UniNet: A Unified Scene Understanding Network and Exploring Multi-Task Relationships through the Lens of Adversarial Attacks

Naresh Kumar Gurulingan*, Elahe Arani*, Bahram Zonooz
Advanced Research Lab, NavInfo Europe, Eindhoven, The Netherlands
{naresh.gurulingan, elahe.arani}@navinfo.eu, bahram.zonooz@gmail.com

Abstract

Scene understanding is crucial for autonomous systems which intend to operate in the real world. Single task vision networks extract information only based on some aspects of the scene. In multi-task learning (MTL), on the other hand, these single tasks are jointly learned, thereby providing an opportunity for tasks to share information and obtain a more comprehensive understanding. To this end, we develop UniNet, a unified scene understanding network that accurately and efficiently infers vital vision tasks including object detection, semantic segmentation, instance segmentation, monocular depth estimation, and monocular instance depth prediction. As these tasks look at different semantic and geometric information, they can either complement or conflict with each other. Therefore, understanding inter-task relationships can provide useful cues to enable complementary information sharing. We evaluate the task relationships in UniNet through the lens of adversarial attacks based on the notion that they can exploit learned biases and task interactions in the neural network. Extensive experiments on the Cityscapes dataset, using untargeted and targeted attacks reveal that semantic tasks strongly interact amongst themselves, and the same holds for geometric tasks. Additionally, we show that the relationship between semantic and geometric tasks is asymmetric and their interaction becomes weaker as we move towards higher-level representations.

1. Introduction

Scene understanding concerns the breakdown of a perceived scene into meaningful attributes which help autonomous systems understand and operate in the real world. With the aid of vision tasks which infer semantic and geometric information from monocular images, autonomous systems can deduce various relevant aspects of the world. While deep learning has provided numerous breakthroughs in single tasks such as object detection [10, 31, 23, 39], semantic segmentation [24, 3, 6], instance segmentation [14, 51], monocular depth estimation [1, 47], and monocular instance depth prediction [5], multi-task learning (MTL) and inter-task relationships have not been fully explored.

A multi-task network could be designed to have either a feature representation shared for all tasks [5, 45] or task-specific feature representations [28]. While both designs benefit from complementary task information, the former provides added benefits such as reduced memory usage and inference time. We propose UniNet, a unified multi-task scene understanding network which is designed to jointly predict five crucial vision tasks - object detection (OD), semantic segmentation (SS), instance segmentation (IS), depth estimation (D), and instance depth prediction (ID). UniNet is based on an asymmetric encoder-decoder architecture, where the encoder is shared between all tasks while the decoder is only used for the lighter tasks. Hence, it provides an acceptable trade-off between accuracy and inference efficiency.

Learning multiple tasks together serves as an inductive bias prioritizing a learned representation that favours all tasks [4]. However, the training procedure must factor in inter-task relationships to effectively search for this representation. Related tasks would promote complementary information sharing while unrelated tasks would interfere with each other, hindering attainment of an optimal shared representation. Therefore, understanding inter-task relationships can provide insights to leverage complementary task information. We consider UniNet as a simple and effective multi-task framework containing a wide variety of vision tasks to study the interaction between different tasks and their relatedness.

Existing works define task relatedness using similarity between features learned by single-task networks [9], using empirical results obtained when tasks are jointly learned [37], and based on domain knowledge such as the alignment of semantic edges and depth discontinuities [40]. Adversarial attacks center around the efficacy of fooling neural networks by making imperceptible changes to the input image.

*Equal contribution.

We hypothesise that these attacks can be used to study task relationships as they exploit learned biases in the neural networks. We therefore study task relationships in a multi-task network through the lens of adversarial attacks.

Using adversarial attacks, we study multi-task relationships and uncover intriguing findings such as - (1) semantic tasks interact strongly amongst themselves and the same holds for geometric tasks, (2) semantic and geometric tasks do not affect each other equally, and (3) while the interaction between semantic and geometric tasks is strong in low-level representations, it becomes weaker in high-level representations. Additionally, we find that intra-task interaction between bounding box classification and regression in object detection promotes a bias towards object shapes. The contributions of this work are summarized as follows:

- We propose UniNet, a unified multi-task scene understanding network designed to infer five tasks which shows competitive performance with existing multi-task approaches while being efficient at inference.
- We uncover inter-task relationships in multi-task networks (UniNet and a modified version of MTI-Net [41]) with the aid of adversarial attacks such as Projected Gradient Descent (PGD) and semantic category hiding. Specifically, semantic tasks and geometric tasks show greater interaction in low-level representation relative to high-level representation.
- We show that intra-task relationship between classification and regression in object detection induces a desirable bias towards object shapes.

2. Related Works

2.1. Single-Task Learning

Existing object detection methods use two-stage [11, 10, 32], one-stage anchor-based [31, 23] or one-stage anchor-free [39] approaches. We use one-stage anchor free FCOS [39] for object detection. Long *et al*. [24] introduced fully convolutional network (FCN) for semantic segmentation. Since then, a number of encoder-decoder based architectures [3, 33, 29, 6] have been proposed. Encoder-decoder based architectures have also been used for supervised depth estimation [1, 47]. UniNet uses an encoder-decoder based architecture for both semantic segmentation and depth estimation. Instance segmentation approaches are generally object detection free [42] or object detection based. The object detection based approaches further differ into region based [14, 19] or encoding based [46, 51] approaches. We use MEInst [51] approach which encodes instance masks using PCA. Chen *et al*. [5] proposed the instance depth prediction task where they predict a single value denoting the median depth of a predicted instance.

