Narrowing the Synthetic-to-Real Gap for Thermal Infrared Semantic Image Segmentation Using Diffusion-based Conditional Image Synthesis

Christian Mayr[1,2], Christian Kübler[1], Norbert Haala[2], Michael Teutsch[1]

[1] Hensoldt Optronics GmbH, Germany
{michael.teutsch, christian.kuebler}@hensoldt.net

[2] University of Stuttgart, Germany
norbert.haala@ifp.uni-stuttgart.de

Abstract

Semantic segmentation is the task of assigning a semantic class to each pixel in an image. Due to the high annotation efforts for fully supervised learning of Deep Neural Networks (DNNs) for this task, only rather few comprehensive public datasets exist. This is particularly the case for thermal infrared imagery. To overcome this lack of training data, we propose to utilize conditional image synthesis in the thermal infrared spectrum. Existing semantic segmentation maps are used to condition the image generation process using pretrained text-to-image diffusion models. Therefore, we use the recently published ControlNet and retrain it to synthesize thermal infrared images for given semantic maps. In this way, we can generate large numbers of synthetic images that we can directly use together with the related segmentation map to train reference semantic segmentation approaches in the thermal infrared spectrum. Our experiments demonstrate that we achieve near state-of-the-art performance with pure synthetic training data on the recently published Full-time Multi-modality Benchmark (FMB) dataset and that our trained model shows better generalization ability across datasets. We provide code at https://github.com/HensoldtOptronicsCV/TIRControlNet.

1. Introduction

Semantic segmentation is a subtopic of image segmentation [32]. The task is to assign a semantic class to each pixel in an image. Semantic segmentation is used in several application areas such as automotive [9], remote sensing [63], and medical imaging [40]. In this paper, we focus on automotive applications. While modern deep learning-based approaches provide convincing results, they still require a large amount of training data. This is specifically the case for fully supervised learning schemes, where each pixel in an image has to be annotated [37]. As a result, only rather few comprehensive public datasets exist that provide dense labels for semantic segmentation [9, 33, 43, 62]. However, each of those datasets contains just a few thousand annotated images, which is by far less compared to public datasets annotated for tasks such as object detection [28] or image classification [42]. That is why recent research aims to reduce the annotation complexity for image segmentation [1, 7, 12, 23]. Procedurally generated synthetic data [38, 41, 58] that comes with precise annotation for free can be an alternative to real data but there is still a domain gap regarding scene and object appearance [36].

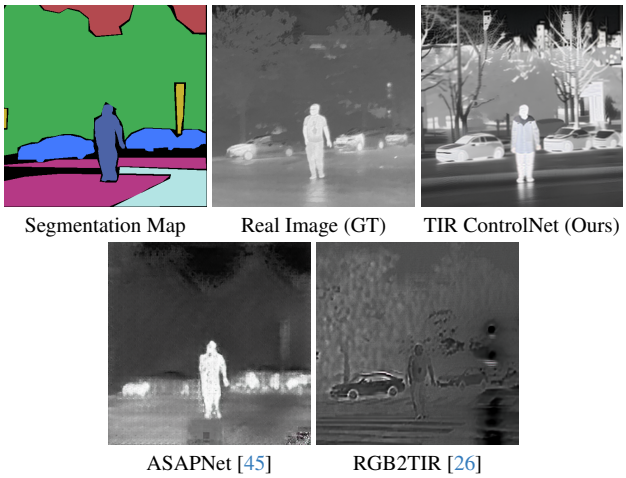


Figure 1. Motivation: in contrast to cGANs for conditional image synthesis via segmentation maps and for VIS-to-TIR image translation, our re-trained segmentation map-guided TIR ControlNet provides more realistic synthetic TIR imagery.

All approaches mentioned so far only concern the Visual-optical (VIS) spectrum. Semantic segmentation, however, is a task highly relevant for the Thermal Infrared (TIR) spectrum as well [52]. Since computer vision in the TIR spectrum is a niche topic in general, only very few literature exists on semantic segmentation approaches and public datasets [22]. To tackle this issue and to avoid any

manual proceducal generation of synthetic TIR imagery [4], we utilize methods for conditional image synthesis in this paper. Diffusion models more and more replace Generative Adversarial Networks (GANs) in image synthesis [10]. With the recently published ControlNet [69], we can condition diffusion model-based image synthesis. ControlNet can be transfer-learned to the long-wave TIR spectrum with the recently released Full-time Multi-modality Benchmark (FMB) dataset [29] that contains 1,500 densely annotated TIR training images. Figure 1 shows the high potential of our re-trained TIR ControlNet compared to conditional GANs (cGANs). Thus, the core idea is to use existing semantic maps taken from the FMB dataset as condition to synthesize more than 120,000 different TIR images. These images are then used to train a reference semantic segmentation approach to analyze if we can narrow the synthetic-to-real gap [36] for semantic segmentation in the TIR spectrum. Our contributions can be summarized as follows:

1. We propose a method to synthesize a potentially unlimited amount of TIR imagery based on ControlNet [69] guided by segmentation maps and utilizing the FMB dataset [29].
2. We train a Transformer-based reference algorithm [61] for semantic segmentation on this synthesized data.
3. In extensive experiments, we show that the proposed approach is able to achieve near state-of-the-art performance in TIR semantic image segmentation compared to a model trained on the real FMB dataset and that our trained model provides a better generalization ability.

The remainder of this paper is organized as follows: related work is presented in Section 2. Our proposed method for the conditional synthesis of TIR imagery and its application to semantic segmentation is presented in Section 3. Experimental results are described in Section 4. We conclude in Section 5.

2. Related Work

Semantic segmentation: Image segmentation [32] and semantic segmentation [15, 25] are among the most popular topics in computer vision. In the VIS spectrum, the task developed from traditional machine learning-based approaches [44] to deep learning-based methods utilizing Convolutional Neural Networks (CNNs) [5, 40, 55] and Vision Transformers [11, 48, 61]. Semantic segmentation in the TIR spectrum is by far less researched [22, 52]. Supervised learning techniques for this task generally suffer from rather small datasets that in some cases are not publicly available as well as incomplete or sparse annotations [14, 19, 24, 27, 46]. Most of the aforementioned datasets provide aligned multi-spectral imagery in the VIS and the TIR spectrum to explore semantic segmentation via spectral fusion, which is promising due to the complementary characteristic of the VIS and the TIR spectrum [22, 52, 54]. In this paper, however, we focus on the TIR spectrum only. The few existing literature on pure TIR semantic segmentation is built up on approaches adopted from the VIS spectrum though [22, 35, 65]. Vision Transformers are outperforming CNNs in semantic segmentation and thus they are currently replacing them [53]. Hence, in this paper we will consider reference approaches in the VIS spectrum for semantic segmentation based on Vision Transformers [61] just like previous literature [22, 29].

