FoV-Net: Field-of-View Extrapolation Using Self-Attention and Uncertainty

Liqian Ma[1] Stamatios Georgoulis[2] Xu Jia[3] Luc Van Gool[1,2]

Fig. 1: Given the present and past frames with narrow FoV, we hallucinate the present frame with wide FoV (left: hallucination, middle: ground truth) and predict the associated uncertainty (right). The past frames are omitted for brevity.

***Abstract*—The ability to make educated predictions about their surroundings, and associate them with certain confidence, is important for intelligent systems, like autonomous vehicles and robots. It allows them to plan early and decide accordingly. Motivated by this observation, in this paper we utilize information from a video sequence with a narrow field-of-view to infer the scene at a wider field-of-view. To this end, we propose a temporally consistent field-of-view extrapolation framework, namely FoV-Net, that: (1) leverages 3D information to propagate the observed scene parts from past frames; (2) aggregates the propagated multi-frame information using an attention-based feature aggregation module and a gated self-attention module, simultaneously hallucinating any unobserved scene parts; and (3) assigns an interpretable uncertainty value at each pixel. Extensive experiments show that FoV-Net does not only extrapolate the temporally consistent wide field-of-view scene better than existing alternatives, but also provides the associated uncertainty which may benefit critical decision-making downstream applications. Project page is at http://charliememory.github.io/RAL21_FoV.**

***Index Terms*—Computer Vision for Automation, Deep Learning for Visual Perception, Visual Learning**

I. Introduction

In our pursuit of intelligent machine perception, it is crucial to endow systems, like autonomous vehicles and robots, with an awareness of the scene content beyond their immediately visible field-of-view (FoV). Simply put, the system should be able to hallucinate its surroundings, and associate each prediction with certain confidence, which could help it plan early and decide accordingly. For example, when a moving camera is turning right at a corner in a road-bound scene, the right blind part is completely unobserved but can be reasonably hallucinated. Or when observing a car in a neighboring lane over time, wide FoV synthesis can help to reason about its future position, also beyond its immediately visible FoV. Such wide FoV hallucination ability can benefit vision-based navigation [58], exploration [28], and augmented-reality telerobotics system [40]. To realize this idea, we can draw inspiration from how humans use vision to relate themselves to the world around them. Humans clearly have a situational awareness that goes beyond their narrow FoV. On the one hand, this is grounded in a capability to propagate local scene content from past observations (*e.g.,* anticipate the future position of a previously observed building based on the car's trajectory when driving). On the other hand, it is due to an ability to hallucinate global scene content for unobserved regions based on the scene's context (*e.g.,* the unobserved side views in a driving scene are likely to contain trees if the car crosses a forest area). Most importantly, humans can typically assign a degree of confidence in these judgments to quantify their intuition.

Motivated by these observations, in this paper, we tackle the problem of *FoV extrapolation*. The goal is to leverage information from a video sequence with narrow FoV (including the present and few past frames) in order to infer the (present) scene at a wider FoV (see Fig. 1). There are several challenges associated with this problem. (1) A large image size discrepancy between the input narrow FoV frames and the output wide FoV frame has to be bridged, and the results should be temporally consistent. (2) Some areas in the wide FoV frame may change significantly or even not appear in any of the past narrow FoV frames. For example, far away objects in the past narrow FoV frames need upscaling or novel view synthesis, and some occluded or unobserved regions need to be inpainted. Thus, lots of details need to be hallucinated in the wide FoV frame. (3) There is ambiguity existing between the narrow FoV observations and the wide FoV ground truth. The ambiguity mainly comes from two sources: the unobserved information and possible 3D estimation errors. In particular, the pixels in the wide FoV frame can be roughly divided into four types (see Fig. 2): (a) the observed narrow FoV pixels in the present frame (no ambiguity); (b) the propagated pixels from past frames with accurate propagation (low ambiguity); (c) the propagated pixels from past frames with noisy propagation (medium ambiguity); (d) the unobserved regions (high ambiguity). When the ambiguity is high, strong

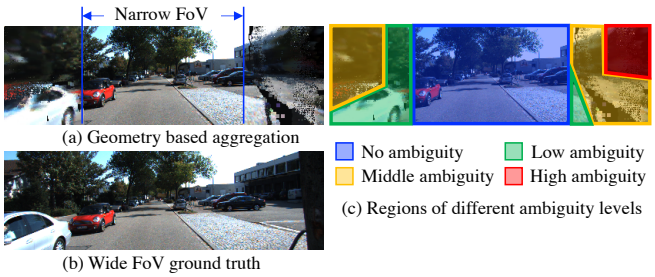
Manuscript received: October 14, 2020; Revised: January 17, 2021; Accepted: February 19, 2021. This paper was recommended for publication by Editor Dana Kulic upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by Toyota Motors Europe.

[1] L. Ma and L. Van Gool are with with the Department of Electrical Engineering, KU Leuven, Belgium {liqian.ma,luc.vangool}@esat.kuleuven.be

[2] S. Georgoulis and L. Van Gool are with the Department of Information Technology and Electrical Engineering, ETH Zürich, Switzerland {georgous,vangool}@vision.ee.ethz.ch

[3] X. Jia is with the School of Artificial Intelligence, Dalian University of Technology, China xjia@dlut.edu.cn

Digital Object Identifier (DOI): see top of this page.



(a) Geometry based aggregation

(b) Wide FoV ground truth

(c) Regions of different ambiguity levels

Fig. 2: Ambiguity illustration. Best viewed in the digital form.

enforcement of pixel reconstruction may mislead the training process. In contrast, perceptual and adversarial losses can be more suitable to predict a plausible outcome.

To address these challenges, we propose a temporally consistent FoV extrapolation framework called *FoV-Net*, which consists of two stages (see Fig. 3). A coordinates generation stage propagates past narrow FoV frames into the present wide FoV frame by leveraging 3D scene information (addressing challenge 1). A frame aggregation stage combines the multi-frame propagated information, simultaneously hallucinating fine details and unobserved scene parts (addressing challenges 2&3). Specifically, in the frame aggregation stage, we introduce an Attention-based Feature Aggregation (AFA) module to better fuse the propagated multi-frame information on the feature level, and a Gated Self-Attention (GSA) module to handle the discussed ambiguities and improve the generation quality (addressing challenges 2&3). Finally, we introduce an uncertainty mechanism to interpret the hallucination uncertainty at each pixel and guide the learning by reducing supervision ambiguity (addressing challenge 3). Such hallucination uncertainty is rarely discussed in the image synthesis field, but is quite important for practical downstream decision-making applications. Fig. 1 gives an example of our FoV extrapolation with its associated uncertainty.