2.2. Multi-Task Learning and Task Relationships

Multi-task learning concerns the joint prediction of multiple tasks. Existing works concentrate on directions such as designing the architecture [22, 45] and designing a loss balancing strategy [15, 22, 13, 7]. Sharing a feature extractor between different task specific heads is a common architecture design approach. Likewise, tasks in UniNet share a feature extractor but segmentation and depth heads additionally share the decoder. Loss balancing is required to balance different loss scales and to encourage complementary information sharing. We use geometric loss strategy [7] for loss balancing.

Standley *et al*. [37] looked at task relationships in a multi-task setting. They took an empirical approach and trained models with several combination of tasks. Tasks in combinations which result in low total loss are considered to have affinities with each other (are related). Other works consider task relationships for transfer learning [49, 9, 35, 36]. Zamir *et al*. [49] studied task transfer relationships using an empirical approach by transferring across different single task networks. Dwivedi *et al*. [9] used similarity between learned features in single task networks using Representation Similarity Analysis (RSA) as task relatedness. Attribution maps have also been used to study transfer relationships on the basis that similar tasks look at similar input image regions [35][36]. Contrary to these works, we study task relationships using adversarial attacks.

2.3. Adversarial Attacks

Szegedy *et al*. [38] showed the existence of perturbations which can fool an image classifier while making no human perceivable changes to the input image. This process of fooling is called adversarial attacks. Fast Gradient Sign Method (FGSM) [12], Iterative FGSM [17], Projected Gradient Descent (PGD) [25], and others have explored stronger or varied attack types.

Arnab *et al*. [2] studied robustness of semantic segmentation models through the use of residual connections, multi-scale processing, and transformations on adversarial images. Metzen *et al*. [27] created universal perturbations which force predicted segmentation of different images to resemble a static scene. They also showed that persons can be hidden from segmentation predictions. Likewise, Wong *et al*. [43] hid instances from depth predictions. Xie *et al*. [44] proposed Dense Adversary Generation (DAG) intended to change the class of a set of predictions in object detection and semantic segmentation.

Klingner *et al*. [16] observed that the adversarial robustness of semantic segmentation improved when trained with self-supervised depth estimation. Mao *et al*. [26] posited that multi-task learning reduces the ease of attaining perturbations which attack all tasks, thereby providing inherent

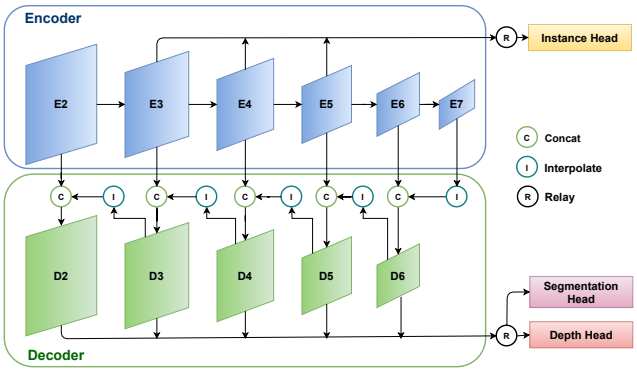


Figure 1. UniNet architecture: The encoder feature maps E3, E4, and E5 (blue) are relayed (R) to the instance head (yellow) which includes predictions of object detection, instance segmentation and instance depth. The decoder features D2 to D6 (green) are relayed to both the semantic segmentation (purple) and depth estimation head (red). Best viewed in color.

robustness. Contrary to these works, we study multi-task relationships instead of robustness.

3. Unified Network (UniNet)

UniNet is a multi-task network designed for scene understanding and infers tasks such as object detection, semantic segmentation, instance segmentation, depth estimation, and instance depth prediction. Object detection [10, 31, 23, 39] and instance segmentation [14, 51] detect and localize each occurrence of countable objects in the scene such as persons and cars. On the contrary, semantic segmentation [24, 3, 6] is a per-pixel classification task which does not provide instance information. All three tasks provide abstract semantic information about the scene. In monocular depth estimation [1, 47], the depth of every input image pixel relative to the camera frame is predicted. Instance depth prediction [5] involves predicting a single depth value for every countable object instance in the scene. Both depth estimation and instance depth prediction provide geometric information about the scene. Collectively, these tasks provide a comprehensive understanding of the scene.

3.1. Architecture

UniNet is based on asymmetric encoder-decoder architecture and is designed with a focus on efficient inference. In addition to the encoder and decoder, UniNet comprises of an instance head, a segmentation head and a depth head (Figure 1). The decoder is shared between both the segmentation and the depth heads while the instance head branches from the encoder. The encoder features E2 to E7 are successively extracted from the input image. The decoder feature D6 is obtained by applying a residual block on the channel-wise concatenation of E6 and bi-linearly up-sampled E7. Decoder features D5, D4, D3 and D2 are also obtained in a similar fashion. E3, E4 and E5 are passed to the instance head while D2, D3, D4, D5 and D6 are passed to the semantic segmentation and depth estimation head.

Figure 2 shows the different heads in UniNet. The instance head (Figure 2a) represents a collection of three instance tasks namely object detection, instance segmentation and instance depth prediction. A Feature Pyramid Network (FPN) [20] is used to obtain features P3, P4 and P5 from E3, E4 and E5, respectively. Each location in P3, P4 and P5 serves as a point anchor, based on which classification logits, bounding box, and centerness values as in FCOS [39], encoded instance masks as in MEInst [51] and median depth are predicted in a dense manner. In the semantic segmentation head (Figure 2b), the channels of D2, D3, D4, D5 and D6 are first reduced to 64 and bi-linearly interpolated to $1/4^{th}$ of the image resolution. All the resultant feature maps are concatenated and provided as input to a convolution layer whose output is bi-linearly interpolated to image resolution. Finally, a convolution layer predicts the segmentation logits. The depth estimation head which predicts pixel-wise depth is architecturally similar to the semantic segmentation head, but the weights are not shared.