Image synthesis: Synthesizing photo-realistic images based on meta-information such as text prompts, semantic maps, depth, etc. is a popular task in computer vision nowadays often referred to as guided or conditional image synthesis [13, 66, 67]. In recent years, deep learning-based approaches for photo-realistic image synthesis evolved from Variational Autoencoders (VAEs) [17] to Generative Adversarial Networks (GANs) [18, 49] and most lately to diffusion models [10]. While conditional GANs used to be the state-of-the-art in image synthesis for many years [18, 30, 45, 49, 56], they are now more and more replaced by guided diffusion models [10, 47, 57, 69]. One reason is that diffusion models, particularly latent diffusion models, are much easier to be trained and utilized compared to GANs [39]. Thermal infrared image synthesis is a niche topic that is often limited to certain tasks and domains [8, 21, 31, 68]. Within this niche, image-to-image translation (also known as style transfer) is the most popular image synthesis method generating corresponding synthetic TIR images for given VIS images [20, 26, 34, 68]. It is usually intended to overcome the lack of training data for deep learning-based approaches in the TIR spectrum. Conditional GANs are typically used for this synthesis. To the best of our knowledge, no literature exists today on utilizing diffusion models for TIR image synthesis guided by semantic maps. However, latent diffusion models are known to synthesize image not with highest precision and texturization [39]. The question remains as to how important this is for TIR images, which are inherently less textured than VIS images [52].

3. Methodology

The methodology is visualized in Fig. 2. For the conditional synthesis of TIR images via segmentation maps, we utilize ControlNet [69] together with its Stable Diffusion backbone. Since ControlNet in its pre-trained version is not able to synthesize TIR images, we perform a re-training using the FMB dataset [29] in Stage 1. After this re-training, we can use different *seeds* to synthesize TIR images with strongly different appearances even for the same segmentation map. This enables us to generate a synthetic training dataset for semantic segmentation that is about ten times larger than the reference dataset FMB. With this large synthetic training dataset, we train a reference approach for se-

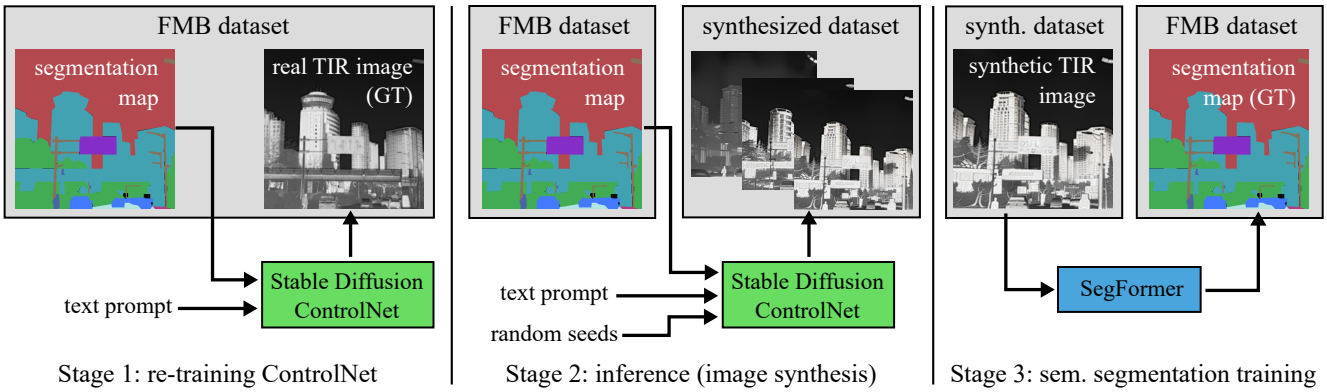


Figure 2. Overview of the methodology. Stage 1: we re-train ControlNet [69] to synthesize TIR imagery by using the real TIR dataset FMB [29] together with its densely annotated segmentation maps. Stage 2: during inference of the re-trained ControlNet, we synthesize TIR images using the real segmentation maps from FMB and random seeds to vary the appearance of the synthetic TIR imagery. In this way, we can generate many differently looking TIR images for the same segmentation map. Stage 3: with this large synthetic TIR dataset and the real segmentation maps, we train a reference approach for semantic segmentation called SegFormer [61] aiming to outperform an *oracle*, which is SegFormer trained on the real TIR images of the FMB dataset. We always use the same text prompt: 'urban automotive scene containing vegetation, vehicles, road, buildings, and persons'.

mantic segmentation [61] aiming to bridge the synthetic-to-real (syn2real) gap. This means that we perform similarly on the FMB test dataset with the reference approach trained on pure synthetic data compared to an *oracle*, which is the same reference approach trained on the real training data of the FMB dataset. In the remainder of this section, we describe the re-training of ControlNet for TIR image synthesis and the training of the reference approach for semantic segmentation in more detail.

3.1. Conditional Thermal Image Synthesis

Diffusion Models progressively denoise a normally distributed random variable to learn a data distribution that reverts the process of a fixed Markov chain [39]. In this way, they can generate photo-realistic images from random noise [16]. ControlNet [69] aims to gain control over the diffusion process based on given control signals such as segmentation maps, depth maps, edge maps, or even sketches. The control signals are injected directly into the diffusion process during inference, using a trainable copy of its diffusion backbone *Stable Diffusion*. This connection via so-called *zero convolutions* allows the model to ingest conditioning inputs, guiding the sampling process of the diffusion model and enabling the generation of images aligned with the given conditions. The architecture of ControlNet preserves the quality and capabilities of the pre-trained Stable Diffusion model while enabling the learning of diverse conditional controls at the same time. ControlNet can take a text prompt to define the search space for image synthesis via text input. However, this text prompt is not mandatory [69]. As we expect faster convergence of the training process, we always prompt the model with a fixed text prompt that describes our domain of interest: 'urban automotive scene containing vegetation, vehicles, road, buildings, and persons'. We utilize ControlNet without any modification of its architecture or training algorithm. Our re-trained version of ControlNet for conditional synthesis of TIR imagery is called TIR ControlNet.