II. Related Work

Video-based image synthesis. This problem exists in various forms in the literature, including video inpainting, video extrapolation, novel view synthesis, future video prediction, video-to-video synthesis, *etc*. Video inpainting [24, 55, 32, 4, 5, 15] aims to hallucinate the missing pixels through warping or generate the missing pixels conditioned on the neighboring (in spatial or temporal dimensions) visible pixels. The typical setting is to utilize the past, present, and future frames to inpaint the missing pixels in the present frame, all within narrow FoV. Video extrapolation [34, 9, 64, 19] usually adopts 2D or 3D geometry-based image warping and stitching techniques to blend the observed pixels of adjacent narrow FoV frames in order to extend the FoV, but totally ignores any unobserved pixels and object view changes. Novel view synthesis [43, 52, 21, 8, 7, 14] aims to generate images of a given object or scene from different viewpoints by blending the observed pixels, as well as hallucinating a few missing pixels mainly for dis-occlusion. When applied to scenes, it is heavily reliant on highly accurate multi-view geometry to produce good results. Future video prediction [11, 3, 36, 38, 16, 35] focuses on hallucinating future frames conditioned on the past and present frames, all within narrow FoV. Undoubtedly, this task entails higher uncertainty in the predictions. Video-to-video synthesis [56, 6, 48, 1] mainly transfers the appearance while preserving the structure of the input (*e.g.,* semantic maps). Thus the input and output are usually well aligned and within narrow FoV. Unlike the existing forms, our goal is to infer the present scene at a wider FoV, including the observed and unobserved pixels, conditioned on the past and present narrow FoV observations. Note that, when the camera is moving forward, most out-of-view regions of the present narrow FoV frame are actually future predictions as far as observations in the past frames are concerned. While if the camera is turning around at a corner, part out-of-view regions may be totally unobserved before, which makes the problem more challenging. The object size and view may change significantly, leading to large unobserved regions. To the best of our knowledge, none of the existing video-based image synthesis forms fully covers the needs of our intricate problem.

Attention. The success of self-attention models in natural language processing has inspired various applications in the computer vision field, such as in image recognition [23, 67], image synthesis [65, 2, 44], video prediction [29, 57], and imitation learning [46]. Self-attention can be formulated as locally adaptable convolutional layers with different weights for different types of image regions [67]. In our FoV extrapolation problem, we observe that such local adaptability is essential in terms of hallucination quality, since different image regions have different characteristics (see Fig. 2). To improve the hallucination quality, we propose a novel gated self-attention module, motivated by the success of gated convolution in image inpainting [62] and that of self-attention in image recognition [67]. Note that, the gated self-attention concept is not new and has been introduced in natural language processing [12, 53, 68, 33]. However, in this work, we extend it to the video domain. Additionally, in order to better aggregate the propagated multi-frame information, we propose an Attention-based Feature Aggregation (AFA) module – which is not to be confused with self-attention used above – that makes our framework more robust to propagation errors and improves the generation quality (see Sec. III-B).

Uncertainty estimation. Reasoning about the uncertainty of neural network prediction is essential for practical decision-making applications [37]. Although uncertainty estimation has been proved to be effective for several computer vision tasks, including object detection [20], semantic segmentation [31, 25], depth estimation [45, 61] and optical flow [26], it remains largely unexplored in the image synthesis literature. In this paper, we propose a hallucination uncertainty estimation mechanism, which not only enables the prediction of uncertainty, but also guides the learning by reducing supervision ambiguity. In general, there are different ways to estimate the uncertainty, including empirical estimation, predictive estimation, and Bayesian estimation (see [45, 37] for a comprehensive survey). Among them, the predictive estimation is desirable due to its effectiveness and efficiency, and has been explored in several computer vision tasks [45, 26, 31]. As a positive side effect,