3.2. Multi-Task Objective

Object detection. Object detection is trained using box regression loss $\mathcal{L}_{reg}$, box classification loss $\mathcal{L}_{cls}$, and centerness loss $\mathcal{L}_{cent}$. We use the varifocal loss [50] as $\mathcal{L}_{cls}$. Following FCOS [39], we use GIoU loss and binary cross entropy loss as $\mathcal{L}_{reg}$ and $\mathcal{L}_{cent}$, respectively.

Semantic segmentation. We use a class balanced cross entropy loss ($\mathcal{L}_{seg}$) as the semantic segmentation loss.

Instance segmentation. We use the Mean Squared Error (MSE) between predictions and encoded ground truth masks ($\mathcal{L}_{is}$). The Principal Component Analysis (PCA) parameters obtained using training data is used to encode and decode instance masks following MEInst [51].

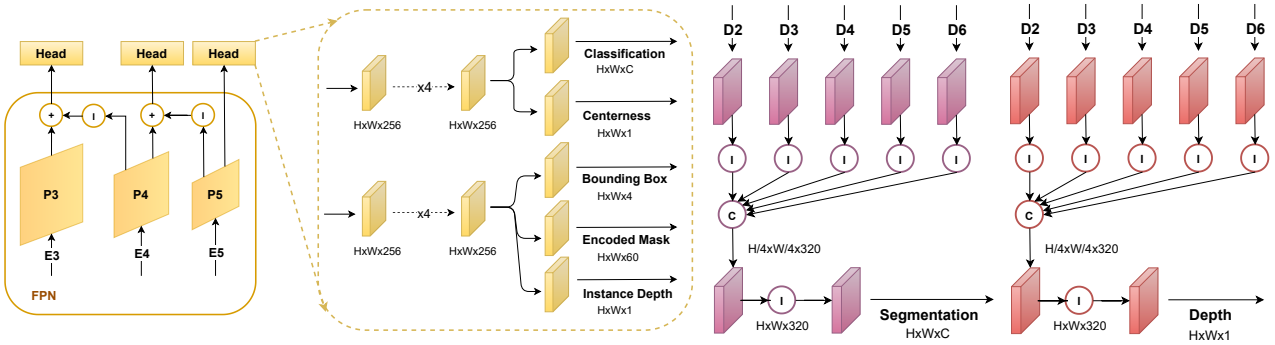
Depth estimation. The Root Mean Square Error (RMSE) between ground truth and prediction depth maps ($\mathcal{L}_{depth}$) is used as the the depth loss.

Instance depth prediction. We use the l_1 loss between predicted median depth and ground truth median depth ($\mathcal{L}_{id}$).

MTL Loss. The MTL loss ($\mathcal{L}_{MTL}$) is the combination of all losses used to train UniNet. To balance the different losses involved, we first use fixed weights λ_is to scale the losses and then use the geometric loss strategy [7]. In general, we observed that this balancing strategy provides an overall improvement in multi-task performance.

$$\mathcal{L}_{\text{MTL}} = \prod_i \sqrt[n]{\lambda_i \mathcal{L}_i} \tag{1}$$

n refers to the total number of losses required to be minimized and $i \in \{reg, cls, cent, seg, is, depth, id\}$.



(a) Instance Head (b) Segmentation Head (c) Depth Head

Figure 2. Architecture of the heads. (a) depicts the FCOS [39] based instance head. For each HxW feature map location, a classification logit, distance to the instance center, a bounding box, an encoded instance mask and a median instance depth is predicted. (b) and (c) depict the semantic segmentation and depth estimation head, respectively. **H, W** in (b) and (c) refer to the height and width of the original image. **C** refers to the number of classes.

4. Multi-Task relationships through Adversarial Attacks

Adversarial attacks exploit vulnerabilities in learned neural network representations with the intent of forcing wrong predictions. In general, imperceptible perturbations obtained by optimizing an attack loss is added to the input image. For untargeted attacks, the same loss used to train the network is used as the attack loss. For targeted attack, the attack loss involves a target which has been modified in a certain manner to force a specific inference. For example, all "persons" in the ground truth can be modified as "road" forcing the network to predict "persons" as "road".

Intuitively, these attacks leverage biases in the learned representation, resulting in perturbations which can fool the network. In multi-task networks, the architecture provides an inductive bias which enforces a feature space to be shared among tasks. This shared feature space facilitates interaction between the tasks which can be exploited by attacks. These intuitions present adversarial attacks as an intriguing front to study and evaluate task relationships in multi-task networks. We use the following attack methods.