3.2. Dataset Preparation and ControlNet Training

Since the pre-trained ControlNet is able to synthesize VIS images but not TIR images, we re-train the model with TIR imagery aiming to transfer its capabilities to the thermal infrared spectrum. The authors recommend to use about 50,000 training samples for the re-training of ControlNet. No publicly available TIR dataset fulfills this requirement at this time. The Freiburg Thermal dataset [54] is a candidate as it comes with about 20,000 training images. However, several observations made us not use this dataset for re-training: (1) the 20,000 training images were collected within just eight acquired videos leading to a limited diversity, (2) the ground truth segmentation maps are automatically generated and thus not fully densely annotated, and (3) the TIR images are provided with a bit depth of 16 bit, which raises the need for additional tone mapping [51] thus reducing the reproducibility of the results. Instead, we utilize the recently published FMB dataset [29]. The dataset is multi-spectral with aligned VIS and TIR images, but in this paper, we only use the thermal spectrum. The FMB just provides 1,500 images, but highly diverse scenes are shown and it comes with high-quality dense annotations for semantic segmentation. The data is split into 1,220 images for training and 280 for testing. The image resolution is 800×600 pixels. Image synthesis via ControlNet, however,

Dataset	Unlabeled	Road	Sidewalk	Building	T-Light	T-Sign	Vegetation	Sky	Person	Car	Truck	Bus	Motorcycle	Bicycle	Pole	Images
FMB train	1.94%	13.45%	1.57%	15.04%	0.10%	0.25%	24.02%	35.86%	0.33%	5.15%	0.37%	0.34%	0.05%	0.00%	1.53%	1,220
FMB test	1.41%	11.33%	1.60%	19.62%	0.18%	0.42%	23.49%	32.53%	0.64%	5.84%	0.38%	0.85%	0.04%	0.00%	1.67%	280
FMB-aug train	1.71%	10.90%	1.24%	16.00%	0.10%	0.27%	23.98%	37.88%	0.33%	5.13%	0.35%	0.32%	0.06%	0.00%	1.73%	12,200
FMB-aug test	0.90%	9.05%	1.05%	20.37%	0.24%	0.60%	21.26%	36.65%	0.79%	5.91%	0.36%	0.97%	0.02%	0.00%	1.83%	280

Table 1. Pixel-wise class distribution of the original FMB dataset and the augmented (FMB-aug) dataset. While for each image of the FMB train dataset five crops of 512×512 pixels are created to produce the FMB-aug train dataset (needed for re-training of ControlNet), the FMB-aug test dataset is created by cropping just the image center with a resolution of 512×512 pixels. After augmentation, each image has a resolution of 512×512 pixels, which is exactly the output resolution of ControlNet. In this way, we avoid any image resizing. By pre-calculating and fixing the data augmentation, we achieve reproducibility for the experiments on image synthesis in Section 4.

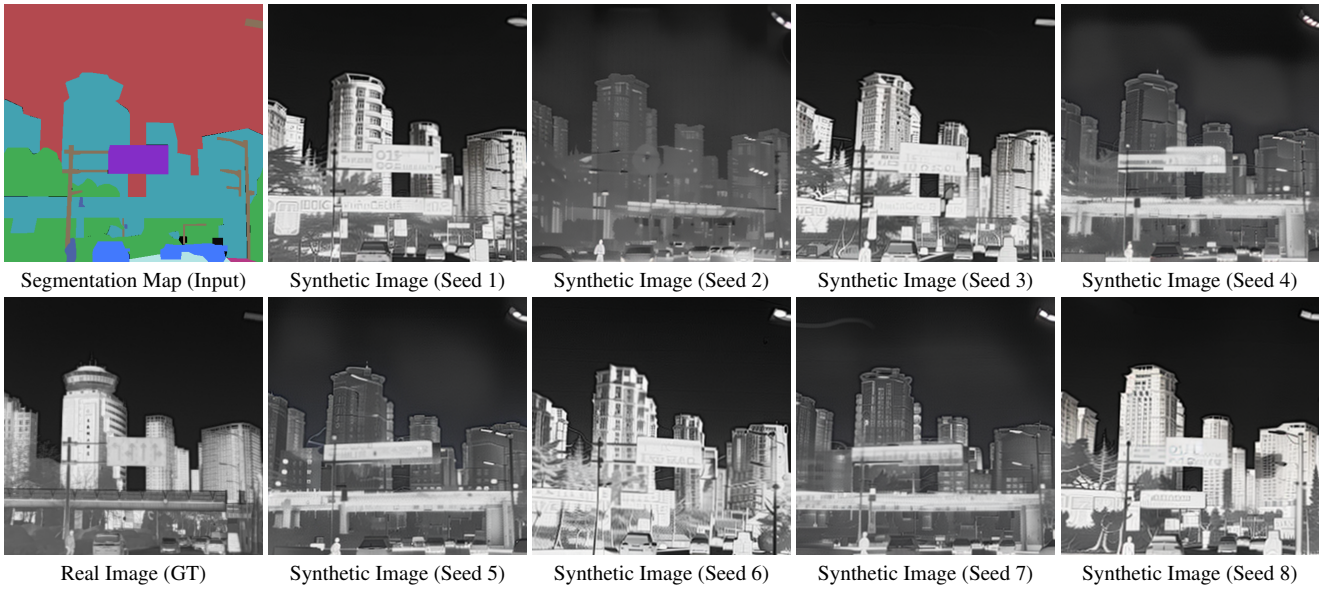


Figure 3. On the left is the reference segmentation map (input) and the related real image (ground truth). The remaining eight images show the influence of different seeds during image synthesis using ControlNet [69] re-trained by us with the FMB dataset [29]. The seed is used to create the random noise that is 'denoised' during the diffusion process. For all images shown, the diffusion *steps* value was set to 70.

comes with a fixed output resolution of 512 × 512 pixels. We avoid to introduce any bias of image resizing by keeping the image resolution constant throughout our methodology and experiments. This enables us to perform data augmentation techniques such as cropping and flipping to extend the amount of training data coming from the FMB dataset. Cropping is performed by cutting tiles of 512×512 pixels from the center and each corner of each 800×600 pixels image. Each crop is flipped as well. In this way, we generate 12,200 training samples. An overview of the pixel-wise class distribution of the original and the augmented FMB (FMB-aug) dataset is shown in Table 1. By pre-calculating and fixing the data augmentation, we achieve reproducibility for the experiments on image synthesis in Section 4. To avoid any bias introduced by the text prompt, we always use the same prompt as already mentioned in the previous subsection. During training (see Fig. 2 Stage 1), we use the FMB dataset's real segmentation maps as input and the real TIR images as learning objective. The sudden convergence [69] occurred somewhere between 60,000 and 120,000 training images.

During inference (see Fig. 2 Stage 2), we can synthesize as many images as we like by using a segmentation map as input together with a random seed that initializes the random noise for the diffusion process. With the *steps* parameter, we can control the number of denoising steps. Some example images generated from the same segmentation map using different random seed values are shown in Fig. 3. Using the real 12,200 segmentation maps of the FMB-aug dataset, we generate 122,000 synthetic TIR images for to train semantic segmentation (see Fig. 2 Stage 3). Although we do have only 12,200 training samples for the re-training of ControlNet instead of the recommended 50,000, the diversity of the generated TIR images in Fig. 3 indicates that we do not have a strong dataset bias in our image synthesis.

