2022.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, 2021.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021.

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015.

Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *CVPR*, 2023.

Hanrong Ye and Dan Xu. Taskprompter: Spatial-channel multi-task prompting for dense scene understanding. In *ICLR*, 2023.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. In *NeurIPS*, 2023.

Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *NeurIPS*, 2023a.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *ICCV*, 2023b.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022.