4.1. Projected Gradient Descent (PGD)

PGD [25] is an untargeted attack where the input image is updated to maximize a given loss. The multi-task loss used to train UniNet is composed of seven individual losses. We group these losses into semantic loss in Eq. 2 and geometric loss in Eq. 3. Each of the individual losses, the multi-task loss (Eq. 1), the semantic loss, and the geometric loss are all used as the PGD attack objective to study the multi-task relationships.

$$\mathcal{L}_{\text{semantic}} = \mathcal{L}_{reg} + \mathcal{L}_{cls} + \mathcal{L}_{seg} + \mathcal{L}_{is}, \tag{2}$$

$$\mathcal{L}_{\text{geometric}} = \mathcal{L}_{depth} + \mathcal{L}_{id}. \tag{3}$$

4.2. Dense Adversary Generation (DAG)

DAG [44] is an untargeted attack aimed at attaining perturbations which force the predictions of a target set to be different from the ground truth class. For object detection, we take the bounding boxes predicted in all locations of the detection head as the target set. For semantic segmentation, all the pixels are part of the target set. However, we modify DAG to perform targeted attacks. Specifically, we use the same algorithm as in [44] but modify the π function as in Eq. 4, where c_1 and c_2 are any two valid prediction classes. Essentially, the attack aims at swapping two selected prediction classes.

$$\pi(c_1) = c_2, \pi(c_2) = c_1. \tag{4}$$

4.3. Semantic Category Hiding

Metzen *et al*. [27] attack semantic segmentation with the objective of hiding a semantic category such as "persons" from a segmentation prediction. All "person" pixels are replaced with the nearest "not person" pixels and is used as the target in the attack objective. The objective also includes a weight term which prioritizes retaining "non person" pixels or hiding "person" pixels. We use the same attack objective to hide persons from predicted segmentation maps without any weight term. Wong *et al*. [43] hide instances from depth maps with the help of ground truth instance masks. However, since we jointly predict segmentation and depth, we hide pixels in the depth map which correspond to "person" pixels in predicted segmentation map.

5. Experiments

5.1. Setup

Dataset. The Cityscapes dataset [8] consists of images from driving scenes captured in European cities. The

Method	Cityscapes							NYUv2						
	SS (mIoU)↑	D (RMSE)↓	MAC (G)	#P (M)	FPS	Energy (J)	Δ_{MTL} (%)↑	SS (mIoU)↑	D (RMSE)↓	MAC (G)	#P (M)	FPS	Energy (J)	Δ_{MTL} (%)↑
Single task [40]	64.96	5.802	35	20	73	2.6	+0.00	40.59	50.27	20	20	121	2.0	+0.00
MTL-baseline [40]	65.14	5.890	37	25	67	2.8	-0.62	40.33	48.51	22	25	109	2.2	+1.43
PADNet [45]	73.86	5.680	300	23	16	3.7	+7.90	38.95	52.81	176	23	24	3.3	-4.55
PADNet† [45]	73.47	5.630	**137**	**18**	**27**	**3.1**	+8.03	38.03	53.90	**80**	**18**	**45**	**2.9**	-6.76
MTI-Net [41]	**76.68**	**5.129**	433	99	10	5.0	**+14.82**	**43.27**	**45.43**	254	99	13	4.8	**+8.12**
MTI-Net† [41]	75.90	5.163	235	53	19	3.9	+3.93	42.50	46.63	137	53	23	3.9	+5.97
Cross-stitch [28]	65.31	5.743	69	40	36	**3.4**	+0.78	40.09	48.49	41	40	59	4.0	+1.15
MTAN [22]	65.70	5.862	42	26	51	3.8	+0.05	40.04	48.34	25	26	81	**2.9**	+1.24
UniNet	**74.49**	**5.379**	**37**	**18**	**48**	3.5	**+10.98**	**40.90**	**47.08**	**22**	**18**	**81**	**2.9**	**+3.55**

Table 1. Comparing UniNet with state-of-the-art multi-task models on Cityscapes and NYUv2 datasets. Δ_{MTL} represents the multi-task performance metric proposed in [40]. MTI-Net and PADNet use additional auxiliary tasks. We also include results for MTI-Net and PADNet without auxiliary tasks (represented by †).

Task	Five tasks	Single task	OD+IS	OD+ID	SS+D
OD (mAP^b)	**38.93±0.14**	38.28±0.72	38.41±0.35	37.85±0.24	-
SS (mAP^m)	73.85±0.15	**74.68±0.37**	-	-	74.49±0.43
IS (mIoU)	22.96±0.09	-	**23.72±0.16**	-	-
D (RMSE)	5.52±0.02	**5.26±0.01**	-	-	5.38±0.02
ID (l_1 loss)	8.29±0.07	-	-	**8.25±0.27**	-

Table 2. UniNet results on different task combinations: all five tasks together, single task (three of the five tasks, since the other two tasks IS and ID cannot be trained without object detection). Therefore, OD+IS and OD+ID results are reported as proxy single tasks. We also include the commonly used combination of SS+D.

dataset consists of 2975, 500, and 1525 images in the training, validation and test sets, respectively. The images are of resolution 1024×2048. The best fitting bounding box of each ground truth instance polygon is used for bounding box regression.

The NYUv2 [34] is an indoor dataset with image resolution of 480×640 consisting of 795 and 654 training and validation images, respectively.

Training details. The encoder features E2 to E5 are obtained using DLA34 (Deep Layer Aggregation) [48] and the features E6 and E7 are obtained using VoVNet19 [18]. For all models, DLA34 is initialized with COCO [21] pretrained weights unless otherwise mentioned. We train all combinations of multi-task models and single task models for a total of 140 epochs with a learning rate of 0.0001 and stepwise schedule where the learning rate is dropped by a factor of 10 at steps 98 and 126. The input images are resized to resolution 512×1024.

Task evaluation. Object detection and instance segmentation are evaluated at a resolution of 1024×2048. Semantic segmentation, depth estimation and instance depth prediction are evaluated at a resolution of 512×1024. For depth evaluation, we ignore all pixel locations where the ground truth depth is greater than 80 or less than 1e-3. All results are reported on the validation set for both datasets.