3.3. Semantic Segmentation Training

For our experiments on semantic segmentation, we utilize SegFormer [61] since it is based on a Transformer architecture and since it was recently used for relevant related work [22, 29]. SegFormer is a rather simple but effective approach consisting of a hierarchical Tranformer encoder generating both high-resolution coarse features and low-resolution fine features as well as a lightweight decoder based on Multi-Layer Perceptron (MLP) fusing the multi-level features. The training itself is inspired by [29]: they train SegFormer with 15,000 iterations and a batch size of 8 processing 80,000 training images. We train SegFormer for 10 epochs on the FMB-aug dataset with a batch size of 8 and 12,200 training images per epoch leading to 15,250 iterations. In contrast to other related work [2, 6], in which synthetic training data is used, we avoid any domain adaptation techniques.

4. Experiments and Results

The experiments are set up to analyze three key aspects: (1) the robustness of TIR image synthesis to relevant hyperparameters via ablation studies, (2) the exploration of the syn2real gap on the reference dataset FMB, and (3) the generalization ability of the proposed approach when applied to other datasets for TIR semantic segmentation. FMB [29] is used as main dataset for the experiments. It is larger compared to other recently published datasets [19] and it contains dense labels in comparison to other related datasets [14]. For all experiments we consider SegFormer [61] as reference approach for sematic segmentation. SegFormer uses a ViT-based architecture as backbone and it was used in recent related work as reference approach as well [22]. For the quantitative evaluation, we use the two most common measure in semantic segmentation: the mean Intersection-over-Union (mIoU) and the average class accuracy known as mean Accuracy (mAcc). Intersection-over-Union and accuracy are calculated for each class separately, and then averaged over all classes providing global values for mAcc and mIoU. Both measures follow the principle *the higher the better*.

4.1. Ablation Studies

During TIR image synthesis using ControlNet, we basically discovered only one relevant and potentially impactful hyperparameter: the *steps*. This parameter determines the number of diffusion steps, *i.e.* the number of denoising steps within the Markov process of Stable Diffusion. The assumption is that a larger number of denoising steps leads to a higher level of texturation in the synthesized image. Hence, this parameter is highly relevant for narrowing the syn2real gap: fine texture information is often considered as one of the major reasons for the appearance gap between synthetic and real data [36] and Stable Diffusion is known to synthesize image not with highest level of texturization [39]. The default parameter for steps in ControlNet is 50. As we qualitatively discovered that 50 steps seem to be too small for TIR image synthesis (see Fig. 4), we start our ablation study with 50 moving on to 70 and 100 steps. SegFormer as the reference approach for semantic segmentation is left unchanged in its hyperparameters. For each value of the steps parameter, we generate 10,000 synthetic images using ControlNet. Then, we train one SegFormer model for each resulting synthetic training dataset an evaluate it on the FMB test dataset. We train for 10 epochs with a batch size of 8. The results are shown in Table 2. While the quantitative evaluation does not provide a clear indication for the best choice of the steps value, the qualitative evaluation show that at least 70 steps is a good value for TIR image synthesis using our re-trained ControlNet. Hence, we set the steps parameter to 70 to save some time during image synthesis since 70 steps take about 3.9 seconds per image, while 100 steps take about 7 seconds.

Steps	mIoU ↑	mAcc ↑
50	**47.5**	**57.2**
70	47.2	57.1
100	**47.5**	**57.2**

Table 2. Influence of Stable Diffusion and ControlNet's *steps* parameter on the SegFormer performance regarding mIoU and mAcc. There is no clear tendency.

4.2. Syn2real Gap on the FMB Dataset

To explore the syn2real gap when training the task of semantic segmentation with the synthesized images and the real images taken from the FMB dataset, we chose the following approach: we train SegFormer on the real FMB-aug training dataset for 10 epochs as mentioned before in Section 3.3. This approach is called *oracle* and it serves as baseline for in-domain fully supervised learning. Our approach is using training images synthesized by TIR ControlNet based on the 12,200 segmentation maps taken from the FMB-aug dataset. To unleash the potential of diffusion-based synthetic training data generation, we create 122,000 training images from those 12,200 segmentation maps. SegFormer is trained for one epoch on these 122,000 training images. Interestingly, we did not observe an increasing performance with a larger number of epochs. We assume that the model may overfit to the rather small number of different segmentation maps.

We compare this approach to several state-of-the-art methods for conditional image synthesis: we generate 12,200 training images from the related FMB-aug segmentation maps using the conditional GAN ASAPNet [45] that

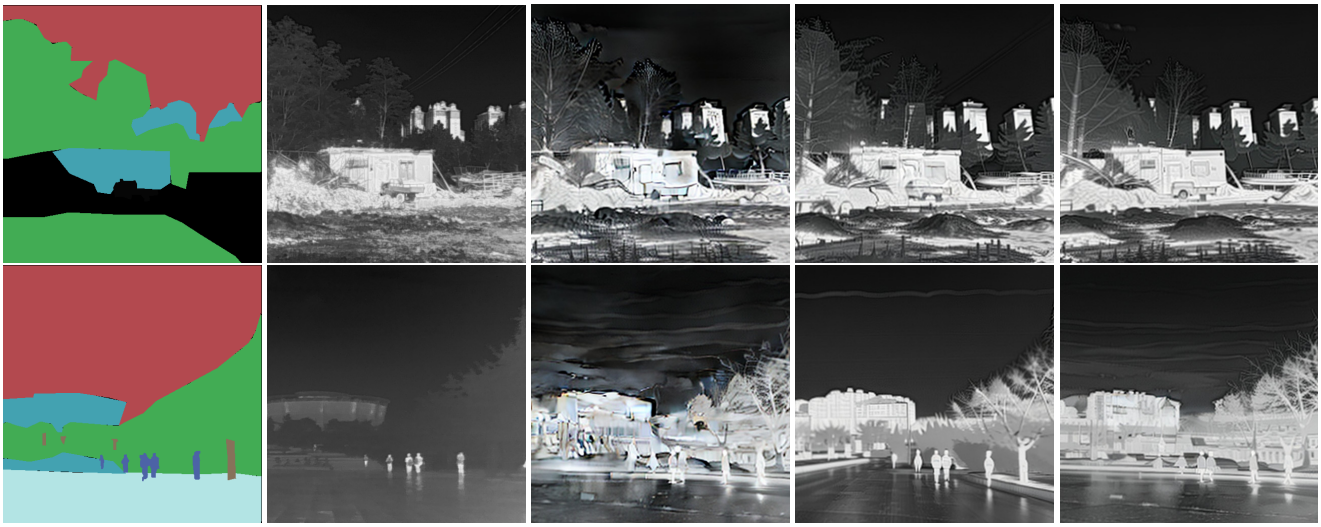


Figure 4. Ablation study to showcase the inference step influence on the 'denoising' process. 70 and 100 denoising steps qualitatively perform best for conditional TIR image synthesis with our re-trained TIR ControlNet.