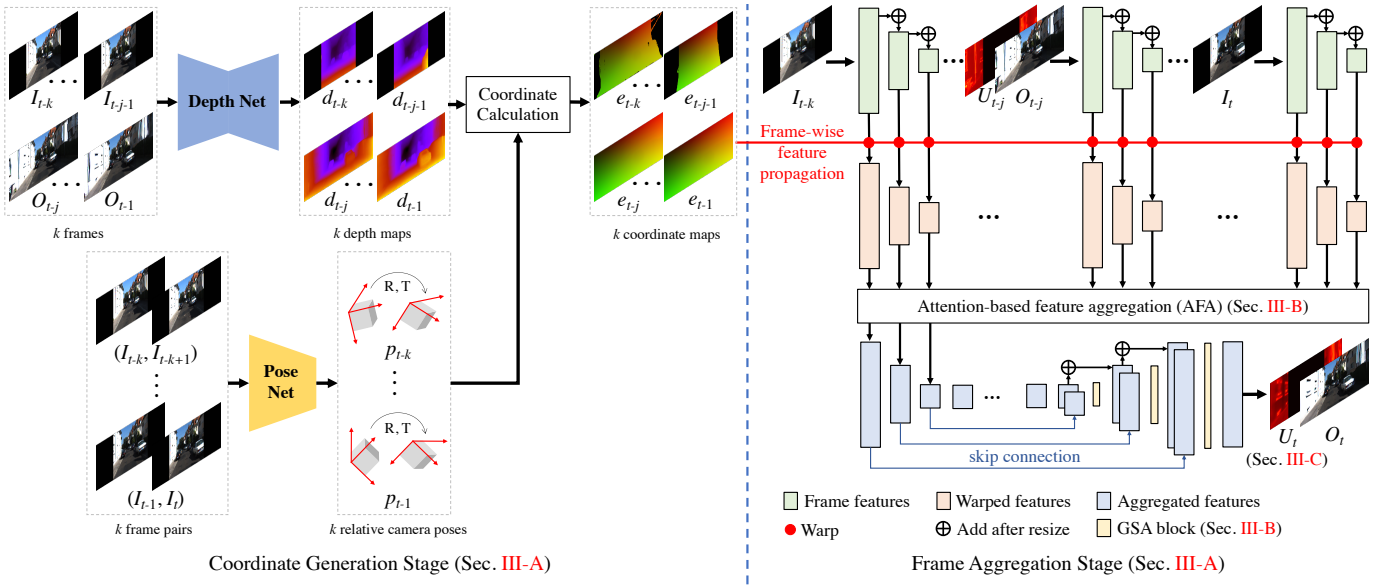


Fig. 3: The proposed *FoV-Net* framework. Left: the coordinate generation stage (Sec. III-A), which estimates the scene-level rigid flow, governed by the camera motion, and uses it to calculate the coordinates (*i.e.,* pixel displacements). Right: the frame aggregation stage (Sec. III-A), which utilizes the generated coordinates to propagate past frames information on a multi-scale feature level – denoted as red dots – and then aggregates the propagated features with an Attention-based Feature Aggregation module (Sec. III-B). To synthesize the final result O_t, a U-Net architecture is adopted to in/out-paint the missing regions, where a Gated Self-Attention module (Sec. III-B) is introduced to handle different ambiguities for better generation quality. Concurrently, an uncertainty map U_t is jointly estimated to interpret the hallucination uncertainty at each pixel and guide the learning by reducing supervision ambiguity (Sec. III-C). For I_t, we use the identical coordinates as e_t, namely, the features are not changed after warping. The coordinate e_i is also used in the GSA blocks to warp the past hidden states, but we omit the arrows here. The discriminator networks for adversarial losses are also omitted for clarity. Due to FoV-Net's recurrent nature, note that previous outputs $\{O_{t-i}\}_{i=1,\dots,j}$ and $\{U_{t-i}\}_{i=1,\dots,j}$ become future inputs $\{I_{t-i}\}_{i=1,\dots,j}$ for temporal coherency purposes. Best viewed in the digital form.

when integrated with the training objective in our problem, the predictive uncertainty can naturally weight the loss functions spatially to reduce the supervision ambiguity.

III. FoV-Net framework

A. System overview

Given a present narrow FoV frame I_t and k past narrow FoV frames $\{I_{t-i}\}_{i=1,\dots,k}$ ($k = 5$ in our experiment), our goal is to synthesize the present wide FoV frame O_t – close to ground truth W_t – and predict the hallucination uncertainty U_t. In addition, the adjacent synthesized results O_{t-1} and O_t should be temporally consistent. To achieve these, we propose a two-stage recurrent framework (Fig. 3) consisting of a coordinate generation stage and a frame aggregation stage, coupled with a hallucination uncertainty mechanism (Sec. III-C). The coordinate generation stage is designed to estimate the scene-level rigid flow, governed by the camera motion, and use it to generate coordinates (*i.e.,* pixel positions) in order to spatially propagate information from the past narrow FoV frames. The frame aggregation stage is designed to aggregate the past narrow FoV frames $I_{t-k}, \dots, I_{t-1}$ and present narrow FoV frame I_t into one wide FoV image, as well as hallucinate the unobserved missing regions. To enforce temporal coherency, we use a simple recurrent feed-forward strategy: replace the narrow FoV inputs $\{I_{t-i}\}_{i=1,\dots,j}$ with the previous outputs $\{O_{t-i}\}_{i=1,\dots,j}$, and feed the corresponding previous uncertainty $\{U_{t-i}\}_{i=1,\dots,j}$ by channel-wise concatenation ($j = 2$ in our settings). We analyze each stage of our framework below.

Coordinate generation stage, that builds upon Monodepth2 [18], consists of a depth network $\mathcal{D}_{\theta_\mathcal{D}}$ and a relative camera pose network $\mathcal{P}_{\theta_\mathcal{P}}$[1]. During training, $\mathcal{P}$ takes a pair of two temporally adjacent frames as input and outputs the relative camera pose, and $\mathcal{D}$ takes one frame of the pair as input and outputs its depth. During inference, however, we do not have access to the wide FoV frame which is required by the inverse warping operation [27] in order to propagate the pixels between two frames. To address this, we design a forward warping strategy to propagate the past narrow FoV frames to the present wide FoV frame. We first utilize the depth maps from the past narrow FoV frames to calculate the rigid flow $f^{rig}_{t\to i}(\hat{e}_{t\to i})$ from the present frame I_t to the past frame I_i, using Eq. 1.

$$f^{rig}_{t\to i}(\hat{e}_{t\to i}) = K\mathrm{T}_{i\to t}\mathcal{D}_i(c_i)K^{-1}c_i - c_i, \tag{1}$$

where K denotes the camera intrinsic matrix, $\mathrm{T}_{i\to t}$ denotes the relative camera pose, and c_i denotes homogeneous coordinates of pixels in frame I_i. Then, using the calculated flow, we compute the spatial mapping, *i.e.,* coordinate map $\hat{e}_{t\to i}$, that

[1]The subscripts $\{\theta_\mathcal{D}, \theta_\mathcal{P}, \theta_\mathcal{A}, \theta_\mathcal{Q}, \theta_\mathcal{T}\}$ are network parameters. We regularly omit the subscripts for brevity.

spatially matches the pixel positions of present frame I_t to the corresponding ones of past frame I_i. Finally, we reverse this correspondence, *i.e.,* $e_i = \text{reverse}(\hat{e}_{t\to i})$, which now corresponds to the spatial positions from the past frame I_i to the present frame I_t. This coordinate map e_i will be used to propagate features in the frame aggregation stage using bilinear sampling.

Frame aggregation stage is designed to aggregate the past narrow FoV frames $I_{t-k},\dots,I_{t-1}$ and present narrow FoV frame I_t as well as previous wide FoV results $\{O_{t-i}\}_{i=1,\dots,j}$, $\{U_{t-i}\}_{i=1,\dots,j}$, into one wide FoV image, simultaneously hallucinating unobserved missing regions. It contains aggregation network $\mathcal{A}_{\theta_\mathcal{A}}$, image discriminator network $\mathcal{Q}_{\theta_\mathcal{Q}}$, and temporal discriminator network $\mathcal{T}_{\theta_\mathcal{T}}$. The aggregation network $\mathcal{A}$ first extracts a residual multi-scale feature pyramid (with $N=3$ levels) from each frame using an encoder with shared weights, and propagates the multi-scale features using the computed coordinates e_i. Then, an Attention-based Feature Aggregation (AFA) module (Sec. III-B) aggregates the propagated features. To synthesize the final result based on the aggregated features, a U-Net decoder is designed, where we introduce the Gated Self-Attention (GSA) module (Sec. III-B) to adaptively handle ambiguities and improve the generation quality.

B. Self-attention mechanism

Attention-based Feature Aggregation (AFA). In Fig. 4, to aggregate the propagated features, each set of propagated multi-scale features maps (Fig. 3 right) are fed into a convolution layer followed by softmax normalization to predict frame-wise spatial attention maps (*i.e.,* one channel attention map for each frame). Then, the propagated feature maps are multiplied by the attention maps and summed across all frames. This attention-based aggregation module can learn to select the useful features among these frames to address the issues caused by depth/pose errors and frame inconsistency.