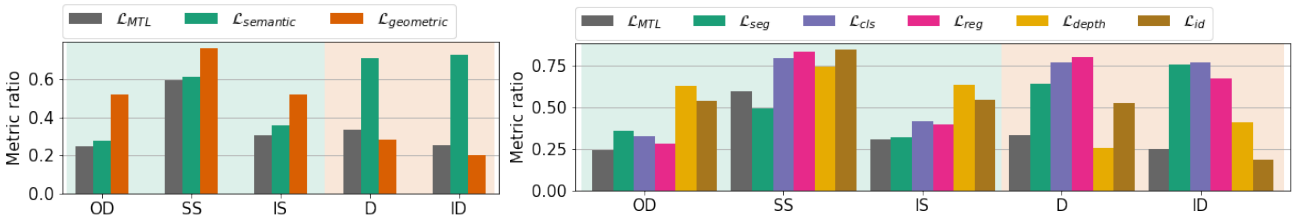
Inference efficiency. To determine inference efficiency, we measure the MAC, number of parameters, inference speed (FPS) and energy consumed by models. FPS and energy are reported on a single NVIDIA RTX-2080 Ti GPU. Following [30] to measure inference energy, we run 500 forward passes with mini-batch size 1 of randomly generated "images" and report the average energy per image. FPS is also measured in a similar fashion using images from the validation set.

Attack evaluation. We use metric ratio to evaluate adversarial attacks. For any given task, metric ratio is the fraction of performance retained after adversarial attack with respect to the performance before attack.

5.2. Multi-Task Performance

Table 1 compares UniNet with state-of-the-art (SOTA) methods in both Cityscapes and NYUv2 datasets on SS+D task. The SOTA methods, the single task networks and the MTL-baseline are trained with the DLA34 backbone using the code provided by [40]. Similar training hyperparameters are used across all methods. In both SS and D, MTI-Net achieves the best overall multi-task performance followed by UniNet. UniNet requires the least amount of computation (lowest MAC) and is also the fastest network (highest FPS). The energy consumption of UniNet is low and is only rivaled by PADNet and MTAN. However, both PADNet and MTAN lag behind UniNet in performance. Overall, UniNet can serve as an attractive baseline to study multi-task relationships due to its simplicity and optimum trade-off between performance and inference efficiency.

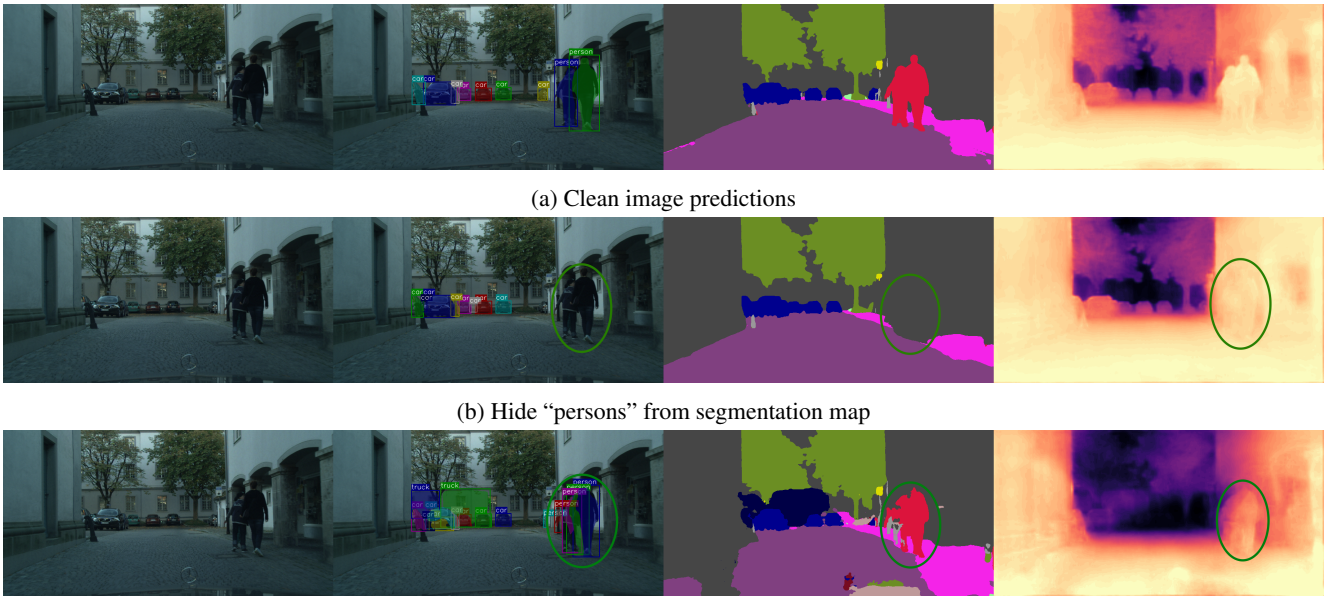
Table 2 shows the results obtained using UniNet. As IS and ID cannot be trained alone, we consider the two task combinations OD+IS and OD+ID as proxy single tasks. Results for further task combinations are provided in Table S1.



(a) Multi-task, semantic and geometric loss

(b) Multi-task and single task loss

Figure 3. PGD attacks using different losses with ϵ=1 as l_∞ perturbation bound. Each bar represents a PGD attack conducted on UniNet with a specific loss (indicated with the bar color). The green shaded and red shaded region contains the metric ratios of semantic tasks and geometric tasks, respectively. Best viewed in color.