we re-trained on the FMB dataset. Furthermore, we generate 12,200 training images from the VIS images of the multi-spectral FMB dataset using the most recent VIS-to-IR image translation approach RGB2TIR [26]. This method was trained on other datasets [50, 64] but we were not able to re-train it on the FMB dataset as the related GitHub repository does not provide code for that. Since we are limited by the number of segmentation maps and the number of real VIS images taken from the FMB dataset, we cannot generate more than 12,200 training images by using the related work mentioned before [26, 45]. SegFormer is trained on these training datasets similarly like on the real FMB-aug dataset for 10 epochs. The test set consists of 280 images. The image resolution is 800×600 pixels, so we crop a tile of size 512×512 pixels at the image center to avoid any image resizing. The results of the quantitative evaluation are shown in Table 3 with a detailed analysis of the individual classes in Table 4.

SegFormer [61] Training Dataset	**Data Type**	**mIoU ↑**	**mAcc ↑**
FMB-aug (Oracle)	Real	51.0	61.5
RGB2TIR Translation [26]	Synthetic	34.7	43.9
ASAPNet [45]	Synthetic	43.1	57.6
TIR ControlNet (Ours)	Synthetic	**47.8**	**57.8**

Table 3. TIR ControlNet as conditional image synthesis method is able to outperform the conditional GANs RGB2TIR and ASAPNet on the task of semantic segmentation using SegFormer. The relative performance gap compared to the oracle is less than 7 %.

Our oracle is better compared to the SegFormer of the original paper [29]. However, here we use images with a different resolution of 512×512 and not all classes are considered in the quantitative evaluation in the original paper. Our proposed TIR ControlNet is able to synthesize images that provide a better training dataset for SegFormer compared to other state-of-the-art conditional image synthesis methods based on cGANs. The relative performance gap compared to the oracle is less than 7 %. A qualitative assessment of the image synthesis quality is given in Fig. 5. Furthermore, a qualitative evaluation of the performance in semantic segmentation is provided in Fig. 6.

4.3. Generalization Ability

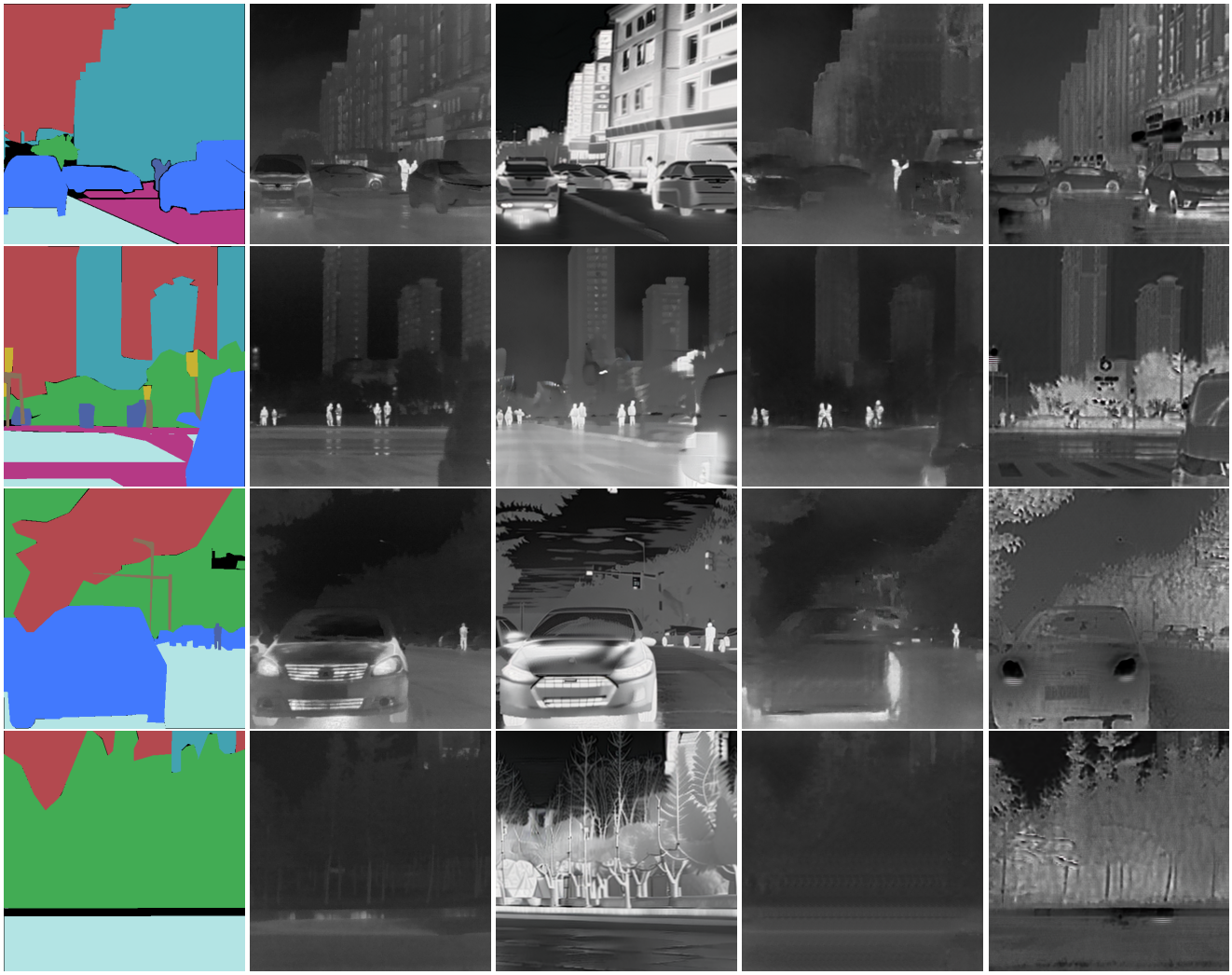
The generalization ability is analyzed using the Freiburg Thermal dataset [54]. We take the same SegFormer models as presented in Section 4.2 and apply them to the test data of the Freiburg Thermal. This test set consists of 64 images with manually annotated ground truth for semantic segmentation. The image resolution is 1920×650 pixels, so we crop a tile of size 512×512 pixels at the image center to avoid any image resizing. Since the TIR images are provided with a bit depth of 16 bit, we perform tone mapping[1] to generate 8 bit images in order to be compliant with the TIR image data format of the FMB dataset.