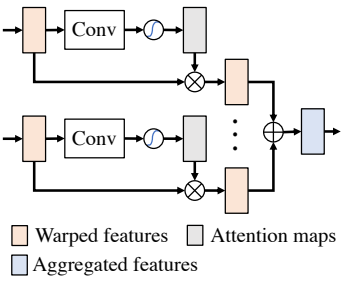


Fig. 4: AFA module. Same color coding as Fig. 3

Gated Self-Attention (GSA). To make our model adaptable to observations with different ambiguity levels, we encompass self-attention and gating mechanisms to construct a Gated Self-Attention (GSA) module. Here, we adopt a patch-wise self-attention block introduced in [67], which efficiently computes local attention weights that vary over spatial coordinates and channels instead of sharing weights to convolve the whole feature maps like conventional CNN. It has the form:

$$\mathbf{y}_i = \sum_{j\in\mathcal{R}(i)} \alpha(\mathbf{x}_{\mathcal{R}(i)})_j \odot \beta(\mathbf{x}_j), \qquad (2)$$
$$\text{where} \quad \alpha(\mathbf{x}_{\mathcal{R}(i)}) = \gamma(\delta(\mathbf{x}_{\mathcal{R}(i)}))).$$

The function β produces the feature vectors $\beta(\mathbf{x}_j)$ that are weighted summarized by the adaptive weight vectors $\alpha(\mathbf{x}_{\mathcal{R}(i)})_j$. The tensor $\mathbf{x}_{\mathcal{R}(i)}$ is the patch of feature vectors in a 7×7 footprint $\mathcal{R}(i)$. $\alpha(\mathbf{x}_{\mathcal{R}(i)})_j$ is the attention vector at location j in tensor $\alpha(\mathbf{x}_{\mathcal{R}(i)})$, corresponding spatially to the vector x_j in $\mathbf{x}_{\mathcal{R}(i)}$. Functions β and γ are mappings implemented via one convolution layer, respectively. The function δ combines the feature vectors x_j from the patch $\mathbf{x}_{\mathcal{R}(i)}$ implemented via a concatenation operation $\delta(\mathbf{x}_{\mathcal{R}(i)})) = [\phi(\mathbf{x}_i), [\psi(\mathbf{x}_j)]_{\forall j\in\mathcal{R}(i)}]$, where ϕ and ψ are mappings implemented via one convolution layer, respectively. $\odot$ denotes the Hadamard product. To reduce the impact of vanishing gradients, we wrap the self-attention block in a residual structure, *i.e.,* $\mathbf{z} = \text{Conv}_r(\mathbf{y}) + \mathbf{x}$. We then equip our self-attention block with a gating mechanism to learn to control the information flows of different ambiguities, formulated as:

$$\mathbf{g} = \text{sigmoid}(\text{Conv}_g(\mathbf{z})) \odot \tanh(\text{Conv}_a(\mathbf{z})), \qquad (3)$$

where $\text{sigmoid}(\text{Conv}_g(\mathbf{z}))$ and $\tanh(\text{Conv}_a(\mathbf{z})$ denote the gate and feature activation, respectively. Conv_r, Conv_g, Conv_a are 2D convolution layers, and the subscriptions stand for residual, gate, and attention, respectively.

C. Uncertainty mechanism

We design an uncertainty mechanism to not only predict the interpretable hallucination uncertainty, but also guide the learning by reducing supervision ambiguity. We draw inspiration from prior work where predicting the data-dependent uncertainty helped in tempering the training objective, *e.g.,* by attenuating the effect from erroneous labels in [26], or by automatically balancing the loss weighting in [31]. To realize this idea, one may design a heuristic weighting map, like the spatially discounted reconstruction loss [63], but such approaches are ad-hoc and cannot adapt to different scenes automatically. Instead, we opt to jointly learn an uncertainty map during training, which serves as a probabilistic interpretation of our model. We build upon predictive estimation [42], and infer the mean and variance of the distribution $p(O_t|I_{t-i},\mathbb{D})$, where $i=0,\dots,K$, and $\mathbb{D}$ denotes the whole dataset. The network is trained by log-likelihood maximization (*i.e.,* negative log-likelihood minimization) and the distribution can be modelled as Laplacian (*i.e.,* corresponding to L1 loss) or Gaussian (*i.e.,* corresponding to L2 loss) respectively. The negative log-likelihood formulation is:

$$L_{1Log} = \frac{\|\mu(x) - x^*\|_1}{\sigma(x)} + \log\sigma(x), \qquad (4)$$

where $\mu(x)$ and $\sigma(x)$ are the network outputs, encoding mean and variance of the distribution. Here, we adopt L1 loss for pixel level reconstruction and reformulate Eq. 4 integrating uncertainty U_t as:

$$L_{1U}^{\theta_A} = \mathbb{E}\left[\left(\frac{\|O_t - W_t\|_1}{U_t}\right)\odot M + \|O_t - W_t\|_1 \odot (1-M) + U_t\right], \qquad (5)$$

where O_t, W_t are the predicted wide FoV RGB image and the ground truth RGB image (3-channel), respectively. U_t is the estimated hallucination uncertainty map (1-channel).

M denotes the mask for out-of-narrow-FoV regions. Within the narrow FoV region $(1-M)$, the L1 is not weighted by uncertainty U_t, as newthe present narrow FoV region has been observed. Note that, to make the uncertainty term U_t more interpretable and stabilize the training process, we constrain U_t in the range (0,1) with a sigmoid function and modify the regularization term from $\log U_t$ to U_t for gradient stabilization. We found that such modification also leads to better performance. Additionally, in our recurrent framework, the previous predicted uncertainty $\{U_{t-i}\}_{i=1,\dots,j}$ are also used in future input to act as a confidence signal.

D. Losses

Coordinate generation losses. Following [18], the objective in this stage is a masked photometric loss $\nu \odot L_{photo}$ and an edge-aware smoothness loss L_{smooth}, summarized below:

$$L_{CG}^{\theta_D,\theta_P} = \nu \odot L_{photo}^{\theta_D,\theta_P} + \lambda_s L_{smooth}^{\theta_D,\theta_P}, \tag{6}$$

$$L_{photo}^{\theta_D,\theta_P} = \min_{t'} pe(I_t, I_{t'\to t}), \tag{7}$$

$$pe(I_a, I_b) = \frac{\alpha}{2}(1-\text{SSIM}(I_a, I_b)) + (1-\alpha)\|I_a - I_b\|_1,$$

$$L_{smooth}^{\theta_D,\theta_P} = |\nabla_x d_t^*| e^{-|\nabla_x I_t|} + |\nabla_y d_t^*| e^{-|\nabla_y I_t|}, \tag{8}$$

where $\nu = [\min_{t'} pe(I_t, I_{t'\to t})]$ acts as an auto-mask for suppressing the effect of objects moving at similar speeds to the camera, and [] is the Iverson bracket. $d^* = d_t/\overline{d_t}$ is the mean-normalized inverse depth. SSIM is the structural similarity [59]. We use hyper-parameters $\alpha = 0.85$ and $\lambda_s = 0.001$ as in [18].