(a) Clean image predictions

(b) Hide "persons" from segmentation map

(c) Hide "persons" from depth map

Figure 4. Semantic category hiding with ϵ=2 as l_∞ perturbation bound. (a) shows predictions on clean image. Hiding "persons" from segmentation map in (b) also hides "persons" from all task predictions. However, in (c) similar effect is not observed when hiding "persons" from depth map. Starting from the left, the columns show the input image, the instance predictions, the predicted segmentation map and the predicted depth map. Best viewed in color.

5.3. Semantic and Geometric PGD Attacks

PGD can be used to attack UniNet using different losses as discussed in Section 4.1. We experiment with l_∞ bound, varied ϵ values with step size α=1. The number of attack iterations is determined by $\min(\epsilon + 4, \lceil 1.25\epsilon \rceil)$. In this section, we evaluate the effect of attacking with multi-task, semantic, and geometric losses. Given that the generated perturbation is expected to increase the attack loss, intuitively, the affected image features are important for the task which concerns the attack loss. Detailed results of adversarial attack on different task combinations and single tasks are provided in Table S2.

Figure 3a highlights the metric ratios of all tasks under multi-task loss (MTL loss), semantic loss and geometric loss attacks. In the green region, semantic tasks OD, SS and IS are least affected by geometric loss. As indicated in the figure, semantic tasks retain most of their performance in relation to clean performance under geometric attack. Likewise, geometric tasks are least affected by semantic attacks (red region). This finding indicates that semantic and geometric tasks look at different image features.

5.4. PGD Attacks with Individual Losses

In this section, we analyze the effects of attacking with individual task losses as illustrated in Figure 3b. MTL loss only affects object detection and instance segmentation more than their corresponding individual losses. This supports the notion that the MTL loss cannot effectively fool all tasks inline with Mao *et al*. [26]. Additional observations include (1) all tasks lose the most performance when attacked with the corresponding individual loss in compari-

Tasks	Clean	Hide from		Hide from (rel.)	
		SS	D	SS	D
OD (mAP^b)	35.23	8.95	13.85	0.25	0.39
IS (mAP^m)	16.46	2.83	5.38	0.17	0.33
SS (mIoU)	77.75	10.39	49.31	0.13	0.63
D (RMSE)	5.02	9.15	35.20	0.55	0.14
ID (abs rel.)	0.16	0.29	1.05	0.54	0.15

Table 3. Results on semantic hiding of "person" class. The first three columns show the actual metric values while the last two columns show metric ratios which is the performance retained by each task after attack with respect to clean performance.

son with other individual losses. (2) Semantic segmentation is the least affected task regardless of which loss is used. This indicates that semantic segmentation is the most robust task. We leave a more rigorous study of this observation as future work.

5.5. Semantic and Geometric Task Relationships

In Section 5.3, we observed with the aid of PGD attacks that the semantic and geometric tasks rely on different image features. In this section, we aim to study the relationships between semantic and geometric tasks. These tasks interact in the shared representation and this interaction is learned based on inter-task relationships during training. Semantic category hiding is a targeted attack in which the attack objective is formulated with the goal of hiding certain categories of objects from predictions. Given that the hiding can either be performed on the predicted segmentation map or the depth map, the effect of one prediction based attack on the other provides an opportunity to analyze semantic and geometric task relationships.

Figure 4 shows the visualizations of predictions (a) before attack, (b) and (c) after attack to remove "persons" from segmentation map (segmentation attack) and depth map (depth attack), respectively. In both attacks, the person region is filled using the surrounding building, road, and sidewalk pixels. The segmentation attack in Figure 4b is able to effectively hide "persons" from all predictions, including depth estimation, which is a geometric task. However, the depth attack in Figure 4c is able to blend "persons" depth into the scene but doesn't completely affect semantic task predictions. This result shows that the semantic and geometric task do not affect each other equally.

5.5.1 Role of Representation Levels

To understand the role of representation levels in inter-task relationships, we analyze (i) hiding "persons" from segmentation map (Figure 4b) and (ii) hiding "persons" from depth map (Figure 4c). We combine both analyses and present the final inference.

(i) Figure 4b shows that the segmentation attack is able to fool all tasks resulting in "persons" being hidden from all predictions. Given this observation, we infer that the segmentation attack likely disrupts all levels of representation.

(ii) Figure 4c shows that the depth attack is able to fool the geometric task. Both semantic segmentation and object detection are able to predict "persons" in the right region but the shape of the "person" masks and bounding boxes are affected. We infer that the depth attack has disrupted the low-level representation affecting both depth estimation and "person" shape in semantic tasks. From this we infer that both semantic and geometric tasks strongly interact in low-level representation. However, at higher level representations, this interaction becomes weaker giving semantic tasks the room to still identify the class of "persons".

In summary, (i) indicates that semantic and geometric tasks interact in all levels, and (ii) shows that the interaction is stronger in low-level representations. Thus, we conclude the semantic and geometric tasks have stronger interactions in low-level representations which becomes weaker in high-level representations. Furthermore, we observe that there is an asymmetric relationship between semantic and geometric tasks, that is the semantic tasks have stronger influence on geometric tasks in comparison to the influence of geometric tasks on semantic tasks.

The segmentation IoU of "person" class and depth RMSE of "person" regions in segmentation ground truth, before and after attack are shown in Table 3. Each of the attacks has the most effect on the corresponding task as is indicated by the numbers highlighted in blue. While depth estimation retains only 55% of its performance under segmentation attack, semantic segmentation retains 63% under depth attack.

5.6. Intra-task Relationship Induces Shape Bias

We use the modified DAG attack discussed in Section 4.2 to swap "person" and "car" classes in the bounding box predictions of object detection. We analyze the intra-task relationship in object detection between the classification (**cls**) and regression (**reg**) objectives. Notably, the two objectives are handled by two different branches in the UniNet.