The results are shown in Table 5. Our SegFormer model trained with the synthetic dataset generated by the TIR ControlNet clearly outperforms the other approaches regarding mIoU and mAcc. We did not fine-tune any approach to the Freiburg Thermal dataset including the SegFormer trained on the real FMB-aug dataset. A qualitative evaluation is

[1] https://github.com/KABIR-VERMA/Tone-Mapping-HDR-Images

SegFormer [61]	Road		Sidewalk		Building		T-Light		T-Sign		Vegetation		Sky		Person		Car		Truck		Bus		Motorcycle		Pole	
Training Dataset	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc
FMB-aug	87.8	96.2	28.6	32.3	82.7	90.6	19.5	20.8	68.0	74.3	81.6	90.5	94.7	98.1	72.1	78.3	78.2	93.1	13.9	14.9	36.3	41.2	20.6	31.9	30.6	37.0
RGB2TIR [26]	78.9	91.3	12.1	15.7	67.4	88.4	12.2	12.9	2.3	2.3	57.6	66.1	89.7	96.0	63.5	74.3	67.6	85.4	2.8	3.8	12.5	12.9	4.3	5.5	14.3	15.5
ASAPNet [45]	83.7	92.2	30.1	46.8	73.0	80.6	18.3	20.1	58.8	72.3	73.2	88.1	89.2	93.8	59.5	65.0	75.0	89.3	1.7	2.9	56.6	70.0	10.6	54.4	25.5	36.5
ControlNet (Ours)	87.2	94.7	24.1	27.7	80.5	89.2	13.1	13.4	67.8	71.4	80.0	91.0	93.3	98.0	64.7	66.7	81.0	89.3	10.0	21.2	15.3	17.7	28.1	43.3	24.2	27.4

Table 4. Detailed quantitative evaluation for the individual classes of the FMB dataset. The mIoU and mAcc values can be found in Table 3.



Segmentation Map (Input) | Real Image (GT) | TIR ControlNet (Ours) | ASAPNet [45] | RGB2TIR Translation [26]

Figure 5. Qualitative assessment showcasing the image synthesis quality of the approaches used in this work. We compare our proposed re-trained TIR ControlNet with two conditional GANs: ASAPNet [45] for conditional image synthesis using segmentation maps as guidance and RGB2TIR translation network [26]. The segmentation maps together with the related real images are taken from the FMB dataset.

provided in Fig. 7, which visually confirms the quantitative results reported in Table 5.

5. Conclusion

Leveraging most recent research findings [29, 69], we used Stable Diffusion and ControlNet together with the FMB dataset to train a method that performs realistic conditional TIR image synthesis. Semantic maps provide the guidance for the image generation process. In this way, we were able to generate large amounts of synthetic training data for semantic image segmentation. With this data, we trained a reference approach called SegFormer. This approach trained on purely synthetic data achieved near state-of-the-art performance compared to SegFormer trained on the real FMB training dataset. Furthermore, it outperformed other conditional image synthesis approaches such as a cGAN guided by semantic maps and a cGAN

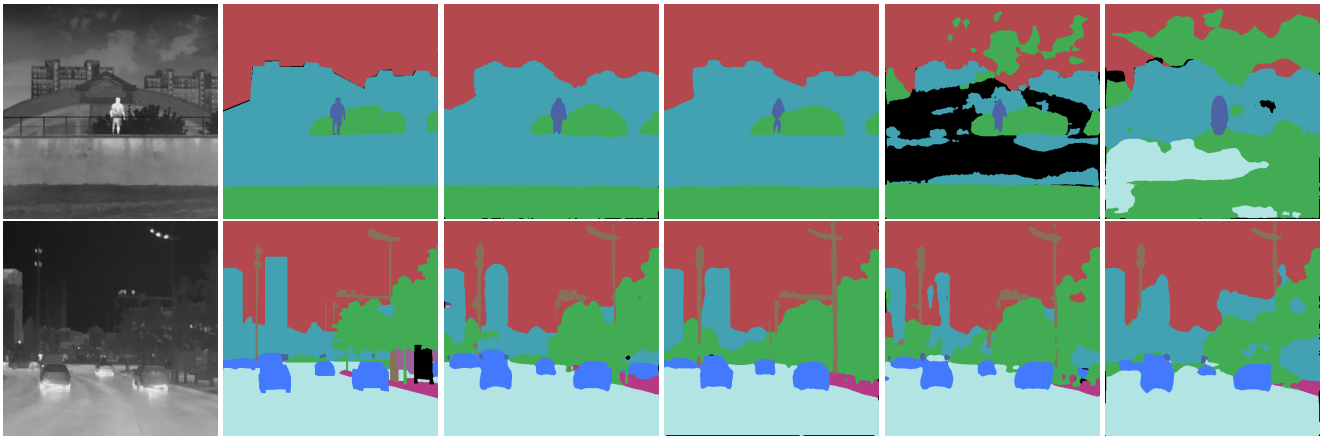


Figure 6. Qualitative evaluation for semantic segmentation on the FMB dataset. We compare the SegFormer network [61] trained on real data (FMB-aug) also called *oracle* with SegFormer trained on different synthetically generated training datasets. Our proposed SegFomer trained with data synthesized by TIR ControlNet called *SegFormer (Ours)* provides the best synthetic training data for SegFormer as it produces the most precise segmentation maps deviating only slightly from the Ground Truth and the oracle.

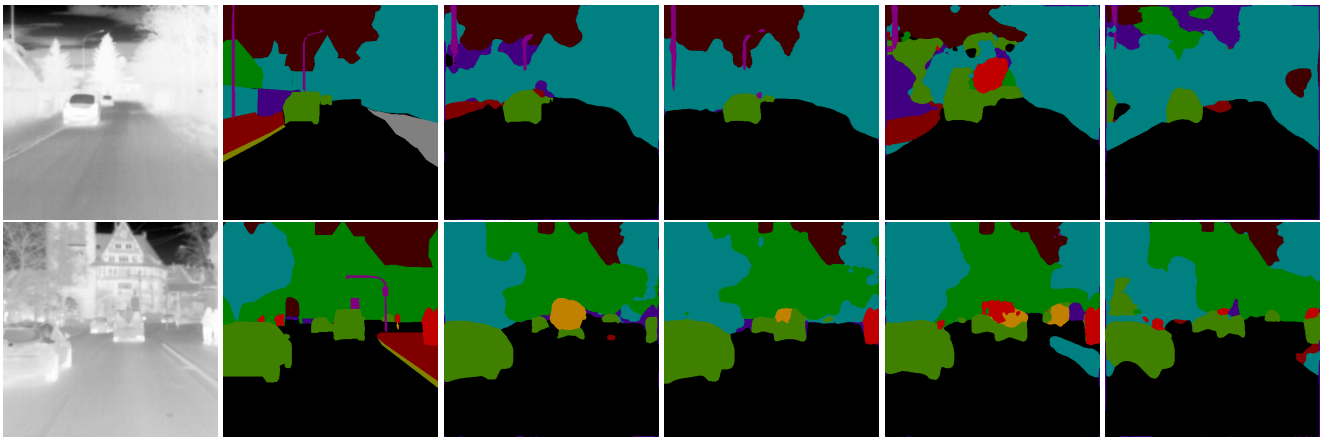


Figure 7. Qualitative evaluation of the generalization ability for semantic segmentation on the Freiburg Thermal test dataset. We compare the SegFormer network [61] trained on real data (FMB-aug, not Freiburg Thermal) with SegFormer trained on different synthetically generated training datasets. Our proposed SegFomer trained with data synthesized by TIR ControlNet called *SegFormer (Ours)* provides the best synthetic training data for SegFormer as it produces the most precise segmentation maps compared to the Ground Truth.