Frame aggregation losses. The objective of the frame aggregation stage has four loss terms: uncertainty-aware L1 reconstruction loss $L_{1U}^{\theta_A}$, perceptual reconstruction loss $L_{perc}^{\theta_A}$ [66], adversarial loss $L_{adv}^{\theta_A,\theta_Q}$ [41], and temporal adversarial loss $L_{advT}^{\theta_A,\theta_T}$. The reconstruction losses $L_{1U}^{\theta_A}$ and $L_{perc}^{\theta_A}$ are used to regress the output towards the target ground truth, while the adversarial losses $L_{adv}^{\theta_A,\theta_Q}$ and $L_{advT}^{\theta_A,\theta_T}$ are used to encourage image photo-realism and temporal coherence. The formulations are as follows:

$$\begin{aligned} L_{FA}^{\theta_A,\theta_Q,\theta_T} &= \lambda_1 L_{1U}^{\theta_A}(O_t, W_t) + \lambda_2 L_{perc}^{\theta_A}(O_t, W_t) \\ &+ L_{adv}^{\theta_A,\theta_Q}(O_t, W_t) + L_{advT}^{\theta_A,\theta_T}(O_t, W_t), \end{aligned} \tag{9}$$

$$L_{perc}^{\theta_A} = \mathbb{E}\big[\|\phi(O_t) - \phi(W_t)\|_2^2 \odot (1+M))\big], \tag{10}$$

$$\begin{aligned} L_{adv}^{\theta_A,\theta_Q} &= \mathbb{E}\big[(Q(W_t, M))^2 \odot (1+M)\big] \\ &+ \mathbb{E}\big[(Q(O_t, M)-1)^2 \odot (1+M)\big], \end{aligned} \tag{11}$$

$$\begin{aligned} L_{advT}^{\theta_A,\theta_T} &= \mathbb{E}\big[(T(\{W_{t-i}\}_{i=0,\dots,j}, M))^2 \odot (1+M)\big] \\ &+ \mathbb{E}\big[(T(\{O_{t-i}\}_{i=0,\dots,j}, M)-1)^2 \odot (1+M)\big], \end{aligned} \tag{12}$$

where $\lambda_1 = 3$, $\lambda_2 = 10$. ϕ is a VGG network [50].

IV. Experiments

To evaluate FoV-Net w.r.t. its FoV extrapolation capabilities, we provide both qualitative and quantitative results. For more video results and training/implementation details, please visit the supplementary materials.

Dataset. Our method is evaluated on two widely used datasets: raw KITTI sequences [17] using the split from Eigen *et al.* [13], and Cityscapes sequences [10]. To reduce the redundant information in videos, we downsample the frame rate to 1/2 and 1/3 for KITTI and Cityscapes sequences. Therefore, we have 39350/4382 and 59526/14992 train/val frames on KITTI and Cityscapes, respectively. During testing, we prepare each video sample with 10 and 5 target frames for KITTI and Cityscapes. Totally, there are 635 (6350 frames) and 1525 test video samples (7625 frames) in KITTI and Cityscapes, respectively. We use both monocular left and right camera sequences for training and validation, but only use left camera data for testing. In each forward pass, we use 6 successive frames ($k = 5$ past + 1 present) as input. The narrow FoV ratio is set to 0.5 in our experiments, *i.e.,* all 6 frames are cropped 25% on both left and right sides to mimic the narrow FoV.

Metrics. For quantitative evaluation, four image quality metrics are used: Structural Similarity (SSIM) [59], Learned Perceptual Image Patch Similarity (LPIPS) [66], Fréchet Inception Distance (FID) [22], and Fréchet Video Distance (FVD) [54]. SSIM and LPIPS are used to evaluate the similarity between the result and the ground truth. FID and FVD are used for realistic appearance evaluation on the image level and video level, respectively. Thus, FVD can reflect both appearance realism and temporal coherence of the results. For SSIM, higher scores are better. For LPIPS, FID, and FVD, lower scores are better. We use VGG [50] pre-trained on ImageNet as the feature extractor of LPIPS. To evaluate how significant the modeled hallucination uncertainties are, we use sparsification plots and Area Under the Sparsification Error (AUSE, the lower, the better) which quantify how close the estimate is to the oracle uncertainty, as in [45]. More details are given in "Uncertainty results".

A. Comparisons

Baselines. As no prior work is directly comparable to our setting, we compare against geometric method Mono [18], flow-based video prediction method VoxelFlow [36], and video completion method LGTSM [5], the closest alternatives. We also evaluate the Mono+LGTSM (Mono-LGTSM) setting, where pixels of past frames are first propagated by Mono before fed into LGTSM.

Hallucination Results. Qualitative comparisons with other alternatives are provided in Fig. 5. We observe that our method can synthesize more realistic and perceptually appealing results. For example, our method can produce hallucinations with less distortion in the 2nd column result, and better preserve the object appearance with less artifacts in the 3rd and 4th column results. This trend is also reflected from the quantitative measurements (SSIM, LPIPS, and FID scores) in Tab. I. Our method generates sharper and more photo-realistic results than VF [36] and LGTSM [5], but the latter have a higher SSIM, probably due to blur, something also argued in [30, 49, 39]. While LPIPS is generally more consistent with human perception and has been wildly adopted in recent image synthesis works [47, 5] Furthermore, our method alleviates the flickering via the simple recurrent strategy which can retain temporal consistency. Note that, the hallucination of unobserved