In the case where there is no relationship between **cls** and **reg**, the DAG attack can be expected to only affect the class of bounding boxes. The results before and after DAG attack is presented in Figure 5. Plot (a) shows that DAG attack is effective as the "person" and "car" mAP has dropped considerably relative to clean performance. In plot (b), we see that the classification loss has increased as expected. However, the regression loss in plot (c) also increases suggesting that DAG has exploited the learned interaction between **cls** and **reg** which is a result of their relationship.

In the Cityscapes dataset, "car" boxes are frequently horizontal or approximately square. Figure 6 shows the visualisation of instance predictions (a) on clean image and (b)

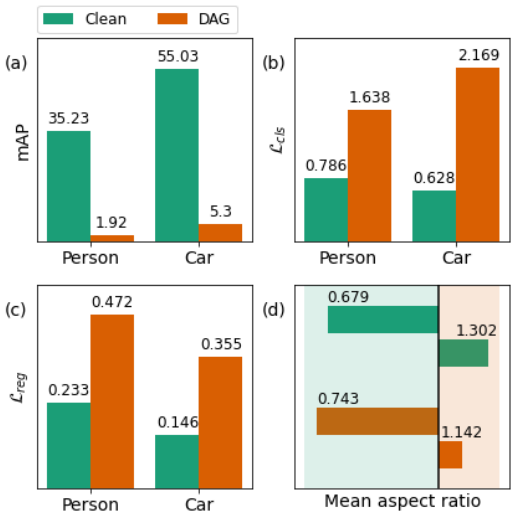


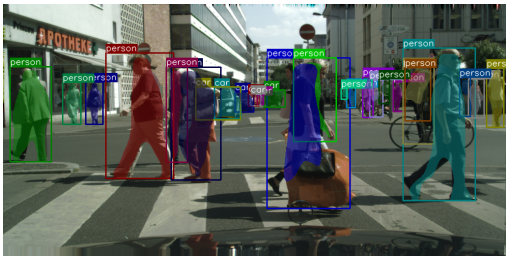
Figure 5. Effect of DAG attack on object detection. Each plot compares before and after DAG attack results specific to the "person" and "car" class. Plots (a), (b) and (c) show the mAP, classification loss and regression loss, respectively. Plot (d) shows the mean aspect ratio of "person" class (green shaded region) and "car" class (red shaded region).

DAG attacked image. We see that the "person" bounding boxes have switched from vertically oriented boxes to either horizontally oriented (green oval region) or approximately square boxes (yellow oval region). This observation suggests that DAG has exploited a shape bias in the learned representation. This shape bias has likely been induced by the relationship between **cls** and **reg** leveraging a similar bias in the dataset during training. As "cars" can generally be expected to have such shapes in the real world, this shape bias is desirable.

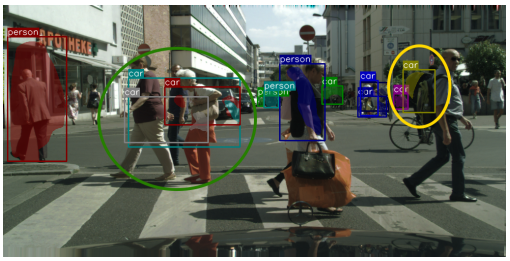
To further evaluate the observed shape bias, we consider the mean aspect ratio of predicted "person" and "car" bounding boxes across the validation set as a proxy indicator of shape. Figure 5 plot (d) shows the mean aspect ratios of "person" and "car" predictions on clean and DAG attacked images. Given that the attack only swaps the two classes, one would expect the mean aspect ratio after attack to switch regions. However, we see that "person" and "car" mean aspect ratios remain below 1 and above 1, respectively. This shows that the aspect ratios have also swapped along with the class, reinforcing the finding that a shape bias exists in the learned representation induced by the intra-task interaction between **cls** and **reg**.

5.7. Ablation study

To check whether the findings are architecture independent, we choose MTI-Net because (i) it has the best performance in segmentation and depth, and (ii) its architecture is quite different from UniNet. We add the instance head to



(a) Predictions on clean image.



(b) Predictions on DAG generated adversarial image.

Figure 6. Shape bias induced by intra-task relationship in object detection. OD and IS predictions visualized on (a) clean image and (b) DAG generated adversarial image. In (b), "car" box predictions on "persons" have horizontal aspect ratio despite that the attack objective only includes box classification. The visualization only includes "car" and "person" predictions. Best viewed in color.

MTI-Net and refer to this new architecture as MTI-Net++. This network is then attacked using all three adversarial attacks revealing findings in line with that obtained using UniNet. These results are provided in the supplementary material and suggest that the conclusions made regarding multi-task relationships are architecture independent.

6. Conclusions

We presented an efficiently designed unified scene understanding network, UniNet. We demonstrated that it has competitive performance with existing works and provides a comprehensive understanding of a scene with the aid of five vision tasks. We introduced adversarial attacks as an exploratory lens to understand and obtain insights about multi-task relationships. With the aid of semantic category hiding, we showed that semantic and geometric tasks have an asymmetric relationship meaning that semantic tasks have a stronger effect on geometric tasks. We also showed that their strong interaction in low-level representations becomes weaker as we move towards high-level representations. With targeted DAG attack, we studied the effects of swapping two instance classes in object detection and showed that intra-task relationship between classification and regression induces a desirable bias towards object shapes in the learned representation. Adversarial attacks thus provide an interesting front to study multi-task relationships.