SegFormer [61] Training Dataset	**Data Type**	**mIoU ↑**	**mAcc ↑**
FMB-aug	Real	42.4	57.4
RGB2TIR Translation [26]	Synthetic	29.5	44.2
ASAPNet [45]	Synthetic	36.9	55.4
TIR ControlNet (Ours)	Synthetic	**49.5**	**59.2**

Table 5. Analysis of the model's generalization ability using the Freiburg Thermal dataset. Our proposed TIR ControlNet as conditional image synthesis method clearly outperforms all other methods on the task of semantic segmentation using SegFormer.

for VIS-to-TIR image translation. Our methods also shows better generalization ability when applied to the Freiburg Thermal dataset. Future work should consider to synthesize TIR videos via temporal consistency [3]. And as an alternative to ControlNet, recent approaches synthesize ground truth together with the generated image, which can be another option for future work [59, 60].

References

[1] Sharat Agarwal, Saket Anand, and Chetan Arora. Reducing Annotation Effort by Identifying and Labeling Contextually Diverse Classes for Semantic Segmentation Under Domain

Shift. In *IEEE WACV*, 2023. 1

[2] Roberto Alcover-Couso, Juan C. SanMiguel, Marcos Escudero-Viñolo, and Alvaro Garcia-Martin. On exploring weakly supervised domain adaptation strategies for semantic segmentation using synthetic data. *Multimedia Tools Appl.*, 82(23):35879–35911, 2023. 5

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. In *IEEE CVPR*, 2023. 8

[4] Francesco Bongini, Lorenzo Berlincioni, Marco Bertini, and Alberto Del Bimbo. Partially Fake it Till you Make It: Mixing Real and Fake Thermal Images for Improved Object Detection. In *29th ACM International Conference on Multimedia (MM)*, 2021. 2

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 834–848, 2018. 2

[6] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning Semantic Segmentation from Synthetic Data: A Geometrically Guided Input-Output Adaptation Approach. In *IEEE CVPR*, 2019. 5

[7] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-Supervised Instance Segmentation. In *IEEE CVPR*, 2022. 1

[8] Sheng-Yang Chiu, Yu-Chee Tseng, and Jen-Jee Chen. Low-Resolution Thermal Sensor-Guided Image Synthesis. In *IEEE WACV Workshops*, 2023. 2

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, U. Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *IEEE CVPR*, 2016. 1

[10] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*, 2021. 2

[11] Jian Ding, Nan Xue, Gui-Song Xia, Bernt Schiele, and Dengxin Dai. HGFormer: Hierarchical Grouping Transformer for Domain Generalized Semantic Segmentation. In *IEEE CVPR*, 2023. 2

[12] Marta Fernandez-Moreno, Bo Lei, Elizabeth A. Holm, Pablo Mesejo, and Raul Moreno. Exploring the trade-off between performance and annotation complexity in semantic segmentation. *Engineering Applications of Artificial Intelligence*, 123, 2023. 1

[13] Stanislav Frolov, Tobias Hinz, Federico Raue, Jörn Hees, and Andreas Dengel. Adversarial text-to-image synthesis: A review. *Neural Networks*, 144:187–209, 2021. 2

[14] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017. 2, 5

[15] Shijie Hao, Yuan Zhou, and Yanrong Guo. A Brief Survey on Semantic Segmentation with Deep Learning. *Neurocomputing*, 406:302–321, 2020. 2

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020. 3

[17] Huaibo Huang, Zhihang Li, Ran He, Zhenan Sun, and Tieniu Tan. IntroVAE: Introspective Variational Autoencoders for Photographic Image Synthesis. In *NeurIPS*, 2018. 2

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE CVPR*, 2017. 2

[19] Wei Ji, Jingjing Li, Cheng Bian, Zongwei Zhou, Jiaying Zhao, Alan L. Yuille, and Li Cheng. Multispectral Video Semantic Segmentation: A Benchmark Dataset and Baseline. In *IEEE CVPR*, 2023. 2, 5

[20] My Kieu, Lorenzo Berlincioni, Leonardo Galteri, Marco Bertini, Andrew D. Bagdanov, and Alberto del Bimbo. Robust pedestrian detection in thermal imagery using synthesized images. In *International Conference on Pattern Recognition (ICPR)*, 2021. 2

[21] Vladimir V. Kniaz, Jiri Hladuvka Vladimir A. Knyaz, Walter G. Kropatsch, and Vladimir Mizginov. ThermalGAN: Multimodal Color-to-Thermal Image Translation for Person Re-Identification in Multispectral Dataset. In *ECCV Workshops*, 2018. 2

[22] Zülfiye Kütük and Görkem Algan. Semantic Segmentation for Thermal Images: A Comparative Survey. In *IEEE CVPR Workshops*, 2022. 1, 2, 5

[23] Hyeokjun Kweon, Sung-Hoon Yoon, and Kuk-Jin Yoon. Weakly Supervised Semantic Segmentation via Adversarial Learning of Classifier and Reconstructor. In *IEEE CVPR*, 2023. 1

[24] Xin Lan, Xiaojing Gu, and Xingsheng Gu. MMNet: Multimodal multi-stage network for RGB-T image semantic segmentation. *Applied Intelligence*, 52(5):5817–5829, 2022. 2

[25] Fahad Lateef and Yassine Ruichek. Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, 338:321–348, 2019. 2

[26] Dong-Guw Lee, Myung-Hwan Jeon, Younggun Cho, and Ayoung Kim. Edge-guided Multi-domain RGB-to-TIR image Translation for Training Vision Tasks with Challenging Labels. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 1, 2, 6, 7, 8

[27] Chenglong Li, Wei Xia, Yan Yan, Bin Luo, and Jin Tang. Segmenting Objects in Day and Night: Edge-Conditioned CNN for Thermal Image Semantic Segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32 (7):3069–3082, 2021. 2