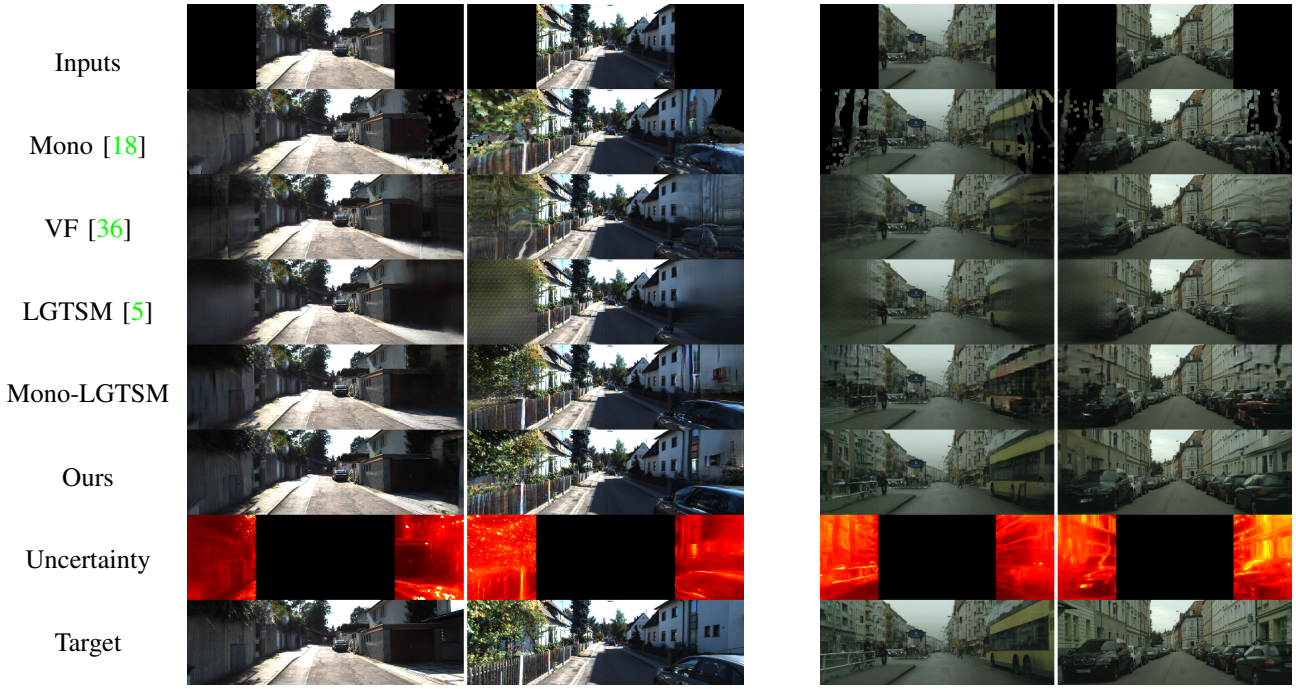


Fig. 5: The qualitative comparisons on KITTI (column 1-2) and Cityscapes (column 3-4). The video version is included in our supplementary PDF.

TABLE I: The quantitative results on KITTI and Cityscapes.

	KITTI				Cityscapes			
Model	SSIM↑	LPIPS↓	FID↓	FVD↓	SSIM↑	LPIPS↓	FID↓	FVD↓
Mono [18]	0.6803	0.2975	31.14	269.0	0.6444	0.3375	44.90	552.8
VF [36]	0.7184	0.3049	33.51	360.8	0.7844	0.2936	38.78	491.5
LGTSM [5]	**0.7369**	0.2905	52.57	495.7	**0.7959**	0.2845	72.86	572.9
Mono-LGTSM	0.7028	0.2798	18.95	201.6	0.7752	0.2866	51.18	457.8
Ours	0.7162	**0.2294**	**10.94**	**82.71**	0.7539	**0.2220**	**9.27**	**203.4**

TABLE II: Ablation study results on KITTI.

Model	SSIM↑	LPIPS↓	FID↓	FVD↓
Base	0.7020	0.2407	12.26	99.75
w/o AFA	0.7016	0.2377	11.25	91.92
w/o GSA	0.7160	0.2386	11.94	98.67
w/o U	**0.7200**	0.2312	11.37	89.70
w/o 3D	0.6830	0.2618	16.71	171.1
w/o Recur	0.7145	0.2331	11.11	109.5
Ours	0.7162	**0.2294**	**10.94**	**82.71**

regions is improved as more surrounding observations become available in later frames. The FVD score is also consistent with such observations. In addition to blur and distortion, Mono [18] and VF [36] suffer from missing pixels. LGTSM [5] cannot propagate the information to the out-of-view regions correctly and thus performs poorly. When using Mono and LGTSM together, we can combine their merits and arrive at better results. This indicates that both 3D cues and hallucination capability are key to the FoV extrapolation problem. FoV-Net gets the best of both worlds, and further addresses ambiguity, leading to the best outcome.

Uncertainty results. To evaluate the significance of estimated uncertainty, we adopt a pixel-wise metric Mean-Square-Error (MSE) to sort all pixels in each hallucinated wide FoV image in order of descending uncertainty. Then, we iteratively remove a subset of pixels in the out-of-view regions (*i.e.,* 5% in our experiments) and compute MSE on the remaining to plot a curve that is supposed to shrink if the uncertainty properly encodes the hallucinated image's errors (see Fig. 7). An ideal sparsification (oracle) is obtained by sorting pixels in descending order of the MSE magnitude. In contrast, a random uncertainty is to remove the pixels randomly each time. Besides, we observe that there is usually high uncertainty in the edge part. Therefore, we also construct an edge uncertainty baseline which adopts an estimated soft edge map [60] to approximate the hallucination uncertainty. As we use 0.5 as our narrow view ratio, the curves decrease to zero when 50% of pixels are extracted. The AUSE scores on KITTI are 0.0049, 0.0105, 0.0149, and 0.0197 for oracle, uncertainty, edge, and random settings, respectively. The AUSE scores on Cityscapes are 0.0021, 0.0042, 0.0070, and 0.0080 for oracle, uncertainty, edge, and random settings, respectively. As shown in Fig. 7, our method can successfully estimate the hallucination uncertainty which reasonably indicates the prediction errors. For example, in the 2nd column result of Fig. 5, the uncertainty is high on the thin pole and car edges, as the depth estimation is noisy there. In the 4th column result of Fig. 5, the uncertainty is high on the car edges and unobserved regions (black holes in the Mono result).

Ablation study is performed on KITTI for each component of FoV-Net. The ablative settings are as follows. **Base:** removing the AFA module, GSA module, and the uncertainty mechanism. **w/o AFA:** using temporal average pooling to replace the AFA module. **w/o GSA:** removing the GSA modules. **w/o**