References

[1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv e-prints*, abs/1812.11941, 2018. 1, 2, 3

[2] A. Arnab, O. Miksik, and P. Torr. On the robustness of semantic segmentation models to adversarial attacks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 888–897, 2018. 2

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 1, 2, 3

[4] R. Caruana. Multitask learning: A knowledge-based source of inductive bias. In *ICML*, 1993. 1

[5] L. Chen, Zeng Yang, J. Ma, and Zheng Luo. Driving scene perception network: Real-time joint detection, depth estimation and semantic segmentation. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1283–1291, 2018. 1, 2, 3

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 833–851, Cham, 2018. Springer International Publishing. 1, 2, 3

[7] Sumanth Chennupati, Ganesh Sistu, S. Yogamani, and S. Rawashdeh. Multinet++: Multi-stream feature aggregation and geometric loss strategy for multi-task learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1200–1210, 2019. 2, 3

[8] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[9] Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12379–12388, 2019. 1, 2

[10] Ross B. Girshick. Fast r-cnn. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 1, 2, 3

[11] Ross B. Girshick, J. Donahue, Trevor Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 2

[12] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 2

[13] Michelle Guo, A. Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *ECCV*, 2018. 2

[14] Kaiming He, Georgia Gkioxari, P. Dollár, and Ross B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 1, 2, 3

[15] Alex Kendall, Yarin Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018. 2

[16] Marvin Klingner, Andreas Bär, and T. Fingscheidt. Improved noise and attack robustness for semantic segmentation by using multi-task training with self-supervised depth estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1299–1309, 2020. 2

[17] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 2

[18] Youngwan Lee, Joong-won Hwang, Sangrok Lee, Yuseok Bae, and Jongyoul Park. An energy and gpu-computation efficient backbone network for real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 5

[19] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13903–13912, 2020. 2

[20] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 3

[21] Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[22] Shikun Liu, Edward Johns, and A. Davison. End-to-end multi-task learning with attention. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1871–1880, 2019. 2, 5

[23] W. Liu, Dragomir Anguelov, D. Erhan, Christian Szegedy, S. Reed, Cheng-Yang Fu, and A. Berg. Ssd: Single shot multibox detector. *ArXiv*, abs/1512.02325, 2016. 1, 2, 3

[24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 1, 2, 3

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2, 4

[26] Chengzhi Mao, Amogh Gupta, V. Nitin, Baishakhi Ray, Shuran Song, Junfeng Yang, and Carl Vondrick. Multitask learning strengthens adversarial robustness. In *ECCV*, 2020. 2, 6

[27] J. H. Metzen, Mummadi Chaithanya Kumar, T. Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. *2017 IEEE International*

Conference on Computer Vision (ICCV), pages 2774–2783, 2017. 2, 4

[28] I. Misra, Abhinav Shrivastava, A. Gupta, and M. Hebert. Cross-stitch networks for multi-task learning. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3994–4003, 2016. 1, 5

[29] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and E. Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *ArXiv*, abs/1606.02147, 2016. 2

[30] Martín Pi Puig, Laura Cristina De Giusti, Marcelo Naiouf, and Armando Eduardo De Giusti. Gpu performance and power consumption analysis: A dct based denoising application. In *XXIII Congreso Argentino de Ciencias de la Computación (La Plata, 2017).*, 2017. 5

[31] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017. 1, 2, 3

[32] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015. 2

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 2

[34] N. Silberman, Derek Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5

[35] J. Song, Yixin Chen, X. Wang, Chengchao Shen, and Mingli Song. Deep model transferability from attribution maps. In *NeurIPS*, 2019. 2

[36] Jie Song, Yixin Chen, Jingwen Ye, Xinchao Wang, Chengchao Shen, Feng Mao, and Mingli Song. Depara: Deep attribution graph for deep knowledge transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[37] Trevor Scott Standley, A. Zamir, Dawn Chen, L. Guibas, Jitendra Malik, and S. Savarese. Which tasks should be learned together in multi-task learning? In *ICML*, 2020. 1, 2

[38] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 2

[39] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 9627–9636, 2019. 1, 2, 3, 4

[40] Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Revisiting multi-task learning in the deep learning era, 2020. 1, 5

[41] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. *ECCV2020*, 2020. 2, 5

[42] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. SOLO: Segmenting objects by locations. In *Proc. Eur. Conf. Computer Vision (ECCV)*, 2020. 2

[43] Alex Wong, Safa Cicek, and Stefano Soatto. Targeted adversarial perturbations for monocular depth prediction. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8486–8497. Curran Associates, Inc., 2020. 2, 4

[44] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and A. Yuille. Adversarial examples for semantic segmentation and object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1378–1387, 2017. 2, 4

[45] D. Xu, Wanli Ouyang, X. Wang, and N. Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018. 1, 2, 5

[46] Wenqiang Xu, Haiyang Wang, Fubo Qi, and Cewu Lu. Explicit shape encoding for real-time instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[47] W. Yin, Yifan Liu, Chunhua Shen, and Y. Yan. Enforcing geometric constraints of virtual normal for depth prediction. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5683–5692, 2019. 1, 2, 3

[48] F. Yu, Dequan Wang, and Trevor Darrell. Deep layer aggregation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2403–2412, 2018. 5

[49] Amir R. Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018. 2

[50] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sünderhauf. Varifocalnet: An iou-aware dense object detector. In *CVPR*, 2021. 3

[51] Rufeng Zhang, Zeyong Tian, Chunhua Shen, Mingyu You, and Y. Yan. Mask encoding for single shot instance segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10223–10232, 2020. 1, 2, 3