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 1

[29] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive Feature Learning and a Full-time Multi-modality Benchmark for Image Fusion and Segmentation. In *IEEE ICCV*, 2023. 2, 3, 4, 5, 6, 7

[30] Ming-Yu Liu, Xun Huang, Jiahui Yu, Ting-Chun Wang, and Arun Mallya. Generative Adversarial Networks for Image and Video Synthesis: Algorithms and Applications. *Proceedings of the IEEE*, 109(5):839–862, 2021. 2

[31] Neelu Madan, Mia Sandra Nicole Siemon, Magnus Kaufmann Gjerde, Bastian Starup Petersson, Arijus Grotuzas, Malthe Aaholm Esbensen, Ivan Adriyanov Nikolov, Mark Philip Philipsen, Kamal Nasrollahi, and Thomas B. Moeslund. ThermalSynth: A Novel Approach for Generating Synthetic Thermal Human Scenarios. In *IEEE WACV Workshops*, 2023. 2

[32] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(07): 3523–3542, 2022. 1, 2

[33] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In *IEEE ICCV*, 2017. 1

[34] Mehmet Akif Özkanoglu and Sedat Ozer. InfraGAN: A GAN architecture to transfer visible images to infrared domain. *Pattern Recognition Letters*, 155:69–76, 2022. 2

[35] Karen Panetta, Shreyas Kamath, Srijith Rajeev, and Sos S. Agaian. FTNet: Feature Transverse Network for Thermal Image Semantic Segmentation. *IEEE Access*, 9:145212–145227, 2021. 2

[36] Viraj Prabhu, David Acuna, Andrew Liao, Rafid Mahmood, Marc T. Law, Judy Hoffman, Sanja Fidler, and James Lucas. Bridging the Sim2Real gap with CARE: Supervised Detection Adaptation with Conditional Alignment and Reweighting, 2023. 1, 2, 5

[37] Aneesh Rangnekar, Christopher Kanan, and Matthew Hoffman. Semantic Segmentation With Active Semi-Supervised Learning. In *IEEE WACV*, 2023. 1

[38] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for Data: Ground Truth from Computer Games. In *ECCV*, 2016. 1

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *IEEE CVPR*, 2022. 2, 3, 5

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. 1, 2

[41] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *IEEE CVPR*, 2016. 1

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 1

[43] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The Adverse Conditions Dataset with Correspondences for Semantic Driving Scene Understanding . In *IEEE ICCV*, 2021. 1

[44] Florian Schroff, Antonio Criminisi, and Andrew Zisserman. Object Class Segmentation using Random Forests. In *BMVC*, 2008. 2

[45] Tamar Rott Shaham, Michael Gharbi, Richard Zhang, Eli Shechtman, and Tomer Michaeli. Spatially-Adaptive Pixelwise Networks for Fast Image Translation. In *IEEE CVPR*, 2021. 1, 2, 5, 6, 7, 8

[46] Shreyas S. Shivakumar, Neil Rodrigues, Alex Zhou, Ian D. Miller, Vijay Kumar, and Camillo J. Taylor. PST900: RGB-Thermal Calibration, Dataset and Segmentation Network. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2020. 2

[47] Jaskirat Singh, Stephen Gould, and Liang Zheng. High-Fidelity Guided Image Synthesis with Latent Diffusion Models. In *IEEE CVPR*, 2023. 2

[48] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for Semantic Segmentation. In *IEEE ICCV*, 2021. 2

[49] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Jürgen Gall, Bernt Schiele, and Anna Khoreva. OASIS: Only Adversarial Supervision for Semantic Image Synthesis. *International Journal of Computer Vision (IJCV)*, 130:2903–2923, 2022. 2

[50] Teledyne FLIR Systems. FREE Teledyne FLIR Thermal Dataset for Algorithm Training, 2021. [Online; accessed 2024-03-15]. 6

[51] Michael Teutsch, Simone Sedelmaier, Sebastian Moosbauer, Gabriel Eilertsen, and Thomas Walter. An Evaluation of Objective Image Quality Assessment for Thermal Infrared Video Tone Mapping. In *IEEE CVPR Workshops*, 2020. 3

[52] Michael Teutsch, Angel D. Sappa, and Riad I. Hammoud. *Computer Vision in the Infrared Spectrum: Challenges and Approaches*. Springer, 2022. 1, 2

[53] Hans Thisanke, Chamli Deshan, Kavindu Chamith, Sachith Seneviratne, Rajith Vidanaarachchi, and Damayanthi Herath. Semantic segmentation using Vision Transformers: A survey. *Engineering Applications of Artificial Intelligence*, 126, 2023. 2

[54] Johan Vertens, Jannik Zürn, and Wolfram Burgard. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. 2, 3, 6

[55] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2021. 2

[56] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *IEEE CVPR*, 2018. 2

[57] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic Image Synthesis via Diffusion Models, 2022. 2

[58] Magnus Wrenninge and Jonas Unger. Synscapes: A Photorealistic Synthetic Dataset for Street Scene Parsing, 2018. 1

[59] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua

Shen. DatasetDM: Synthesizing Data with Perception Annotations Using Diffusion Models. In *NeurIPS*, 2023. 8

[60] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. DiffuMask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *IEEE ICCV*, 2023. 8

[61] 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. 2, 3, 5, 6, 7, 8

[62] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *IEEE CVPR*, 2020. 1

[63] Xiaohui Yuan, Jianfang Shi, and Lichuan Gu. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169, 2021. 1

[64] Seungsang Yun, Minwoo Jung, Jeongyun Kim, Sangwoo Jung, Younghun Cho, Myung-Hwan Jeon, Giseop Kim, and Ayoung Kim. STheReO: Stereo Thermal Dataset for Research in Odometry and Mapping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022. 6

[65] Ying Zang, Bo Yu, Longjiao Yu, Dongsheng Yang, and Qingshan Liu. Far-Infrared Object Segmentation Focus on Transmission of Overall Semantic Information. *IEEE Access*, 8:182564–182579, 2020. 2

[66] Yu Zeng, Zhe Lin, Jianming Zhang, Qing Liu, John Collomosse, Jason Kuen, and Vishal M. Patel. SceneComposer: Any-Level Semantic Image Synthesis. In *IEEE CVPR*, 2023. 2

[67] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, Lingjie Liu, Adam Kortylewski, Christian Theobalt, and Eric Xing. Multimodal Image Synthesis and Editing: The Generative AI Era. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15098–15119, 2023. 2

[68] Lichao Zhang, Abel Gonzalez-Garcia, Joost van de Weijer, Martin Danelljan, and Fahad Shahbaz Khan. Synthetic data generation for end-to-end thermal infrared tracking. *IEEE Transactions on Image Processing*, 28(4):1837–1850, 2018. 2

[69] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *IEEE ICCV*, 2023. 2, 3, 4, 7