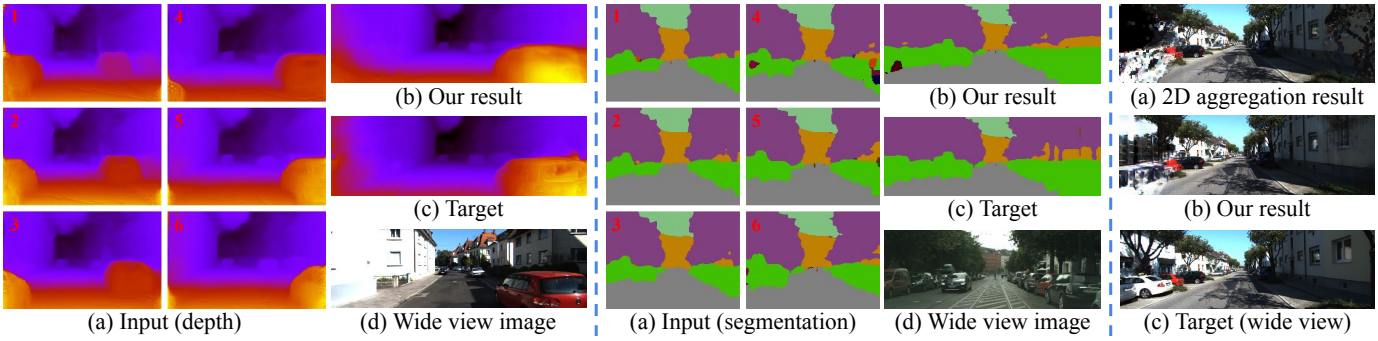
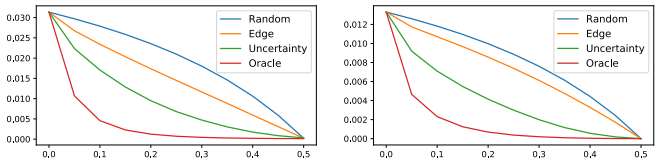


Fig. 6: Left: depth extrapolation. Middle: semantic segmentation extrapolation. Right: failure case. Zoom in for more details.



(a) KITTI (b) Cityscapes

Fig. 7: The sparsification plots. The x-axis denotes the fraction of removed pixels, and the y-axis shows the MSE on the remaining pixels. MSE converges to 0 after removing all pixels in the out-of-view regions. Zoom in for more details.

U: removing the uncertainty mechanism. **w/o 3D:** removing the coordinate generation stage and the feature propagation operation in the frame aggregation network. **w/o Recur:** removing temporal modeling, *i.e.,* no recurrent feed-forward and temporal discriminator $\mathcal{T}$. The quantitative results are in Tab. II. When equipped with our AFA and GSA modules, the results are perceptually closer to the wide FoV targets (in terms of LPIPS score decreasing), as well as more photo-realistic (in terms of FID/FVD scores decreasing) and temporal consistency (in terms of FVD score decreasing). Besides, our interpretable uncertainty mechanism could improve the performance moderately, indicating its ability to guide the learning by reducing supervision ambiguity. In addition, the 3D cues and temporal modeling are also important to good performance.

B. Extended Applications

In this section, we provide several extensions (also see supplementary material). We further apply our method to depth extrapolation (Fig. 6 left) and semantic extrapolation (Fig. 6 middle). Results show that our method can benefit other vision perception tasks as well, which are important for robotic motion planning and navigation.

Depth extrapolation. To extrapolate the wide FoV depth, the input will be the narrow FoV depth maps. Our frame aggregation network is then trained with the scale-invariant depth regression loss [13]. Both the input and target depth maps are estimated by our depth network D.

Semantic segmentation extrapolation. To infer the wide FoV semantic segmentation, the input is the narrow FoV segmentation maps, and our frame aggregation network is then trained with a cross-entropy loss. Both the input and target segmentation maps are estimated by HRNet [51].

V. Limitations and Conclusions

While FoV-Net has achieved good results, there remain a plethora of avenues for future work. First, our FoV-Net may be not robust to fast moving objects. For example, in Fig. 6 right, the blurred left region is due to the white moving car which causes serious propagation errors. One potential solution is to extend the FoV extrapolation in the 3D space with a multi-sensor system and model these moving objects explicitly. Second, in this work, we focus on extrapolating the present frame to a limited wide view. While 360°FoV extrapolation and future prediction could be more helpful in some cases. We plan to extend our method both spatially and temporally to enable 360°FoV future prediction. In conclusion, we present FoV-Net to tackle the *FoV extrapolation* problem. Our framework propagates and aggregates the information observed from the past and current narrow FoV frames to generate the current wide FoV, as well as predicts the hallucination uncertainty. We take a significant step to endow machines with hallucination ability, and we believe such an ability can benefit the robotics community.

References

[1] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh. “Recycle-gan: Unsupervised video retargeting”. In: *ECCV*. 2018.
[2] A. Brock, J. Donahue, and K. Simonyan. “Large scale gan training for high fidelity natural image synthesis”. In: *ICLR*. 2019.
[3] W. Byeon, Q. Wang, R. Kumar Srivastava, and P. Koumoutsakos. “Contextvp: Fully context-aware video prediction”. In: *ECCV*. 2018.
[4] Y.-L. Chang, Z. Y. Liu, K.-Y. Lee, and W. Hsu. “Free-form video inpainting with 3d gated convolution and temporal patchgan”. In: *ICCV*. 2019.
[5] Y.-L. Chang, Z. Y. Liu, K.-Y. Lee, and W. Hsu. “Learnable gated temporal shift module for deep video inpainting”. In: *BMVC*. 2019.
[6] D. Chen, J. Liao, L. Yuan, N. Yu, and G. Hua. “Coherent online video style transfer”. In: *ICCV*. 2017.
[7] X. Chen, J. Song, and O. Hilliges. “NVS Machines: Learning Novel View Synthesis with Fine-grained View Control”. In: *arXiv preprint arXiv:1901.01880* (2019).
[8] I. Choi, O. Gallo, A. Troccoli, M. H. Kim, and J. Kautz. “Extreme view synthesis”. In: *ICCV*. 2019.
[9] J. Choi and I. S. Kweon. “Deep Iterative Frame Interpolation for Full-frame Video Stabilization”. In: *ACM TOG* 39.1 (2020), pp. 1–9.
[10] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, et al. “The Cityscapes Dataset for Semantic Urban Scene Understanding”. In: *CVPR*. 2016.
[11] E. Denton and R. Fergus. “Stochastic video generation with a learned prior”. In: *ICML*. 2018.

[12] B. Dhingra, H. Liu, Z. Yang, W. W. Cohen, and R. Salakhutdinov. "Gated-attention readers for text comprehension". In: *ACL*. 2017.

[13] D. Eigen, C. Puhrsch, and R. Fergus. "Depth map prediction from a single image using a multi-scale deep network". In: *NeurIPS*. 2014.

[14] J. Flynn, M. Broxton, P. Debevec, M. DuVall, G. Fyffe, et al. "DeepView: View synthesis with learned gradient descent". In: *CVPR*. 2019.

[15] C. Gao, A. Saraf, J.-B. Huang, and J. Kopf. "Flow-edge Guided Video Completion". In: *ECCV*. 2020.

[16] H. Gao, H. Xu, Q.-Z. Cai, R. Wang, F. Yu, et al. "Disentangling propagation and generation for video prediction". In: *ICCV*. 2019.

[17] A. Geiger, P. Lenz, and R. Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite". In: *CVPR*. 2012.

[18] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. "Digging into self-supervised monocular depth estimation". In: *ICCV*. 2019.

[19] H. Guo, S. Liu, T. He, S. Zhu, B. Zeng, et al. "Joint video stitching and stabilization from moving cameras". In: *IEEE transactions on image processing (TIP)* 25.11 (2016), pp. 5491–5503.

[20] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang. "Bounding box regression with uncertainty for accurate object detection". In: *CVPR*. 2019.

[21] P. Hedman, J. Philip, T. Price, J.-M. Frahm, G. Drettakis, et al. "Deep blending for free-viewpoint image-based rendering". In: *ACM TOG* 37.6 (2018), pp. 1–15.

[22] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *NeurIPS*. 2017.

[23] H. Hu, Z. Zhang, Z. Xie, and S. Lin. "Local relation networks for image recognition". In: *ICCV*. 2019.

[24] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf. "Temporally coherent completion of dynamic video". In: *ACM TOG* 35.6 (2016), pp. 1–11.

[25] P.-Y. Huang, W.-T. Hsu, C.-Y. Chiu, T.-F. Wu, and M. Sun. "Efficient uncertainty estimation for semantic segmentation in videos". In: *ECCV*. 2018.

[26] E. Ilg, O. Cicek, S. Galesso, A. Klein, O. Makansi, et al. "Uncertainty estimates and multi-hypotheses networks for optical flow". In: *ECCV*. 2018.

[27] M. Jaderberg, K. Simonyan, and A. Zisserman. "Spatial transformer networks". In: *NeurIPS*. 2015.

[28] D. Jayaraman and K. Grauman. "Learning to look around: Intelligently exploring unseen environments for unknown tasks". In: *CVPR*. 2018.

[29] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool. "Dynamic filter networks". In: *NeurIPS*. 2016.

[30] J. Johnson, A. Alahi, and L. Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution". In: *ECCV*. 2016.

[31] A. Kendall, Y. Gal, and R. Cipolla. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics". In: *CVPR*. 2018.

[32] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon. "Deep video inpainting". In: *CVPR*. 2019.

[33] T. Lai, Q. H. Tran, T. Bui, and D. Kihara. "A Gated Self-attention Memory Network for Answer Selection". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2019.

[34] S. Lee, J. Lee, B. Kim, K. Kim, and J. Noh. "Video Extrapolation Using Neighboring Frames". In: *ACM TOG* 38.3 (2019), pp. 1–13.

[35] W. Liu, A. Sharma, O. Camps, and M. Sznaier. "Dyan: A dynamical atoms-based network for video prediction". In: *ECCV*. 2018.

[36] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala. "Video Frame Synthesis using Deep Voxel Flow". In: *ICCV*. 2017.

[37] A. Loquercio, M. Segu, and D. Scaramuzza. "A general framework for uncertainty estimation in deep learning". In: *IEEE Robotics and Automation Letters (RAL)* 5.2 (2020), pp. 3153–3160.

[38] W. Lotter, G. Kreiman, and D. Cox. "Deep predictive coding networks for video prediction and unsupervised learning". In: *arXiv preprint arXiv:1605.08104* (2016).

[39] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, et al. "Pose guided person image generation". In: *NeurIPS*. 2017.

[40] Z. Makhataeva and H. A. Varol. "Augmented Reality for Robotics: A Review". In: *Robotics* 9.2 (2020), p. 21.

[41] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, et al. "Least squares generative adversarial networks". In: *ICCV*. 2017.

[42] D. A. Nix and A. S. Weigend. "Estimating the mean and variance of the target probability distribution". In: *ICNN*. 1994.

[43] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg. "Transformation-grounded image generation network for novel 3d view synthesis". In: *CVPR*. 2017.

[44] N. Parmar, A. Vaswani, J. Uszkoreit, Ł. Kaiser, N. Shazeer, et al. "Image transformer". In: *ICML*. 2018.

[45] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia. "On the uncertainty of self-supervised monocular depth estimation". In: *CVPR*. 2020.

[46] K. Ramachandruni, M. Babu, A. Majumder, S. Dutta, and S. Kumar. "Attentive Task-Net: Self Supervised Task-Attention Network for Imitation Learning using Video Demonstration". In: *IEEE Int. Conf. on Robotics and Automation (ICRA)*. 2020.

[47] Y. Ren, X. Yu, J. Chen, T. H. Li, and G. Li. "Deep image spatial transformation for person image generation". In: *CVPR*. 2020.

[48] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, et al. "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network". In: *CVPR*. 2016.

[49] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, et al. "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network". In: *CVPR*. 2016.

[50] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *ICLR*. 2015.

[51] K. Sun, B. Xiao, D. Liu, and J. Wang. "Deep high-resolution representation learning for human pose estimation". In: *CVPR*. 2019.

[52] S.-H. Sun, M. Huh, Y.-H. Liao, N. Zhang, and J. J. Lim. "Multi-view to Novel view: Synthesizing novel views with Self-Learned Confidence". In: *ECCV*. 2018.

[53] Q. H. Tran, G. Haffari, and I. Zukerman. "A generative attentional neural network model for dialogue act classification". In: *ACL*. 2017.

[54] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, et al. "Towards Accurate Generative Models of Video: A New Metric & Challenges". In: *arXiv preprint arXiv:1812.01717* (2018).

[55] C. Wang, H. Huang, X. Han, and J. Wang. "Video inpainting by jointly learning temporal structure and spatial details". In: *AAAI*. 2019.

[56] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, et al. "Video-to-video synthesis". In: *NeurIPS*. 2018.

[57] X. Wang, R. Girshick, A. Gupta, and K. He. "Non-local neural networks". In: *CVPR*. 2018.

[58] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, et al. "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation". In: *CVPR*. 2019.

[59] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[60] S. Xie and Z. Tu. "Holistically-nested edge detection". In: *ICCV*. 2015.

[61] N. Yang, L. von Stumberg, R. Wang, and D. Cremers. "D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry". In: *arXiv preprint arXiv:2003.01060* (2020).

[62] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, et al. "Free-form image inpainting with gated convolution". In: *ICCV*. 2019.

[63] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, et al. "Generative image inpainting with contextual attention". In: *CVPR*. 2018.

[64] F. Zhang and F. Liu. "Parallax-tolerant image stitching". In: *CVPR*. 2014.

[65] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. "Self-attention generative adversarial networks". In: *ICML*. 2019.

[66] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". In: *CVPR*. 2018.

[67] H. Zhao, J. Jia, and V. Koltun. "Exploring Self-attention for Image Recognition". In: *CVPR*. 2020.

[68] Y. Zhao, X. Ni, Y. Ding, and Q. Ke. "Paragraph-level neural question generation with maxout pointer and gated self-attention networks". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2018.